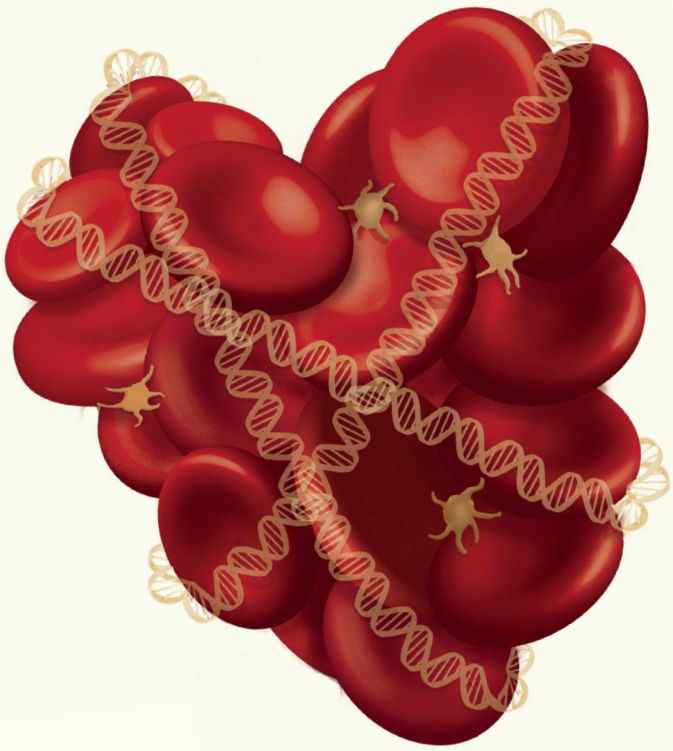
Hemostasis and Cardiovascular Disease

a molecular epidemiology approach



Paul Stefan de Vries

Hemostasis and Cardiovascular Disease
a molecular epidemiology approach

Paul Stefan de Vries

ACKNOWLEDGMENTS

The work presented in this thesis was conducted at the Cardiovascular Group of the Department of Epidemiology, Erasmus MC, Rotterdam, the Netherlands.
All of the studies described in this thesis involved the Rotterdam Study, which is supported by the Erasmus MC and the Erasmus University Rotterdam, the Netherlands Organization for Scientific Research (NWO), the Netherlands Organization for Health Research and Development (ZonMw), the Dutch Heart Foundation, the Research Institute for Diseases in the Elderly (RIDE), the Ministry of Education, Culture, and Science, the Ministry of Health Welfare and Sports, the European Commission (DG XII), and the municipality of Rotterdam.
The contribution of the inhabitants, general practitioners and pharmacists of the Ommoord district to the Rotterdam Study are gratefully acknowledged.

Publication of this thesis was kindly supported by the Department of Epidemiology and the Erasmus University Rotterdam. Financial support by the Dutch Heart Foundation for the publication of this thesis is gratefully acknowledged.

Cover design by Remco Wetzels | remcowetzels.nl
Layout and printing by Optima Grafische Communicatie | www.ogc.nl

ISBN 978-94-6169-799-8

Hemostasis and Cardiovascular Disease
a molecular epidemiology approach

Hemostase en hart- en vaatziekten
een moleculaire epidemiologie aanpak

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
woensdag 20 januari 2016 om 9:30 uur

Door

Paul Stefan de Vries

geboren te Amsterdam

Erasmus University Rotterdam

Promotiecommissie

Promotor:	prof.dr. O.H. Franco
Overige leden:	prof.dr. A.G. Uitterlinden Dr. M.P.M. de Maat prof.dr. H. Snieder
Copromotor:	Dr. A. Dehghan

Paranimfen:	Symen Ligthart Ivo van Wijk

To Lised

And to my parents

TABLE OF CONTENTS

Chapter 1

General introduction

Despite improvements in prevention and treatment, coronary heart disease (CHD) remains the leading cause of death.[1] CHD refers to the buildup of atherosclerotic plaques in the coronary arteries and the accompanying narrowing of the arteries, which may result in a myocardial infarction. Whether or not a myocardial infarction actually occurs depends on many factors: the extent of atherosclerotic plaques,[2-4] the stability of the plaques,[4,5] the narrowing of the artery,[6] and the intensity of the thrombotic response to plaque rupture.[7] The larger the blood clot, the higher the chance of obstructing the flow of blood through the coronary arteries. Indeed, several types of antithrombotic medication, including aspirin, are effective at reducing the risk and severity of myocardial infarctions.[8]

The formation of pathogenic blood clots resulting in myocardial infarctions is driven by the same mechanisms that work to stop bleeding: damaged blood vessels are constricted to limit blood flow, a platelet plug forms, and the coagulation cascade is set off, resulting in the formation of a fibrin mesh. Together, these three mechanisms cooperate to stop bleeding, achieving hemostasis. The coagulation cascade is an intricate pathway involving many proteins (Figure 1). Fibrin is formed when fibrinogen is cleaved by thrombin. Thrombin, in turn, first needs to be formed from prothrombin through cleavage by factor X. Factor X can be activated either through the intrinsic pathway by factor VIII or the extrinsic pathway by factor VII. Platelet adhesion and aggregation in turn depends mainly on von Willebrand factor (VWF) and fibrinogen (Figure 2). More recently, another protein called ADAMTS13 has been found to decrease the activity of VWF in platelet adhesion and aggregation.[9,10]

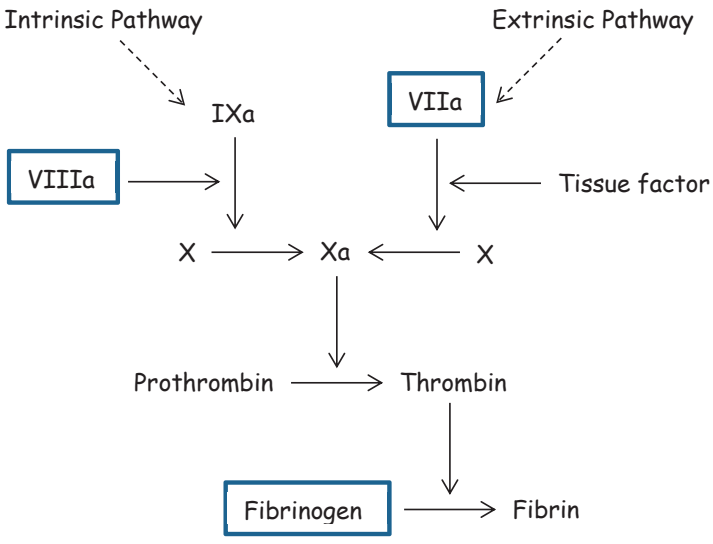


Figure 1. The coagulation cascade

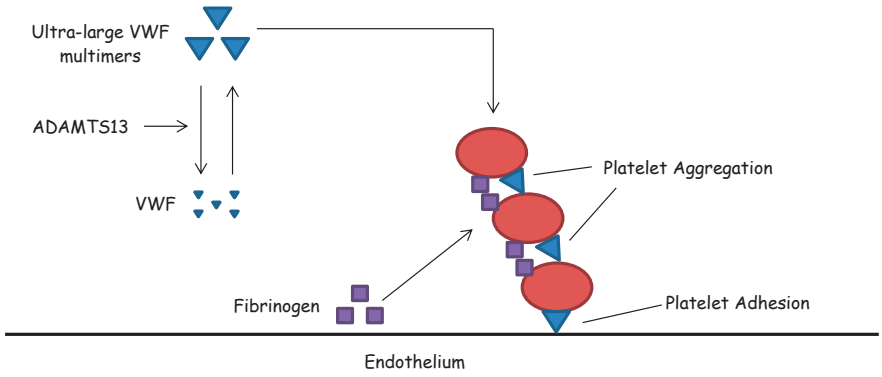


Figure 2. Platelet aggregation and adhesion.

CLINICAL IMPLICATIONS OF HEMOSTATIC FACTORS

Differences among individuals in the level and activity of the proteins involved in hemostasis partly determine differences in their ability to provoke clotting or stop bleeding. While abnormally low levels of many of these proteins can cause bleeding disorders such as von Willebrand's disease,[11-14] high levels may promote thrombosis and thereby contribute to cardiovascular events such as myocardial infarction and stroke.[15,16] Fibrinogen, VWF, factor VII, and factor VIII are all associated with an increased risk of incident coronary heart disease according to large population-based cohort studies.[17-22] On the other hand, ADAMTS13 levels and activity have been associated with a reduced risk of coronary heart disease in case-control studies.[23,24] Whether these associations reflect causation is unclear, partly because the levels of these proteins in the blood can change in response to a diverse set of factors. Fibrinogen, for example, is highly increased during the acute-phase response.[25] Thus, regardless of its essential function in hemostasis it is more closely correlated to inflammatory markers such as C-reactive protein as to other hemostatic factors. Another example is VWF, whose levels are higher in individuals with endothelial dysfunction.[26] In the case of ADAMTS13, there is evidence supporting its antithrombotic effect, but its association with other risk factors of CHD remains unexplored.

GENETICS OF HEMOSTATIC FACTORS

Over the past decade the standard approach to identify genetic variants that affect phenotypes has been large-scale genome-wide association (GWA) studies.[27] The strength of GWA studies lies in their hypothesis-free approach, interrogating mil-

lions of genetics variants rather than a select few. Different studies use different genotyping arrays to measure hundreds of thousands to a few million single nucleotide polymorphisms (SNPs) in their participants. The number of overlapping SNPs among different arrays is generally low, making it difficult to simply combine the results of several GWA studies. This challenge was overcome by using the correlation structure between SNPs to impute a set of 2.5 million SNPs, regardless of which SNPs were genotyped. A reference panel from which these correlations can be obtained was made available by the HapMap project.[28] GWA studies based on HapMap have discovered 23 genetic loci for fibrinogen,[29-31] 5 loci for factor VII, 5 loci for factor VIII, and 8 loci for VWF.[32] No previous GWA studies of ADAMTS13 have been performed, but several variants within the ADAMTS13 gene are known to affect ADAMTS13 levels and activity.[33]

GENETICS OF CHD

Similarly, the largest GWA study of CHD identified 46 susceptibility loci.[34] Furthermore, the authors of the study put forward a set of 152 variants independently associated with CHD at a false discovery rate of 5%. While for many phenotypes, such as hemostatic factors, the primary aim of performing a GWA study is to uncover new biology, for GWA studies of clinical outcomes an additional aim is to improve risk prediction. This is particularly relevant for CHD, as across the world risk prediction programs are implemented to identify individuals at a high risk of CHD so that preventive strategies can be initiated, including lifestyle interventions such as smoking cessation, and ultimately drug interventions with lipid-lowering, antihypertensive, or antithrombotic medication. Many studies have thus been performed testing whether genetic variants for CHD found through GWA studies improve risk prediction of incident CHD.[35-39] So far, these studies indicate that genetic variants are of little or no benefit to CHD risk prediction.

PROGRESS IN GENETIC EPIDEMIOLOGY

One limitation of HapMap-based GWA studies is that they only investigate common SNPs.[40] They do not cover low-frequency and rare variants, and they do not cover variants other than SNPs, such as large structural variants and small indels. The creation of improved reference panels is thus the first of several developments that are underway that could potentially transform the field of genetic epidemiology. These include population-specific reference panels such as Genomes of the Netherlands

and UK10K,[41] but also cosmopolitan reference panels such as the 1000 Genomes Project.[42] These reference panels are based on sequences of hundreds to thousands of individuals, and thus provide more information on rare variants than HapMap.

Second, new genotyping arrays have been designed that measure mainly non-synonymous variants in the protein-coding exonic regions of the genome.[43] While exonic regions comprise only a small percentage of the genome, these genotyping arrays are based on the assumption that variants within them have the highest potential for inducing phenotypic variation.

Third, rather than genotyping known variants, it is now feasible to sequence the exons or even the whole genome.[44] The main advantage is that sequencing also allows access to rare variants not covered in the reference panel, including population-specific variants. Additionally, even when low-frequency variants are accessible through imputation, they often have a low imputation quality. Effectively this is a type of measurement error that reduces the power to detect associations. This is not an issue with sequencing as all variants are directly measured.

Fourth, studies are increasingly measuring dynamic aspects of genomics, such as gene expression. While the amino acid sequence of a protein is encoded by genetic variants that do not change, gene expression levels are regulated by transcription factors, microRNAs, methylation, DNA accessibility, and other epigenetic factors. The levels of these factors, and hence gene expression levels, can change in response to the environment. Vitamin D, for example, is either obtained through the diet or produced in response to sun exposure. Vitamin D then activates Vitamin D receptor, a transcription factor that regulates the expression of over 200 genes.[45] Genetic variants can also affect gene expression levels, for example by affecting the level or activity of transcription factors or microRNAs.[46,47] Thus, besides measuring expression levels themselves, these interactions can also be captured by studying genetic variants known or suspected to affect gene expression levels.

AIM OF THIS THESIS

The aim of this thesis was to study hemostatic and genetic risk factors of cardiovascular disease. To improve our understanding of how hemostatic factors are related to cardiovascular disease we studied the genetic epidemiology of these factors using several novel approaches. For ADAMTS13 we also studied associations with cardiovascular risk factors, given that these associations remain largely unexplored for this new marker.

OUTLINE OF THIS THESIS

Chapter 2 focuses on genetic association studies of proteins involved in hemostasis. In Chapter 2.1 we perform a GWA study, based on 1000G imputation, of circulating fibrinogen concentration in over 120,000 individuals. To be able to adequately examine the benefit of using 1000G imputation over HapMap imputation, in Chapter 2.2 we perform a head to head comparison of these two methods using circulating fibrinogen concentration as an example phenotype. We then further examine the genetics of fibrinogen, but also factor VII, factor VIII, and VWF, using study designs especially suited for the identification of rare variants. In Chapter 2.3 we performed an exome-wide study using genotypes obtained from the Ilumina Exome Chip. In Chapter 2.4 we performed a similar study using exome sequencing. In Chapter 2.5 we combine the GWA study and exome chip approaches to study both common and rare genetic variants associated with ADAMTS13 activity.

In Chapter 3 we further characterize the novel hemostatic factor ADAMTS13 by examining its association with cardiovascular risk factors. In Chapter 3.1 we explored the association of ADAMTS13 activity with kidney function decline, and in Chapter 3.2 we examine the association of ADAMTS13 activity with incident type 2 diabetes.

In Chapter 4 we investigate coronary heart disease and the underlying atherosclerosis directly. In Chapter 4.1 we evaluate the incremental predictive value of genetic risk scores in the risk prediction of incident coronary heart disease. In Chapter 4.2 we systematically investigate the association of microRNA seed sequence variants with cardiovascular risk factors and disease. The seed sequence is the region of microRNAs that is used to bind to target genes. Genetic variants in the seed sequence of a microRNA can therefore lead to a loss or gain of target genes, and alter the expression of these genes. In Chapter 4.3 we perform a transcriptome-wide association study of carotid intima media thickness, aiming to identify genes that are differentially expressed in the presence of atherosclerosis.

Finally, in Chapter 5 we give an overview of the main findings of this thesis, examine the implications of the results, and discuss methodological issues that came to light.

REFERENCES

1. Lozano R, Naghavi M, Foreman K, et al. Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010. Lancet. 2012;380:2095-2128.
2. Kavousi M, Elias-Smale S, Rutten JH, et al. Evaluation of newer risk markers for coronary heart disease risk classification: a cohort study. Ann Intern Med. 2012;156:438-444.
3. Stein JH, Korcarz CE, Post WS. Use of carotid ultrasound to identify subclinical vascular disease and evaluate cardiovascular disease risk: summary and discussion of the American Society of Echocardiography consensus statement. Prev Cardiol. 2009;12:34-38.
4. Criqui MH, Denenberg JO, Ix JH, et al. Calcium density of coronary artery plaque and risk of incident cardiovascular events. JAMA. 2014;311:271-278.
5. Naghavi M, Libby P, Falk E, et al. From vulnerable plaque to vulnerable patient: a call for new definitions and risk assessment strategies: Part I. Circulation. 2003;108:1664-1672.
6. Frobert O, van't Veer M, Aarnoudse W, Simonsen U, Koolen JJ, Pijls NH. Acute myocardial infarction and underlying stenosis severity. Catheter Cardiovasc Interv. 2007;70:958-965.
7. White HD, Chew DP. Acute myocardial infarction. Lancet. 2008;372:570-584.
8. Antithrombotic Trialists Collaboration, Baigent C, Blackwell L, et al. Aspirin in the primary and secondary prevention of vascular disease: collaborative meta-analysis of individual participant data from randomised trials. Lancet. 2009;373:1849-1860.
9. Fujikawa K, Suzuki H, McMullen B, Chung D. Purification of human von Willebrand factor-cleaving protease and its identification as a new member of the metalloproteinase family. Blood. 2001;98:1662-1666.
10. Gerritsen HE, Robles R, Lammle B, Furlan M. Partial amino acid sequence of purified von Willebrand factor-cleaving protease. Blood. 2001;98:1654-1661.
11. Lenting PJ, Casari C, Christophe OD, Denis CV. von Willebrand factor: the old, the new and the unknown. J Thromb Haemost. 2012;10:2428-2437.
12. Lapecorella M, Mariani G, International Registry on Congenital Factor VIID. Factor VII deficiency: defining the clinical picture and optimizing therapeutic options. Haemophilia. 2008;14:1170-1175.
13. Verbruggen B, Meijer P, Novakova I, Van Heerde W. Diagnosis of factor VIII deficiency. Haemophilia. 2008;14 Suppl 3:76-82.
14. Peyvandi F. Epidemiology and treatment of congenital fibrinogen deficiency. Thromb Res. 2012;130 Suppl 2:S7-11.
15. Golec DB. Fibrinogen and thrombophilia. Clin Lab Sci. 2001;14:269-271.
16. Previtali E, Bucciarelli P, Passamonti SM, Martinelli I. Risk factors for venous and arterial thrombosis. Blood Transfus. 2011;9:120-138.
17. Emerging Risk Factors C, Kaptoge S, Di Angelantonio E, et al. C-reactive protein, fibrinogen, and cardiovascular disease prediction. N Engl J Med. 2012;367:1310-1320.
18. Willeit P, Thompson A, Aspelund T, et al. Hemostatic factors and risk of coronary heart disease in general populations: new prospective study and updated meta-analyses. PLoS One. 2013;8:e55175.
19. Folsom AR, Wu KK, Rosamond WD, Sharrett AR, Chambless LE. Prospective study of hemostatic factors and incidence of coronary heart disease: the Atherosclerosis Risk in Communities (ARIC) Study. Circulation. 1997;96:1102-1108.

20. Tzoulaki I, Murray GD, Lee AJ, Rumley A, Lowe GD, Fowkes FG. Relative value of inflammatory, hemostatic, and rheological factors for incident myocardial infarction and stroke: the Edinburgh Artery Study. Circulation. 2007;115:2119-2127.
21. Tracy RP, Arnold AM, Ettinger W, Fried L, Meilahn E, Savage P. The relationship of fibrinogen and factors VII and VIII to incident cardiovascular disease and death in the elderly: results from the cardiovascular health study. Arterioscler Thromb Vasc Biol. 1999;19:1776-1783.
22. Ruddock V, Meade TW. Factor-VII activity and ischaemic heart disease: fatal and non-fatal events. QJM. 1994;87:403-406.
23. Maino A, Siegerink B, Lotta LA, et al. Plasma ADAMTS-13 levels and the risk of myocardial infarction: an individual patient data meta-analysis. J Thromb Haemost. 2015;13:1396-1404.
24. Sonneveld MA, de Maat MP, Leebeek FW. Von Willebrand factor and ADAMTS13 in arterial thrombosis: a systematic review and meta-analysis. Blood Rev. 2014;28:167-178.
25. Davalos D, Akassoglou K. Fibrinogen as a key regulator of inflammation in disease. Semin Immunopathol. 2012;34:43-62.
26. Mannucci PM. von Willebrand factor: a marker of endothelial damage? Arterioscler Thromb Vasc Biol. 1998;18:1359-1362.
27. Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. Am J Hum Genet. 2012;90:7-24.
28. International HapMap Consortium. The International HapMap Project. Nature. 2003;426:789-796.
29. Danik JS, Pare G, Chasman DI, et al. Novel loci, including those related to Crohn disease, psoriasis, and inflammation, identified in a genome-wide association study of fibrinogen in 17 686 women: the Women's Genome Health Study. Circ Cardiovasc Genet. 2009;2:134-141.
30. Dehghan A, Yang Q, Peters A, et al. Association of novel genetic Loci with circulating fibrinogen levels: a genome-wide association study in 6 population-based cohorts. Circ Cardiovasc Genet. 2009;2:125-133.
31. Sabater-Lleal M, Huang J, Chasman D, et al. Multiethnic meta-analysis of genome-wide association studies in >100 000 subjects identifies 23 fibrinogen-associated Loci but no strong evidence of a causal association between circulating fibrinogen and cardiovascular disease. Circulation. 2013;128:1310-1324.
32. Smith NL, Chen MH, Dehghan A, et al. Novel associations of multiple genetic loci with plasma levels of factor VII, factor VIII, and von Willebrand factor: The CHARGE (Cohorts for Heart and Aging Research in Genome Epidemiology) Consortium. Circulation. 2010;121:1382-1392.
33. Tseng SC, Kimchi-Sarfaty C. SNPs in ADAMTS13. Pharmacogenomics. 2011;12:1147-1160.
34. CARDIoGRAMplusC4D Consortium, Deloukas P, Kanoni S, et al. Large-scale association analysis identifies new risk loci for coronary artery disease. Nat Genet. 2013;45:25-33.
35. Brautbar A, Pompeii LA, Dehghan A, et al. A genetic risk score based on direct associations with coronary heart disease improves coronary heart disease risk prediction in the Atherosclerosis Risk in Communities (ARIC), but not in the Rotterdam and Framingham Offspring, Studies. Atherosclerosis. 2012;223:421-426.
36. Hughes MF, Saarela O, Stritzke J, et al. Genetic markers enhance coronary risk prediction in men: the MORGAM prospective cohorts. PLoS One. 2012;7:e40922.
37. Paynter NP, Chasman DI, Pare G, et al. Association between a literature-based genetic risk score and cardiovascular events in women. JAMA. 2010;303:631-637.

38. Thanassoulis G, Peloso GM, Pencina MJ, et al. A genetic risk score is associated with incident cardiovascular disease and coronary artery calcium: the Framingham Heart Study. Circ Cardiovasc Genet. 2012;5:113-121.
39. Ganna A, Magnusson PK, Pedersen NL, et al. Multilocus genetic risk scores for coronary heart disease prediction. Arterioscler Thromb Vasc Biol. 2013;33:2267-2272.
40. Zuk O, Schaffner SF, Samocha K, et al. Searching for missing heritability: designing rare variant association studies. Proc Natl Acad Sci U S A. 2014;111:E455-464.
41. Genome of the Netherlands Consortium. Whole-genome sequence variation, population structure and demographic history of the Dutch population. Nat Genet. 2014;46:818-825.
42. 1000 Genomes Project Consortium, Abecasis GR, Auton A, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491:56-65.
43. Grove ML, Yu B, Cochran BJ, et al. Best practices and joint calling of the HumanExome BeadChip: the CHARGE Consortium. PLoS One. 2013;8:e68095.
44. Wang Q, Lu Q, Zhao H. A review of study designs and statistical methods for genomic epidemiology studies using next generation sequencing. Front Genet. 2015;6:149.
45. Ramagopalan SV, Heger A, Berlanga AJ, et al. A ChIP-seq defined genome-wide map of vitamin D receptor binding: associations with disease and evolution. Genome Res. 2010;20:1352-1360.
46. Westra HJ, Peters MJ, Esko T, et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. Nat Genet. 2013;45:1238-1243.
47. Huan T, Rong J, Liu C, et al. Genome-wide identification of microRNA expression quantitative trait loci. Nat Commun. 2015;6:6601.

Chapter 2

Genetic association studies of hemostatic factors

Chapter 2.1

Genome-wide association study of circulating fibrinogen concentration

Manuscript based on this chapter

Paul S. de Vries, Daniel I. Chasman, Maria Sabater-Lleal, Ming-Huei Chen, Jennifer E. Huffman, Maristella Steri, Weihong Tang, Alexander Teumer, Riccardo E. Marioni, Vera Grossmann, Jouke J. Hottenga, Stella Trompet, Martina Müller-Nurasyid, Jing Hua Zhao, Jennifer A. Brody, Marcus E. Kleber, Xiuqing Guo, Jie Jin Wang, Paul L. Auer, John R. Attia, Lisa R. Yanek, Tarunveer S. Ahluwalia, Jari Lahti, Cristina Venturini, Toshiko Tanaka, Lawrence F. Bielak, Peter K. Joshi, Ares Rocanin-Arjo, Ivana Kolcic, Pau Navarro, Lynda M. Rose, Christopher Oldmeadow, Helene Riess, Johanna Mazur, Saonli Basu, Anuj Goel, Qiong Yang, Mohsen Ghanbari, Gonneke Willemsen, Ann Rumley, Edoardo Fiorillo, Anton J. M. de Craen, Anne Grotevendt, Robert Scott, Kent D. Taylor, Graciela E. Delgado, Jie Yao, Annette Kifley, Charles Kooperberg, Rehan Qayyum, Lorna M. Lopez, Tina L. Berentzen, Katri Räikkönen, Massimo Mangino, Stefania Bandinelli, Patricia A. Peyser, Sarah Wild, David-Alexandre Trégouët, Alan F. Wright, Jonathan Marten, Tatijana Zemunik, Alanna C. Morrison, Bengt Sennblad, Geoffrey Tofler, Moniek P. M. de Maat, Eco J. C. de Geus, Gordon D. Lowe, Magdalena Zoledziewska, Naveed Sattar, Harald Binder, Uwe Völker, Melanie Waldenberger, Kay-Tee Khaw, Barbara McKnight, Jie Huang, Nancy S. Jenny, Elizabeth G. Holliday, Lihong Qi, Mark G. McEvoy, Diane M. Becker, John M. Starr, Antti-Pekka Sarin, Pirro G. Hysi, Dena G. Hernandez, Min A. Jhun, Harry Campbell, Anders Hamsten, Fernando Rivadeneira, Wendy L. McArdle, P. Eline Slagboom, Tanja Zeller, Wolfgang Koenig, Bruce M. Psaty, Talin Haritunians, Jingmin Liu, Aarno Palotie, André G. Uitterlinden, David J. Stott, Albert Hofman, Oscar H. Franco, Ozren Polasek, Igor Rudan, Pierre-Emmanuel Morange, James F. Wilson, Sharon L. R. Kardia, Luigi Ferrucci, Tim D. Spector, Johan G. Eriksson, Torben Hansen, Ian J. Deary, Lewis C. Becker, Rodney J. Scott, Paul Mitchell, Winfried März, Nick J. Wareham, Annette Peters, Andreas Greinacher, Philipp S. Wild[*], J. Wouter Jukema, Dorret I. Boomsma, Caroline Hayward, Francesco Cucca, Russell Tracy, Hugh Watkins, Alex P. Reiner, Aaron R. Folsom, Paul M. Ridker, Christopher J. O'Donnell, Nicholas L. Smith, David P. Strachan[*], and Abbas Dehghan[*].

[*]contributed equally to this study as senior authors.

A meta-analysis of 120,246 individuals identifies 18 new loci for fibrinogen concentration
Human Molecular Genetics. 2015; Epub ahead of print.

ABSTRACT

Background: Genome-wide association studies have previously identified 23 genetic loci associated with circulating fibrinogen concentration. These studies used HapMap imputation and did not examine the X chromosome. 1000 Genomes imputation provides better coverage of uncommon variants, and includes indels.

Methods: We conducted a genome-wide association analysis of 34 studies imputed to the 1000 Genomes Project reference panel and including ~120,000 participants of European ancestry (95,806 participants with data on the X chromosome). Approximately 10.7 million SNPs and 1.2 million indels were examined.

Results: We identified 41 genome-wide significant fibrinogen loci of which 18 were newly identified. There were no genome-wide significant signals on the X chromosome. The lead variants of 5 significant loci were indels. We further identified 6 additional independent signals, including 3 rare variants, at two previously characterized loci: *FGB* and *IRF1*.

Conclusions: The new loci emphasize the importance of STAT3 to fibrinogen regulation, and highlight new inflammatory pathways.

INTRODUCTION

Fibrinogen is a coagulation factor crucial to clot formation, and an active regulator of the inflammatory response.[1] It is a strong and established predictor of cardiovascular disease, autoimmune disorders, and cancer.[1-5] Circulating fibrinogen concentration has a moderate heritability of 34% to 46%.[6-8] Previous genome-wide association studies (GWAS) have highlighted genetic loci involved in inflammatory pathways such as the acute-phase response and interleukin 1 and 6 signaling as main determinants of fibrinogen concentration.[9-13]

The variance in fibrinogen concentration explained by genetic loci identified in these previous GWAS is less than one tenth of its estimated heritability.[11] It is therefore likely that part of the heritability stems from genetic variants that are not well tagged by the single nucleotide polymorphisms (SNPs) found in HapMap, including further common, uncommon, and rare SNPs, and other types of variants such as insertions or deletions (indels). Additionally, part of the heritability could be explained by variants on the X chromosome, which has not previously been interrogated.

To better interrogate the full range of genetic variants, including those with low minor allele frequency that may have been poorly tagged by HapMap variants, we performed a meta-analysis of 34 GWAS imputed using 1000 Genomes Project reference panels,[14] including the X chromosome. We performed a joint/conditional analysis to identify additional independent signals within known and new loci associated with plasma fibrinogen concentration.

METHODS

Study sample

This meta-analysis was conducted within the framework of the Cohorts for Heart and Aging Research in Genetic Epidemiology (CHARGE) consortium.[15] The study sample consists of 34 studies with 120,246 individuals of European ancestry. 12 studies with 25,453 participants were not included in the previous fibrinogen GWAS.[11] Fibrinogen concentration was measured in citrated or EDTA plasma samples using a variety of methods including the Clauss method, immunonephelometric methods, immunoturbidimetric methods, and prothrombin time derived methods as described in **Supplemental Table 1** and the **Supplemental Methods**, which further describe the studies. All studies were approved by appropriate research ethics committees and all respondents signed informed consent prior to participation.

Genotyping and imputation

Genotyping, pre-imputation quality control, imputation, and analysis methods are presented in **Supplemental Table 2**. All studies imputed variant dosages using reference panels from the 1000 Genomes Project using MACH or IMPUTE.[14,16-18] The phase I version 3 reference panel was used by all studies except two, which used the phase I version 2 reference panel. Before meta-analysis, we excluded variants with MACH imputation quality < 0.3 or IMPUTE imputation quality < 0.4, and variants with effective minor allele count (minor allele count × imputation quality) < 10. These filters were applied at the level of individual studies. Because we wanted to focus only on those variants that passed these filters in a large proportion of the studies, we additionally excluded variants with a total sample size of less than half of the maximum sample size at the meta-analysis level.

Autosomal association analysis

Plasma fibrinogen concentration was converted to g/L and natural-log transformed. All studies adjusted for age and sex. When necessary, analyses were also adjusted for study-specific covariates, such as center or case/control status. In family studies, linear mixed models were used to account for family structure. Analyses were adjusted for principal components to account for population structure and cryptic relatedness. These adjustments are shown in **Supplemental Table 2**. To account for remaining stratification, we applied a genomic control correction to the results of each of the studies before meta-analysis. We used an inverse-variance model with fixed effects implemented in METAL to meta-analyze association results.[19] Heterogeneity was assessed using I^2 and corresponding *P*-values.

As proposed by Huang et al, variants with *P*-values lower than 2.5×10^{-8} were considered genome-wide significant (based on a Bonferroni correction for 2,000,000 tests).[20] Significant variants were assigned to loci in order of ascending *P*-value. A variant was assigned to a new locus when there were no significant variants within 500 kb of it belonging to a previously defined locus. Variants were annotated to genes using ANNOVAR version 2013Mar07.[21]

X-chromosome association analysis

Of the 120,246 participants, 95,806 had imputed data on the X chromosome. Dosages of variants on the X chromosome were coded as [0,2] in men and [0,1,2] in women. This way one allele in men has the same value as two alleles in women. Thus, we assume full inactivation of one of the two X chromosomes in women. Variants in the pseudo-autosomal region were excluded. Analyses of the X chromosome were stratified by sex in each study, and the studies then were meta-analyzed separately for men and women using an inverse-variance model with fixed effects.[19] We then

combined the sex-specific meta-analysis results for variants on the X chromosome using both an inverse variance weighted model with fixed effects and a sample-size weighted model based on *P*-values and effect direction. The sample-size weighted model does not take the effect size into account, and thus may work better when there are different effects in men and women,[22,23] as can happen when there is incomplete inactivation in women.

Conditional analysis

Some loci may harbor multiple independent variants that affect fibrinogen.[11,24] To putatively identify these jointly significant variants, we used an approximate method for conditional and joint analysis using meta-analysis summary statistics implemented in GCTA.[25,26] The method consists of a genome-wide stepwise selection procedure selecting variants according to their conditional *P*-values and, after the model has been optimized, the estimation of the joint effects of the selected variants. This method depends on a reference panel to estimate linkage disequilibrium patterns between variants. We used best-guess imputation for variants with imputation quality > 0.3 in 5,733 unrelated individuals from the Rotterdam Study as the reference panel.[27] A description of the Rotterdam Study is given in the **Supplemental Methods**.

Functional annotation

For each locus, we searched the National Human Genome Research Institute GWAS catalog for genome-wide significant associations with other traits within 100kb of the lead variant.[28] We used the Blood eQTL browser, a publicly available database, to examine whether any lead variants, or their most correlated HapMap proxy (with $R^2 > 0.8$), were associated with expression levels of nearby genes in blood. Results from the blood eQTL browser are based on non-transformed peripheral blood samples from 5,311 individuals with replication in 2,775 individuals.[29] For each lead SNP and its highly correlated neighbors (with $R^2 > 0.9$), we used HaploReg V2 to determine the level of conservation, association with gene expression in a range of tissues including the liver, and any overlap with ENCODE transcription factor binding sites, and DNAse-hypersensitive, promoter , and enhancer regions in various cell types.[30,31] Furthermore, we determined the overlap of these SNPs with microRNAs and microRNA binding sites (see **Supplemental Methods**).[32-34]

Variance explained

In the Women's Genome Health Study, the largest contributor to the meta-analysis, we computed a weighted genetic risk score based on the lead variants at each genome-wide significant locus, as well as any jointly significant variants identified in the conditional analysis.[35] A description of the Women's Genome Health Study is

given in the **Supplemental Methods**. Beta coefficients from the genome-wide association meta-analysis including all studies were used as weights, except in loci with multiple jointly significant variants. For variants at these loci, joint beta coefficients were obtained from the conditional analysis. The genetic risk score was computed as the sum of the weighted variants dosages. The variance in fibrinogen concentration explained was estimated using a linear regression model. Additionally, for any loci with jointly significant variants we compared the variance explained by the lead variant to the variance explained by the jointly significant variants. We were not able to directly compare our estimate of the variance explained to previous estimates, as these had been computed in different populations and were adjusted for age and sex. Thus, we re-calculated the variance explained without adjustment for age and sex. For this we used HapMap-imputed dosages of the independently associated SNPs reported by Sabater-Lleal et al.[11] Since the variance explained is estimated on the basis of imperfectly imputed dosages, we expect our estimates to be slightly lower than if they were based on measured genotypes.

RESULTS

Autosomal meta-analysis

Participant characteristics in each study are shown in **Supplemental Table 1**, covariates adjusted for by each study are shown in **Supplemental Table 2**, and genomic inflation factors are shown in **Supplemental Table 3**. The meta-analysis of the autosomes included 9,492,263 SNPs and 841,128 indels, of which 4,354 SNPs and 420 indels at 41 loci were genome-wide significant. Of these, 18 loci are new signals (**Table 1**), while 23 have been associated with fibrinogen concentration by previous GWAS (**Table 2**). Among genome-wide significant variants, 14 of 4,354 were rare (MAF ≤ 0.01), and a further 477 were uncommon (0.01 < MAF ≤ 0.05). The lead variants of known locus *SNX13*, and novel loci *ATXN2L*, *GYS2*, *GIMAP4*, and *IFT122* were indels. Separate QQ plots of all autosomal variants, common variants, uncommon variants, rare variants, SNPs, and indels are shown in **Supplemental Figure 1**. A Manhattan plot of all autosomal variants is shown in **Supplemental Figure 2**. Additionally, a Manhattan plot highlighting rare and uncommon variants is shown in **Supplemental Figure 3**. Heterogeneity I^2 and *P*-values are shown in **Supplemental Table 4**. Only rs7439150 at the fibrinogen gene cluster showed significant heterogeneity (I^2: 50.0, *P*-value: 0.0004). Regional plots are shown in **Supplemental Figure 4**, and forest plots are shown in **Supplemental Figure 5**. Associations with rare variants were found at the two most robust fibrinogen loci: the fibrinogen gene cluster and the *IRF1* locus (lead variant annotated to *C5orf56*). Associations with uncommon variants were also

Table 1. Association of the lead variants at 18 newly identified loci with natural-log transformed plasma fibrinogen concentration (g/L).

Locus	Variant	Position	Closest Gene	eQTL	NSYN variants	A1/A2	Frequency	β	*P*-value
2p25.3	rs7588285	3648186	*COLEC11*			C/G	0.20	0.0074	1.2×10^{-08}
3p25.3	rs62246343	9543642	*LHFPL4*			T/C	0.17	0.0071	2.2×10^{-08}
3q21.1	rs1976714	122864771	*PDIA5*			T/G	0.35	-0.0055	2.3×10^{-08}
3q21.3	3:129228166	129228166	*IFT122*	*RPL32P3*		D/R	0.10	0.009	1.0×10^{-08}
7p14.2	rs2710804	36084529	*EEPD1*			C/T	0.37	0.0055	2.9×10^{-09}
7q36.1	7:150289652	150289652	*GIMAP4*	*GIMAP4*		D/R	0.21	-0.0073	9.3×10^{-11}
8p23.1	rs7012814	9173358	*LOC157273*			A/G	0.47	0.0060	2.1×10^{-10}
9q22.2	rs3138493	92219260	*GADD45G*	*SEMA4D*		T/C	0.48	-0.0054	2.5×10^{-09}
10q23.31	rs2250644	91008879	*LIPA*			T/C	0.33	0.0054	2.2×10^{-08}
10q26.13	rs2420915	122840277	*MIR5694*	*WDR11*		A/G	0.09	-0.0094	5.2×10^{-09}
11p12	rs7934094	43505707	*TTC17*			G/T	0.22	-0.0083	2.5×10^{-13}
12p12.1	12:21703935	21703935	*GYS2*			R/D	0.37	0.0062	8.4×10^{-09}
12q24.12	rs7310615	111865049	*SH2B3*	*SH2B3*	*SH2B3*	C/G	0.50	-0.0069	1.5×10^{-13}
15q15.1	rs56702977	42671308	*CAPN3*	*ZFP106*		A/G	0.13	0.0080	2.1×10^{-09}
16p11.2	16:28845027	28845027	*ATXN2L*	*TUFM*		D/R	0.39	0.0061	7.7×10^{-10}
16q22.2	rs1035560	72032730	*PKD1L3*	*HP*		C/T	0.40	0.0064	2.6×10^{-12}
17q21.2	rs7224737	40289364	*RAB5C*	*STAT3*	*HSPB9*	A/G	0.24	0.0061	6.1×10^{-09}
19q13.33	rs73058052	50099422	*PRR12*	*IRF3*	*PRRG2*	T/C	0.16	0.0074	2.0×10^{-08}

Abbreviations: eQTL indicates the gene with the strongest significant association between its expression levels in blood and the lead variant or its proxy. NSYN variants indicates genes containing nonsynonymous variant correlated to the lead variant ($R^2 > 0.9$). A1 indicates the coded allele. A2 indicates the other allele. Frequency is the frequency of the coded allele. β indicates the β coefficient adjusted for age, sex, population structure, and study-specific covariates, such as center or case/control status. The β coefficient can be interpreted as the ln(g/L) change in fibrinogen per 1 unit change in the dosage of the coded allele.

Table 2. Association of the lead variants at 23 known loci with natural-log transformed plasma fibrinogen concentration (g/L).

Locus	Variant	Position	Closest Gene	eQTL	NSYN variants	A1/A2	Frequency	β	*P*-value
1p31.3	rs1892534	66105944	*LEPR*			T/C	0.38	-0.0073	4.3×10^{-15}
1q21.3	rs61812598	154420087	*IL6R*		*IL6R*	A/G	0.39	-0.0115	2.7×10^{-36}
1q44	rs10157379	247605599	*NLRP3*	*NLRP3*		C/T	0.38	-0.0103	6.3×10^{-29}
2q12	rs1558643	102731691	*IL1R1*			T/C	0.40	0.0058	3.1×10^{-10}
2q13	rs6734238	113841030	*IL1F10*	*IL1RN*		G/A	0.41	0.0106	6.7×10^{-30}
2q34	rs715	211543055	*CPS1*		*CPS1*	C/T	0.32	-0.0082	4.3×10^{-16}
2q37.3	rs59104589	242237902	*HDLBP*	*STK25*		T/C	0.34	-0.0083	8.2×10^{-19}
3q22.2	rs9840812	135843162	*PPP2R3A*	*PCCB*		C/T	0.23	0.0117	1.7×10^{-27}
4p16.3	rs59950280	3452345	*HGFAC*			A/G	0.34	0.0075	1.7×10^{-12}
4q31.3	rs7439150	155481541	*FGB*		*FBG*	A/G	0.20	0.0313	9.5×10^{-181}
5q31.1	rs2057655	131807624	*C5orf56*	*SLC22A4*		A/G	0.21	-0.0203	1.8×10^{-73}
7p21.1	7:17904452	17904452	*SNX13*			R/D	0.48	0.0067	1.3×10^{-13}
7p15.3	rs71520386	22853521	*TOMM7*			T/C	0.20	0.0066	5.1×10^{-09}
8q24.3	rs11780978	145034852	*PLEC*	*GRINA*		A/G	0.40	0.0059	5.5×10^{-10}
10q21.3	rs7916868	64988931	*JMJD1C*			A/T	0.49	0.0089	1.6×10^{-22}
11q12.2	rs11230201	59996994	*MS4A6A*	*MS4A6A*		G/C	0.41	-0.0057	4.5×10^{-10}
12q13.12	rs2731439	51060350	*DIP2B*	*DIP2B*		T/C	0.36	-0.0064	8.7×10^{-12}
14q24.1	rs367677	69273090	*ZFP36L1*			G/A	0.22	0.0077	1.8×10^{-12}
15q21.2	rs12913259	51014716	*SPPL2A*			T/C	0.30	-0.0068	2.3×10^{-12}
16q12.2	rs11859517	53181247	*CHD9*			T/C	0.29	-0.0074	8.9×10^{-14}
20q13.12	rs1800961	43042364	*HNF4A*		*HNF4A*	T/C	0.03	-0.0170	1.2×10^{-10}
21q22.2	rs9808651	40466468	*PSMG1*			A/G	0.27	-0.0095	2.5×10^{-20}
22q13.33	rs75347843	51112361	*SHANK3*	*ARSA*		A/G	0.19	0.0084	1.8×10^{-10}

Abbreviations: eQTL indicates the gene with the strongest significant association between its expression levels in blood and the lead variant or its proxy. NSYN variants indicates genes containing nonsynonymous variant correlated to the lead variant ($R^2 > 0.9$). A1 indicates the coded allele. A2 indicates the other allele. Frequency is the frequency of the coded allele. β indicates the β coefficient adjusted for age, sex, population structure, and study-specific covariates, such as center or case/control status. The β coefficient can be interpreted as the ln(g/L) change in fibrinogen per 1 unit change in the dosage of the coded allele.

Table 3. Joint/conditional association of 8 variants at 2 loci with natural-log transformed plasma fibrinogen concentration (g/L).

Locus	Variant	Position	Closest Gene	Annotation	A1/A2	Frequency	β	*P*-value	Joint β	Joint *P*-value
4q31.3	rs7439150	155481541	*FGB*	intergenic	A/G	0.205	0.0313	9.5×10^{-181}	0.0259	1.9×10^{-92}
4q31.3	rs150768229	155488301	*FGB*	intronic	C/A	0.009	-0.0458	6.4×10^{-12}	-0.0385	9.3×10^{-09}
4q31.3	rs6054	155489608	*FGB*	NSYN	T/C	0.005	-0.1228	2.4×10^{-53}	-0.1222	4.9×10^{-52}
4q31.3	rs148685782	155533035	*FGG*	NSYN	C/G	0.005	-0.2239	1.2×10^{-87}	-0.2179	4.0×10^{-82}
4q31.3	rs76289367	155546159	*FGG*	intergenic	G/T	0.148	0.0263	2.0×10^{-76}	0.0109	1.6×10^{-11}
5q31.1	rs12777	131671662	*SLC22A4*	SYN	G/C	0.044	0.0240	9.3×10^{-27}	0.0207	6.9×10^{-21}
5q31.1	5:131786964	131786964	*C5orf56*	ncRNA	I/R	0.015	-0.0543	2.5×10^{-14}	-0.0428	2.0×10^{-09}
5q31.1	rs2057655	131807624	*C5orf56*	ncRNA	A/G	0.207	-0.0203	1.8×10^{-73}	-0.0188	1.9×10^{-64}

Abbreviations: A1 indicates the coded allele. A2 indicates the other allele. Frequency is the frequency of the coded allele. NSYN indicates a nonsynonymous exonic variant. SYN indicates a synonymous exonic variant. β indicates the β coefficient adjusted for age, sex, population structure, and study-specific covariates, such as center or case/control status. Joint β indicates the β coefficient of the jointly significant variants, adjusted for the above and for each other. All β coefficients can be interpreted as the ln(g/L) change in fibrinogen per 1 unit change in the dosage of the coded allele.

found at these loci, as well as at *SPPL2A* and *HNF4A*. At one known locus (*SNX13*) and four new loci (*IFT122*, *GIMAP4*, *GYS2*, and *ATXN2L*) the lead variant was an indel. At each of these loci there were also SNPs in linkage disequilibrium with the indel that reached genome-wide significance. *CD300LF* was the only previously identified locus that was not represented among our significant results. The previously reported lead variant in *CD300LF*, rs10512597 (*P*-value: 1.8×10^{-7}), had a smaller effect size (β: -0.006 ln(g/L)) than was previously reported (β: -0.008 ln(g/L)). There was no strong evidence of heterogeneity (I^2: 22.7, *P*-value: 0.11).

Conditional analysis

Two loci (fibrinogen gene cluster and *IRF1*) harbored multiple jointly significant variants (**Table 3**). Forest plots of the additional variants discovered through conditional analysis are shown in **Supplemental Figure 6**, and their heterogeneity I^2 and *P*-values are shown in **Supplemental Table 4**. At the fibrinogen gene cluster, five variants were jointly significant: the lead variant rs7439150, an additional common variant rs76289367, and three rare variants, rs150768229, rs6054, and rs148685782. rs148685782 showed significant heterogeneity (I^2 = 65.0, *P*-value = 0.0004). At the *IRF1* locus three variants were jointly significant: the lead variant, rs2057655, and two uncommon variants, rs12777 and 5:131786964. Of the secondary signals, rs12777 is in strong linkage disequilibrium with a previously associated SNP, rs1242111 (R^2=0.8), while 5:131786964 is a new independent signal (R^2 = 0.0). The uncommon variants near *SPPL2A* were not significant in the conditional analysis. The uncommon lead variant rs141272690 was only marginally significant in the primary analysis (*P*-value = 1.89×10^{-8}), so that even a small correlation with the lead common variant rs12913259 (R^2 = 0.02) raised the *P*-value above the threshold in the conditional analysis.

X-chromosome meta-analysis

The meta-analysis of the X chromosome included 251,747 SNPs and 26,448 indels. There were no genome-wide significant variants detected on the X chromosome. This was true in both sex-specific meta-analyses, and in the combined meta-analyses, irrespective of whether the sex-specific results were combined using inverse-variance weighted meta-analysis or sample size based meta-analyses. QQ plots and Manhattan plots for the X chromosome are shown in **Supplemental Figure 7 and 8**.

Functional annotation

Genome-wide significant associations with other traits were found for 28 out of the 41 loci, of which 10 were associated with cholesterol levels, 7 were associated with C-reactive protein, and 5 were associated with platelet count (**Supplemental Table 5**). Out of the 41 lead variants, 20 were associated with blood expression levels of one

or more neighboring genes (**Supplemental Table 6**). Notably, rs1035559 at 16q22.2 was exclusively associated with *HP* expression levels ($P = 9.8\times10^{-198}$), and rs7224737 at 17q21.2 was exclusively associated with *STAT3* expression levels ($P = 5.4\times10^{-12}$). Out of the 41 lead variants, 36 were available in HaploReg V2. Detailed annotation of these variants as well as 457 correlated SNPs is shown in **Supplemental Table 7**. Eight of these SNPs are predicted to influence the binding of miRNAs to transcripts of their host gene. Further information about these SNPs and their effect on miRNA binding is shown in **Supplemental Table 8**. Of these eight SNPs, two were lead variants. First, the fibrinogen decreasing minor allele of lead variant rs715 in the 3′-UTR of *CPS1* is predicted to create a miRNA binding site for miR-3154. Second, the fibrinogen increasing minor allele of lead variant rs6224634 in the 3′-UTR of *LHFPL4* is predicted to disrupt the binding site of miR-6761-3p. In both cases predicted successful miRNA-target gene binding is associated with lower fibrinogen concentration.

Variance explained

In the Women's Genome Health Study, the lead variant at the fibrinogen gene cluster explained 0.8% of the variance, and all five jointly significant variants together explained 1.6% of the variance. At 5q31.1 the lead variant explained 0.2% of the variance, while all three jointly significant variants together explained 0.3% of the variance. The 47 independently significant variants at 41 loci explained 3.0% of the variance in circulating fibrinogen concentration. The variance explained by the 23 previously identified loci was 2.6%.

DISCUSSION

We identified 18 new autosomal loci associated with circulating fibrinogen concentration in individuals of European ancestry, increasing the variance explained from 2.6% to 3.0%. The small increase in the variance explained relative to the large number of new loci is suggestive of a highly polygenic genetic architecture. At two loci (fibrinogen gene cluster and *IRF1* locus) rare or uncommon variants were jointly significant alongside common lead variants. In five cases the lead variant at an associated locus was an indel. There were no significant associations on the X chromosome: this may be result of issues specific to the X chromosome rather than the absence of relevant signals. The most important issue is that the X chromosome is generally poorly covered by genotyping arrays.[36]

Four of the 18 new loci implicate inflammatory pathways not previously linked to fibrinogen. First, the septin gene family is represented at two significant loci: *SEPT7* at 7p14.2 and *SEPT2* at 2q37.3. Proteins from the septin gene family form cage-like struc-

tures around bacteria to facilitate autophagy.[37] The link between these processes and fibrinogen concentration is unclear. Second, our results also implicate genes from the GIMAP family, which are structurally similar to septins.[38] The signal at 7q36.1 appears to be driven by one or more genes from a cluster of eight GIMAP genes, and the lead variant is associated with blood expression levels of four of these. Through their involvement in lymphocyte maturation, these genes influence lymphocyte counts and diversity, and thereby also the inflammatory response.[39] Finally, the lead variant at 16q22.2 is strongly associated with blood expression levels of the neighboring *HP* (*P*-value ≤ 9.8×10^{-198}), the gene encoding haptoglobin. Like fibrinogen, haptoglobin is an acute-phase reactant. The association of rs1035560 with fibrinogen suggests that besides sharing upstream regulators, haptoglobin itself may be involved in the regulation of circulating fibrinogen.

Six of the new loci appear to be closely related to STAT3, a transcription factor working downstream of IL-6 that upregulates the expression of fibrinogen and other acute-phase proteins.[40] At 17q21.2, lead variant rs7224737 (175 kb from *STAT3*) was associated with *STAT3* blood expression levels ($P = 5.4\times10^{-12}$). At 9q22.2, the lead variant rs3138493 lies upstream of *GADD45G*. This gene is expressed in the liver, where it has been shown to inhibit the Tyr705 phosphorylation of STAT3.[41] As Tyr705 phosphorylation of STAT3 allows it to dimerize and move into the nucleus, it is essential for the upregulation of STAT3 targets like the fibrinogen genes. At 10q26.13, the lead variant rs2420915 is an intergenic SNP close to *FGFR2*. Over-expression of *FGFR2*, or the related *FGFR1* is required for the Tyr705 phosphorylation of STAT3.[41] At 19q13.33, the lead variant rs73058052 is associated with blood expression levels of *IRF3*. After activation in response to viral infection, *IRF3* enables the expression of type I interferons *INFA* and *INFB*, leading to the upregulation of *STAT3*.[42,43] Furthermore, our results point towards two SH2B adaptor proteins implicated in STAT3 signaling. At 12q24.12, the lead variant rs7310615 was associated with blood expression levels of *SH2B3*. Using immortalized B lymphoblastoid cell lines, a loss of the SH2B3 protein was accompanied by increased STAT3 phosphorylation.[44] At 16p11.2, lead variant 16:28845027 lies close to *SH2B1*. The β variant of SH2B1 appears to form a complex with STAT3, allowing STAT3 to cross through the membrane into the nucleus as an alternative to STAT3 dimerization.[45] Collectively, these findings suggest that a wide range of disturbances to STAT3 may affect circulating fibrinogen concentration.

In addition to STAT3, our results highlight HNF4A, another transcription factor known to regulate fibrinogen gene expression. The association between lead variant rs1800961 and circulating fibrinogen has been previously been described by Wassel et al and Hufman et al.[12,46] rs1800961 is a nonsynonymous coding variant that has been shown to decrease *HNF4A* expression in vitro.[47]

The majority of rare and uncommon variants associated with fibrinogen concentration were found at loci with common variant signals. Only the signal at *HNF4A* was led by an uncommon variant, and no signals were led by rare variants. Conditional analysis suggests that there are two secondary signals at the *IRF1* locus led by uncommon variants, and three secondary signals near the fibrinogen gene cluster led by rare variants. The uncommon variants that were significant near *SPPL2A* were not significant in the conditional analysis, but the linkage disequilibrium with the lead common variant was very low. Our results suggest that common and rare variant signals are often independent of each other, and do not support the hypothesis that associations with common variants are synthetic associations merely reflecting linkage disequilibrium with rare variants.[48,49]

Absolute effect sizes of significant variants ranged from 0.005 to 0.033 ln(g/L) among common variants, 0.013 to 0.087 ln(g/L) among uncommon variants, and 0.036 to 0.254 ln(g/L) among rare variants. Despite their small effect size, common variants have helped discover biologically relevant fibrinogen loci. Therefore, the complete lack of overlap between the effect sizes of significant common and rare variants suggests that further rare variants with smaller effect sizes are likely to exist at important and possibly unknown fibrinogen loci. While the rare variants with large effects we found were limited to the two most important fibrinogen loci, rare variants with moderate effects may be more widespread.

When considering not only the primary signal at the fibrinogen gene cluster, but also the four additional signals the variance explained by the locus doubles from 0.8% to 1.6%. Two of these additional signals are driven by rare non-synonymous exonic variants (rs6054 and rs148685782) with very large effect sizes (β=-0.12 and β =-0.21 ln(g/L) respectively). The association between rs6054 and fibrinogen has been described earlier in a candidate gene stud,[12] and rs148685782 (also known as γAla82Gly) has previously been reported as a causal variant for mild congenital hypofibrinogenaemia.[50-52] Furthermore, in a previous study we examined exome-wide genotypes using exome arrays and identified independent associations of both rs6054 and rs148685782 with fibrinogen.[46] In the present study, however, two further variants, rs140473879 and rs149234484, are in strong linkage disequilibrium with rs148685782 and tag this signal. These variants are intergenic, but each changes several regulatory motifs. Thus, the identification of rs148685782 as a causal variant is not conclusive.

Strengths of this study include the use of a large ethnically homogenous sample, and coverage of previously unexamined uncommon and rare variants, indels, and variants on the X chromosome. At the same time, the lack of ethnic heterogeneity may also be a limitation, as including different ethnicities can help narrow down the association signal to a smaller region.[53] This study has other limitations that should

be acknowledged. To most effectively use the available data, we used all 34 studies in the discovery sample.[54] The results have thus not been replicated. Nevertheless, the consistent association of these loci across the 34 studies and the strict Bonferroni correction enforcing a 5% false discovery rate ensure that essentially all of the loci represent true associations. A second limitation is that an approximation based on meta-analysis summary data was used to identify additional independently associated variants at the identified loci rather than a stepwise conditional analysis using individual-level data. Different methods were used to measure plasma fibrinogen across the studies: EDTA or citrate plasma samples were used, and a variety of assays were used.[55] While the association between fibrinogen and cardiovascular disease has previously been shown to be independent of assay type, the genetic etiology of fibrinogen may differ across assay types.[56] However, to minimize the impact on our results, studies that used multiple assays to measure fibrinogen performed their analyses stratified by the assay.

Finally, our ability to attribute these signals to causal genes remains limited. For each locus we reported the gene closest to the lead variant, but proximity alone is not strong evidence that a gene is the underlying causal gene. Thus, we also reported the genes whose expression levels in blood were most strongly associated with the lead variant, and we reported genes with nonsynonymous exonic variants in high linkage disequilibrium with the lead variant. Based on blood expression levels, some signals were characterized by a single promising candidate causal gene, but other signals were associated with either no candidate causal genes, or more than one. Furthermore, genetic variants can have effects on the expression of multiple genes across different tissues, and these effects can be tissue specific.

We identified 41 loci that collectively explain 3% of the variance in plasma fibrinogen concentration. Of these loci, 18 had not been identified previously through GWAS. The new loci emphasize the importance of STAT3 to fibrinogen regulation, and highlight several new potential pathways that should be experimentally confirmed. The use of 1000 Genomes Project imputation increased our ability to assess the role of uncommon variants, resulting in an in depth characterization of the two most important fibrinogen loci.

Supplement available online at:
http://hmg.oxfordjournals.org/

REFERENCES

1. Davalos D, Akassoglou K. Fibrinogen as a key regulator of inflammation in disease. Semin Immunopathol. 2012;34:43-62.
2. Seebacher V, Polterauer S, Grimm C, et al. The prognostic value of plasma fibrinogen levels in patients with endometrial cancer: a multi-centre trial. Br J Cancer. 2010;102:952-956.
3. Yapijakis C, Bramos A, Nixon AM, Ragos V, Vairaktaris E. The interplay between hemostasis and malignancy: the oral cancer paradigm. Anticancer Res. 2012;32:1791-1800.
4. Emerging Risk Factors Collaboration, Kaptoge S, Di Angelantonio E, et al. C-reactive protein, fibrinogen, and cardiovascular disease prediction. N Engl J Med. 2012;367:1310-1320.
5. Fibrinogen Studies Collaboration, Danesh J, Lewington S, et al. Plasma fibrinogen level and the risk of major cardiovascular diseases and nonvascular mortality: an individual participant meta-analysis. JAMA. 2005;294:1799-1809.
6. de Lange M, Snieder H, Ariens RA, Spector TD, Grant PJ. The genetics of haemostasis: a twin study. Lancet. 2001;357:101-105.
7. Souto JC, Almasy L, Borrell M, et al. Genetic determinants of hemostasis phenotypes in Spanish families. Circulation. 2000;101:1546-1551.
8. Neijts M, van Dongen J, Kluft C, Boomsma DI, Willemsen G, de Geus EJ. Genetic architecture of the pro-inflammatory state in an extended twin-family design. Twin Res Hum Genet. 2013;16:931-940.
9. Danik JS, Pare G, Chasman DI, et al. Novel loci, including those related to Crohn disease, psoriasis, and inflammation, identified in a genome-wide association study of fibrinogen in 17 686 women: the Women's Genome Health Study. Circ Cardiovasc Genet. 2009;2:134-141.
10. Dehghan A, Yang Q, Peters A, et al. Association of novel genetic Loci with circulating fibrinogen levels: a genome-wide association study in 6 population-based cohorts. Circ Cardiovasc Genet. 2009;2:125-133.
11. Sabater-Lleal M, Huang J, Chasman D, et al. Multiethnic meta-analysis of genome-wide association studies in >100 000 subjects identifies 23 fibrinogen-associated Loci but no strong evidence of a causal association between circulating fibrinogen and cardiovascular disease. Circulation. 2013;128:1310-1324.
12. Wassel CL, Lange LA, Keating BJ, et al. Association of genomic loci from a cardiovascular gene SNP array with fibrinogen levels in European Americans and African-Americans from six cohort studies: the Candidate Gene Association Resource (CARe). Blood. 2011;117:268-275.
13. Baumert J, Huang J, McKnight B, et al. No evidence for genome-wide interactions on plasma fibrinogen by smoking, alcohol consumption and body mass index: results from meta-analyses of 80,607 subjects. PLoS One. 2014;9:e111156.
14. 1000 Genomes Project Consortium, Abecasis GR, Auton A, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491:56-65.
15. Psaty BM, O'Donnell CJ, Gudnason V, 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. 2009;2:73-80.
16. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. 2009;5:e1000529.
17. Li Y, Willer C, Sanna S, Abecasis G. Genotype imputation. Annu Rev Genomics Hum Genet. 2009;10:387-406.
18. Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet Epidemiol. 2010;34:816-834.

19. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics. 2010;26:2190-2191.
20. Huang J, Ellinghaus D, Franke A, Howie B, Li Y. 1000 Genomes-based imputation identifies novel and refined associations for the Wellcome Trust Case Control Consortium phase 1 Data. Eur J Hum Genet. 2012;20:801-805.
21. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38:e164.
22. Magi R, Lindgren CM, Morris AP. Meta-analysis of sex-specific genome-wide association studies. Genet Epidemiol. 2010;34:846-853.
23. Magi R, Morris AP. GWAMA: software for genome-wide association meta-analysis. BMC Bioinformatics. 2010;11:288.
24. Gusev A, Bhatia G, Zaitlen N, et al. Quantifying missing heritability at known GWAS loci. PLoS Genet. 2013;9:e1003993.
25. Yang J, Ferreira T, Morris AP, et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nat Genet. 2012;44:369-375, S361-363.
26. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet. 2011;88:76-82.
27. Hofman A, Brusselle GG, Darwish Murad S, et al. The Rotterdam Study: 2016 objectives and design update. Eur J Epidemiol. 2015;30:661-708.
28. Welter D, MacArthur J, Morales J, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. 2014;42:D1001-1006.
29. Westra HJ, Peters MJ, Esko T, et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. Nat Genet. 2013;45:1238-1243.
30. Ward LD, Kellis M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. Nucleic Acids Res. 2012;40:D930-934.
31. Encode Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489:57-74.
32. Ghanbari M, de Vries PS, de Looper H, et al. A genetic variant in the seed region of miR-4513 shows pleiotropic effects on lipid and glucose homeostasis, blood pressure, and coronary artery disease. Hum Mutat. 2014;35:1524-1531.
33. Ghanbari M, Franco OH, de Looper H, Hofman A, Erkeland S, Dehghan A. Genetic Variations in miRNA Binding Sites Affect miRNA-Mediated Regulation of Several Genes Associated with Cardiometabolic Phenotypes. Circ Cardiovasc Genet. 2015.
34. Ghanbari M, Sedaghat S, de Looper HW, et al. The association of common polymorphisms in miR-196a2 with waist to hip ratio and miR-1908 with serum lipid and glucose. Obesity (Silver Spring). 2015;23:495-503.
35. Ridker PM, Chasman DI, Zee RY, et al. Rationale, design, and methodology of the Women's Genome Health Study: a genome-wide association study of more than 25,000 initially healthy american women. Clin Chem. 2008;54:249-255.
36. Wise AL, Gyi L, Manolio TA. eXclusion: toward integrating the X chromosome in genome-wide association analyses. Am J Hum Genet. 2013;92:643-647.
37. Mostowy S, Bonazzi M, Hamon MA, et al. Entrapment of intracytosolic bacteria by septin cage-like structures. Cell Host Microbe. 2010;8:433-444.
38. Schwefel D, Frohlich C, Eichhorst J, et al. Structural basis of oligomerization in septin-like GTPase of immunity-associated protein 2 (GIMAP2). Proc Natl Acad Sci U S A. 2010;107:20299-20304.

39. Ciucci T, Bosselut R. Gimap and T cells: a matter of life or death. Eur J Immunol. 2014;44:348-351.
40. Zhang L, Yang Z, Ma A, et al. Growth arrest and DNA damage 45G down-regulation contributes to Janus kinase/signal transducer and activator of transcription 3 activation and cellular senescence evasion in hepatocellular carcinoma. Hepatology. 2014;59:178-189.
41. Dudka AA, Sweet SM, Heath JK. Signal transducers and activators of transcription-3 binding to the fibroblast growth factor receptor is activated by receptor amplification. Cancer Res. 2010;70:3391-3401.
42. Hiscott J, Pitha P, Genin P, et al. Triggering the interferon response: the role of IRF-3 transcription factor. J Interferon Cytokine Res. 1999;19:1-13.
43. Schindler C, Levy DE, Decker T. JAK-STAT signaling: from interferons to cytokines. J Biol Chem. 2007;282:20059-20063.
44. Perez-Garcia A, Ambesi-Impiombato A, Hadler M, et al. Genetic loss of SH2B3 in acute lymphoblastic leukemia. Blood. 2013;122:2425-2432.
45. Chang YJ, Chen KW, Chen CJ, et al. SH2B1beta interacts with STAT3 and enhances fibroblast growth factor 1-induced gene expression during neuronal differentiation. Mol Cell Biol. 2014;34:1003-1019.
46. Huffman JE, de Vries PS, Morrison AC, et al. Rare and low-frequency variants and their association with plasma levels of fibrinogen, FVII, FVIII, and vWF. Blood. 2015;126:e19-29.
47. Ek J, Rose CS, Jensen DP, et al. The functional Thr130Ile and Val255Met polymorphisms of the hepatocyte nuclear factor-4alpha (HNF4A): gene associations with type 2 diabetes or altered beta-cell function among Danes. J Clin Endocrinol Metab. 2005;90:3054-3059.
48. Dickson SP, Wang K, Krantz I, Hakonarson H, Goldstein DB. Rare variants create synthetic genome-wide associations. PLoS Biol. 2010;8:e1000294.
49. Wray NR, Purcell SM, Visscher PM. Synthetic associations created by rare variants do not explain most GWAS results. PLoS Biol. 2011;9:e1000579.
50. Brennan SO, Fellowes AP, Faed JM, George PM. Hypofibrinogenemia in an individual with 2 coding (gamma82 A-->G and Bbeta235 P-->L) and 2 noncoding mutations. Blood. 2000;95:1709-1713.
51. Ivaskevicius V, Jusciute E, Steffens M, et al. gammaAla82Gly represents a common fibrinogen gamma-chain variant in Caucasians. Blood Coagul Fibrinolysis. 2005;16:205-208.
52. Wyatt J, Brennan SO, May S, George PM. Hypofibrinogenaemia with compound heterozygosity for two gamma chain mutations - gamma 82 Ala-->Gly and an intron two GT-->AT splice site mutation. Thromb Haemost. 2000;84:449-452.
53. DIAbetes Genetics Replication and Meta-analysis (DIAGRAM) Consortium, Asian Genetic Epidemiology Network Type 2 Diabetes (AGEN-T2D) Consortium, South Asian Type 2 Diabetes (SAT2D) Consortium, et al. Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility. Nat Genet. 2014;46:234-244.
54. Skol AD, Scott LJ, Abecasis GR, Boehnke M. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. Nat Genet. 2006;38:209-213.
55. Skeppholm M, Wallen NH, Blomback M, Kallner A. Can both EDTA and citrate plasma samples be used in measurements of fibrinogen and C-reactive protein concentrations? Clin Chem Lab Med. 2008;46:1175-1179.
56. Peters SA, Woodward M, Rumley A, Koenig W, Tunstall-Pedoe H, Lowe GD. Direct comparisons of three alternative plasma fibrinogen assays with the von Clauss assay in prediction of cardiovascular disease and all-causes mortality: the Scottish Heart Health Extended Cohort. Br J Haematol. 2013;162:392-399.

Chapter 2.2

Comparison of HapMap and 1000 genomes imputation

Manuscript based on this chapter

Paul S. de Vries, Maria Sabater-Lleal, Daniel I. Chasman, Stella Trompet, Tarunveer S. Ahluwalia, Alexander Teumer, Marcus E. Kleber, Ming-Huei Chen, Jie Jin Wang, John R. Attia, Riccardo E. Marioni, Maristella Steri, Lu-Chen Weng, Rene Pool, Vera Grossmann, Jennifer A. Brody, Cristina Venturini, Toshiko Tanaka, Lynda M. Rose, Christopher Oldmeadow, Johanna Mazur, Saonli Basu, Mattias Frånberg, Qiong Yang, Symen Ligthart, Jouke J. Hottenga, Ann Rumley, Antonella Mulas, Anton J. M. de Craen, Anne Grotevendt, Kent D. Taylor, Graciela E. Delgado, Annette Kifley, Lorna M. Lopez, Tina L. Berentzen, Massimo Mangino, Stefania Bandinelli, Alanna C. Morrison, Anders Hamsten, Geoffrey Tofler, Moniek P. M. de Maat, Harmen H. M. Draisma, Gordon D. Lowe, Magdalena Zoledziewska, Naveed Sattar, Philipp S. Wild, Uwe Völker, Barbara McKnight, Jie Huang, Elizabeth G. Holliday, Mark G. McEvoy, John M. Starr, Pirro G. Hysi, Dena G. Hernandez, Weihua Guan, Fernando Rivadeneira, Wendy L. McArdle, P. Eline Slagboom, Tanja Zeller, Bruce M. Psaty, André G. Uitterlinden, Eco J. C. de Geus, David J. Stott, Lackner J. Karl, Albert Hofman, Oscar H. Franco, Jerome I. Rotter, Luigi Ferrucci, Tim D. Spector, Ian J. Deary, Winfried März, Andreas Greinacher, Harald Binder, Francesco Cucca, Dorret I. Boomsma, Hugh Watkins, Weihong Tang, Paul M. Ridker, J. Wouter Jukema, Rodney J. Scott, Paul Mitchell, Torben Hansen, Christopher J. O'Donnell, Nicholas L. Smith, David P. Strachan, and Abbas Dehghan.

Comparison of the use of HapMap and 1000 Genomes reference panels in a large-scale genome-wide association study.

Submitted.

ABSTRACT

Background: Many consortia conducting genome-wide association (GWA) studies are now using the more computationally intensive 1000 Genomes Project reference panel (1000G) for imputation with the expectation that this will lead to the discovery of additional associated loci that would have remained undetected with the HapMap project reference panel (HapMap). This expectation has not yet been tested in any large-scale GWA dataset comprising the same set of individuals.

Methods: In order to assess the performance improvement of 1000G imputation over HapMap in identifying associated loci, we compared the results derived from the two reference panels using our GWA study of circulating fibrinogen concentration comprising 91,953 individuals.

Results: While 29 loci were identified in both the HapMap and 1000G GWA studies, we identified six additional signals using 1000G imputation. However, one locus identified in the HapMap GWA study was not significant in the 1000G GWA study. Furthermore, among the loci that were significant in both the HapMap and 1000G GWA studies, five loci were over one order of magnitude more significant in the 1000G GWA study, compared to two in the HapMap GWA study. When using a stricter Bonferroni correction for the 1000G GWA study (P-value < 2.5×10^{-8}), there were 4 loci significant only in the HapMap GWA study, 5 loci significant only in the 1000G GWA study, and 26 overlapping loci.

Conclusions: 1000G imputation enables the identification of additional loci compared to HapMap imputation, but this may be accompanied by a higher type 1 error rate. When the significance threshold is adjusted accordingly, the difference between the two reference panels is less pronounced.

INTRODUCTION

Most genome-wide association (GWA) studies to date have used their genotyped single nucleotide polymorphisms (SNPs) to impute about 2.5 million SNPs detected in the HapMap Project (HapMap), including mostly common SNPs with a minor allele frequency (MAF) of over 5%.[1-13] HapMap imputation made the meta-analysis of studies that used different genotyping arrays with low overlap, and the interrogation of most common SNPs possible.[1] However, low-frequency and rare variation is generally not covered.[14] Similarly, genetic variation other than SNPs, such as small insertion-deletions (indels) and large structural variants are not included in HapMap-based imputed projects, contributing to possible sources of missing heritability.

In contrast, the more recently released Phase 1 version 3 of the 1000 Genomes Project (1000G) is based on a larger set of individuals, and comprises nearly 40 million variants including 1.4 million indels.[15] 1000G allows the interrogation of most common and low-frequency variants (MAF > 1%), and some rare variants (MAF < 1%) that were previously not covered.[16] 1000G imputation thus has several perceived benefits, but given that the denser 1000G imputation comes at the cost of an increased computational and analytical burden, it is important to examine the observed benefits. While several GWA studies using 1000G imputation have been published or are in progress, their sample size differs from the previous GWA studies using HapMap imputation, making comparison difficult. Therefore, with the aim of evaluating the benefits of using 1000G imputation in GWA studies compared to HapMap imputation, we carried out a GWA study of a quantitative trait, circulating fibrinogen concentration, using both HapMap and 1000G imputed data on a single set of the same 91,953 individuals.

METHODS

Population

The sample for both the HapMap and 1000G GWA studies consists of 22 studies including the same 91,953 European-ancestry participants. The sample is largely a subset of the sample used in our previous work, and when possible the same analyses were used in this project.[17,18] However, to ensure that only the same individuals were used, one or both of the analyses was rerun using only overlapping individuals when necessary. All studies were approved by appropriate research ethics committees and all respondents signed informed consent prior to participation.

Genotyping and imputation

Studies imputed dosages of genetic variants using reference panels from the 1000 genomes project with MACH[19,20] or IMPUTE.[21] Studies imputed variant dosages using phase 2 reference panels from the HapMap project with MACH,[19,20] IMPUTE,[21] or BIMBAM.[22] We excluded variants with MACH imputation quality < 0.3, IMPUTE/BIMBAM imputation quality < 0.4, or MAF < 0.01 from each study.

Fibrinogen measurement

Fibrinogen concentration was measured in citrated or EDTA plasma samples using a variety of methods including the Clauss method, immunonephelometric methods, immunoturbidimetric methods, and other functional methods. Fibrinogen concentration was measured in g/L and natural-log transformed.

Genome-wide association analysis

All analyses were adjusted for age and sex, and study specific covariates such as center or case/control status. In family studies, linear mixed models were used to account for family structure. Some studies adjusted the analysis for principle components to account for population structure and cryptic relatedness. Some studies used a different number of principle components in the HapMap and 1000G analyses. We applied a genomic control correction to the results of each of the studies before meta-analysis to remove any remaining genomic inflation. The genomic inflation factor used in this correction was calculated separately in the HapMap and 1000G analyses for each study. We meta-analyzed the results using an inverse-variance model with fixed effects implemented in METAL.[23] Loci were defined as the 500 Kb area on either side of lead variants (the variant with the smallest *P*-value). Build 36 positions of HapMap SNPs were converted to build 37 using the UCSC genome browser (http://genome.ucsc.edu/cgi-bin/hgLiftOver). Variants were annotated to genes using ANNOVAR version 2013Mar07. At the meta-analysis level, the imputation quality of each variant was defined as the sample-size weighted mean imputation quality across the studies, not including studies where the variant was filtered out.

Comparison of HapMap and 1000G

When a locus was significant in both the HapMap and 1000G GWA studies we defined it as an overlapping locus. When a locus was significant in only one of the two analyses we defined it as a non-overlapping locus. To compare the strength of association in the HapMap and 1000G GWAS, we identified loci with *P*-value differences of 1 order of magnitude or greater (for example: from 5×10^{-8} compared to 5×10^{-9} or less).

For each significant locus we used two approaches to assess the relationship between lead variants from HapMap and 1000G. First, we determined whether or not the more significant of the two lead variants or a good proxy (linkage disequilibrium $R^2 > 0.8$) was included in the analysis of the other reference panel. If so, we examined its association in the other reference panel. Thus, if a locus was more significant in the 1000G GWA study, we checked whether the 1000G lead variant or a proxy was included in the HapMap GWA study. Second, we examined the correlation R^2 between HapMap and 1000G lead variants in the form of imputed genotype dosages. This was done in 5966 individuals from the Rotterdam Study.

Sensitivity analysis

First, we compared the results of the HapMap and 1000G GWA studies when applying a stricter Bonferroni-corrected *P*-value threshold of 2.5×10^{-8} to the 1000G GWA study. This threshold was suggested by Huang et al to keep the type 1 error rate at 5% when using 1000G data.[24] Second, we repeated the analysis without using genomic control corrections. Third, we repeated the analysis in 34,098 participants using only the 10 studies that used the same imputation and analysis software as well as the same covariates for the HapMap and 1000G GWA studies.

RESULTS

Baseline characteristics of the participants for each of the included studies are shown in **Supplemental Table 1**. The HapMap GWA study included 2,749,429 SNPs, and the 1000G GWA study included 10,883,314 variants. Using a genome-wide significance threshold of 5×10^{-8}, a total of 1,210 SNPs across 30 loci were associated with circulating fibrinogen concentration in the HapMap GWA study compared with 4,096 variants across 35 loci in the 1000G GWA study (**Supplemental Figures 1 and 2**). Of these loci, six were associated only in the 1000G GWA study and one was associated only in the HapMap GWA study, while 29 were overlapping (**Figure 1A**). The main results for both overlapping and non-overlapping loci are summarized in **Figure 1**. The HapMap and 1000G lead variants of non-overlapping loci are described in **Table 1**, and leads variants of overlapping loci are described in **Table 2**. Among significant loci, the correlation coefficient of the beta coefficients, *P*-values, and imputation qualities of HapMap and 1000G lead variants were 0.925, 0.998, and 0.435 respectively (**Supplemental Figure 3**).

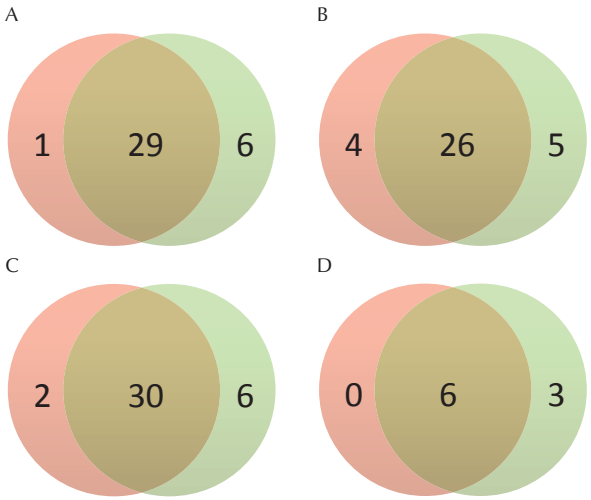


Figure 1. Venn diagram of the number of loci significant using HapMap (left circle) and 1000G (right circle) imputation in A) the main analysis, B) the sensitivity analysis applying a significance threshold of 2.5×10^{-8} to the 1000G GWA analysis, C) the sensitivity analysis without using genomic control corrections, and D) the sensitivity analysis excluding studies that used different imputation software, analysis software, or covariates in the HapMap and 1000G GWA analyses.

Non-overlapping loci

The lead variants for non-overlapping loci always differed between the HapMap and 1000G GWA studies, and all *P*-value differences were greater than 1 order of magnitude (for example: from 5×10^{-8} to 5×10^{-9} or less). Differences between HapMap and 1000G imputation for the seven non-overlapping loci are summarized in **Figure 2**.

Regional plots of the six loci significant only in the 1000G GWA study are shown in **Figure 3**. For four of these six loci, the correlation R^2 between imputed dosages of HapMap and 1000G lead variants was less than 0.8 (**Supplemental Table 2**). None of the 1000G lead variants among these four loci were included in the HapMap GWA study, and neither were any good proxies.

A regional plot of the 6p21.3 locus, which was significant only in the HapMap GWA study, is shown in **Figure 4**. The lowest *P*-value at the locus was 8.5×10^{-9} in the HapMap GWA study compared to 7.9×10^{-6} in the 1000G GWA study. The correlation R^2 between imputed dosages of the HapMap and 1000G lead variants was 0.07. The HapMap lead SNP was included in the 1000G GWA study under a different name, rs114339898, but the imputation quality was only high enough for inclusion in 7 of the studies.

Overlapping loci

The lead variants of eight of the 29 overlapping loci were the same for the HapMap and 1000G GWA studies. *P*-value differences between the HapMap and 1000G GWA studies were often small: they were smaller than or equal to one order of magnitude for 22 loci. *P*-values differed by more than one order of magnitude for seven loci. Five of these loci were more significant in the 1000G GWA study (2q37.3,

Table 1. Non-overlapping loci that were significant in either the HapMap or 1000G GWA studies

	HapMap					1000G				
Locus	Lead Variant	Beta	*P*-value	MAF	Imputation Quality	Lead Variant	Beta	*P*-value	MAF	Imputation Quality
Significant in 1000G										
1q42.13	rs10489615	0.0052	8.3×10^{-07}	0.38	0.97	rs10864726	0.0059	1.1×10^{-08}	0.40	0.96
3q21.1	rs16834024	0.0173	1.4×10^{-07}	0.03	0.79	rs1976714	0.0064	7.5×10^{-09}	0.35	0.89
4p16.3	rs2699429	0.0060	1.3×10^{-07}	0.43	0.87	rs59950280	0.0080	2.5×10^{-11}	0.34	0.80
7p15.3	rs1029738	0.0057	3.2×10^{-07}	0.30	1.00	rs61542988	0.0065	3.1×10^{-08}	0.25	0.98
8p23.1	rs7004769	0.0062	1.4×10^{-06}	0.20	1.00	rs7012814	0.0061	8.0×10^{-09}	0.47	0.91
11q12.2	rs7935829	0.0056	5.6×10^{-08}	0.40	0.99	rs11230201	0.0060	3.0×10^{-09}	0.41	0.99
Significant in HapMap										
6p21.3	rs12528797	0.0095	8.5×10^{-09}	0.11	0.98	rs116134220	0.0082	7.9×10^{-06}	0.49	0.89

Abbreviations: HapMap refers to the GWA study using imputation based on the HapMap project. 1000G refers to the GWA study using imputation based on the 1000 Genomes Project. Variants were coded according to the fibrinogen increasing allele. MAF refers to minor allele frequency.

4q31.3, 10q21.3, 12q24.12, and 21q22.2), while two of these loci were more significant in the HapMap GWA study (5q31.1 and 8q24.3).

Among the five overlapping loci with lower *P*-values in the 1000G GWA study, the correlation R^2 between imputed dosages of lead variants from HapMap and 1000G was higher than 0.8 for 4 loci, but was 0.68 for the 12q24.12 locus (**Supplemental Table 4**). There was no good proxy of the 1000G lead variant at the 12q24.12 locus included in the HapMap GWA study.

The 5q31.1 and 8q24.3 loci had lower *P*-values in the HapMap GWA study. The correlation R^2 between imputed dosages from HapMap and 1000G was almost perfect for 5q31.1, but was 0.75 for 8q24.3. The HapMap lead variant of the 8q24.3 locus was also included in the 1000G GWA study. These differences between HapMap and 1000G imputation for the 29 overlapping loci are summarized in **Figure 5**.

Sensitivity analyses

Because more independent variants are included in the 1000G GWA study, it may not be fair to use the conventional genome-wide significance threshold of 5×10^{-8}.[24,25] When we restricted the significant loci from the 1000G GWA study to just those with a *P*-value below 2.5×10^{-8}, there were 4 loci significant only in the HapMap GWA study, 5 loci significant only in the 1000G GWA study, and 26 overlapping loci (**Figure 1B**). Three loci that were significant using both HapMap and 1000G imputation thus became non-significant when the stricter significance threshold was applied to the 1000G results.

Genomic inflation factors to correct for genomic control were calculated separately for the HapMap and 1000G analyses of each study. Thus, differences in the genomic inflation factors could explain some of the differences between the HapMap and 1000G results. When we repeated the HapMap and 1000G GWA study without applying genomic control corrections, 2 loci were associated only with circulating fibrinogen concentration in the HapMap GWA study, 6 were only associated in the 1000G GWA study, and 30 were associated in both GWA studies (**Figure 1C**).

For practical reasons, not all of the studies used the same imputation software, analysis software, or covariates for the HapMap and 1000G analyses. Specifically, fewer studies used principal components in the HapMap GWA study. When we restricted the analysis to those studies that used the same imputation software, analysis software, and covariates in the HapMap and 1000G GWA studies, 3 loci were associated only in the 1000G GWA study, and 6 were associated in both the HapMap and the 1000G GWA studies (**Figure 1D**). No loci were associated only in the HapMap GWA study.

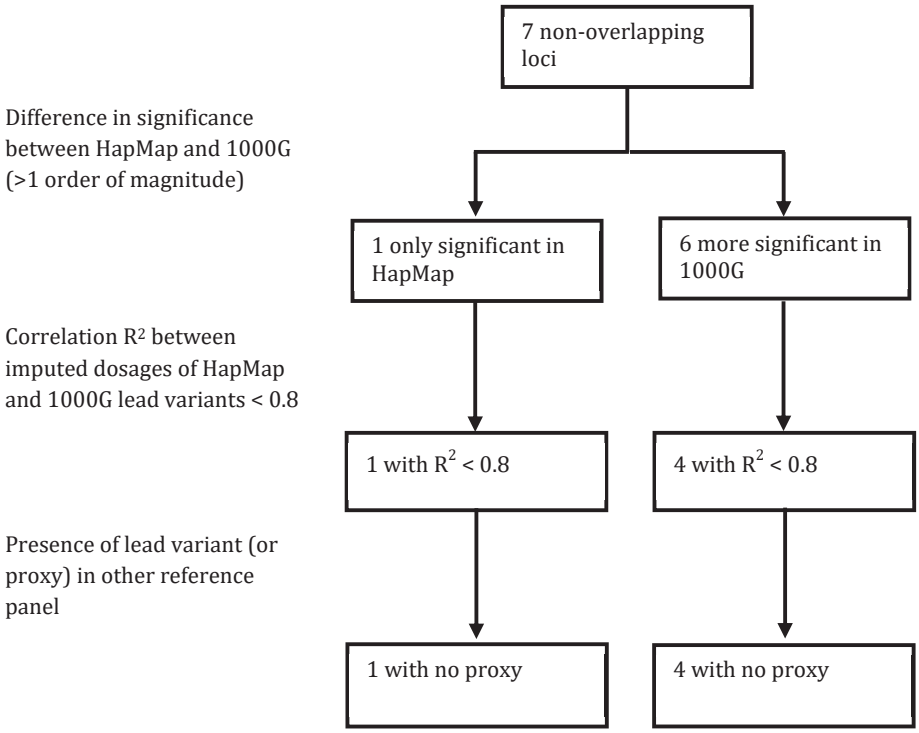


Figure 2. Summary of the differences between HapMap and 1000G imputation for the seven non-overlapping loci.

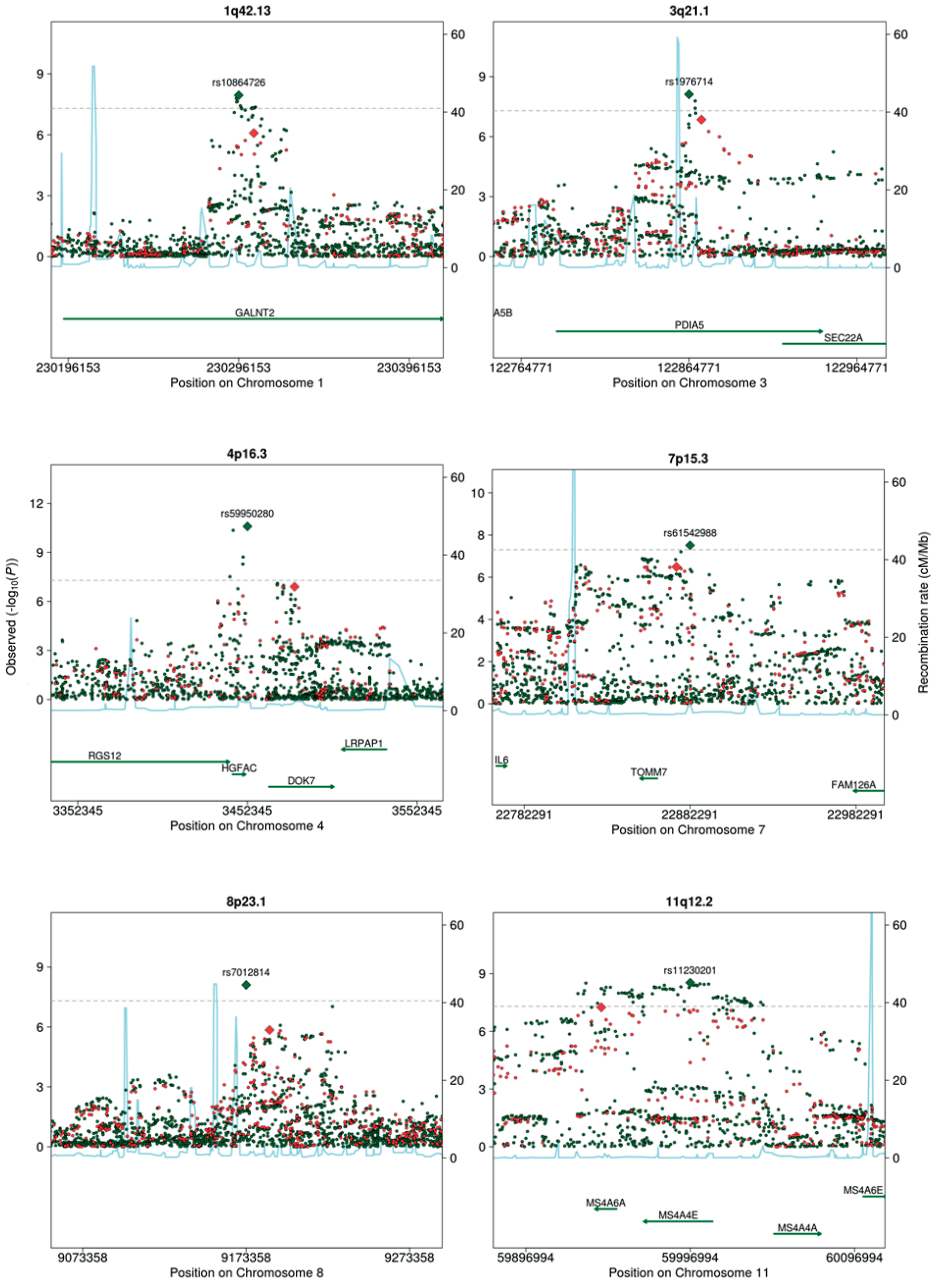


Figure 3. Regional plots of non-overlapping loci that were more significantly associated with fibrinogen in the 1000G GWA study, including variants from both the HapMap (red) and 1000G (green) GWA studies.

Table 2. Overlapping loci that were significant in both the HapMap and 1000G GWA studies

	HapMap					1000G				
Locus	Lead Variant	Beta	*P*-value	MAF	Imputation Quality	Lead Variant	Beta	*P*-value	MAF	Imputation Quality
1p31.3	rs4655582	0.0069	4.8×10^{-11}	0.38	0.98	rs2376015	0.0075	5.1×10^{-12}	0.35	0.91
1q21.3	rs8192284	0.0115	8.9×10^{-29}	0.40	0.97	rs61812598	0.0114	1.8×10^{-28}	0.39	0.99
1q44	rs12239046	0.0103	9.7×10^{-21}	0.38	0.99	rs12239046	0.0102	9.8×10^{-22}	0.38	0.99
2q12	rs1558643	0.0066	5.8×10^{-10}	0.40	0.99	rs1558643	0.0063	6.0×10^{-10}	0.40	0.98
2q13	rs6734238	0.0106	1.7×10^{-23}	0.41	0.99	rs6734238	0.0106	3.7×10^{-24}	0.41	1.00
2q34	rs715	0.0092	9.1×10^{-14}	0.32	0.92	rs715	0.0082	1.7×10^{-13}	0.32	0.89
2q37.3	rs1476698	0.0075	4.2×10^{-12}	0.36	1.00	rs59104589	0.0081	2.4×10^{-14}	0.34	0.98
3q22.2	rs548288	0.0113	6.6×10^{-21}	0.24	0.99	rs150213942	0.0117	3.1×10^{-21}	0.23	0.95
4q31.3	rs2227401	0.0311	4.7×10^{-134}	0.21	0.95	rs72681211	0.0313	1.3×10^{-142}	0.20	0.99
5q31.1	rs1012793	0.0208	4.4×10^{-60}	0.21	0.98	rs1012793	0.0207	1.0×10^{-58}	0.20	0.98
7p21.1	rs10950690	0.0071	9.9×10^{-12}	0.48	0.94	rs12699921	0.0071	1.3×10^{-12}	0.47	0.98
7q14.2	rs2710804	0.0061	9.3×10^{-09}	0.38	0.98	rs2710804	0.0057	4.3×10^{-08}	0.38	0.99
7q36.1	rs13226190	0.008	2.2×10^{-10}	0.21	0.99	rs13234724	0.0076	1.6×10^{-09}	0.21	0.99
8q24.3	rs7464572	0.0066	2.4×10^{-09}	0.40	0.98	rs11136252	0.0056	4.6×10^{-08}	0.42	0.96
9q22.2	rs7873907	0.006	5.4×10^{-09}	0.50	0.96	rs3138493	0.006	3.5×10^{-09}	0.48	0.98
10q21.3	rs10761756	0.0093	5.4×10^{-20}	0.48	1.00	rs7916868	0.0097	1.2×10^{-21}	0.49	0.97
11p12	rs7937127	0.0083	2.3×10^{-10}	0.18	0.99	rs7934094	0.0081	2.9×10^{-10}	0.22	0.90
12q13.12	rs1521516	0.0072	3.0×10^{-11}	0.36	1.00	12:51042486	0.0073	4.9×10^{-12}	0.36	0.98
12q24.12	rs3184504	0.0066	1.1×10^{-10}	0.49	0.97	rs4766897	0.009	3.8×10^{-12}	0.34	0.64
14q24.1	rs194741	0.0092	8.3×10^{-14}	0.25	0.95	rs194714	0.0086	3.7×10^{-13}	0.25	0.97
15q15.1	rs1703755	0.0088	1.8×10^{-09}	0.14	0.96	rs8026198	0.009	5.9×10^{-10}	0.15	0.93
15q21.2	rs12915052	0.0069	2.4×10^{-10}	0.31	1.00	rs11630054	0.0067	3.3×10^{-10}	0.34	0.99
16q12.2	rs12598049	0.0074	3.0×10^{-11}	0.32	0.99	rs6499550	0.007	8.2×10^{-11}	0.32	0.98
16q22.2	rs11864453	0.0057	4.6×10^{-08}	0.40	0.99	rs1035560	0.0058	1.2×10^{-08}	0.40	0.99
17q21.2	rs7224737	0.0073	2.2×10^{-09}	0.23	0.99	rs7224737	0.0068	5.2×10^{-09}	0.24	1.00
17q25.1	rs10512597	0.0078	2.2×10^{-08}	0.18	0.94	rs35489971	0.0077	1.6×10^{-08}	0.18	0.94
20q13.12	rs1800961	0.0183	6.8×10^{-09}	0.03	0.95	rs1800961	0.0178	1.7×10^{-09}	0.03	0.99
21q22.2	rs4817986	0.0091	1.9×10^{-14}	0.28	0.95	rs9808651	0.0093	5.4×10^{-16}	0.28	0.94
22q13.33	rs6010044	0.0074	2.5×10^{-08}	0.20	0.89	rs75347843	0.0082	4.3×10^{-08}	0.19	0.76

Abbreviations: HapMap refers to the GWA study using imputation based on the HapMap project. 1000G refers to the GWA study using imputation based on the 1000 Genomes Project. Variants were coded according to the fibrinogen increasing allele. MAF refers to minor allele frequency.

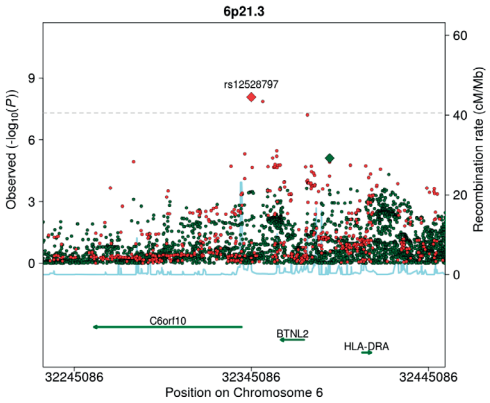


Figure 4. Regional plot of 6p21.3, a non-overlapping locus that was more significantly associated with fibrinogen in the HapMap GWA study, including variants from both the HapMap (red) and 1000G (green) GWA studies.

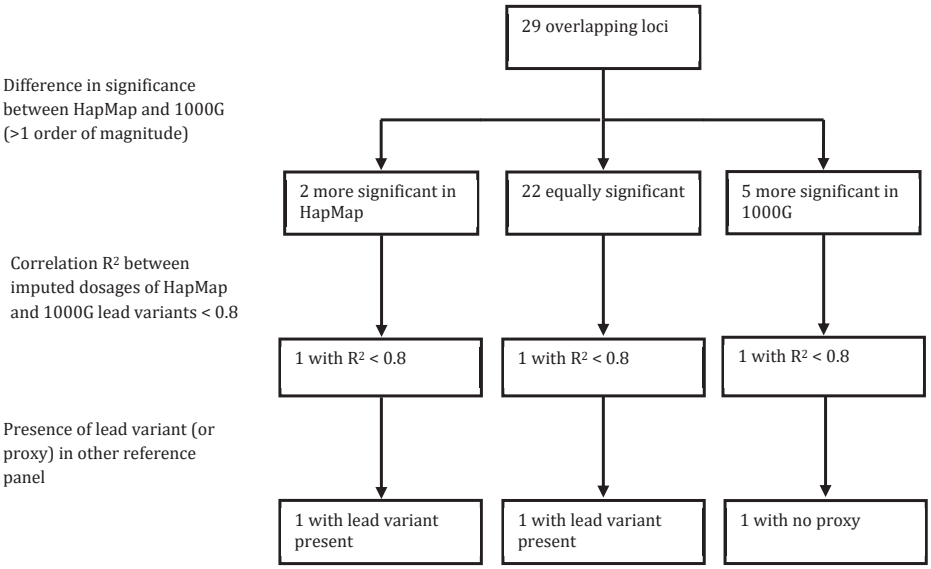


Figure 5. Summary of the differences between HapMap and 1000G imputation for the 29 overlapping loci.

DISCUSSION

In our fibrinogen GWA study of 91,953 individuals, using 1000G imputation instead of HapMap imputation led to the identification of six additional fibrinogen loci, suggesting an improvement in the detection of associated signals. Nevertheless, there was also one locus that was only identified when using HapMap imputation, and the advantage of 1000G imputation was attenuated when using a more stringent Bonferroni correction for the 1000G GWA study The inclusion of indels in the 1000G GWA study did not lead to the identification of any new loci. Only one locus in our

1000G GWA study was led by an indel, and it was in strong linkage disequilibrium with a SNP present in HapMap.

While this is the first study of the impact of HapMap and 1000G imputation on genome-wide associations using the exact same individuals at the level of a large-scale consortium, four previous studies have addressed this question on a smaller scale. In the Wellcome Trust Case Control Consortium, Huang et al re-analyzed GWA studies of 7 diseases (bipolar disorder, coronary artery disease, Crohn's disease, hypertension, rheumatoid arthritis, type 1 and 2 diabetes) with 1000G imputation, and found two novel loci: one for type 1 diabetes and one for type 2 diabetes.[24] For each disease the sample consisted of 2000 cases and 3000 controls. A more conservative genome-wide significance threshold of 2.5×10^{-8} was used in the 1000G GWA studies, while the MAF threshold was the same at 1%. The second study was a 1000G imputed GWA study of around 2000 cases of venous thrombosis and 2400 controls.[26] Using a conservative *P*-value threshold of 7.4×10^{-9}, but no MAF threshold, Germain et al identified an uncommon variant at a novel locus that was not identified in the HapMap GWAS.[26] Third, the National Cancer Institute Breast and Prostate Cancer Cohort Consortium found no new loci by applying 1000G imputation to their existing dataset of 2800 cases and 4500 controls.[27,28] The conventional genome-wide significance threshold of 5×10^{-8} was used, but no MAF threshold was used. Fourthly, Wood et al compared HapMap and 1000G imputation for a total of 93 quantitative traits in 1210 individuals from the InCHIANTI study.[29] Using a significance threshold of 5×10^{-8} for both the HapMap and 1000G GWA studies, they found 20 overlapping associations, 13 associations that were only significant using 1000G imputation, and 1 association that was only significant using HapMap imputation. For the association only significant in HapMap, the *P*-value difference between HapMap and 1000G lead variants was less than 1 order of magnitude. When the authors lowered their significance threshold to 5×10^{-11} to reflect the number of tests being done in analyses multiple traits, 9 associations remained significant based on HapMap imputation and 11 associations remained significant based on 1000G imputed.

All four of these comparison studies used an earlier 1000 genomes reference panel. The present study adds further to the literature as it is based on the widely implemented Phase 1 Version 3 of 1000G. Crucially, the large sample size allowed us to examine differences at many non-overlapping and overlapping loci, and improved the generalizability of our results, as ongoing GWA studies are often also large.

Two further studies with different approaches also provide insight. First, Springelkamp et al found a novel locus using 1000G imputation even though the sample size was smaller than the previous HapMap GWA study.[30,31] The same genome-wide significance (5×10^{-8}) and MAF (1%) thresholds were used. The lowest *P*-value at the locus was 1.9×10^{-8}. Because different individuals were included in these GWA studies,

the difference between HapMap and 1000G may partially be explained by sampling variability. Second, Shin et al identified 299 SNP-metabolite associations based on HapMap imputation, and reexamined the associated loci using 1000G imputation in the same individuals.[32] They found that HapMap and 1000G imputation yielded similar *P*-values and variance explained for all loci but one. For that locus, 1000G imputation led to a much stronger association, increasing the variance explained from 10% to 16%, and decreasing the *P*-value from 8.8×10^{-113} to 7.7×10^{-244}. Although Shin et al did not compare loci identified using HapMap and 1000G, their results do support our finding that large differences in association are possible, albeit not present at every locus. These studies, along with the present study, suggest that signals not previously found in HapMap GWA studies can be found in 1000G GWAS using the same sample size.

In this study we demonstrate that, although 1000G imputation was more effective at identifying associated loci overall, HapMap imputation can outperform 1000G imputation for specific loci. The 6p21.3 locus, corresponding to the major histocompatibility complex (MHC), was significant in the HapMap GWA study but not in the 1000G GWA study. The MHC is highly polymorphic and hosts many repetitive sequences, making it difficult to genotype and sequence.[33-35] The HapMap reference panel was based largely on the genotyping of variants that were known at that time, whereas the 1000G reference panel is based entirely on low-coverage sequencing. This may explain the rather large discrepancy between HapMap and 1000G at this locus.

Differences in associations when GWA studies are based on different participants can be explained by sampling variability, even with the same sample size. Thus, by using exactly the same participants in the HapMap and 1000G comparisons in the present project, we rule out both statistical power and sampling variability as possible explanations for differences between the HapMap and 1000G GWA studies. Nevertheless, some differences were not controlled for and thus remain as potential alternative explanations.

First, genomic control corrections were applied to the results of each of the studies before meta-analysis, separately for the HapMap and 1000G GWA studies. As a result, for any given study, there could be differences between the correction applied to the HapMap GWA analysis and to the 1000G GWA analysis. As these differences do not appear to differ systematically between the HapMap and 1000G GWA analyses in our study, the genomic control corrections are unlikely to explain our results. The results from our sensitivity analysis were concordant with this interpretation: when no genomic control corrections were applied there were 6 loci only significant in the 1000G GWA study compared to 2 loci only significant in the HapMap GWA study.

The second difference between the HapMap and 1000G GWA studies that may explain our results is that in the 1000G GWA study more studies adjusted for principal components. This difference reflects common practice, as population stratification is suspected to have a stronger influence on variants with lower MAF, and 1000G includes more of these.[36] However, the adjustments are applied to variants across the spectrum of minor allele frequencies, which may have influenced our results. Thirdly, some studies used different software for HapMap and 1000G imputation (**Supplemental Table 1**). The imputation quality metrics used by IMPUTE and MACH are different, and this has traditionally been dealt with by applying different imputation quality thresholds: 0.3 for MACH and 0.4 for IMPUTE.[5,37] Thus, in studies that used different imputation software for the HapMap and 1000G GWA studies, the filtering of variants can be expected to differ. There may, additionally, be real differences in imputation quality. Finally, some studies used different analysis software. When we restricted our analysis to only those studies that used the same covariates, analysis software, and imputation software for the HapMap and 1000G GWA studies, we found similar differences between the HapMap and 1000G GWA studies: 3 loci were only significant in the 1000G GWA study, while all loci significant in the HapMap GWA study were also significant in the 1000G GWA study. This suggests that differences in imputation software, analysis software, and covariates do not fully explain the observed difference between the HapMap and 1000G GWA studies.

1000G GWA studies may include more independent statistical tests than HapMap GWA studies.[24,25] Thus, while a *P*-value threshold of 5×10^{-8}, correcting for 1 million independent tests, maintains the type I error rate at 5% for HapMap GWA studies, this may not be the case for 1000G GWA studies. Using 1000G pilot data, Huang et al estimated that 2 million independent tests were being done, and thus suggested a *P*-value threshold of 2.5×10^{-8}.[24] In this study we used a *P*-value threshold of 5×10^{-8} for both the HapMap and 1000G GWA studies, in accordance with the majority of published 1000G GWA studies.[30,38-41] When we used the threshold of 2.5×10^{-8}, the difference between the HapMap and 1000G GWA studies became smaller. Thus, while we expect 1000G imputation may lead to novel findings using the conventional genome-wide significance threshold, the same thing may not be expected when using stricter, and perhaps more appropriate thresholds. In other words, using the traditional significance threshold also for 1000G may increase the power, but also the type 1 error rate.

In this study we only examined variants with a MAF of greater than 1%. This restriction was common practice for HapMap GWA studies, but given the improved coverage of rare variants in 1000G, this may not remain the case for 1000G GWAS. Different MAF thresholds have been used in published 1000G GWAS, but many have used 1%.[24,26,27,30,31,38-42] Therefore, an advantage of 1000G not illustrated by this study may

be the identification of rare variants, at new loci or as secondary signals at known loci. The advantage of 1000G imputation will then in part depend on the importance and impact of rare variants in the trait being studied, as well as the distribution of these variants. Rare and uncommon variants are often clustered in genes with previously associated common variants, limiting the new biology accessed through their identification.[43] This appears to be the case for fibrinogen concentration as well.[17,44]

In conclusion, we show that the reference panel used in GWA studies can have a large impact on the statistical power for common variants, although our results do not support the expectation that 1000G imputation always outperforms HapMap imputation, as we found one locus that appeared to be better covered in HapMap. Using 1000G imputation did lead to more associated loci than using HapMap imputation, but this advantage was attenuated when using a stricter *P*-value threshold for the 1000G GWA study. This may have broader implications: while more extensive reference panels have improved coverage, the penalty to the significance threshold for including further variants may outweigh these gains, especially if the additional variants are poorly or moderately imputed.

REFERENCES

1. International HapMap Consortium. The International HapMap Project. Nature. 2003;426:789-796.
2. Wood AR, Esko T, Yang J, et al. Defining the role of common variation in the genomic and biological architecture of adult human height. Nat Genet. 2014;46:1173-1186.
3. Global Lipids Genetics C, Willer CJ, Schmidt EM, et al. Discovery and refinement of loci associated with lipid levels. Nat Genet. 2013;45:1274-1283.
4. Smith NL, Huffman JE, Strachan DP, et al. Genetic predictors of fibrin D-dimer levels in healthy adults. Circulation. 2011;123:1864-1872.
5. Locke AE, Kahali B, Berndt SI, et al. Genetic studies of body mass index yield new insights for obesity biology. Nature. 2015;518:197-206.
6. International Consortium for Blood Pressure Genome-Wide Association Studies, Ehret GB, Munroe PB, et al. Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. Nature. 2011;478:103-109.
7. Huang J, Sabater-Lleal M, Asselbergs FW, et al. Genome-wide association study for circulating levels of PAI-1 provides novel insights into its regulation. Blood. 2012;120:4873-4881.
8. Huang J, Huffman JE, Yamakuchi M, et al. Genome-wide association study for circulating tissue plasminogen activator levels and functional follow-up implicates endothelial STXBP5 and STX2. Arterioscler Thromb Vasc Biol. 2014;34:1093-1101.
9. Estrada K, Styrkarsdottir U, Evangelou E, et al. Genome-wide meta-analysis identifies 56 bone mineral density loci and reveals 14 loci associated with risk of fracture. Nat Genet. 2012;44:491-501.
10. CARDIoGRAMplusC4D Consortium, Deloukas P, Kanoni S, et al. Large-scale association analysis identifies new risk loci for coronary artery disease. Nat Genet. 2013;45:25-33.
11. Dehghan A, Dupuis J, Barbalic M, et al. Meta-analysis of genome-wide association studies in >80 000 subjects identifies multiple loci for C-reactive protein levels. Circulation. 2011;123:731-738.
12. Smith NL, Chen MH, Dehghan A, et al. Novel associations of multiple genetic loci with plasma levels of factor VII, factor VIII, and von Willebrand factor: The CHARGE (Cohorts for Heart and Aging Research in Genome Epidemiology) Consortium. Circulation. 2010;121:1382-1392.
13. Loth DW, Artigas MS, Gharib SA, et al. Genome-wide association analysis identifies six new loci associated with forced vital capacity. Nat Genet. 2014;46:669-677.
14. Zuk O, Schaffner SF, Samocha K, et al. Searching for missing heritability: designing rare variant association studies. Proc Natl Acad Sci U S A. 2014;111:E455-464.
15. 1000 Genomes Project Consortium, Abecasis GR, Auton A, et al. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491:56-65.
16. Zheng HF, Rong JJ, Liu M, et al. Performance of genotype imputation for low frequency and rare variants from the 1000 genomes. PLoS One. 2015;10:e0116487.
17. de Vries PS, Chasman DI, Sabater-Lleal M, et al. A meta-analysis of 120,246 individuals identifies 18 new loci for fibrinogen concentration. Hum Mol Genet. 2015.
18. Sabater-Lleal M, Huang J, Chasman D, et al. Multiethnic meta-analysis of genome-wide association studies in >100 000 subjects identifies 23 fibrinogen-associated Loci but no strong evidence of a causal association between circulating fibrinogen and cardiovascular disease. Circulation. 2013;128:1310-1324.

19. Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet Epidemiol. 2010;34:816-834.
20. Li Y, Willer C, Sanna S, Abecasis G. Genotype imputation. Annu Rev Genomics Hum Genet. 2009;10:387-406.
21. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. 2009;5:e1000529.
22. Servin B, Stephens M. Imputation-based analysis of association studies: candidate regions and quantitative traits. PLoS Genet. 2007;3:e114.
23. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics. 2010;26:2190-2191.
24. Huang J, Ellinghaus D, Franke A, Howie B, Li Y. 1000 Genomes-based imputation identifies novel and refined associations for the Wellcome Trust Case Control Consortium phase 1 Data. Eur J Hum Genet. 2012;20:801-805.
25. Li MX, Yeung JM, Cherny SS, Sham PC. Evaluating the effective numbers of independent tests and significant *P*-value thresholds in commercial genotyping arrays and public imputation reference datasets. Hum Genet. 2012;131:747-756.
26. Germain M, Saut N, Oudot-Mellakh T, et al. Caution in interpreting results from imputation analysis when linkage disequilibrium extends over a large distance: a case study on venous thrombosis. PLoS One. 2012;7:e38538.
27. Machiela MJ, Chen C, Liang L, et al. One thousand genomes imputation in the National Cancer Institute Breast and Prostate Cancer Cohort Consortium aggressive prostate cancer genome-wide association study. Prostate. 2013;73:677-689.
28. Schumacher FR, Berndt SI, Siddiq A, et al. Genome-wide association study identifies new prostate cancer susceptibility loci. Hum Mol Genet. 2011;20:3867-3875.
29. Wood AR, Perry JR, Tanaka T, et al. Imputation of variants from the 1000 Genomes Project modestly improves known associations and can identify low-frequency variant-phenotype associations undetected by HapMap based imputation. PLoS One. 2013;8:e64343.
30. Springelkamp H, Iglesias AI, Cuellar-Partida G, et al. ARHGEF12 influences the risk of glaucoma by increasing intraocular pressure. Hum Mol Genet. 2015.
31. Hysi PG, Cheng CY, Springelkamp H, et al. Genome-wide analysis of multi-ancestry cohorts identifies new loci influencing intraocular pressure and susceptibility to glaucoma. Nat Genet. 2014;46:1126-1130.
32. Shin SY, Fauman EB, Petersen AK, et al. An atlas of genetic influences on human blood metabolites. Nat Genet. 2014;46:543-550.
33. Major E, Rigo K, Hague T, Berces A, Juhos S. HLA typing from 1000 genomes whole genome and whole exome illumina data. PLoS One. 2013;8:e78410.
34. Hosomichi K, Jinam TA, Mitsunaga S, Nakaoka H, Inoue I. Phase-defined complete sequencing of the HLA genes by next-generation sequencing. BMC Genomics. 2013;14:355.
35. de Bakker PI, Raychaudhuri S. Interrogating the major histocompatibility complex with high-throughput genomics. Hum Mol Genet. 2012;21:R29-36.
36. Mathieson I, McVean G. Differential confounding of rare and common variants in spatially structured populations. Nat Genet. 2012;44:243-246.
37. Shungin D, Winkler TW, Croteau-Chonka DC, et al. New genetic loci link adipose and insulin biology to body fat distribution. Nature. 2015;518:187-196.
38. Verhaaren BF, Debette S, Bis JC, et al. Multi-Ethnic Genome-Wide Association Study of Cerebral White Matter Hyperintensities on MRI. Circ Cardiovasc Genet. 2015.

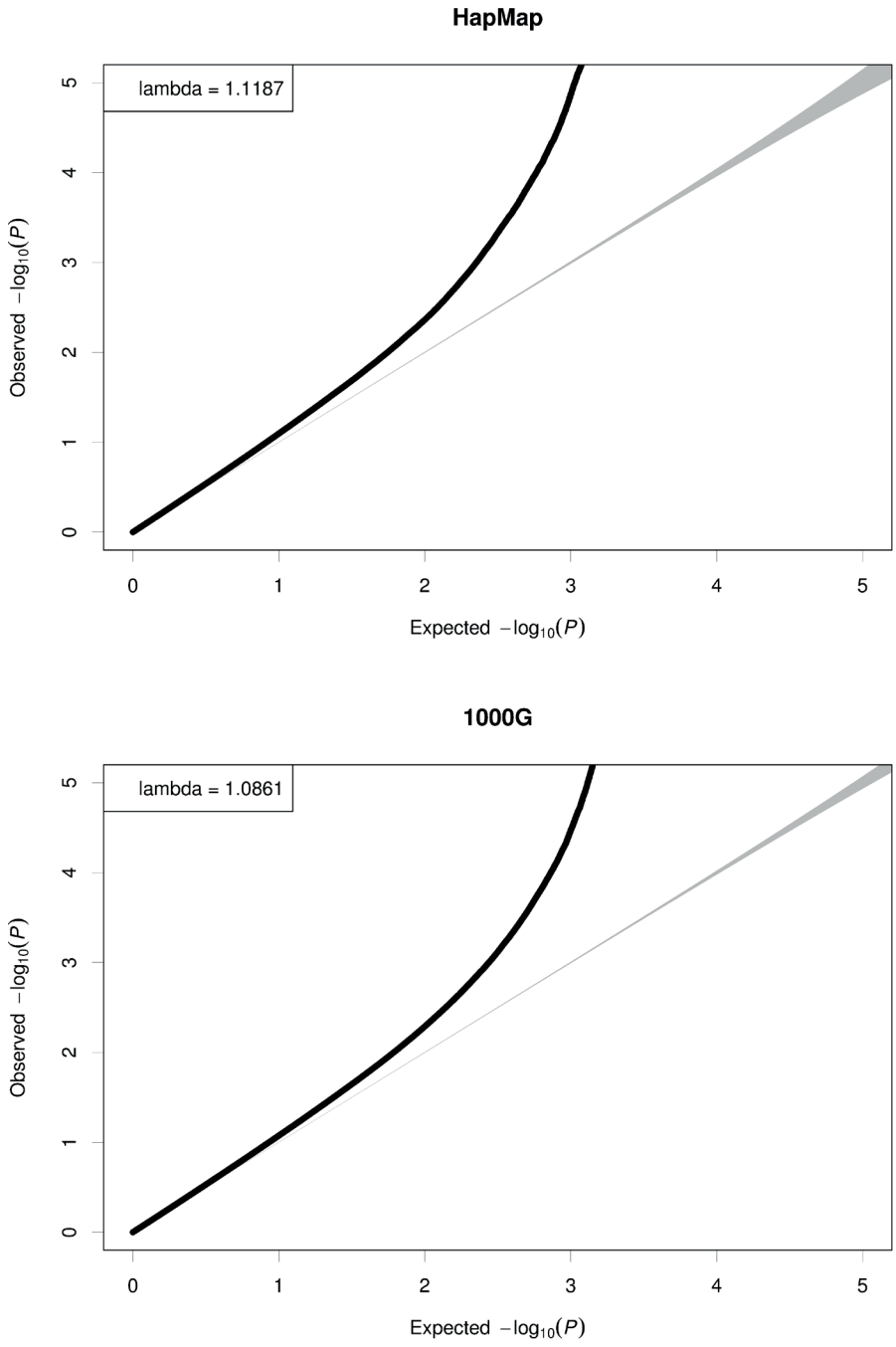
39. Geller F, Feenstra B, Carstensen L, et al. Genome-wide association analyses identify variants in developmental genes associated with hypospadias. Nat Genet. 2014;46:957-963.
40. Feenstra B, Pasternak B, Geller F, et al. Common variants associated with general and MMR vaccine-related febrile seizures. Nat Genet. 2014;46:1274-1282.
41. Germain M, Chasman DI, de Haan H, et al. Meta-analysis of 65,734 Individuals Identifies TSPAN15 and SLC44A2 as Two Susceptibility Loci for Venous Thromboembolism. Am J Hum Genet. 2015;96:532-542.
42. Surakka I, Horikoshi M, Magi R, et al. The impact of low-frequency and rare variants on lipid levels. Nat Genet. 2015.
43. Panagiotou OA, Evangelou E, Ioannidis JP. Genome-wide significant associations for variants with minor allele frequency of 5% or less--an overview: A HuGE review. Am J Epidemiol. 2010;172:869-889.
44. Huffman JE, de Vries PS, Morrison AC, et al. Rare and low-frequency variants and their association with plasma levels of fibrinogen, FVII, FVIII, and vWF. Blood. 2015.

Supplemental Table 1. Characteristics of the included studies and their participants.

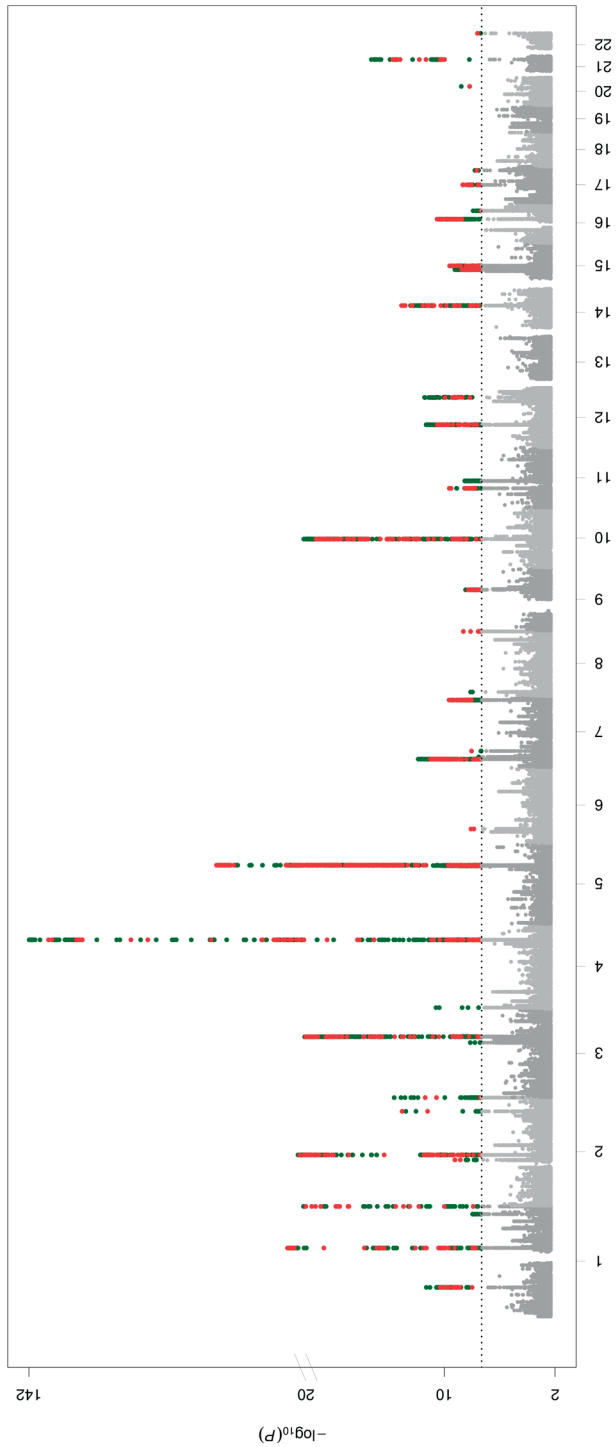
Study	N	Age - Mean (SD)	Male (%)	Fibrinogen- Mean (SD), g/l	BMI (kg/m2) - Mean (SD)	Current Smoker (%)	Coronary Heart Disease (%)	Venous Thrombosis (%)	Type 2 Diabetes (%)
ARIC	8801	54.2 (5.7)	46.9%	3.0 (0.6)	26.97 (4.82)	24.6%	4.9%	2.0%	8.6%
B58C	6085	45.2 (0.4)	49.7%	3.0 (0.6)	27.35 (4.85)	23.5%	NA	NA	1.5%
BMES	2446	66.9(9.2)	42.9%	3.6 (0.9)	25.5 (9.5)	9.6%	NA	NA	10.5%
CHS	3224	72.3 (5.4)	38.9%	3.2 (0.6)	26.3 (4.4)	11.4%	0.0%	4.3%	13.7%
FHS	7022	46.6 (11.5)	46.1%	3.2 (0.7)	27.0 (5.2)	18.9%	10.8%	NA	4.8%
GHS I	2743	55.6 (10.9)	51.0%	3.6 (0.8)	27.20 (4.76)	18.4%	4.5%	4.2%	7.4%
GHS II	1148	55.0 (10.9)	50.0%	3.6 (0.8)	27.26 (4.90)	21.0%	4.5%	3.5%	7.9%
GOYA-Male	1447	45.6 (7.9)	100.0%	3.1 (0.8)	30.4 (6.6)	51.0%	2.6%	NA	5.3%
HCS	2108	66.3 (7.5)	50.0%	3.3 (0.6)	28.78(4.9)	7.7%	10.3%	NA	10.4%
InCHIANTI	1196	68.4 (15.4)	44.4%	3.5 (0.8)	27.17 (4.14)	18.8%	15.6%	NA	11.2%
LBC1921	486	76.1 (0.6)	42.4%	3.6 (0.9)	26.19 (4.11)	6.6%	NA	NA	4.9%
LBC1936	989	69.6 (0.8)	50.8%	3.3 (0.6)	27.83 (4.42)	12.6%	NA	NA	7.7%
LURIC	3057	62.7 (10.6)	70.0%	4.0 (1.1)	27.46 (4.03)	23.1%	79.1%	6.1%	40.4%
NTR	3348	48.0 (14.4)	37.8%	2.8 (0.7)	25.46 (4.04)	30.0%	NA	NA	2.9%
PROCARDIS	3489	61.9 (7.4)	75.5%	3.9 (0.9)	28.3 (4.87)	50.3%	100.0%	NA	15.2%
PROSPER-PHASE	5096	75.3 (3.3)	48.1%	3.6 (0.7)	26.82 (4.19)	26.5%	44.6%	0.0%	10.3%
RS-I-1	2430	70.5 (8.8)	36.3%	2.8 (0.7)	26.5 (3.9)	22.9%	8.4%	NA	11.5%
RS-I-3	2074	71.8 (7.0)	46.0%	4.0 (0.9)	26.8 (3.9)	15.9%	10.8%	NA	14.0%
RS-II	2102	64.8 (8.0)	45.5%	3.9 (0.9)	27.24 (3.98)	19.6%	6.5%	NA	11.7%
SardiNIA	4543	43.2 (17.7)	43.8%	3.3 (0.7)	25.32 (4.61)	19.5%	3.3%	NA	4.5%
SHIP	3841	48.8 (16.1)	48.5%	3.0 (0.7)	27.23 (4.76)	31.4%	5.1%	1.0%	8.2%
TwinsUK	1198	49.1 (12.6)	4.7%	3.0 (0.8)	26.06(4.95)	3.8%	1.2%	NA	31.9%
WGHS	23080	54.2 (7.1)	0.0%	3.6 (0.8)	25.9 (5.0)	11.6%	NA	2.7%	2.5%

Supplemental Table 2. Correlation between the lead variants from the HapMap and 1000G GWA studies.

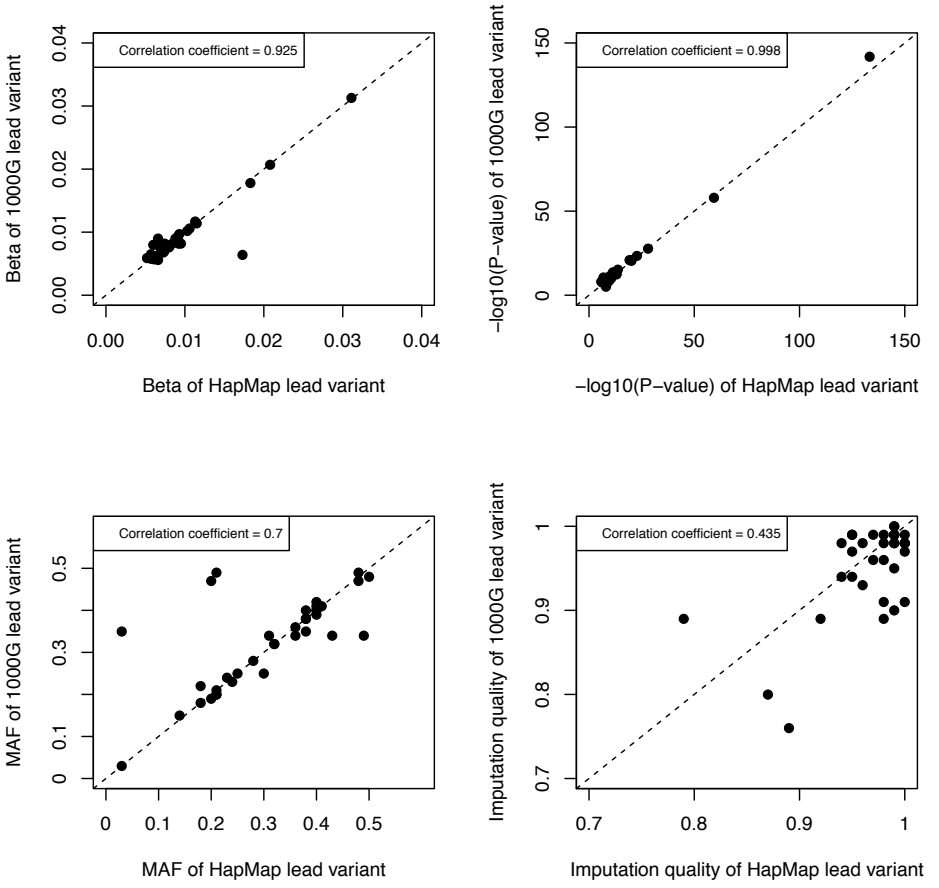
Locus	Lead Variant$_{HapMap}$	Lead Variant$_{1000G}$	Overlapping	Correlation of Imputed Dosages
1p31.3	rs4655582	rs2376015	Yes	0.97
1q21.3	rs8192284	rs61812598	Yes	1
1q42.13	rs10489615	rs10864726	No	0.95
1q44	rs12239046	rs12239046	Yes	1
2q12	rs1558643	rs1558643	Yes	0.97
2q13	rs6734238	rs6734238	Yes	1
2q34	rs715	rs715	Yes	0.91
2q37.3	rs1476698	rs59104589	Yes	0.96
3q21.1	rs16834024	rs1976714	No	0.02
3q22.2	rs548288	rs150213942	Yes	0.82
4p16.3	rs2699429	rs59950280	No	0.18
4q31.3	rs2227401	rs72681211	Yes	0.97
5q31.1	rs1012793	rs1012793	Yes	0.99
6p21.3	rs12528797	rs116134220	No	0.07
7p21.1	rs10950690	rs12699921	Yes	0.95
7p15.3	rs1029738	rs61542988	No	0.11
7q14.2	rs2710804	rs2710804	Yes	1
7q36.1	rs13226190	rs13234724	Yes	0.99
8p23.1	rs7004769	rs7012814	No	0.16
8q24.3	rs7464572	rs11136252	Yes	0.75
9q22.2	rs7873907	rs3138493	Yes	0.92
10q21.3	rs10761756	rs7916868	Yes	0.9
11p12	rs7937127	rs7934094	Yes	0.38
11q12.2	rs7935829	rs11230201	No	0.96
12q13.12	rs1521516	12:51042486	Yes	1
12q24.12	rs3184504	rs4766897	Yes	0.68
14q24.1	rs194741	rs194714	Yes	0.98
15q15.1	rs1703755	rs8026198	Yes	0.93
15q21.2	rs12915052	rs11630054	Yes	0.81
16q12.2	rs12598049	rs6499550	Yes	0.99
16q22.2	rs11864453	rs1035560	Yes	0.99
17q21.2	rs7224737	rs7224737	Yes	1
17q25.1	rs10512597	rs35489971	Yes	1
20q13.12	rs1800961	rs1800961	Yes	1
21q22.2	rs4817986	rs9808651	Yes	0.99
22q13.33	rs6010044	rs75347843	Yes	0.93



Supplemental Figure 1. Quantile-Quantile (QQ) plots comparing the HapMap and 1000G GWA studies.



Supplemental Figure 2. Manhattan plot comparing the HapMap (red) and 1000G (green) GWA studies.*
*Associations from the HapMap GWA study were plotted on top of associations from the 1000G GWA study, and were thus given priority when competing for space in the figure.



Supplemental Figure 3. Comparison of lead variants of the HapMap and 1000G GWA studies of significant loci.

Chapter 2.3

Exome array study of hemostatic factors

Manuscript based on this chapter

Jennifer E. Huffman, Paul S. de Vries, Alanna C. Morrison, Maria Sabater-Lleal, Tim Kacprowski, Paul L. Auer, Jennifer A. Brody, Daniel I. Chasman, Ming-Huei Chen, Xiuqing Guo, Li-An Lin, Riccardo E. Marioni, Martina Müller-Nurasyid, Lisa R. Yanek, Nathan Pankratz, Megan L. Grove, Moniek P.M. de Maat, Mary Cushman, Kerri L. Wiggins, Lihong Qi, Bengt Sennblad, Sarah E. Harris, Ozren Polasek, Helene Riess, Fernando Rivadeneira, Lynda M. Rose, Anuj Goel, Kent D. Taylor, Alexander Teumer, André G. Uitterlinden, Dhananjay Vaidya, Jie Yao, Weihong Tang, Daniel Levy, Melanie Waldenberger, Diane M. Becker, Aaron R. Folsom, Franco Giulianini, Andreas Greinacher, Albert Hofman, Chiang-Ching Huang, Charles Kooperberg, Angela Silveira, John M. Starr, Konstantin Strauch, Rona J. Strawbridge, Alan F. Wright, Barbara McKnight, Oscar H. Franco, Neil Zakai, Rasika A. Mathias, Bruce M. Psaty, Paul M. Ridker, Geoffrey H. Tofler, Uwe Völker, Hugh Watkins, Myriam Fornage, Anders Hamsten, Ian J. Deary, Eric Boerwinkle, Wolfgang Koenig, Jerome I. Rotter, Caroline Hayward, Abbas Dehghan, Alex P. Reiner, Christopher J. O'Donnell, and Nicholas L. Smith for the CHARGE Hemostasis Working Group.

Rare and low-frequency variants and their association with plasma levels of fibrinogen, FVII, FVIII, and vWF.

Blood. 2015; 126(11):e19-29.

ABSTRACT

Background: Fibrinogen, coagulation factor VII (FVII), factor VIII (FVIII), and its carrier von Willebrand factor (VWF) play key roles in hemostasis. Previously identified common variants explain only a small fraction of the trait heritabilities and additional variation may be explained by associations with rarer variants with larger effects.

Methods: The aim of this study was to identify low-frequency (minor allele frequency [MAF] ≥0.01 and <0.05) and rare (MAF <0.01) variants that influence plasma concentrations of these 4 hemostatic factors by meta-analyzing exome chip data from up to 76,000 participants of 4 ancestries.

Results: We identified 12 novel associations of low-frequency (n=2) and rare (n=10) variants across the fibrinogen, FVII, FVIII, and VWF traits that were independent of previously identified associations. Novel loci were found within previously reported genes and had effect sizes much larger than and independent of previously identified common variants. In addition, associations at *KCNT1*, *HID1,* and *KATNB1* identify new candidate genes related to hemostasis for follow-up replication and functional genomic analysis.

Conclusions: Newly identified low-frequency and rare-variant associations accounted for modest amounts of trait variance and therefore are unlikely to increase predicted trait heritability but provide new information to understanding individual variation in hemostasis pathways.

Introduction

Fibrinogen, coagulation factor VII (FVII), factor VIII (FVIII) and its carrier protein von Willebrand factor (VWF) play key roles in hemostasis. Plasma levels of these hemostatic factors are associated with risk of arterial and venous thrombosis, and fibrinogen is also a marker of inflammation.[1-6] Previous genome-wide association studies (GWAS) interrogated mainly common genetic variation and identified variants of modest effect across these phenotypes[4, 7-14] with the largest studies identifying 23 loci for fibrinogen,[9] 5 each for FVII[13] and FVIII[13] and 8 for VWF[13]. Nonetheless, the associated variants still explain little of the trait heritabilities.[9, 12, 15] An additional proportion of the missing heritability may be attributed to association with rare variants, which are not captured by the conventional genome-wide marker arrays or imputation panels that have been used for GWAS.[15] In addition, the investigation of rare genetic variation is important to understanding individual variation in the biology underlying hemostasis pathways.

The aim of this study was to identify low-frequency and rare variants, analyzed individually or at the level of the gene, that influence plasma concentrations of fibrinogen, FVII, FVIII, and VWF. To this end, we meta-analyzed phenotype-genotype associations of low-frequency (minor allele frequency [MAF] = 0.01-0.05) and rare (MAF<0.01) exonic variants within 76,000 individuals of European, African, Hispanic, or East-Asian ancestry from 16 studies within the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium.[16] We restricted our analyses to variants which were predicted to alter the coding sequence of the gene product in order to enhance the likelihood of identifying causal variants and to reduce the multiple testing burden.

METHODS

Setting and participating cohorts

This study was organized within the CHARGE Consortium Hemostasis Working Group and included 16 cohorts of European (EUR), African (AFR), East-Asian (ASI), or Hispanic (HIS) ancestry. Descriptions and ancestry composition of participating cohorts are found in the **Supplemental Information (Sections I&II)**.

Hemostatic factors

Hemostasis phenotypes included plasma measures of fibrinogen, FVII, FVIII, and VWF. Fibrinogen (g/l) was available in all 16 studies; FVII activity (% or IU/ml*100) and FVII antigen (% or IU/ml*100) were available in 7 studies; FVIII activity (% or IU/

Table 1. Study participant characteristics and phenotype assay or measure.

Cohort	Ancestry	N	% Female	Age (yrs), mean	Trait mean (SD)	Assay/ Measure
Fibrinogen (g/l)						
ARIC[49]	EUR	10,757	53.1	54.3	2.90 (1.21)	Clauss
	AFR	3,643	61.9	53.5	3.13 (1.23)	
CARDIA[50]	EUR	2,041	52.5	30.5	2.51 (1.23)	immunonephelometry
	AFR	1,709	56.9	29.4	2.66 (1.23)	
CHS[51]	EUR	4,034	56.2	72.8	3.13 (1.22)	Clauss
	AFR	757	62.2	72.7	3.35 (1.23)	
FHS[52, 53]	EUR	6,711	54.3	46.0	3.24 (0.68)	Clauss
GeneSTAR[54]	EUR	1,091	51.2	41.2	3.51 (0.98)	modified Clauss
	AFR	641	61.9	40.6	3.80 (1.12)	
KORA S4[55, 56]	EUR	2,687	53.1	47.9	2.60 (0.58)	immunonephelometry
Korcula[57]	EUR	748	64.3	56.4	4.55 (1.52)	Clauss
LBC1921[58, 59]	EUR	466	57.4	79.1	3.59 (0.86)	Clauss
LBC193[58, 60]	EUR	973	49.2	69.6	3.27 (0.63)	Clauss
MESA[61]	EUR	2,483	52.1	62.7	3.35 (0.7)	immunonephelometry on the BNII nephelometer
	AFR	1,638	53.8	62.2	3.60 (0.79)	
	ASI	764	50.8	62.4	3.29 (0.61)	
	HIS	1,431	51.5	61.0	3.59 (0.75)	
PROCARDIS[62]	EUR	1,404	36.8	60.9	4.06 (0.96)	immunonephelometric
RS-I-1[63-65]	EUR	1,114	59.0	70.2	2.70 (1.26)	Prothrombin time
RS-I-3[63-65]	EUR	972	46.7	72.4	3.96 (0.89)	Prothrombin time
SCARF[66]	EUR	697	17.5	53.2	3.47 (0.79)	immunonephelometric
SHIP[67]	EUR	5,940	52.3	47.9	2.99 (0.71)	Clauss
WGHS[7, 68]	EUR	22,411	100	54.7	3.59 (0.78)	Mass-based immunoturbidimetric assay
WHI[69-71]	EUR	1,204	100	69.6	3.06 (0.86)	Clauss
Factor VII						
ARIC	EUR	10,544	52.9	54.3	118.3(26.7)	clotting assay (% activity)
	AFR	3,574	61.9	53.6	116.7 (28.4)	
CARDIA	EUR	997	52.5	30.6	83.7 (21.5)	clotting assay (% activity)
	AFR	637	55.6	29.2	84.2 (26.2)	
CHS	EUR	4,063	56.2	72.8	125.9 (29.5)	clotting assay (% activity)
	AFR	760	62.1	72.6	113.0 (26.4)	
FHS	EUR	2,620	55.3	53.9	100.3 (16.3)	ELISA (% antigen)
RS-I	EUR	670	59.0	70.6	107.5 (19.1)	clotting assay (% activity)
SCARF	EUR	698	17.5	53.2	139.9 (35.8)	ELISA (% antigen)
WHI	EUR	809	100	69.9	146.0 (52.5)	clotting assay (% activity)

Table 1. (continued)

Cohort	Ancestry	N	% Female	Age (yrs), mean	Trait mean (SD)	Assay/ Measure
Factor VIII						
ARIC	EUR	10,708	53.0	54.3	124.1 (30.6)	clotting assay (% activity)
	AFR	3,618	61.7	53.5	144.8 (41.7)	
CARDIA	EUR	998	52.6	30.6	89.8 (31.7)	clotting assay (% activity)
	AFR	632	55.6	29.2	103.5 (38.7)	
CHS	EUR	4,009	56.2	72.8	120.8 (36.7)	clotting assay (% activity)
	AFR	191	63.9	72.6	138.3 (43.9)	
MESA	EUR	2,483	52.1	62.7	156.9 (64.6)	clotting assay (% activity)
	AFR	1,638	53.8	62.2	178.0 (74.6)	
	ASI	764	7.7	62.4	157.9 (57.2)	
	HIS	1,418	51.5	61.0	161.8 (63.4))	
RS-I	EUR	1,832	52.0	68.6	115.7 (46.1)	clotting assay (% activity)
von Willebrand Factor						
ARIC	EUR	10,736	53.1	54.3	110.7 (39.1)	ELISA (% antigen)
	AFR	3,625	61.8	53.5	131.4 (51.1)	
CARDIA	EUR	1,002	52.6	30.6	89.9 (36.4)	ELISA (% antigen)
	AFR	636	55.7	29.2	94.3 (44.4)	
FHS	EUR	2,621	55.3	53.9	125.3 (45.0)	ELISA (% antigen)
GeneSTAR	EUR	991	52.5	42.6	78.7 (46.1)	ELISA (% antigen)
	AFR	582	62.2	42.6	76.8 (42.5)	
LBC1921	EUR	150	57.3	86.6	149.7 (45.9)	ELISA (% antigen)
LBC1936	EUR	706	47.9	72.5	122.6 (37.8)	ELISA (% antigen)
MESA	EUR	443	54.7	62.7	135.2 (54.5)	ELISA (% antigen)
	AFR	193	64.8	62.2	156.1 (64.8)	
RS-I	EUR	1,587	49.9	73.1	135.9 (54.1)	ELISA (% antigen)

EUR=European; AFR=African; ASI = East-Asian; HIS=Hispanic; SD=standard deviation. Full cohort descriptions can be found in the Supplemental Material. FVII was measured as % antigen for FHS and SCARF and % activity for all other studies.

ARIC = Atherosclerosis Risk in Communities Study; CARDIA = The Coronary Artery Risk Development in Young Adults Study; CHS = Cardiovascular Health Study; FHS = Framingham Heart Study; GeneSTAR = Genetic Study of Atherosclerosis Risk; KORA S4 = Kooperative Gesundheitsforschung in der Region Augsburg; Korcula = CROATIA-Korcula Study; LBC1921 = Lothian Birth Cohort 1921; LBC1936 = Lothian Birth Cohort 1936; MESA = Multi-Ethnic Study of Atherosclerosis; PROCARDIS = Precocious Coronary Artery Disease Study; RS-I = Rotterdam Study; SCARF = Stockholm Coronary Artery Risk Factors; SHIP = The Study of Health in Pomerania; WGHS = Women's Genome Health Study; WHI = Women's Health Initiative.

ml*100) was available in 5 studies; and VWF antigen (% or IU/ml*100) was available in 8 studies. Methods used by each study are noted in **Table 1**.

Genotype calling and quality control

Fourteen studies were genotyped using the HumanExome BeadChip v1.0 (Illumina, Inc., San Diego, CA) whereas one was genotyped using v1.1 and another using v1.2 of the BeadChip. Variant calling and quality control procedures are described in the **Supplemental Information (Section IIIa)** and in previously published articles.[17, 18] Prior to analysis, individual studies recoded variants to additive coding using the minor allele derived from the CHARGE joint calling.

Statistical analysis

Each study natural-log transformed fibrinogen measures. For untransformed FVII, FVIII, or VWF, participants with values 3 standard deviations above or below the population mean were removed prior to cohort-level analysis. Study-specific regression analyses were adjusted for sex, age, study design variables, and population substructure using principal components. MAF thresholds (described below) were defined using the ancestry-specific allele frequencies derived from the CHARGE joint calling.[17] Variant annotation was performed centrally within CHARGE using dbNSFP v2.0.[19, 20] All association analyses were performed using the R package seqMeta (http://cran.r-project.org/web/packages/seqMeta/index.html). A table detailing the genotyping chip and version of statistical software used by each study is found in **Supplemental Table S1**.

We investigated low-frequency and rare variants individually using standard single-variant association analyses. From among the functional variants on the array (defined as missense, stop-gain, stop-loss, or splice site changes), we selected variants with a MAF <5% and an expected minor allele count (eMAC) of greater than or equal to 5 in the total meta-analysis sample for single-variant association of autosomal chromosomes. Since commonly occurring variation on the X chromosome had not previously been investigated for some of the phenotypes, no upper MAF threshold was used when testing for associated variants on this chromosome. The Y and mitochondrial chromosomes were not interrogated. Bonferroni-corrected *P*-value thresholds of statistical significance were based on the number of single-variant tests performed and varied by ancestry: 2.5×10^{-7} (all cohorts - ALL), 2.6×10^{-7} (EUR+AFR cohorts), 2.9×10^{-7} (EUR-only), 3.3×10^{-7} (AFR-only), 1.7×10^{-6} (ASI-only) and 4.7×10^{-7} (HIS-only) (see **Supplemental Information Section IIIb**).

Analytical methods that aggregate the effect of multiple rare variants across a gene were used to test for association. This resulted in a *P*-value for a gene rather than for a single-variant. Both unidirectional and random effects tests were used; the former

is more powerful when rare variant effects within a region are in the same direction and the latter is more powerful when rare variants affect a phenotype in opposite directions or when many variants have null effects.

All gene-based tests were again restricted to include only functional single nucleotide variants. Random effects (Sequence Kernel Association Test [SKAT][21]) and unidirectional[22] (T5) gene tests were performed using only variants with a MAF <5%. The T5 burden was defined as the total number of rare alleles among variants in the gene with a MAF <5%.[23] All genes were required to contain more than 1 variant to be included in the analysis and have a cumulative minor allele frequency greater than the frequency such that the meta-analysis sample size would have an eMAC of 5. A Bonferroni-corrected, gene-based *P*-value threshold of 1.9×10^{-6} was used for gene-based tests (0.05/26,965 genes).

Meta-analyses of single-variant and gene-based analyses were performed using seqMeta v1.3. The primary analysis was to meta-analyze all ancestries together with a secondary set of ancestry-specific analyses performed to complement and inform the results of the primary analysis. All significant non-synonymous variants were re-annotated using an updated version of dbNSFP (v.3.0).[19, 20, 24, 25]

To test for independence of the new discoveries from variants previously demonstrated to be associated with the phenotype at that locus, conditional analyses were performed and meta-analyzed. These analyses were undertaken by EUR and AFR ancestry cohorts only and in some cases, the SNPs conditioned on differed between ancestry groups, generally due to the conditional SNP being monomorphic in 1 population. A description of conditional analyses undertaken is included in **Supplemental Table S3**.

RESULTS

Single-variant and gene-based tests for all 4 hemostatic factors identified significantly associated loci for all phenotypes. The Q-Q plots for all association analyses are found in **Supplemental Figures S1-S3**. Functional annotations for all significant non-synonymous single variants can be found in **Supplemental Table S2**.

Fibrinogen

Exome array genotyping and fibrinogen measures were available from 76,316 participants across 16 cohorts and 4 ancestry groups.

Table 2. Single-Variant Meta-Analysis Results for Hemostatic Factors Fibrinogen, Factor VII, Factor VIII and von Willebrand Factor*.

Variant	AA Change	Gene	Ancestry	N	MAF	BETA	*P*-value
Fibrinogen							
rs201909029 (new)	K178N (K148N)	*FGB*	ALL	76,316	7.7×10^{-04}	-0.139	$\mathbf{3.2\times10^{-13}}$
			EUR	65,733	8.8×10^{-04}	-0.139	$\mathbf{5.2\times10^{-13}}$
			AFR	8,388	6.0×10^{-05}	-0.163	4.3×10^{-01}
			ASI	764	0	NA	NA
			HIS	1,431	3.5×10^{-04}	-0.117	5.5×10^{-01}
rs6054	P265L (P235L)	*FGB*	ALL	76,316	4.2×10^{-03}	-0.111	$\mathbf{1.8\times10^{-43}}$
			EUR	65,733	4.7×10^{-03}	-0.111	$\mathbf{3.7\times10^{-42}}$
			AFR	8,388	1.2×10^{-03}	-0.104	2.6×10^{-02}
			ASI	764	1.3×10^{-03}	-0.130	3.0×10^{-01}
			HIS	1,431	0	NA	NA
rs145051028 (new)	S245F (S219F)	*FGG*	ALL	76,316	1.6×10^{-04}	-0.239	$\mathbf{4.8\times10^{-09}}$
			EUR	65,733	0	NA	NA
			AFR	8,388	1.5×10^{-03}	-0.239	$\mathbf{4.8\times10^{-09}}$
			ASI	764	0	NA	NA
			HIS	1,431	0	NA	NA
rs148685782	A108G (A82G)	*FGG*	ALL	76,316	3.3×10^{-03}	-0.238	$\mathbf{9.2\times10^{-152}}$
			EUR	65,733	3.8×10^{-03}	-0.239	$\mathbf{2.3\times10^{-150}}$
			AFR	8,388	4.2×10^{-04}	-0.165	3.4×10^{-02}
			ASI	764	0	NA	NA
			HIS	1,431	3.5×10^{-04}	-0.347	7.7×10^{-02}
rs10479001	A225V	*PDLIM4*	ALL	76,316	5.5×10^{-02}	0.013	$\mathbf{1.3\times10^{-08}}$
			EUR	65,733	4.5×10^{-02}	0.018	$\mathbf{4.3\times10^{-11}}$
			AFR	8,388	1.4×10^{-01}	-0.001	8.3×10^{-01}
			ASI	764	0	NA	NA
			HIS	1,431	5.5×10^{-02}	0.019	2.3×10^{-01}
rs1800961	T117I T139I T169I	*HNF4A*	ALL	76,316	2.7×10^{-02}	-0.020	$\mathbf{2.3\times10^{-10}}$
			EUR	65,733	3.0×10^{-02}	-0.020	$\mathbf{5.5\times10^{-10}}$
			AFR	8,388	5.9×10^{-03}	0.012	5.5×10^{-01}
			ASI	764	1.1×10^{-02}	-0.031	4.8×10^{-01}
			HIS	1,431	4.2×10^{-02}	-0.038	4.0×10^{-02}
rs151272083 (new)	R865Q R877Q R891Q R910Q	*KCNT1*	ALL	76,316	2.2×10^{-03}	0.007	5.3×10^{-01}
			EUR	65,733	2.4×10^{-03}	0.017	1.3×10^{-01}
			AFR	8,388	7.2×10^{-04}	-0.330	$\mathbf{2.7\times10^{-07}}$
			ASI	764	0	NA	NA
			HIS	1,431	0	NA	NA

Table 2. (continued)

Variant	AA Change	Gene	Ancestry	N	MAF	BETA	*P*-value
rs141869748 (new)	I193T I421T	*HID1*	ALL	76,316	1.6×10^{-04}	-0.216	4.2×10^{-07}
			EUR	65,733	0	NA	NA
			AFR	8,388	1.3×10^{-03}	-0.252	**4.0×10^{-08}**
			ASI	764	0	NA	NA
			HIS	1,431	1.1×10^{-03}	0.008	9.4×10^{-01}
Factor VII							
rs150525536 (new)	R117Q R70Q R139Q	*F7*	ALL	25,372	9.5×10^{-04}	-31.44	**1.8×10^{-17}**
			EUR	20,401	9.8×10^{-05}	-13.92	2.2×10^{-01}
			AFR	4,971	4.4×10^{-03}	-33.56	**9.7×10^{-18}**
rs121964926 (new)	R342Q R295Q R364Q	*F7*	ALL	25,372	1.2×10^{-03}	-25.02	**1.3×10^{-14}**
			EUR	20,401	4.2×10^{-04}	-0.52	9.3×10^{-01}
			AFR	4,971	4.4×10^{-03}	-38.08	**2.8×10^{-21}**
rs3093248 (new)	E423K E376K E445K	*F7*	ALL	25,372	7.5×10^{-04}	-22.00	2.8×10^{-07}
			EUR	20,401	2.5×10^{-05}	-62.77	2.3×10^{-02}
			AFR	4,971	3.7×10^{-03}	-20.99	1.3×10^{-06}
Factor VIII							
rs7962217	G2705R	*VWF*	ALL	28,291	4.6×10^{-02}	5.16	**2.5×10^{-13}**
			EUR	20,030	5.5×10^{-02}	4.84	**4.0×10^{-11}**
			AFR	6,079	1.6×10^{-02}	8.58	7.9×10^{-03}
			ASI	764	7.2×10^{-03}	17.63	3.0×10^{-01}
			HIS	1,418	5.8×10^{-02}	10.21	2.7×10^{-02}
rs41276738 (new)	R854Q	*VWF*	ALL	28,291	4.0×10^{-03}	-16.89	**2.2×10^{-13}**
			EUR	20,030	5.3×10^{-03}	-15.96	**9.2×10^{-12}**
			AFR	6,079	9.9×10^{-04}	-49.57	3.8×10^{-04}
			ASI	764	0	NA	NA
			HIS	1,418	1.1×10^{-03}	-19.47	5.5×10^{-01}
rs141041254 (new)	E2377K	*STAB2*	ALL	28,291	8.7×10^{-04}	26.81	**2.1×10^{-08}**
			EUR	20,030	1.2×10^{-03}	28.06	**7.6×10^{-09}**
			AFR	6,079	2.5×10^{-04}	-11.70	6.6×10^{-01}
			ASI	764	0	NA	NA
			HIS	1,418	0	NA	NA
rs1800291	D1260E	*F8*	ALL	28,291	2.7×10^{-01}	-1.73	**8.2×10^{-08}**
			EUR	20,030	1.7×10^{-01}	-2.15	**5.0×10^{-09}**
			AFR	6,079	3.5×10^{-01}	-0.54	4.5×10^{-01}
			ASI	764	4.7×10^{-02}	7.29	1.8×10^{-01}
			HIS	1,418	2.5×10^{-01}	0.28	8.9×10^{-01}

Table 2. (continued)

Variant	AA Change	Gene	Ancestry	N	MAF	BETA	*P*-value
rs142508811 (new)	D413D D410D (predicted to alter splicing)	*KATNB1*	ALL	28,291	2.7×10^{-04}	39.36	4.8×10^{-04}
			EUR	20,030	1.8×10^{-04}	1.08	9.4×10^{-01}
			AFR	6,079	6.6×10^{-04}	86.35	**2.8×10^{-07}**
			ASI	764	0	NA	NA
			HIS	1,418	0	NA	NA
von Willebrand Factor							
rs141041254 (new)	E2377K	*STAB2*	ALL	23,272	8.2×10^{-04}	33.65	**2.4×10^{-07}**
			EUR	18,236	9.9×10^{-04}	35.21	**1.1×10^{-07}**
			AFR	5,036	2.0×10^{-04}	-11.56	7.5×10^{-01}

*Table reports only SNPs that were still significant after conditional analyses. AA Change = amino acid change of SNP; amino acid position in brackets is for the mature protein for *FGB* (position-30) and *FGG* (position-26) ** ALL = all ancestries (only EUR+AFR for FVII and vWF); EUR=European-only; AFR=African-only; ASI = East-Asian-only; HIS=Hispanic-only; MAF = minor allele frequency from CHARGE joint calling. SNPs achieving genome-wide significance threshold (p= 2.50×10^{-07} (ALL), 2.88×10^{-07} (EUR), 3.30×10^{-07} (AFR), 1.70×10^{-06} (ASI) and 4.67×10^{-07} (HIS)) are bolded.

Fibrinogen: single-variant testing

Associations for 6 rare or low-frequency variants that exceeded array-wide significance were observed within 4 genes: 2 fibrinogen structural genes, (*FGB* and *FGG*) and 2 other genes, *PDLIM4* and *HNF4A* (**Table 2**, **Supplemental Figure S4**).

Two rare variants within *FGB*, rs6054 (Pro235Leu, MAF=0.0042, p=1.8×10^{-43}) and rs201909029 (Lys148Asn, MAF=0.00077, p=3.2×10^{-13}) were associated with lower fibrinogen levels. Both variants had similar effect sizes (-0.111 and -0.139 ln(g/l)) and the magnitude and direction of the association was similar for both variants in all ancestry groups (**Table 2**). Fibrinogen levels were lower by 10.5% and 13.0%, respectively, per copy of the minor allele when other model factors are fixed (see **Supplemental Information [Section IIIc]**). The rs6054 association has been previously reported[10] but the rs201909029 variant association is new. Two rare variants within *FGG* were also associated with fibrinogen levels: rs148685782 (Ala82Gly, MAF=0.0033, p=9.2×10^{-152}) and rs145051028 (Ser219Phe, MAF=0.00016, p=4.8×10^{-09}). In this study, rs148685782 had an effect size of -0.238 ln(g/l), which translates to a 21.1% lower fibrinogen level per copy of the minor allele. The direction and magnitude of the effect was similar across all ancestry groups where it was polymorphic (**Table 2**). The *FGG* Ala82Gly variant has previously been associated with low plasma fibrinogen levels.[26-28] The rs145051028 variant has an effect size of -0.239 ln(g/l) or a 21.3% lower level of fibrinogen per copy of the minor allele and was only polymorphic in AFR-ancestry cohorts. This association has not been previously reported.

In order to determine if the newly and previously identified associations within the fibrinogen gene cluster were independent of one another, 3 separate conditional analysis were undertaken: (1) adjustment for previously associated common variants in *FGB* (rs4220 and rs6056),[10] (2) adjustment for the significant rare variants in *FGG* (rs148685782 and rs145051028 (AFR-only)) and (3) adjustment for the most significant rare variant in *FGB* (rs6054) (**Supplemental Table S3**). Results demonstrated independence of all variants from one another (**Table 4**). In total, the rare variants within the fibrinogen gene cluster explained ~1.3% and ~0.12% of the trait variance in the EUR and AFR populations, respectively. The majority of the variance in the EUR population (~0.9%) was attributed to *FGG* rs148685782.

The association of low-frequency variants within the *PDLIM4* and *HNF4A* genes support prior reported associations. The *PDLIM4* SNP was in high linkage disequilibrium (LD) with a previously reported *IRF1* SNP rs11242111 (R^2=0.85, D'=1 within 1000Genomes Map Pilot 1 v.3, CEU) on chromosome 5[9] and the *HNF4A* SNP, rs1800961, has been previously reported although it was just below the genome-wide significance threshold in that study.[10] The effect size for each was 10-fold smaller than those for *FGB* and *FGG*.

Single variants in *KCNT1* and in *HID1*, located in regions not previously reported to be associated with fibrinogen levels, reached array-wide significance in the exploratory AFR-only analysis of fibrinogen (**Table 2**, **Supplemental Figure S4**). *KCNT1* rs151272083 (MAF=0.00072, $p=2.7\times10^{-07}$) codes for an Arg891Gln change (also reported as the same amino acid change at position 865, 877, or 910 due to transcriptional variation) and was predicted to decrease fibrinogen by 0.330 ln(g/l) or approximately 28.1% per copy of the minor allele in the AFR population. This SNP was also polymorphic in EUR populations but did not reach statistical significance and the estimated effect was 20-fold smaller (β=0.017, p=0.13). *HID1* rs141869748 (Ile421Thr / Ille193Thr, MAF=0.0013, $p=4.0\times10^{-08}$) was associated with 0.252 ln(g/l) lower fibrinogen (22.3% decrease per copy of the minor allele) in the AFR population. This SNP was monomorphic in the EUR and ASI populations and its estimated effect in the HIS population, although small, was not in the same direction despite a similar MAF (MAF=0.0011, β=0.008, p=0.94).

When we further explored these characteristics of the novel associations in the AFR population we found no evidence for heterogeneity across studies (p_{HET}=0.07 (rs151272083) and 0.91 (rs141869748), **Supplemental Figure S5**) and we confirmed that carriers of the variant allele in AFR cohorts had lower mean plasma fibrinogen levels than non-carriers (**Supplemental Table S5**). The variants explained approximately 0.7% (rs151272083) and 0.4% (rs141869748) of the trait variance.

Table 3. Gene-Based Test Meta-Analysis Results for Hemostatic Factors Fibrinogen, Factor VII, Factor VIII and von Willebrand Factor.

			P-value	
Gene	**Ancestry**	**N**	**SKAT5**	**T5**
Fibrinogen				
	ALL	76,316	**1.25×10^{-45}**	**5.59×10^{-32}**
	EUR	65,733	**2.03×10^{-44}**	**1.16×10^{-36}**
FGB	AFR	8,388	4.50×10^{-01}	5.60×10^{-01}
	ASI	764	3.00×10^{-01}	2.98×10^{-01}
	HIS	1,431	9.37×10^{-01}	9.39×10^{-01}
	ALL	76,316	**6.90×10^{-99}**	**7.25×10^{-31}**
	EUR	65,733	**2.49×10^{-111}**	**1.35×10^{-61}**
FGG	AFR	8,388	**2.82×10^{-09}**	3.18×10^{-04}
	ASI	764	NA	NA
	HIS	1,431	5.65×10^{-01}	8.18×10^{-01}
Factor VII				
	ALL	25,372	**6.24×10^{-35}**	**2.36×10^{-37}**
F7	EUR	20,401	6.71×10^{-05}	**8.21×10^{-07}**
	AFR	4,971	**1.83×10^{-35}**	**3.03×10^{-32}**
Factor VIII				
	ALL	28,291	**5.10×10^{-18}**	**5.71×10^{-30}**
	EUR	20,030	**1.90×10^{-13}**	**1.61×10^{-17}**
ABO	AFR	6,079	1.91×10^{-03}	3.44×10^{-04}
	ASI	764	8.37×10^{-01}	9.56×10^{-01}
	HIS	1,418	3.48×10^{-01}	2.89×10^{-02}
	ALL	28,291	**5.21×10^{-21}**	**1.61×10^{-06}**
	EUR	20,030	**2.20×10^{-07}**	1.47×10^{-04}
VWF	AFR	6,079	8.13×10^{-03}	4.09×10^{-01}
	ASI	764	1.41×10^{-01}	8.01×10^{-01}
	HIS	1,418	2.27×10^{-01}	4.07×10^{-01}
	ALL	28,291	**3.49×10^{-07}**	2.56×10^{-03}
	EUR	20,030	**6.49×10^{-07}**	5.83×10^{-03}
STAB2	AFR	6,079	1.44×10^{-01}	8.23×10^{-02}
	ASI	764	1.78×10^{-01}	9.55×10^{-02}
	HIS	1,418	9.13×10^{-01}	3.09×10^{-01}
von Willebrand Factor				
	ALL	23,272	**4.07×10^{-19}**	**3.69×10^{-29}**
ABO	*EUR*	18,236	**2.84×10^{-13}**	**4.17×10^{-18}**
	AFR	5,036	2.89×10^{-03}	3.01×10^{-04}
	ALL	23,272	**2.99×10^{-07}**	8.07×10^{-03}
STAB2	*EUR*	18,236	**1.53×10^{-06}**	1.66×10^{-01}
	AFR	5,036	7.24×10^{-04}	6.46×10^{-02}

Genes achieving genome-wide significance ($p<1.85\times10^{-06}$) are bolded. N=number of participants included in analysis; ALL = all ancestries (only EUR+AFR for FVII and vWF); EUR=European-only; AFR=African-only; ASI = East-Asian-only; HIS=Hispanic-only.

Fibrinogen: gene-based testing

SKAT and T5 tests yielded gene-level associations with all 4 genes described above: *FGB, FGG, PDLIM4*, and *HNF4A (***Table 3**). Gene-based testing did not identify other genes contributing to plasma-level variation in fibrinogen.

Factor VII

Exome array genotyping and coagulation FVII measures were available from 25,372 participants across 7 studies comprised of EUR and AFR participants.

FVII: single-variant testing

Five exome-wide significant coding rare-variant associations were observed in *F7* as well as nearby genes *MCF2L* and *PROZ*. When conditioning on a common, previously-reported coding variant rs6046 in *F7,*[13] 3 previously unreported rare variants within *F7* remained exome-wide significant whereas the variants in *MCF2L* and *PROZ* were no longer significant (**Table 4**). The minor alleles of *F7* variants rs150525536 (Arg117Gln, MAF=0.0010, p_{cond}= 1.0×10^{-22}), rs121964926 (Arg342Gln, MAF=0.0015, p_{cond}= 1.5×10^{-14}), and rs3093248 (Glu423Lys, MAF=0.00085, p_{cond} = 1.4×10^{-07}) were all associated with significantly lower plasma FVII levels (**Table 2**, **Supplemental Figure S4**). The three variants explained ~0.06% of the trait variance in EUR participants and 4.5% of the trait variance in AFR participants. For all identified variants, the MAF was lower in EUR than in AFR population but the direction of effect was the same even if the magnitude varied (**Table 2**). Sensitivity analyses that removed the 2 studies with FVII antigen rather than activity measured did not impact the findings.

FVII: gene-based testing

SKAT and T5 tests yielded gene-level associations with *F7* (**Table 3**). No other gene was associated with plasma levels of FVII.

Factor VIII and von Willebrand factor

As reported by our prior GWAS, association results for plasma levels of FVIII and VWF were similar so will be presented together.[13] FVIII measures were available from 28,291 participants from 5 cohorts across all ancestry groups while VWF was available in 23,272 EUR and AFR participants from 8 cohorts.

FVIII and VWF: single-variant testing

Genome-wide significant rare and low-frequency variants are presented in **Table 2** and cluster plots for the associated SNPs are found in **Supplemental Figure S4**. Five novel low-frequency and rare variant associations were found for FVIII and VWF levels, most within loci with previous FVIII/VWF associations.[13]

Low-frequency variant rs7962217 (Gly2705Arg, MAF=0.046, $p = 2.5\times10^{-13}$) and rare variant rs41276738 (Arg854Gln, MAF=0.0040, $p = 2.2\times10^{-13}$) in *VWF* were significantly associated with lower plasma levels of FVIII but not VWF ($p = 0.96$, $p = 0.03$, respectively). Only the association of rs7962217 has been reported previously[29] and conditioning on the most significant common *VWF* variants associated with FVIII levels (rs1063856 and rs62643635[13]) did not materially alter these results (**Table 4**). Ancestry-specific analyses yielded effects with the same direction and similar magnitudes although the MAF varied by up to 2 orders of magnitude (**Table 2**).

A single rare variant in *STAB2*, rs141041254 (Glu2377Lys, MAF=0.00087), was significantly associated with FVIII ($p=2.1\times10^{-08}$) and VWF levels ($p=2.4\times10^{-07}$) and the new signal remained unchanged when adjusting for rs2271637, the most highly associated *STAB2* common-variant on the array. In the 2 ancestries in which the variant was polymorphic (AFR and EUR), the direction and the magnitude of the effect diverged (**Table 2**). This association has not been reported previously.

For FVIII and VWF levels, 11 significant single-variant associations were observed with rare or low-frequency variants within *ABO* and surrounding genes on chromosome 9. However, after conditioning on common variants tagging the major ABO blood types (A1, A2, B, & O), none of the 11 associations identified in this region remained. A description of these conditional analyses is presented in the **Supplemental Information (Section IIId)** and **Supplemental Table S4**.

In exploratory analyses and for the FVIII phenotype only, there was a significant association with a common variant on the X-chromosome in *F8*, the gene encoding FVIII. This coding variant, rs1800291 (Asp1260Glu, MAF=0.27, $p=8.2\times10^{-08}$) had a MAF and effect direction that varied across ancestry groups (**Table 2**).

For the FVIII phenotype only, a rare variant in *KATNB1*, a gene not previously associated with FVIII levels, achieved array-wide significance in the AFR population. This variant, rs142508811, was rare in both EUR and AFR populations and monomorphic in ASI and HIS; the estimated effect size was 80-fold larger in AFR than EUR populations. Across the studies with AFR populations, there was no evidence of heterogeneity ($p_{HET}=0.74$) and a forest plot for these associations are presented in **Supplemental Figure S5**. Levels of FVIII in carriers of the variant allele had a higher mean FVIII than non-carriers (**Supplemental Table S5**).

For the FVIII phenotype, the 5 variants explained approximately 0.9% of the phenotype variation in both EUR and AFR populations. For the VWF phenotype, the *STAB2* variant explained 0.2% and 0% in EUR and AFR populations, respectively.

FVIII and VWF: gene-based testing

For FVIII levels, *ABO*, *VWF,* and *STAB2* yielded gene-wide significant associations with SKAT testing while *ABO* and *VWF* were significant with T5 testing (**Table 3**). For VWF

Table 4. Single-Variant Test Meta-Analysis Results for Conditional Analyses of Hemostatic Factors Fibrinogen, Factor VII, Factor VIII and von Willebrand Factor.

			P-value			
Variant (Gene)	**Ancestry**	**N**	**UNCOND**	**COND1**	**COND2**	**COND3**
Fibrinogen						
rs201909029 *(FGB)*	ALL	46,841	**1.97×10^{-10}**	**1.35×10^{-09}**	**2.27×10^{-10}**	**3.44×10^{-10}**
	EUR	40,091	**2.69×10^{-10}**	**1.83×10^{-09}**	**3.10×10^{-10}**	**4.68×10^{-10}**
	AFR	6,750	4.25×10^{-01}	4.24×10^{-01}	4.21×10^{-01}	4.25×10^{-01}
rs6054 *(FGB)*	ALL	46,841	**1.00×10^{-41}**	**6.72×10^{-39}**	**2.67×10^{-42}**	
	EUR	40,091	**4.86×10^{-41}**	**3.40×10^{-38}**	**5.46×10^{-42}**	
	AFR	6,750	7.66×10^{-02}	7.25×10^{-02}	1.97×10^{-01}	
rs145051028 *(FGG)*	ALL	46,841	2.93×10^{-06}	2.67×10^{-06}		2.90×10^{-06}
	EUR	40,091	NA	NA	NA	NA
	AFR	6,750	2.93×10^{-06}	2.67×10^{-06}		2.90×10^{-06}
rs148685782 *(FGG)*	ALL	46,841	**3.24×10^{-144}**	**6.52×10^{-137}**		**2.49×10^{-143}**
	EUR	40,091	**1.03×10^{-143}**	**2.16×10^{-136}**		**8.02×10^{-143}**
	AFR	6,750	9.46×10^{-02}	9.52×10^{-02}		9.43×10^{-02}
Factor VII						
rs150525536 (F7)	ALL	20,549	**8.29×10^{-20}**	**1.02×10^{-22}**		
	EUR	16,338	2.23×10^{-01}	1.20×10^{-01}		
	AFR	4,211	**3.45×10^{-20}**	**7.56×10^{-23}**		
rs121964926 (F7)	ALL	20,549	**5.71×10^{-14}**	**1.49×10^{-14}**		
	EUR	16,338	9.25×10^{-01}	5.80×10^{-01}		
	AFR	4,211	**1.75×10^{-20}**	**1.95×10^{-20}**		
rs3093248 (F7)	ALL	20,549	2.54×10^{-06}	**1.35×10^{-07}**		
	EUR	16,338	NA	NA		
	AFR	4,211	2.54×10^{-06}	**1.35×10^{-07}**		
Factor VIII						
rs7962217 (VWF)	ALL	25,477	**6.60×10^{-11}**	**1.64×10^{-09}**		
	EUR	20,030	**8.69×10^{-10}**	**1.39×10^{-08}**		
	AFR	5,447	1.18×10^{-02}	2.35×10^{-02}		
rs41276738 (VWF)	ALL	25,477	**1.56×10^{-11}**	**9.85×10^{-14}**		
	EUR	20,030	**1.52×10^{-10}**	**1.41×10^{-12}**		
	AFR	5,447	5.96×10^{-03}	3.47×10^{-03}		
rs141041254 (STAB2)	ALL	25,477	**7.37×10^{-09}**	**4.11×10^{-09}**		
	EUR	20,030	**4.03×10^{-09}**	**2.22×10^{-09}**		
	AFR	5,447	9.17×10^{-01}	9.20×10^{-01}		
von Willebrand Factor						
rs141041254 (STAB2)	ALL	22,636	**6.82×10^{-08}**	**3.29×10^{-08}**		
	EUR	18,236	**2.85×10^{-08}**	**1.34×10^{-08}**		
	AFR	4,400	7.46×10^{-01}	7.49×10^{-01}		

SNPs achieving genome-wide significance threshold (p= 2.57×10^{-07} (ALL), 2.88×10^{-07} (EUR), 3.30×10^{-07} (AFR)) are bolded. N*=number of participants included in analysis, only EUR and AFR cohorts were asked to run conditional analyses and not all cohorts participated; ** UNCOND = unadjusted analyses; description of conditional analyses are found in **Supplemental Table S3**. ALL = EUR+AFR; EUR=European-only; AFR=African-only. SNPs where results are shaded grey were conditioned on for that analysis.

levels, *ABO* and *STAB2* yielded gene-wide significant associations with SKAT testing while *ABO* was significant with T5 testing; *VWF* gene did not achieve significance for VWF. No new associations were identified through gene-based testing.

DISCUSSION

We identified 12 novel associations of low-frequency (n=2) and rare (n=10) variants across the fibrinogen, FVII, FVIII, and VWF traits that were independent of previously identified associations. Nine of the variants were within genes previously established as associated with the trait; findings for associations in 3 new candidate loci were detected in those of AFR ancestry, possibly due to monomorphic or much lower frequency characteristics of these variants in all other ancestries. These newly identified associations accounted for modest amounts of the variance explained and suggest that at most a small proportion of the missing heritability can be attributable to them. The gene-based tests did not reveal new loci.

Associations of rare variants with fibrinogen levels were found in gene regions previously associated with fibrinogen by common variant GWAS. The association of *FGB* rare variant rs6054 with lower fibrinogen has been previously reported.[10] While the association of *FGB* rs201909029 is a novel finding in this context, it has been reported in mild hypofibrinogenaemia cases[26] in clinical databases (MERIVALE II)[30] although it has not been reported to cause haemorrhage or thrombosis.[30] The rare *FGG* variant rs148685782 was associated with hypofibrinogenaemia and haemorrhage[26-28] in multiple affected individuals. *FGG* rs145051028, which was associated with fibrinogen levels in AFR cohorts only, has not been reported in clinical databases or population studies. This may be due to the low MAF but also a lack of studies including AFR participants. Conditional analyses showed that the common and rare variant associations across the fibrinogen gene cluster were independent, an observation supported by their low R^2 for the pairwise LD. Within the fibrinogen gene cluster, the 4 significant *FGB* and *FGG* rare variants explained 2 to 4-fold more trait variance than the common *FGB* rs4220 variant,[7, 9, 10, 14, 31] which had an effect size of 0.029 ln(g/l), or a 2.9% higher levels of fibrinogen per copy of the minor allele, in this study.

In exploratory ancestry-stratified analyses, the associations of *KCNT1* and *HID1* with fibrinogen in AFR participants were the only findings that identify new candidate loci influencing fibrinogen regulation. These findings can only be considered hypothesis generating and require replication.

We identified 3 rare coding variants in the FVII protein structural gene (*F7*) associated with plasma levels of FVII, none of which were previously reported in

the epidemiologic literature. rs150525536 was rare in the AFR population and had a 10-fold lower frequency in the EUR population. A previous case-report of this variant was found in a male, EUR ancestry homozygote, with severe FVII deficiency who also carried another *F7* mutation (Arg212Gln).[32] Both mutations were thought to contribute to the phenotype. The mutation reported here is found in the first epidermal growth factor-like domain and is required for binding to tissue factor, its cofactor. It causes reduced binding to tissue factor and reduced clotting ability in a concentration-dependent manner as well as slower activation.[32] Variant rs121964926 was also more common among the AFR population than in the EUR population. It has been observed clinically in both asymptomatic and symptomatic individuals with FVII activity <5% from Germany and France as well as patients with reduced FVII activity from Costa Rica, Venezuela, and the USA.[33] Nothing has been reported regarding clinical consequences of the rs3093248 variant.

The findings for the VWF trait consisted of a subset of the FVIII results. None of the associations between variants within the *ABO* gene region and FVIII/VWF were independent of established ABO blood group alleles. Two rare variants in *VWF* were associated with plasma FVIII levels, rs7962217 and rs41276738. rs7962217 was associated with higher FVIII levels whereas rs41276738 was associated with lower levels and had a similar effect size as that of the strongest genetic predictor of FVIII levels, the O-deletion tagging SNP (rs657152). rs41276738 has been reported in patients with von Willebrand disease type 1[34, 35] and type 2N.[36-43] but the association with VWF levels did not reach exome-wide significance, although its direction was consistent with the direction of effects on FVIII. The *STAB2* variant rs141041254 was associated with higher plasma levels of both FVIII and VWF. The effect size was over 10-fold larger than that reported for the more common *STAB2* variant rs2271637 (β_{FVIII}=1.95%, β_{VWF}=2.47%). A common *F8* coding variant rs1800291 was associated with a much smaller effect on FVIII compared with the *ABO* O-deletion variant. It has been reported previously[29, 44, 45] and in the EAHAD Coagulation Factor Variants Database is annotated as unlikely to be pathogenic. The *KATNB1* rs142508811 variant and FVIII association was restricted to the AFR population, although MAF and direction of effect was similar across the 2 polymorphic populations.

Inferring causality of uncommon and rare variants with a phenotypic expression is challenging and requires strong statistical evidence combined with experimental data.[46] Inferring clinical implications from the causal variants requires additional evidence[47] not available in our approach. In this article, we identified rare variants associated with higher or lower phenotype levels in 4 hemostasis measures. Some of the variants have been found in patients with diseases related to blood clotting and suggest that these genes and their uncommon and rare genetic variation may play a role in a clinical phenotype.[26-28, 32-43] The distribution of the phenotypes within

our research populations were within the extremes of a clinically important range (range = 0.80-11.40 g/l (fibrinogen); 26-441% activity & 2-297% antigen (FVII); 14-500% activity (FVIII); 2-374% antigen (VWF)). Further, the magnitude of difference in the phenotype associated with the variant was mostly modest, although some were larger and were associated with a change equivalent to half the size of the estimated population mean for the phenotype of interest. Therefore, the magnitude of any clinically relevant effects of these variants would be expected to be small to modest. The findings from our study suggest that the contribution of the uncommon and rare variants to complex clinical phenotypes, such as arterial or venous thrombosis or hemorrhagic stroke, should be evaluated in large populations. This article identifies several variants which may be good potential candidates.

We decided *a priori* to use all the phenotype-genotype association data for discovery in order to reduce false negative findings[48] but this approach provided us with no replication setting. Although these candidate variants are now well characterized, the rare allele frequencies will create challenges for replication in the absence of additional large phenotyped populations. However, our findings provide strong rationale for further functional genomic follow-up and some of our observations confirm associations for several rare variants that have been reported in patients with the corresponding congenital clotting factor deficiencies. This investigation of low-frequency and rare variants on the 4 phenotypes was limited to the variants included on the BeadChip. Differing sample sizes of the meta-analysis between phenotypes likely affected our power to detect associations, but this may also be influenced by biological differences. Further, we did not have the statistical power to test for differences in associations across the 4 ancestries. While not an aim of our study, a subsequent effort with this objective would be worthwhile to better understand the genetic architecture of the phenotypes. Lastly while we enriched our variant population with those predicted to be causal, we cannot attribute causality to the variants with novel associations.

The quality of rare variant genotype calling was maximized by the joint clustering performed within CHARGE on thousands of samples.[17] By incorporating individuals of non-European ancestry in the primary analysis, we increased our power to detect association where variants may be more frequent or genetic diversity greater in one ancestry group than another. It also allowed us to broadly look at ancestry-specific gene and rare-variant associations but was vastly underpowered to draw any strong conclusions.

In meta-analyses of 4 hemostatic factors and functionally enriched exonic variants, novel associations of low-frequency and rare variants were identified in 16 studies that included 4 ancestries. Novel variant-associations were found within previously reported genes and had effect sizes that were often independent of and much larger

than previously reported common variants. In addition, rare variant associations at *KCNT1*, *HID1,* and *KATNB1* identify new candidate genes related to hemostasis for follow-up replication and functional genomic analysis.

Supplement available online at:
http://www.bloodjournal.org/content/126/11/e19

REFERENCES

1. Folsom AR. Hemostatic risk factors for atherothrombotic disease: An epidemiologic view. *Thromb Haemost*. 2001;86(1):366-373
2. Folsom AR, Cushman M, Heckbert SR, Ohira T, Rasmussen-Torvik L, Tsai MY. Factor VII coagulant activity, factor VII -670A/C and -402G/A polymorphisms, and risk of venous thromboembolism. *J Thromb Haemost*. 2007;5(8):1674-1678
3. Smith A, Patterson C, Yarnell J, Rumley A, Ben-Shlomo Y, Lowe G. Which hemostatic markers add to the predictive value of conventional risk factors for coronary heart disease and ischemic stroke? The Caerphilly Study. *Circulation*. 2005;112(20):3080-3087
4. Danesh J, Lewington S, Thompson SG, et al. Plasma fibrinogen level and the risk of major cardiovascular diseases and nonvascular mortality: An individual participant meta-analysis. *JAMA*. 2005;294(14):1799-1809
5. Koster T, Blann AD, Briët E, Vandenbroucke JP, Rosendaal FR. Role of clotting factor VIII in effect of von Willebrand factor on occurrence of deep-vein thrombosis. *Lancet*. 1995;345(8943):152-155
6. Spiel AO, Gilbert JC, Jilma B. von Willebrand factor in cardiovascular disease: focus on acute coronary syndromes. *Circulation*. 2008;117(11):1449-1459
7. Danik JS, Pare G, Chasman DI, et al. Novel loci, including those related to Crohn disease, psoriasis, and inflammation, identified in a genome-wide association study of fibrinogen in 17 686 women: The Women's Genome Health Study. *Circ Cardiovasc Genet*. 2009;2(2):134-141
8. Lovely RS, Yang Q, Massaro JM, et al. Assessment of genetic determinants of the association of γ′ fibrinogen in relation to cardiovascular disease. *Arterioscler Thromb Vasc Biol*. 2011;31(10):2345-2352
9. Sabater-Lleal M, Huang J, Chasman D, et al. Multiethnic meta-analysis of genome-wide association studies in >100 000 subjects identifies 23 fibrinogen-associated loci but no strong evidence of a causal association between circulating fibrinogen and cardiovascular disease. *Circulation*. 2013;128(12):1310-1324
10. Wassel CL, Lange LA, Keating BJ, et al. Association of genomic loci from a cardiovascular gene snp array with fibrinogen levels in European Americans and African-Americans from six cohort studies: The Candidate Gene Association Resource (CARe). *Blood*. 2011;117(1):268-275
11. Johnsen JM, Auer PL, Morrison AC, et al. Common and rare von Willebrand factor (VWF) coding variants, VWF levels, and factor VIII levels in African Americans: The NHLBI Exome Sequencing Project. *Blood*. 2013;122(4):590-597
12. Taylor KC, Lange LA, Zabaneh D, et al . A gene-centric association scan for Coagulation Factor VII levels in European and African Americans: The Candidate Gene Association Resource (CARe) Consortium. *Hum Mol Genet*. 2011;20(17):3525-3534
13. Smith NL, Chen MH, Dehghan A, et al. Novel associations of multiple genetic loci with plasma levels of factor VII, factor VIII, and von Willebrand factor: The CHARGE (Cohorts for Heart and Aging Research in Genome Epidemiology) Consortium. *Circulation*. 2010;121(12):1382-1392
14. Dehghan A, Yang Q, Peters A, et al. Association of novel genetic loci with circulating fibrinogen levels: A genome-wide association study in 6 population-based cohorts. *Circ Cardiovasc Genet*. 2009;2(2):125-133
15. Manolio TA, Collins FS, Cox NJ, et al. Finding the missing heritability of complex diseases. *Nature*. 2009;461(7265):747-753

16. Psaty BM, O'Donnell CJ, Gudnason V, 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*. 2009;2(1):73-80
17. Grove ML, Yu B, Cochran BJ, et al. Best practices and joint calling of the HumanExome Bead-Chip: The CHARGE Consortium. *PLoS One*. 2013;8(7):e68095
18. Peloso GM, Auer PL, Bis JC, 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*. 2014;94(2):223-232
19. Liu X, Jian X, Boerwinkle E. dbNSFP v2.0: A database of human non-synonymous SNVs and their functional predictions and annotations. *Hum Mutat*. 2013;34(9):E2393-2402
20. Liu X, Jian X, Boerwinkle E. dbNSFP: A lightweight database of human nonsynonymous SNPs and their functional predictions. *Hum Mutat*. 2011;32(8):894-899
21. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet*. 2011;89(1):82-93
22. Li B, Leal SM. Methods for detecting associations with rare variants for common diseases: Application to analysis of sequence data. *Am J Hum Genet*. 2008;83(3):311-321
23. Morris AP, Zeggini E. An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genet Epidemiol*. 2010;34(2):188-193
24. Dong C, Wei P, Jian X, et al. Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Hum Mol Genet*. 2015;24(8):2125-2137
25. Jian X, Boerwinkle E, Liu X. In silico prediction of splice-altering single nucleotide variants in the human genome. *Nucleic Acids Res*. 2014;42(22):13534-13544
26. Ivaskevicius V, Jusciute E, Steffens M, et al. gammaAla82Gly represents a common fibrinogen gamma-chain variant in Caucasians. *Blood Coagul Fibrinolysis*. 2005;16(3):205-208
27. Brennan SO, Fellowes AP, Faed JM, George PM. Hypofibrinogenemia in an individual with 2 coding (gamma82 A-->G and Bbeta235 P-->L) and 2 noncoding mutations. *Blood*. 2000;95(5):1709-1713
28. Wyatt J, Brennan SO, May S, George PM. Hypofibrinogenaemia with compound heterozygosity for two gamma chain mutations - gamma 82 Ala-->Gly and an intron two GT-->AT splice site mutation. *Thromb Haemost*. 2000;84(3):449-452
29. Tang W, Cushman M, Green D, et al. Gene-centric approach identifies new and known loci for FVIII activity and VWF antigen levels in European Americans and African Americans. *Am. J.Hematol*. 2015;90(6):534-540
30. Maghzal GJ, Brennan SO, George PM. Fibrinogen B beta polymorphisms do not directly contribute to an altered in vitro clot structure in humans. *Thromb Haemost*. 2003;90(6):1021-1028
31. Schmelzer CH, Ebert RF, Bell WR. A polymorphism at B beta 448 of fibrinogen identified during structural studies of fibrinogen Baltimore II. *Thromb Res*. 1988;52(2):173-177
32. Chaing S, Clarke B, Sridhara S, et al. Severe factor VII deficiency caused by mutations abolishing the cleavage site for activation and altering binding to tissue factor. *Blood*. 1994;83(12):3524-3535
33. Herrmann FH, Wulff K, Auerswald G, et al. Factor VII deficiency: Clinical manifestation of 717 subjects from Europe and Latin America with mutations in the factor 7 gene. *Haemophilia*. 2009;15(1):267-280
34. Goodeve A, Eikenboom J, Castaman G, et al. Phenotype and genotype of a cohort of families historically diagnosed with type 1 von Willebrand disease in the European study, Molecular

and Clinical Markers for the Diagnosis and Management of Type 1 von Willebrand Disease (MCMDM-1VWD). *Blood*. 2007;109(1):112-121
35. James PD, Notley C, Hegadorn C, et al. The mutational spectrum of type 1 von Willebrand disease: Results from a Canadian cohort study. *Blood*. 2007;109(1):145-154
36. Corrales I, Catarino S, Ayats J, et al. High-throughput molecular diagnosis of von Willebrand disease by next generation sequencing methods. *Haematologica*. 2012;97(7):1003-1007
37. Schneppenheim R, Brassard J, Krey S, et al. Defective dimerization of von Willebrand factor subunits due to a Cys-> Arg mutation in type IID von Willebrand disease. *Proc Natl Acad Sci U S A*. 1996;93(8):3581-3586
38. Eikenboom JC, Reitsma PH, Peerlinck KM, Briët E. Recessive inheritance of von Willebrand's disease type I. *Lancet*. 1993;341(8851):982-986
39. Mazurier C. von Willebrand disease masquerading as haemophilia A. *Thromb Haemost*. 1992;67(4):391-396
40. Peerlinck K, Eikenboom JC, Ploos Van Amstel HK, et al. A patient with von Willebrand's disease characterized by a compound heterozygosity for a substitution of Arg854 by Gln in the putative factor-VIII-binding domain of von Willebrand factor (vWF) on one allele and very low levels of mRNA from the second vWF allele. *Br J Haematol*. 1992;80(3):358-363
41. Gaucher C, Jorieux S, Mercier B, Oufkir D, Mazurier C. The "Normandy" variant of von Willebrand disease: characterization of a point mutation in the von Willebrand factor gene. *Blood*. 1991;77(9):1937-1941
42. Kroner PA, Friedman KD, Fahs SA, Scott JP, Montgomery RR. Abnormal binding of factor VIII is linked with the substitution of glutamine for arginine 91 in von Willebrand factor in a variant form of von Willebrand disease. *J Biol Chem*. 1991;266(29):19146-19149
43. Cacheris PM, Nichols WC, Ginsburg D. Molecular characterization of a unique von Willebrand disease variant. A novel mutation affecting von Willebrand factor/factor VIII interaction. *J Biol Chem*. 1991;266(21):13499-13502
44. Viel KR, Machiah DK, Warren DM, et al. A sequence variation scan of the coagulation factor VIII (FVIII) structural gene and associations with plasma FVIII activity levels. *Blood*. 2007;109(9):3713-3724
45. Scanavini D, Legnani C, Lunghi B, Mingozzi F, Palareti G, Bernardi F. The factor VIII D1241E polymorphism is associated with decreased factor VIII activity and not with activated protein C resistance levels. *Thromb Haemost*. 2005;93(3):453-456
46. MacArthur DG, Manolio TA, Dimmock DP, et al. Guidelines for investigating causality of sequence variants in human disease. *Nature*. 2014;508(7497):469-476
47. Richards S, Aziz N, Bale S, et al. Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med*. 2015;17(5):405-423
48. Skol AD, Scott LJ, Abecasis GR, Boehnke M. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat Genet*. 2006;38(2):209-213
49. The ARIC Investigators. The Atherosclerosis Risk in Communities (ARIC) study: Design and objectives. *Am J Epidemiol*. 1989;129(4):687-702
50. Friedman GD, Cutter GR, Donahue RP, et al. CARDIA: Study design, recruitment, and some characteristics of the examined subjects. *J Clin Epidemiol*. 1988;41(11):1105-1116
51. Fried LP, Borhani NO, Enright P, et al. The Cardiovascular Health Study: Design and rationale. *Ann Epidemiol*. 1991;1(3):263-276

52. Feinleib M, Kannel WB, Garrison RJ, McNamara PM, Castelli WP. The Framingham Offspring Study. Design and preliminary data. *Prev Med*. 1975;4(4):518-525
53. Kannel WB, Dawber TR, 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*. 1961;55:33-50
54. Vaidya D, Yanek LR, Moy TF, Pearson TA, Becker LC, Becker DM. Incidence of coronary artery disease in siblings of patients with premature coronary artery disease: 10 years of follow-up. *The Am J Cardiol*. 2007;100(9):1410-1415
55. Holle R, Happich M, Lowel H, Wichmann HE, Monica Kora Study Group. KORA--a research platform for population based health research. *Gesundheitswesen*. 2005;67 Suppl 1:S19-25
56. Wichmann HE, Gieger C, Illig T, Monica Kora Study Group. KORA-gen--resource for population genetics, controls and a broad spectrum of disease phenotypes. *Gesundheitswesen*. 2005;67 Suppl 1:S26-30
57. Zemunik T, Boban M, Lauc G, et al. Genome-wide association study of biochemical traits in Korcula Island, Croatia. *Croat Med J*. 2009;50(1):23-33
58. Deary IJ, Gow AJ, Pattie A, Starr JM. Cohort profile: The Lothian Birth Cohorts of 1921 and 1936. *Int J Epidemiol*. 2012;41(6):1576-1584
59. Deary IJ, Whiteman MC, Starr JM, Whalley LJ, Fox HC. The impact of childhood intelligence on later life: following up the Scottish mental surveys of 1932 and 1947. *J Pers Soc Psychol*. 2004;86(1):130-147
60. Deary IJ, Gow AJ, Taylor MD, et al. The Lothian Birth Cohort 1936: A study to examine influences on cognitive ageing from age 11 to age 70 and beyond. *BMC Geriatr*. 2007;7:28
61. Bild DE, Bluemke DA, Burke GL, et al. Multi-ethnic study of atherosclerosis: Objectives and design. *Am J Epidemiol*. 2002;156(9):871-881
62. Clarke R, Peden JF, Hopewell JC, et al. Genetic variants associated with Lp(a) lipoprotein level and coronary disease. *N Engl J Med*. 2009;361(26):2518-2528
63. Hofman A, Breteler MM, van Duijn CM, et al. The Rotterdam Study: 2010 objectives and design update. *Eur J Epidemiol*. 2009;24(9):553-572
64. Hofman A, Darwish Murad S, van Duijn CM, et al. The Rotterdam Study: 2014 objectives and design update. *Eur J Epidemiol*. 2013;28(11):889-926
65. Hofman A, Grobbee DE, de Jong PT, van den Ouweland FA. Determinants of disease and disability in the elderly: the Rotterdam Elderly Study. *Eur J Epidemiol*. 1991;7(4):403-422
66. Samnegard A, Silveira A, Lundman P, et al. Serum matrix metalloproteinase-3 concentration is influenced by MMP-3 -1612 5A/6A promoter genotype and associated with myocardial infarction. *J Intern Med*. 2005;258(5):411-419
67. Völzke H, Alte D, Schmidt CO, et al. Cohort profile: The study of health in Pomerania. *Int J Epidemiol*. 2011;40(2):294-307
68. Ridker PM, Chasman DI, Zee RY, et al. Rationale, design, and methodology of the Women's Genome Health Study: A genome-wide association study of more than 25,000 initially healthy american women. *Clin Chem*. 2008;54(2):249-255
69. The Women's Health Initiative Study Group. Design of the Women's Health Initiative clinical trial and observational study. *Control Clin Trials*. 1998;19(1):61-109
70. Anderson CA, Boucher G, Lees CW, et al. Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. *Nat Genet*. 2011;43(3):246-252
71. Hays J, Hunt JR, Hubbell FA, et al. The Women's Health Initiative recruitment methods and results. *Ann Epidemiol*. 2003;13(9 Suppl):S18-77

Chapter 2.4

Whole-exome sequencing study of hemostatic factors

Manuscript based on this chapter

Nathan Pankratz, Peng Wei, Jennifer A. Brody, Ming-Huei Chen, Paul S. de Vries, Jennifer E. Huffman, Paul L. Auer, Eric Boerwinkle, Mary Cushman, Moniek P.M. de Maat, Aaron R. Folsom, Oscar H. Franco, Richard A. Gibbs, Kelly K. Haagenson, Albert Hofman, Jill M Johnsen, Christie L Kovar, Robert Kraaij, Barbara McKnight, Ginger A. Metcalf, Donna Muzny, Bruce M. Psaty, Weihong Tang, André G. Uitterlinden, Jeroen G.J. van Rooij, Narayanan Veeraraghavan, Abbas Dehghan, Christopher J. O'Donnell, Alex P. Reiner, Nicholas L. Smith, and Alanna C. Morrison on behalf of the CHARGE consortium and ESP.

Whole exome sequencing of 14,000 individuals identifies novel rare variation association with hemostatic factors.

Submitted.

ABSTRACT

Background: Circulating plasma hemostatic factors, such as fibrinogen, coagulation factors VII and VIII, and von Willebrand factor (VWF), are heritable intermediate phenotypes associated with the risk of clinical thrombotic events.

Methods: To identify rare and low-frequency variants associated with these factors, we conducted whole exome sequencing in 10,860 individuals of European ancestry (EA) and 3,529 African Americans (AA) from the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) consortium and the National Heart, Lung, and Blood Institute Exome Sequencing Project (ESP).

Results: We identified single nucleotide variants with genome-wide significant associations for fibrinogen (*FGG*, $p=2\times10^{-28}$), factor VII (*F7*, $p=2\times10^{-261}$), factor VIII (*ABO*, $p=4\times10^{-94}$), and VWF (*ABO*, $p=9\times10^{-115}$). Gene-based tests demonstrated significant associations with rare variation in *FGG* (with fibrinogen, $p=9\times10^{-13}$; two novel variants), *F7* (with factor VII, $p=1\times10^{-72}$; six novel variants), and *VWF* (with factor VIII and/or VWF; $p=3\times10^{-14}$; two novel variants). These ten novel rare variant associations were independent of the known common variants in the same genes and tended to have much larger effect sizes.

Conclusions: These efforts represent the largest integration of whole exome sequence data from two national projects to identify genetic variation associated with plasma hemostatic factors.

INTRODUCTION

Fibrinogen, factor VII (FVII), factor VIII (FVIII) and von Willebrand factor (VWF) are circulating plasma hemostatic factors that have been associated with the development of venous thrombosis or athero-thrombotic cardiovascular disease in human populations.[1,2] Estimates of heritability range from 0.28 to 0.44 for fibrinogen,[3-5] 0.33 to 0.63 for FVII,[3-5] 0.29 to 0.61 for FVIII,[3-5] and 0.32 to 0.75 for VWF.[4,5] Characterization of common and low-frequency variation influencing inter-individual and inter-population differences in circulating fibrinogen, FVII, FVIII, and VWF may lead to improved understanding of the role of hemostasis in inflammation and athero-thrombotic risk, and potentially reveal novel biologic pathways influencing these hemostatic factors.

Recent genome-wide association studies (GWAS) have demonstrated that common polymorphisms with minor allele frequencies (MAF) greater than 0.05 contribute to the heritability of all of these traits[6-11] and that the variants underlying variation in FVIII and VWF heavily overlap.[7] However, the common polymorphisms identified to-date explain only a small proportion of the heritability,[7,12] and the amount of variation that they explain is modest: 12.8% for VWF, 7.7% for FVII, 10.0% for FVIII and <2.0% for fibrinogen.[6,7] This suggests that additional loci or variation within known genes may account for inter-individual variability in these hemostatic factors.

The aim of this study was to characterize rare and low-frequency variants associated with plasma levels of fibrinogen, FVII, FVIII, and VWF by analyzing whole exome sequence data in individuals of European and African ancestry from two large, coordinated exome sequencing projects.

METHODS

Study subjects and hemostatic factor measurements

Exome sequence data for individuals of European ancestry (EA) and African ancestry (AA) came from the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium[13] and from the National Heart, Lung, and Blood Institute Exome Sequencing Project (ESP). Individuals from CHARGE came from four population-based cohorts: the Atherosclerosis Risk in Communities Study (ARIC), the Cardiovascular Health Study (CHS), the Framingham Heart Study (FHS), and the Rotterdam Study (RS). Participants in ESP were sampled from six population-based cohorts – ARIC, CHS, FHS, Coronary Artery Risk Development In Young Adults (CARDIA), the Multi-Ethnic Study of Atherosclerosis (MESA), and the Women's Health Initiative (WHI) – and do not overlap the CHARGE participants.

Detailed descriptions of each of the seven cohorts and the techniques used to measure hemostatic factor levels are provided in previous publications.[14-26] Fibrinogen was available in all seven cohorts, FVII activity in six, and FVIII activity or VWF antigen in five. Plasma levels of fibrinogen were measured in g/L, and FVII, FVIII, and VWF were measured in international units (IU/dL, which are sometimes denoted as a percentage). All participants provided written informed consent as approved by local human-subjects committees.

Exome sequencing and variant calling

DNA from ARIC, CHS and FHS participants from CHARGE were all prepared using the HGSC VCRome 2.1 design[27] (42Mb, NimbleGen), sequenced, and called together. DNA from RS participants were prepared using Roche NimbleGen SeqCap v2 (44Mb), and DNA from ESP participants were prepared using either Roche Nimblegen SeqCap EZ or Agilent SureSelect Human All Exon 50Mb. All samples were paired end sequenced (2×100bp for the CHARGE cohorts and 2×76bp for ESP) using Illumina GAII or HiSeq instruments. After taking into account available hemostatic factor measures, there were 8,859 EA and 2,664 AA from CHARGE and 2,001 EA and 865 AA from ESP. Analyses were conducted using a total of 14,389 unique individuals (10,860 EA and 3,529 AA) across CHARGE and ESP.

Annotation of whole exome sequence variants

To facilitate meta-analysis between CHARGE and ESP, we created a combined variant annotation file including all quality-controlled variant sites observed in CHARGE or ESP. Variants were annotated using ANNOVAR[28] and dbNSFP v2.0 (https://sites.google.com/site/jpopgen/dbNSFP) according to the reference genome GRCh37 and National Center for Biotechnology Information RefSeq. The combined variant information file contained 6,605,975 unique sites, including the 2,706,509 sites that were polymorphic in the samples with hemostatic factors.

Association analyses

Samples with extreme values for hemostatic factors (>3 standard deviations from the mean) were excluded from analyses to prevent spurious associations with rare variants. Final sample sizes for each of the four traits are summarized in **Table 1**. All four traits were natural-log (ln) transformed, and the distributions for all studies was approximately normal following this transformation. Cohort-level analyses were carried out using the seqMeta R package (http://cran.r-project.org/web/packages/seqMeta/). Data from the CHARGE cohorts (ARIC, CHS, FHS, and RS) were each analyzed separately, while the five cohorts that make up ESP were included in a single pooled analysis. Fixed effect inverse-variance weighted meta-analyses of single

variant and gene-based tests were conducted using seqMeta for ancestry-specific results as well as trans-ethnic analyses.

Single variant testing was done for individual variants where the minor allele count (MAC) was at least 40 across cohorts (MAC=40 translates into MAF>0.0014 for this meta-analysis). This was an a priori determined threshold that was designed to reduce the chance for false positive associations caused by extreme phenotypic values as well as to reduce the number of tests performed, thereby increasing power to detect true associations. Within each meta-analysis group (trans-ethnic, EA, AA) we tested for single variant association with hemostatic factor levels by linear regression with an additive genetic model adjusting for age, sex, and race-specific principal components (PCs). The trans-ethnic analysis was the primary approach and an association was considered to be study-wide significant at $p<1.9\times10^{-7}$ for single variants given a Bonferroni correction for testing as many as 270,221 single variant sites with MAC≥40. Conditional analyses were conducted to establish statistical independence among identified variants using a Bonferroni correction for the number of variants in each gene region ($p<0.05$ / # of polymorphic variants tested). Since the goal of conditional analyses is to establish allelic heterogeneity in genes known to be associated with the trait, the MAC requirement was dropped to >5 for all conditional analyses.

Table 1. Number of samples for each study with whole exome sequencing data listed by hemostatic factor

Study	Race	Fibrinogen	FVII	FVIII	vWF	Maximum
Atherosclerosis Risk in Communities Study (ARIC)	EA	5,652	5,527	5,658	5,682	5,682
Cardiovascular Health Study (CHS)	EA	737	742	734	-	742
Framingham Heart Study (FHS)	EA	741	667	-	667	741
Rotterdam Study (RS)	EA	987	263	906[a]	788[a]	906+788[a]
NHLBI Exome Sequencing Project (ESP)[b]	EA	2,001	1,204	1,282	1,181	2,001
EA Subtotal:		10,118	8,403	8,580	8,318	10,860
Atherosclerosis Risk in Communities Study (ARIC)	AA	2,655	2,602	2,659	2,664	2,664
NHLBI Exome Sequencing Project (ESP)[b]	AA	865	644	598	439	865
AA Subtotal:		3,520	3,246	3,257	3,103	3,529
Total:	EA+AA	13,638	11,649	11,837	11,421	14,389

[a] These subsets of individuals are mutually exclusive

[b] ESP consists of non-overlapping samples from ARIC, CHS, FHS, Coronary Artery Risk Development In Young Adults (CARDIA), the Multi-Ethnic Study of Atherosclerosis (MESA), and the Women's Health Initiative (WHI)

We performed gene-based tests that included only rare and low-frequency variants (MAF<0.05) annotated as stop-gain, stop-loss, splicing, missense, or small insertion or deletion sites (indels). Using the seqMeta package, two gene-level tests were performed. The first was a "T5" test where all variants passing the above mentioned filters were summed together to generate a gene burden score.[29,30] The second test was the Sequence Kernel Association Test (SKAT),[31] which analyzes the same variants as the T5 test, but has greater power when effects are in both directions and up-weights the contribution of rarer variants. All gene-level tests were adjusted for the same covariates as the single variant test and required the gene to have a cumulative minor allele count of at least 40, similar to the single variant tests. The trans-ethnic analysis was the primary approach, and an association was considered to be significant at $p<1.5\times10^{-6}$, which is the Bonferroni-corrected significance threshold for two gene-based tests and the 16,848 qualifying genes (i.e., MAC>40).

RESULTS

Participant characteristics

Characteristics of the participating cohorts are summarized in **Supplemental Table 1**. Taking into account sample size, the mean age was 58.2±6.2 years and 56.8% were female. The means (standard deviations) for the hemostatic factors were: 3.1 (0.7) g/L for fibrinogen, 116 (27) IU/dL for FVII, 129 (38) IU/dL for FVIII, and 118 (46) IU/dL for VWF.

Single variant associations with fibrinogen

In single variant analyses, we did not observe significant inflation or deflation of the meta-analysis *P*-values (0.96≤lambdas≤1.06), indicating that there was no serious overcorrection or confounding by any covariate, population substructure, or lab effects. **Table 2** summarizes the significant loci in the trans-ethnic meta-analysis and lists the index variant (the variant with the smallest *P*-value).

Single variant associations with fibrinogen

There were two loci significantly associated with fibrinogen levels, one within the fibrinogen gene cluster (*FGA*, *FGB*, and *FGG*) and the other at *IRF1* (**Table 2**). There were five study-wide significant variants within the fibrinogen gene cluster (Ala82Gly in *FGG*; Tyr345Tyr, Arg448Lys, and Ser159Ser in *FGB*; and Thr312Ala in *FGA*; variants in the gene cluster are annotated using the sequence of the mature, circulating protein). The three *FGB* variants were all in high linkage disequilibrium with one another (R^2=0.99), but all other combinations were not (R^2≤0.01). A rare variant in

FGG, Ala82Gly (EA MAF=0.0036; AA MAF=0.0006; EA+AA $p=2.4\times10^{-28}$), was associated with -0.70 g/L lower levels, on average, for each copy of the minor allele, which is more than one standard deviation from the trait mean. A known common variant within *IRF1* (EA+AA $p=4.2\times10^{-9}$) also reached genome-wide significance (-0.06 g/L per allele). The index variants (the marker with the smallest *P*-value) for *IRF1* and each of the genes in the fibrinogen cluster demonstrated at least a trend ($p<0.05$ in the same direction of effect) in both races.

Conditional analyses involved the 5 variants identified at the fibrinogen gene cluster that met our study-wide significance threshold (Ala82Gly, Tyr345Tyr, Arg448Lys, Ser159Ser, and Thr312Ala). Additionally, a sub-threshold rare variant in *FGB*, Pro176Leu (rs6054; EA MAF=0.004; AA MAF=0.0009; EA+AA $p=4.6\times10^{-6}$) was also included because it was shown to be significantly associated with fibrinogen in a previous study.[12] When the two rare variants (Ala82Gly and Pro235Leu) were included as covariates in the analysis model, the common variants in *FGB* and *FGA* maintained their level of significance (Tyr345Tyr $p<2.1\times10^{-10}$, Thr312Ala $p<2.4\times10^{-9}$). Similarly, when the common variants in *FGB* were included as covariates, the rare variants maintained a similar level of significance (Ala82Gly $p=9.3\times10^{-26}$; Pro235Leu $p=1.2\times10^{-5}$), indicating that the effects are independent. The common Thr312Ala variant in *FGA* represents a third independent effect, as it remained significant when conditioning on either the common *FGB* variants or the rare variants. No additional variation was significantly associated at the *IRF1* locus.

Single variant associations with FVII

There were five loci significantly associated with FVII levels, encompassing known genes *GCKR, ADH4, MS4A6A, PROCR*, and *F7*. All of the index variants in these regions were common. The variant at the *F7* locus had the largest effect, where the index variant, Arg413Gln, decreased FVII values by an average of 17 IU/dL (0.68 standard deviation units) for each copy of the minor allele (EA MAF=0.11; AA MAF=0.12; EA+AA $p=1.8\times10^{-261}$). In addition, there were six rare (MAF<1%) missense variants in *F7* significantly associated with FVII levels. All variants demonstrated significant association within EAs. In AAs, suggestive association in the same direction was observed for variants in *ADH4*, *PROCR*, and *F7* ($p<0.003$), but not for variants in *GCKR* and *MS4A6A* ($p>0.20$). See **Table 2** for full results.

Conditional analyses were conducted at the *F7* locus in order to better understand the contribution of common and rare variation. When the Arg413Gln index variant in *F7* was included as a covariate in the model, signals for two additional common variants in strong linkage disequilibrium with the index SNP ($R^2=0.78$-0.90) were severely attenuated but not abolished. In contrast, all six rare variants maintained their level of significance ($1.8\times10^{-25}<p<5\times10^{-6}$) and were associated with lower FVII

levels between 17 and 50 IU/dL. Minor allele counts for these variants ranged from 11 to 49. Four of these variants were present almost exclusively in AAs (AA MAC≥20; EA MAC≤2), while one was present only in EAs (EA MAC=11), and one was present in both (EA MAC=9; AA MAC=19).

Single variant associations with FVIII

Three loci were significantly associated with FVIII: multiple variants in *ABO* (index variant=rs8176749; EA MAF=0.07; AA MAF=0.16; EA+AA $p=4.3\times10^{-94}$), a variant in *VWF* common only in AAs (His817Gln; rs57950734; EA MAF=0.0001, EA p=0.39; AA MAF=0.10, AA $p=6.0\times10^{-15}$) and a variant in *STAB2* that is more common in EAs (rs7296626; EA MAF=0.06, EA $p=1.3\times10^{-10}$; AA MAF=0.01, AA p=0.07). The index variant at the *ABO* locus tags the O deletion. See **Table 2** for more detail.

When the O deletion at the *ABO* locus was included as a covariate in the model, the *P*-value for a variant that tags the A2 blood type (Pro156Leu) became more extreme (EA conditional model 1 $p=1.4\times10^{-13}$). Pro156Leu was not significant in AAs before (p=0.13) or after (p=0.55) despite being more common in AAs (EA MAF=0.07; AA MAF=0.22). When both variants (Type O and Type A2) were included in the model, then a variant that tags the B blood type that was significant in the unconditional results (Leu266Met; EA MAF=0.07, EA unconditional model $p=1.0\times10^{-44}$; AA MAF=0.16, AA unconditional model $p=4.0\times10^{-52}$) regained significance in AAs (EA conditional model 2 p=0.10; AA conditional model 2 $p=4.8\times10^{-7}$). When all three blood types were included, an uncommon missense variant that tags the O^2 blood group haplotype gained significance in EAs (Gly268Arg; EA MAF=0.02, EA $p=2.9\times10^{-25}$; AA MAF=0.0003, AA p=0.03).

When the His817Gln variant in *VWF* was included as a covariate in the model, a variant common in both AAs and EAs remained significant (Thr789Ala; EA MAF=0.36, EA $p=1.1\times10^{-8}$; AA MAF=0.41, AA $p=3.9\times10^{-6}$). When conditioning on Thr789Ala, a third independent signal that was common only in AAs remained (Arg2185Gln; AA MAF=0.19, AA $p=2.6\times10^{-13}$). In addition, three rare missense variants (MAF<0.01) also remained significant in the trans-ethnic analyses after conditioning on each of the three common variants (Tyr1584Cys $p=3.2\times10^{-13}$; Arg854Gln $p=8.6\times10^{-8}$; and Arg2287Trp $p=8.7\times10^{-6}$). When conditioning on rs7296626 near STAB2, several synonymous variants remained nominally significant (p<0.001) after serial conditional analysis, but did not achieve study-wide significance (Asn1113Asn, Ala1996Ala, Leu80Leu).

Single variant associations with VWF

Five loci were significantly associated with VWF. Three of these variants, the *ABO* O deletion tag (rs8176749), the common *VWF* variant (Thr789Ala) and a synonymous

Table 2. Index variants for the loci with single nucleotide variant associations exceeding study-wide significance ($p<1.9 \times 10^{-7}$)

Trait / Index Variant	Gene	Sig[1]	Function	beta EA+AA[2]	est. effect[3]	effect in SD[4]	*P*-value EA+AA	*P*-value EA	*P*-value AA	MAF[5] EA+AA	MAF[5] EA	MAF[5] AA
Fibrinogen												
rs148685782	*FGG*	5	Ala82Gly	-0.256	-0.70	-1.03	**2.4×10^{-28}**	**4.5×10^{-27}**	0.02	0.003	0.004	0.0006
rs2706379	*IRF1*	1	ncRNA	-0.020	-0.06	-0.09	**4.2×10^{-09}**	**1.6×10^{-07}**	0.01	0.23	0.21	0.28
Factor VII												
rs1260326	*GCKR*	1	Leu446Pro	-0.018	-2.09	-0.08	**6.3×10^{-09}**	**4.3×10^{-09}**	0.38	0.33	0.41	0.14
rs1126670	*ADH4*	4	Pro255Pro	0.019	2.16	0.08	**5.2×10^{-09}**	**4.0×10^{-07}**	0.003	0.27	0.31	0.17
rs12453	*MS4A6A*	2	Leu137Leu	0.017	1.95	0.07	**2.5×10^{-08}**	**3.5×10^{-08}**	0.20	0.34	0.40	0.20
rs6046	*F7*	8	Arg413Gln	-0.157	-17.47	-0.64	**1.8×10^{-261}**	**2.3×10^{-230}**	**7.3×10^{-38}**	0.11	0.11	0.12
rs867186	*PROCR*	4	Ser219Gly	0.054	6.36	0.23	**4.9×10^{-29}**	**1.1×10^{-26}**	0.0002	0.10	0.10	0.09
Factor VIII												
rs8176749	*ABO*	28	Leu310Leu	0.132	17.76	0.46	**4.3×10^{-94}**	**5.7×10^{-46}**	**1.8×10^{-51}**	0.10	0.07	0.16
rs57950734	*VWF*	7	His817Gln	-0.097	-11.98	-0.31	**9.1×10^{-15}**	0.39	**6.0×10^{-15}**	0.03	0.0001	0.10
rs7296626	*STAB2*	2	intronic	0.057	7.51	0.20	**2.5×10^{-11}**	**1.3×10^{-10}**	0.07	0.05	0.06	0.01
von Willebrand Factor												
rs1039084	*STXBP5*	2	Asn436Ser	0.030	3.49	0.08	**1.2×10^{-09}**	**1.0×10^{-07}**	0.003	0.48	0.46	0.44
rs8176741	*ABO*	36	His219His	0.198	24.81	0.54	**8.6×10^{-115}**	**2.3×10^{-60}**	**1.7×10^{-57}**	0.09	0.07	0.16
rs1063856	*VWF*	8	Thr789Ala	0.059	6.81	0.15	**1.8×10^{-30}**	**3.5×10^{-20}**	**1.7×10^{-12}**	0.42	0.36	0.41
rs35102665	*STAB2*	1	Ala1996Ala	0.089	10.83	0.23	**1.1×10^{-09}**	**8.4×10^{-10}**	0.74	0.03	0.04	0.004
rs17564	*STX2*	1	Ser42Thr	-0.037	-4.44	-0.10	**1.7×10^{-12}**	**9.5×10^{-11}**	0.004	0.45	0.35	0.27

[1] sig=number of significant markers ($p<1.9\times10^{-7}$) in the same region

[2] EA=European/European Americans; AA=African Americans; EA+AA=the combine multi-ethnic meta-analysis

[3] All traits were natural-log (ln) transformed, so the effect on the trait in the original units can be estimated using basic algebra

[4] Dividing the estimated effect[4] by the pooled standard deviation in Table 1 gives the effect in standard deviation units

[5] MAF=minor allele frequency

variant in *STAB2* (Ala1996Ala), were also significantly associated with FVIII (see **Table 2** for a comparison). Two variants associated with VWF but not FVIII were in *STXBP5* (Asn436Ser) and *STX2* (Ser42Thr). See **Table 2** for more detail.

The same pattern of associations with *ABO* blood types found to be significant for FVIII was also significant for VWF, with the O deletion having the strongest effect, the A2 group tagged by Pro156Leu (EA $p=6.1\times10^{-12}$; AA p=0.31), the B group tagged by Leu-266Met (EA $p=5.0\times10^{-4}$; AA $p=1.8\times10^{-6}$), and the O^2 group tagged by Gly268Arg (EA $p=1.2\times10^{-26}$; AA p=0.11). Conditional analyses in *STX2* revealed no secondary signal (all 26 variants with MAC>10 had p≥0.08).

After conditioning on the common variant in *VWF* (Thr789Ala), the variant with the next smallest *P*-value was Arg2185Gln (EA MAF=0.002, EA p=0.13; AA MAF=0.19, AA $p=2.7\times10^{-17}$). Despite the high correlation between FVIII and VWF, the variant significantly associated with FVIII after conditioning on both Thr789Ala and Arg2185Gln (His817Gln) was not significant for VWF (p=0.72). Similar to FVIII, additional rare missense variants (Tyr1584Cys, Arg2287Trp, and Ser1486Leu) remained significant after conditioning on the common variants.

Gene-based test results

Results for the gene-based tests (T5 and SKAT) are summarized in **Table 3.** Burden testing revealed significant gene-level associations between fibrinogen levels and *FGG*, as well as between factor VII level and the *F7* gene. Factor VIII and VWF levels were both significantly associated with *VWF* and several genes at the *ABO* locus (*REXO4*, *ADAMTS13*, *SURF2*, *C9orf96*). However, the burden tests for these other genes surrounding *ABO* were no longer significant when conditioning on the variants tagging the common ABO blood types, indicating no evidence of rare functional variants in the region with an independent association with either factor VIII or VWF.

Table 3. Results for gene-based tests of association

				T5 *P*-values			SKAT *P*-values	
Trait	**Gene**	**#variants**	**EA+AA**	**EA**	**AA**	**EA+AA**	**EA**	**AA**
Fibrinogen	*FGG*	78	0.0001	**4.7×10^{-10}**	0.001	**9.1×10^{-13}**	**3.0×10^{-18}**	3.0×10^{-06}
Factor VII	*F7*	115	**1.3×10^{-72}**	**1.1×10^{-19}**	**1.2×10^{-55}**	**2.3×10^{-46}**	**6.0×10^{-19}**	**3.9×10^{-39}**
Factor VIII	*REXO4* (*ABO* locus)	58	**9.7×10^{-07}**	0.03	5.3×10^{-06}	**9.9×10^{-12}**	0.24	**5.7×10^{-12}**
Factor VIII	*VWF*	640	0.0009	1.9×10^{-06}	0.04	**3.2×10^{-14}**	4.3×10^{-06}	1.1×10^{-05}
vWF	*REXO4* (*ABO* locus)	58	0.0002	0.08	0.0004	**8.8×10^{-11}**	0.15	**1.9×10^{-09}**
vWF	*VWF*	640	3.7×10^{-05}	**4.9×10^{-07}**	2.1×10^{-05}	**1.0×10^{-07}**	0.0002	**5.7×10^{-10}**

EA=European/European Americans; AA=African Americans; EA+AA=the combine multi-ethnic meta-analysis; T5=Gene burden test including all functional variants with a minor allele frequency less than 5%; SKAT=Sequence Kernel Association Test; vWF=von Willebrand factor

DISCUSSION

Analysis of exome sequence data allows for the opportunity to identify and analyze coding variation across the full allele frequency spectrum (from common to rare) and to distinguish independent signals from common and rare coding variation within significant loci. Analysis of these four hemostatic factors in a large (n=10,860 EA individuals and 3,529 AA individuals) meta-analysis of multiple multiethnic human population studies indicates that there are new independent signals in known loci that are detectable with sequence data. This study complements a parallel effort that has meta-analyzed the same four hemostatic factors in a larger sample but that was limited to the 250,000 markers on the Illumina HumanExome Beadchip ("exome chip").[32]

Single variant analyses identified associations with fibrinogen at the *FGA*, *FGB*, and *FGG* gene cluster on chromosome 4 that overlap the exome chip study, including the rare *FGG* Ala82Gly mutation that has been reported in a pair of case reports with hypofibrinogenemia.[33,34] In addition, assessment of the significant gene-based test for *FGG* revealed that there were a large number of rare missense or nonsense variants (n=51), including two that were gene-wide significant: the rare Ala82Gly variant identified in the single variant analyses and Ser219Phe, which is only polymorphic in AAs (AA MAF=0.001, AA MAC=8, AA $p=7.4\times10^{-6}$) and which was also significant in the exome chip meta-analysis. The current study also showed that there are independent signals from common (*FGB* Tyr345Tyr and *FGA* Thr312Ala) and rare variation (*FGG* Ala82Gly and *FGB* Pro235Leu) at this locus. Importantly, common and rare variation at this locus may either increase (Tyr345Tyr) or decrease (Ala82Gly, Pro235Leu, Thr312Ala) fibrinogen levels, substantiating the fact that these variants each have important independent contributions to the trait. The contribution from *FGA* (Thr312Ala, not present on the exome chip) is also independent of the common and rare variation in *FGB* and *FGG*. The signal at the *IRF1* gene has also been observed in previous studies.[6,11] In sum, this study identified two novel rare variants in *FGG* (Ala82Gly and Ser219Phe) that have not been previously associated with fibrinogen levels in a population-based study.

This study detected associations between FVII levels and common variation in genes (*GCKR, ADH4, MS4A6A, PROCR*, and *F7*) identified in prior a GWAS[7] and also identified six rare variants in *F7* (all MAF<0.0025) that were independent ($1.8\times10^{-25}<p<5\times10^{-6}$) from the common index variant (Arg413Gln). These variants are not in linkage disequilibrium with each other ($R^2<0.001$) and have not been reported in previous studies. One rare variant is not present on the exome chip (Arg375Trp). These novel rare variants have race-specific contributions: four in AAs (Arg139Gln, Arg364Gln, Ile200Ser, Glu445Lys), one in EAs (Ala354Val), and one in both (Arg-

375Trp). Assessment of the significant gene-based test for *F7* revealed that there were a large number of rare missense or nonsense variants (n=63), including 10 that were gene-wide significant. These include very rare variation not tested in single variant analyses due to their low minor allele count (MAC<40), including an in-frame deletion of three bases present in only five AAs (chr13:113771790:CTGT:C; AA $p=5.8\times10^{-6}$). Only 6 of these 10 significant rare variants and 9 of the 27 variants showing a trend and contributing to the gene-based test are on the exome chip. None of these rare functional variants have been previously associated with FVII levels in a population-based study;[7,8] however, some of these variants are present in the F7 mutation database (http://www.umd.be/F7/W_F7/index.html) and a subset of those have been noted in individuals with FVII deficiency.

The signal for FVIII at the *ABO* locus can be fully explained after taking into account variants tagging the major ABO blood types (A2, B, O, & O^2); however, novel variation was identified at the *VWF* and *STAB2* loci. A previous study[35] reported an association with Pro2039Thr in *STAB2* (chr12:104139034), and here we report an unlinked (R^2=0.003 in the ARIC EAs) novel intronic variant near a splice site of *STAB2* (rs7296626; EA MAF=0.06, EA $p=1.3\times10^{-10}$; AA MAF=0.01, AA p=0.07). Conditional analyses revealed several independent signals at *VWF*, including His817Gln and known common variants Thr789Ala and Arg2185Gln.[36] Importantly, this study identified three rare independent missense variants in VWF, two of which are novel (Tyr1584Cys and Arg854Gln). Both variants were on the exome chip; however, while Arg854Gln reached genome-wide significance in the exome chip analyses, Tyr1584Cys failed genotyping. The Try1584Cys and Arg854Gln signals are driven by the EAs and decreased FVIII levels by 32 and 16 IU/dL, respectively. Assessment of the significant gene-based test for *VWF* revealed that there were a large number of rare functional variants (n=353). These included the four gene-wide significant rare variants driving the gene-based finding: Tyr1584Cys, Arg854Gln, Arg2287Trp (which was only seen in AAs), and Gly2705Arg (which is more common in EAs). In sum, this study confirmed the rare variant associations identified in Johnsen et al.[36] and identified two novel associations with rare variants in *VWF* (Tyr1584Cys, and Arg854Gln) that are associated with FVIII levels.

Among the five loci significantly associated with VWF, there were two common variants that were not associated with FVIII. The Ser42Thr variant in *STX2* is near an intronic *STX2* variant (rs7978987) reported to be associated with VWF in Smith et al.[7] Similarly, Asn436Ser in *STXBP5* is located near synonymous rs9390459 in *STXBP5* in Smith et al.[7] As with FVIII, the association at the *ABO* locus for VWF can be explained after taking into account variants tagging the major ABO blood types. Conditional analyses at VWF revealed independent significant variants, all of which are known[36] except for the novel rare missense variant Tyr1584Cys that was also associated with

FVIII. Assessment of the significant gene-based test for *VWF* revealed that there were a large number of rare functional variants (n=349), including three that were gene-wide significant: Tyr1584Cys, Arg2287Trp, and Ser1486Leu which is only seen in AAs. Associations with Arg2287Trp and Ser1486Leu have been reported previously,[36] while Tyr1584Cys is novel.

In order to maximize power to identify new associations, we used all available samples with whole exome sequencing and these phenotypes in the primary analyses. As a result, there is no available independent replication set. As the cost of sequencing continues to drop, it is likely that other studies with these hemostasis phenotypes will generate sequence data and collaborate. The inclusion of a large number of African Americans has allowed us to identify race-specific variants that would not have been possible with European samples alone. This is one particular advantage over analyses of the exome chip, which was designed primarily using data from white, non-Hispanic samples. In the future, sequencing Asian and Hispanic samples will likely identify additional rare variants associated with hemostatic factors.

Using the largest sample of individuals of both EA and AA descent reported with whole exome sequencing data (total n of over 14,000), we have extended prior studies and identified ten novel associations between rare variants in *FGG*, *F7* and *VWF* and hemostatic factors. While these variants are in genes where common variation is known to be associated with the same trait, the discoveries herein from whole exome sequencing identifies novel independent signals with generally much larger effect sizes than previously reported. This study validates the use of exome sequencing to identify novel variation associated with disease endophenotypes.

REFERENCES

1. Kaptoge S, White IR, Thompson SG, et al. Associations of plasma fibrinogen levels with established cardiovascular disease risk factors, inflammatory markers, and other characteristics: individual participant meta-analysis of 154,211 adults in 31 prospective studies: the fibrinogen studies collaboration. *Am J Epidemiol*. 2007;166(8):867-879.
2. Danesh J, Lewington S, Thompson SG, et al. Plasma fibrinogen level and the risk of major cardiovascular diseases and nonvascular mortality: an individual participant meta-analysis. *JAMA*. 2005;294(14):1799-1809.
3. Freeman MS, Mansfield MW, Barrett JH, Grant PJ. Genetic contribution to circulating levels of hemostatic factors in healthy families with effects of known genetic polymorphisms on heritability. *Arterioscler Thromb Vasc Biol*. 2002;22(3):506-510.
4. Souto JC, Almasy L, Borrell M, et al. Genetic determinants of hemostasis phenotypes in Spanish families. *Circulation*. 2000;101(13):1546-1551.
5. de Lange M, Snieder H, Ariens RA, Spector TD, Grant PJ. The genetics of haemostasis: a twin study. *Lancet*. 2001;357(9250):101-105.
6. Dehghan A, Yang Q, Peters A, et al. Association of novel genetic Loci with circulating fibrinogen levels: a genome-wide association study in 6 population-based cohorts. *Circ Cardiovasc Genet*. 2009;2(2):125-133.
7. Smith NL, Chen MH, Dehghan A, et al. Novel associations of multiple genetic loci with plasma levels of factor VII, factor VIII, and von Willebrand factor: The CHARGE (Cohorts for Heart and Aging Research in Genome Epidemiology) Consortium. *Circulation*. 2010;121(12):1382-1392.
8. Taylor KC, Lange LA, Zabaneh D, et al. A gene-centric association scan for Coagulation Factor VII levels in European and African Americans: the Candidate Gene Association Resource (CARe) Consortium. *Hum Mol Genet*. 2011;20(17):3525-3534.
9. De Visser MC, Sandkuijl LA, Lensen RP, et al. Linkage analysis of factor VIII and von Willebrand factor loci as quantitative trait loci. *J Thromb Haemost*. 2003;1(8):1771-1776.
10. Keightley AM, Lam YM, Brady JN, Cameron CL, Lillicrap D. Variation at the von Willebrand factor (vWF) gene locus is associated with plasma vWF:Ag levels: identification of three novel single nucleotide polymorphisms in the vWF gene promoter. *Blood*. 1999;93(12):4277-4283.
11. Sabater-Lleal M, Huang J, Chasman D, et al. Multiethnic meta-analysis of genome-wide association studies in >100 000 subjects identifies 23 fibrinogen-associated Loci but no strong evidence of a causal association between circulating fibrinogen and cardiovascular disease. *Circulation*. 2013;128(12):1310-1324.
12. Wassel CL, Lange LA, Keating BJ, et al. Association of genomic loci from a cardiovascular gene SNP array with fibrinogen levels in European Americans and African-Americans from six cohort studies: the Candidate Gene Association Resource (CARe). *Blood*. 2011;117(1):268-275.
13. Dehghan A, Yang Q, Peters A, et al. Association of novel genetic Loci with circulating fibrinogen levels: a genome-wide association study in 6 population-based cohorts. *Circ Cardiovasc Genet*. 2009;2(2):125-133.
14. The Atherosclerosis Risk in Communities (ARIC) Study: design and objectives. The ARIC investigators. *Am J Epidemiol*. 1989;129(4):687-702.
15. Fried LP, Borhani NO, Enright P, et al. The Cardiovascular Health Study: design and rationale. *Ann Epidemiol*. 1991;1(3):263-276.
16. Feinleib M, Kannel WB, Garrison RJ, McNamara PM, Castelli WP. The Framingham Offspring Study. Design and preliminary data. *Prev Med*. 1975;4(4):518-525.

17. Kannel WB, Dawber TR, 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*. 1961;55:33-50.
18. Hofman A, Grobbee DE, de Jong PT, van den Ouweland FA. Determinants of disease and disability in the elderly: the Rotterdam Elderly Study. *Eur J Epidemiol*. 1991;7(4):403-422.
19. Hofman A, Breteler MM, van Duijn CM, et al. The Rotterdam Study: 2010 objectives and design update. *Eur J Epidemiol*. 2009;24(9):553-572.
20. Hofman A, Darwish Murad S, van Duijn CM, et al. The Rotterdam Study: 2014 objectives and design update. *Eur J Epidemiol*. 2013;28(11):889-926.
21. Tennessen JA, Bigham AW, O'Connor TD, et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science*. 2012;337(6090):64-69.
22. Fu W, O'Connor TD, Jun G, et al. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature*. 2013;493(7431):216-220.
23. Friedman GD, Cutter GR, Donahue RP, et al. CARDIA: study design, recruitment, and some characteristics of the examined subjects. *J Clin Epidemiol*. 1988;41(11):1105-1116.
24. Design of the Women's Health Initiative clinical trial and observational study. The Women's Health Initiative Study Group. *Control Clin Trials*. 1998;19(1):61-109.
25. Hays J, Hunt JR, Hubbell FA, et al. The Women's Health Initiative recruitment methods and results. *Ann Epidemiol*. 2003;13(9 Suppl):S18-77.
26. Bild DE, Bluemke DA, Burke GL, et al. Multi-ethnic study of atherosclerosis: objectives and design. *Am J Epidemiol*. 2002;156(9):871-881.
27. Bainbridge MN, Wang M, Wu Y, et al. Targeted enrichment beyond the consensus coding DNA sequence exome reveals exons with higher variant densities. *Genome Biol*. 2011;12(7):R68.
28. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*. 2010;38(16):e164.
29. Li B, Leal SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet*. 2008;83(3):311-321.
30. Morris AP, Zeggini E. An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genet Epidemiol*. 2010;34(2):188-193.
31. Wu MC, Lee S, Cai T, et al. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet*. 2011;89(1):82-93.
32. Huffman JE, de Vries PS, Morrison AC, et al. Rare and low-frequency variants and their association with plasma levels of fibrinogen, FVII, FVIII, and vWF. *Blood*. 2015.
33. Brennan SO, Fellowes AP, Faed JM, George PM. Hypofibrinogenemia in an individual with 2 coding (gamma82 A-->G and Bbeta235 P-->L) and 2 noncoding mutations. *Blood*. 2000;95(5):1709-1713.
34. Wyatt J, Brennan SO, May S, George PM. Hypofibrinogenaemia with compound heterozygosity for two gamma chain mutations - gamma 82 Ala-->Gly and an intron two GT-->AT splice site mutation. *Thromb Haemost*. 2000;84(3):449-452.
35. Antoni G, Oudot-Mellakh T, Dimitromanolakis A, et al. Combined analysis of three genome-wide association studies on vWF and FVIII plasma levels. *BMC Med Genet*. 2011;12:102.
36. Johnsen JM, Auer PL, Morrison AC, et al. Common and rare von Willebrand factor (VWF) coding variants, VWF levels, and factor VIII levels in African Americans: the NHLBI Exome Sequencing Project. *Blood*. 2013;122(4):590-597.

Supplemental Table 1. Sample demographics

Study	Race	Age (SD)	%Female	Fibrinogen in g/L (SD)	FVII in IU/dL (SD)	FVIII in IU/dL (SD)	vWF in IU/dL (SD)
ARIC	EA	54.4 (5.7)	52.5%	2.96 (0.63)	118.98 (28.72)	124.47 (33.08)	111.24 (41.13)
ARIC	AA	53.2 (5.8)	63.7%	3.18 (0.67)	117.52 (28.91)	146.42 (44.91)	133.7 (55.48)
CHS	EA	72.9 (5.8)	52.5%	3.24 (0.69)	126.35 (28.61)	124.84 (38.01)	
FHS	EA	53.5 (9.7)	53.5%	3.10 (0.63)	101.24 (16.24)		125.36 (43.72)
RS-I-1	EU	69.2 (4.7)	59.4%	2.74 (0.62)	108.55 (18.55)	116.50 (50.83)	
RS-I-3	EU	71.5 (4.8)	52.5%	3.85 (0.85)			133.25 (72.05)
ESP	EA	59.9 (11.8)	57.2%	3.29 (0.82)	119.45 (32.50)	129.03 (48.36)	116.52 (47.59)
ESP	AA	58.7 (10.3)	69.7%	3.39 (0.82)	120.14 (31.42)	160.35 (67.52)	135.91 (58.35)
Weighted Average		**58.22 (6.18)**	**56.8%**	**3.09 (0.68)**	**115.85 (27.15)**	**128.65 (38.4)**	**117.71 (46.16)**

Chapter 2.5

Genome-wide association study of ADAMTS13 activity

Manuscript based on this chapter

Paul S. de Vries, Johan Boender, Michelle A.H. Sonneveld, Fernando Rivadeneira, M. Arfan Ikram, Hanspeter Rottensteiner, Albert Hofman, André G. Uitterlinden, Frank W.G. Leebeek, Oscar H. Franco, Abbas Dehghan*, and Moniek P.M. de Maat*.

*contributed equally to this study as senior authors.

Genetic variants in the ADAMTS13 and SUPT3H genes are associated with ADAMTS13 activity.

Blood. 2015; 125(25): 3949-55.

ABSTRACT

Background: ADAMTS13 cleaves von Willebrand factor, reducing its prothrombotic activity. The genetic determinants of ADAMTS13 activity remain unclear.

Methods: We performed a genome-wide association study of ADAMTS13 activity in the Rotterdam Study, a population-based cohort study. We used imputed genotypes of common variants in a discovery sample of 3,443 individuals and replication sample of 2,025 individuals. We examined rare exonic variant associations in *ADAMTS13* in 1,609 individuals using an exome array.

Results: rs41314453 in *ADAMTS13* was associated with ADAMTS13 activity in both our discovery (Beta: -20.2%, *P*-value: 1.3×10^{-33}) and replication sample (*P*-value: 3.3×10^{-34}), and explained 3.6-6.5% of the variance. In the combined analysis of our discovery and replication samples, there were two further independent associations at the *ADAMTS13* locus: rs3118667 (Beta: 3.0, *P*-value: 9.6×10^{-21}) and rs139911703 (Beta: -11.6, *P*-value: 3.6×10^{-8}). Additionally, rs10456544 in *SUPT3H* was associated with a 4.2 increase in ADAMTS13 activity (*P*-value: $1.13.6\times10^{-8}$). Finally, we found three independent associations with rare coding variants in *ADAMTS13*: rs148312697 (Beta: -32.2%, *P*-value: 3.7×10^{-6}), rs142572218 (Beta: -46.0%, *P*-value: 3.9×10^{-5}), and rs36222275 (Beta: -13.9%, *P*-value: 2.9×10^{-3}).

Conclusions: We identified rs41314453 as the main genetic determinant of ADAMTS13 activity, and present preliminary for further associations at the *ADAMTS13* and *SUPT3H* loci.

INTRODUCTION

ADAMTS13 (A Disintegrin And Metalloproteinase with ThromboSpondin motifs 13) cleaves ultra large von Willebrand Factor (VWF) into smaller multimers.[1-3] ADAMTS13 thereby greatly reduces the activity of VWF in its role in platelet adhesion and aggregation. Through this effect on VWF, ADAMTS13 has antithrombotic properties.

The role of ADAMTS13 in thrombosis is especially evident in patients with thrombocytopenic thrombotic purpura (TTP), a disorder resulting from a severe deficiency of ADAMTS13: patients with TTP have a wide range of symptoms, including thrombocytopenia and microangiopathy, which may result in stroke, and myocardial infarction.[4] Beyond patients with TTP, we and others recently showed that low ADAMTS13 activity and levels within the normal range are also associated with increased risk of cardiovascular outcomes.[5-9]

These associations between ADAMTS13 activity and arterial thrombosis raise the question of how ADAMTS13 activity is regulated. Several rare single nucleotide polymorphisms (SNPs) in the *ADAMTS13* gene causing TTP have been identified along with a few common variants with more modest effects on ADAMTS13.[10,11] However, it is not known whether these associations are independent of each other, or even whether they exhibit the strongest associations at the locus. Furthermore, the role of genetic variation outside of the ADAMTS13 locus remains unknown. The optimal method to identify genetic determinants is a genome-wide association (GWA) study, with a hypothesis-free approach. To date, no studies on the genetics of ADAMTS13 using this approach have been reported.

Thus, in the Rotterdam Study, a large population-based cohort study, we conducted a GWA study of ADAMTS13 activity, including a conditional analysis to identify multiple independent signals. Additionally, we characterized the *ADAMTS13* gene and any other genes with associated common variants by examining the role of rare variants.

METHODS

Study description and population

The Rotterdam Study is a prospective, population-based cohort study of determinants of several chronic diseases in older adults.[12,13] The first cohort (RS-I), includes 7,983 inhabitants of Ommoord, a district of Rotterdam in the Netherlands, who were 55 years or older. The baseline examination took place between 1990 and 1993. The third visit took place between March 1997 and December 1999, and included 4,797 participants. A second cohort (RS-II) was established between February 2000 and

December 2001, including another 3,011 inhabitants of Ommoord who reached the age of 55 years after the baseline examination of RS-I, and individuals aged 55 years or older who had migrated into the research area. The study was approved by the Medical Ethics Committee of Erasmus University, Rotterdam, the Netherlands, and all included participants gave their written informed consent.

ADAMTS13 measurement

Citrated plasma samples were collected at the third visit of RS-I and the baseline examination of RS-II, and stored at -80°C. Between June and October 2013, we measured ADAMTS13 activity using a kinetic assay based on the Fluorescence Resonance Energy Transfer Substrate VWF 73 (FRETS-VWF73) assay.[14] This assay uses a peptide containing the ADAMTS13 cleavage site of VWF, and thus captures variation in the VWF cleavage rate determined by ADAMTS13 levels and structure, but not by alterations in VWF.

Plasma samples were measured against a reference curve of serial dilutions of normal human plasma defined to have an ADAMTS13 activity of 1 IU/ml, and we express ADAMTS13 activity as a percentage of this. In total, the ADAMTS13 activity of 6,258 participants was measured: 3,791 from RS-I, and 2,467 from RS-II.

Genotyping and imputation

We used two sources of genetic variants: genome-wide SNPs genotyped by the Illumina Infinum II HumanHap550 array or 610 quad array and exome-wide SNPs genotyped by the Illumina HumanExome BeadChip v1.0. We genotyped 6,291 participants from RS-I and 2,157 participants from RS-II using the Illumina Infinium II HumanHap550 or 610 quad arrays. All genotyped participants were of European ancestry based on their self-report. Prior to imputation, genotyped SNPs with a call rate below 98%, a minor allele frequency (MAF) below 1%, or a hardy-weinberg equilibrium *P*-value of less than 1×10^{-6} were excluded. In RS-I 512,849 SNPs remained after filtering and these were used for imputation. In RS-II, 537,405 SNPs were used for imputation. Dosages of 19,537,258 SNPs were imputed in both studies using the Genomes of the Netherlands (GoNL) version 4 reference panel.[15-17] MACH version 1.0.15 was used to perform the imputations. The imputation quality of each SNP defined as the estimated squared correlation of imputed and true genotypes, and ranged from 0 to 1. After imputation, SNPs with a MAF below 0.01 or an imputation quality below 0.3 were excluded. The overlap between participants with ADAMTS13 activity measurements and genotypes was 3,423 in RS-I, and 2,025 in RS-II.

Exonic variants of 3,163 individuals from RS-I were successfully genotyped using the Illumina HumanExome BeadChip v1.0. In 1,609 of these individuals ADAMTS13 was measured. Genotype calling was performed at the University of Texas Health

Science Center in Houston, together with ten other cohorts from the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium.[18] This joint calling in a total of 62,000 individuals was done to improve the calling of rare variants compared to what could be accomplished in RS-I alone. A total of 108,678 SNPs were included after filtering out monomorphic SNPs and SNPs with low call rate.

Common variant analysis

We performed a discovery GWA analysis in RS-I. In the discovery, linear regression, as implemented in ProbABEL version 0.4.3, was used to examine the association of each SNP with ADAMTS13 activity, adjusted for age and sex.[19] SNPs were analysed in the form of genotype dosages (ranging from 0 to 2) using an additive model. A genome-wide significance threshold of 5×10^{-8} was used. Regional plots were created using LocusZoom.[20]

Replication analyses in RS-II were also performed using ProbABEL version 0.4.3. The significance threshold was determined using a Bonferroni correction for the number of SNPs. The variance of ADAMTS13 activity explained by replicating SNPs was examined, with R version 3.1.1. We used HaploReg V2 to browse ENCODE resource to examine the functional implication of these SNPs, along with any correlated SNPs (correlation $R^2 > 0.8$).[21,22]

Lastly, to maximize our power and the accuracy of our effect estimates, we also performed a meta-analysis of RS-I and RS-II. We used an inverse-variance model with fixed effects as implemented in METAL.[23] We applied a genomic control correction to the combined results to account for genomic inflation. To identify secondary signals at significant loci, we then performed a stepwise conditional analysis repeating the GWA analysis adjusted for the most significant variant in each locus (defined as +/- 250KB of the top SNP). This approach was repeated with additional adjustment for secondary signals until no further genome-wide significant signals remained.

Rare variant analysis

In a subset of RS-I participants, we used the exome chip to examine the effect of rare variants. To maximize our power, we included only SNPs within genes that were highlighted in the common variant analysis. Additionally, we only included SNPs that were functional according to the dbNSFP database (missense, stop-gain, stop-loss or splice site) with a MAF below 0.01.[24] We then used the seqMeta package implemented in R to determine the association between the rare variant burden in selected genes and ADAMTS13 activity, and to examine the association of individual SNPs.

This package has previously been described in further detail.[25] Rare variant burden analysis was performed using both a T1 test, and a sequence kernel association test (SKAT).[26] In T1 tests, the sum of rare variant dosages is created for each gene, and associated with the traits of interest. T1 tests are unidirectional: they are more powerful when, within a gene, the effect sizes of rare variants are consistently in the same direction. SKAT is a bidirectional test and is more powerful when the effect direction of rare variants within a gene varies. Single variant analysis was done using score tests. All analyses were adjusted for age, sex, and the independently significant common variants. Additionally, the analyses were adjusted for four ancestry-informative principal components, as rare variants are more susceptible population stratification.[27] Finally, we performed stepwise conditional analysis to determine whether rare variant associations were independent from each other.

Estimation of the heritability

In RS-I, we estimated the proportion of variance of ADAMTS13 activity explained by all SNPs together. First, we constructed a matrix of pairwise genetic relationships based on common (MAF ≥ 0.01) well-imputed (imputation quality > 0.3) SNPs. We excluded one individual from each pair with a pairwise relationship larger than 0.025, reducing the number of included individuals to 2455. We then used a restricted maximum likelihood model to estimate the proportion of variance explained by the genetic relationships. The result can be interpreted as the lower bound of the heritability.[28] The estimated heritability is expected to be lower than the true heritability because it is based on imperfectly imputed SNPs that may in turn be only partially correlated to the underlying causal variants.

We then calculated the variance explained by the combination of independently significant variants using the adjusted R-squared resulting from a linear regression model in R. We did this separately for RS-I and RS-II.

Additionally, to place genetic determinants of ADAMTS13 into a wider context, we estimated the variance of ADAMTS13 activity explained by genome-wide significant SNPs, as well as by baseline characteristics including age, sex, total and high density lipoprotein (HDL) cholesterol, prevalent type 2 diabetes, current smoking, body mass index (BMI), and systolic and diastolic blood pressure. We used the partial.R2 function from the asbio package in R. All variables were included in a single linear regression model, and the resulting partial coefficients of determination indicate the variance explained on top of the other variables in the model.

Table 1. Characteristics of the participants included in the discovery in the Rotterdam Study I (RS-I) and in the replication in the Rotterdam Study II (RS-II).

	RS-I	RS-II
Sample Size	3423	2025
Age (years)	72.4 ±7.0	64.6 ±7.9
Sex (% males)	41.3	45.1
ADAMTS13 activity (%)	89.5 ±17.4	95.0 ±17.6
BMI (kg/m^2)	26.8 ±3.9	27.2 ±4.0
Current smoking (%)	15.8	19.5
Total cholesterol (mmol/L)	5.8 ±1.0	5.8 ±1.0
HDL cholesterol (mmol/L)	1.4 ±0.4	1.4 ±0.4
Systolic blood pressure (mmHg)	143.3 ±21.0	143.1 ±21.3
Diastolic blood pressure (mmHg)	75.2 ±11.2	78.9 ±10.8
Prevalent Type 2 Diabetes (%)	14.1	11.5

Abbreviations: BMI refers to body mass index, and HDL refers to high-density lipoprotein. Continuous variables are summarized by their mean ± standard deviation.

RESULTS

Discovery in RS-I and replication in RS-II

Participant characteristics are shown in **Table 1**. Participants in RS-I were older (mean age = 72.4 years old, standard deviation = ±7.0) than participants in RS-II (mean age = 64.6 years old, standard deviation = ±7.9). The mean ADAMTS13 was 89.5% in RS-I and 95.0% in RS-II, with a range of 5% to 198% across the two cohorts. After removing rare and poorly imputed SNPs, 8,237,900 SNPs were included in the discovery GWA analysis, of which 329 were significantly associated with ADAMTS13 activity (**Supplemental Figure 1 and 2**). All of these SNPs mapped to the *ADAMTS13* locus. The minor allele of the lead SNP, rs41314453, was associated with a 20.2% decrease in ADAMTS13 activity (*P*-value = 1.3×10^{-33}). The signal was successfully replicated in RS-II: the minor allele of rs41314453 was associated with a 23.5% decrease in ADAMTS13 activity (*P*-value = 3.3×10^{-34}).

Combined analysis of RS-I and RS-II

In the combined analysis of RS-I and RS-II rs41314453 was also the lead variant at the *ADAMTS13* locus (**Table 2** and **Figure 1A**). There was one genome-wide significant SNP outside of the *ADAMTS13* locus: rs10456544, an intronic SNP in the *SUPT3H* gene (**Table 2** and **Figure 1B**). The minor allele was associated with a 4.2% increase in ADAMTS13 activity. After adjustment for rs41314453 and rs10456544, there were no significant variants remaining at the *SUPT3H* locus, but there was a second signal at the *ADAMTS13* locus. The minor allele of lead variant rs3118667 was associated

with 3.0% increase in ADAMTS13 activity (**Table 2**). When additionally adjusting for rs3118667, there was a third genome-wide significant signal at the *ADAMTS13* locus. The minor allele of the lead variant, rs139911703, was associated with an 11.6% decrease in ADAMTS13 activity (**Table 2**).

Table 2. Association of common variants with ADAMTS13 activity in the combined analysis of RS-I and RS-II.

SNP Name	Chromosome	Position*	Gene	Effect / Other Allele	Frequency	Imputation Quality	Beta	*P*-value
Adjusted for age, sex, and principal components 1-4								
rs41314453	9	136,307,825	*ADAMTS13*	T/C	1.88%	0.84	-21.7	1.2×10^{-63}
rs10456544	6	45,181,694	*SUPT3H*	A/T	7.11%	0.69	4.2	1.1×10^{-8}
Additional adjustment for rs41314453 and rs10456544								
rs3118667	9	136,291,063	*ADAMTS13*	C/T	47.09%	0.93	3.0	9.6×10^{-21}
Additional adjustment for rs3118667								
rs139911703	9	136,081,887	*OBP2B*	A/G	1.10%	0.52	-11.6	3.6×10^{-8}

Abbreviations: SNP refers to single nucleotide polymorphism. Frequency refers to the frequency of the effect allele as a percentage. Beta refers to the beta coefficient, and should be interpreted as the change in ADAMTS13 activity (%) per 1 allele increase in the effect allele. *The DNA position is coded according the build 37.

Rare variant analyses

There were 11 functional SNPs with MAF < 0.01% in *ADAMTS13* and 4 in *SUPT3H*. For single variant analysis, we thus used a *P*-value threshold of 0.0033. Three rare variants were associated with ADAMTS13 activity: rs148312697 (Beta = -32.8, *P*-value = 3.6×10^{-6}, Frequency = 0.16%), rs142572218 (Beta = -46.0, *P*-value = 3.9×10^{-5}, Frequency = 0.06%), and rs36222275 (Beta = -14.7, *P*-value = 2.2×10^{-3}, Frequency = 0.34%). The association of these variants was independent of the three associated common variants in *ADAMTS13* (**Table 3**), and stepwise conditional analysis suggests that the associations are also independent of each other (**Supplemental Table 1**).

The spread across the functional domains of ADAMTS13 of these associated rare nonsynonymous variants, as well as the associated common nonsynonymous variant (rs41314453), is shown in **Figure 2**. None of the rare variants in *SUPT3H* was significantly associated to ADAMTS13 activity.

Although we only examined two genes, we used a *P*-value threshold of 0.013 to adjust for doing both SKAT and T1 tests. The 11 variants in ADAMTS13 had a cumulative minor allele frequency of 1.1%. Rare variant burden in *ADAMTS13* was associated with ADAMTS13 activity according to both the T1 (*P*-value = 5.7×10^{-8}) and SKAT test (*P*-value = 1.5×10^{-6}). These associations remained significant after adjusting for the

Table 3. Association of rare non-synonymous exonic variants in the *ADAMTS13* gene with ADAMTS13 activity, adjusted for common variants rs41314453, rs3118667, and rs139911703.

SNP Name	Amino Acid Change	Position*	Exon	Effect / Other Allele	Frequency	Beta	*P*-value
rs148312697	Asp187His	136,291,338	6	C/G	0.16%	-32.1	3.3×10^{-6}
rs142572218	Arg1060Trp	136,319,670	24	T/C	0.06%	-46.7	1.8×10^{-5}
rs36222275	Gly982Arg	136,314,986	23	A/G	0.34%	-13.3	4.4×10^{-3}

Abbreviations: SNP refers to single nucleotide polymorphism. Frequency refers to the frequency of the effect allele. Beta refers to the beta coefficient, and should be interpreted as the change in ADAMTS13 activity (%) per 1 allele increase in the effect allele. *The DNA position is coded according the build 37, and refers to the position on chromosome 9.
Adjustments: Age, sex, principal components 1-4, rs41314453, rs3118667, and rs139911703.

three associated common variants in *ADAMTS13* (**Supplemental Table 2**). When we additionally adjusted the burden tests for the three rare SNPs in a stepwise manner, the association diminished with each step, and finally lost significance upon adjustment for all three rare SNPs (**Supplemental Table 2**). The rare variant burden in *SUPT3H* was not associated to ADAMTS13 activity according to the T1 (*P*-value = 0.5) and SKAT tests (*P*-value = 0.7).

Estimation of the heritability

The variance of ADAMTS13 activity explained by all SNPs in RS-I was 35.2% (*P*-value = 0.009), which can be interpreted as the lower bound of the heritability. The variance explained by the four independently significant common SNPs was 5.8-8.2%. The variance of ADAMTS13 activity explained by each of the four independently significant common SNPS on top of other baseline characteristics is shown in **Supplemental Table 3**. This table also shows the variance explained by other baseline characteristics. The variance explained by rs41314453 (3.6-6.5%) is comparable to the variance explained by age (3.9-6.5%) as well as the variance explained by sex (4.5-6.7%). The variance explained by rs3118667 (1.3-2.1%) is comparable to the variance explained by current smoking (1.5-1.7%). Because the estimates for SNPs are based on imputed dosages rather than directly measured genotypes, the actual variance explained by the SNPs is likely to be higher.

DISCUSSION

In this first-ever GWA study of ADAMTS13 activity, we robustly identified rs41314453 within the *ADAMTS13* gene as the main genetic determinant of ADAMTS13 activity in both our discovery and replication cohort, explaining between 3.6 and 6.5 percent

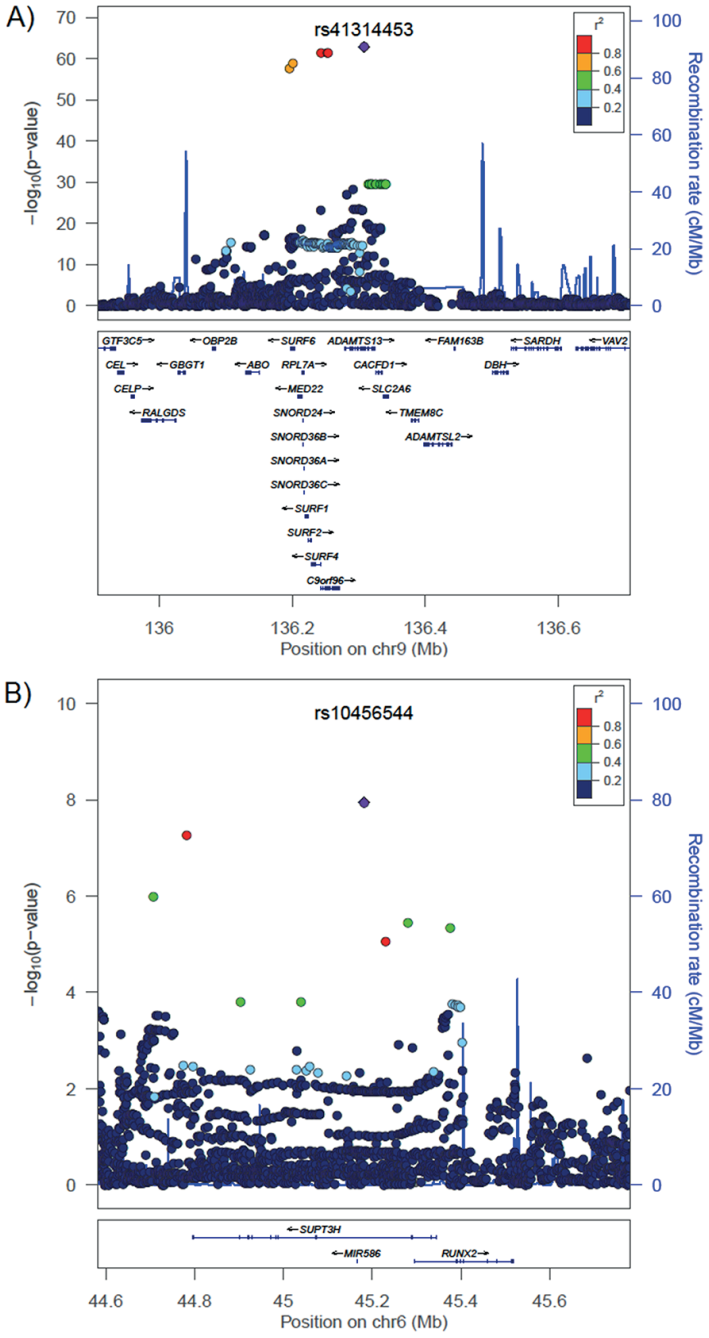


Figure 1. Regional plots of the association between ADAMTS13 activity and A) the *ADAMTS13* locus and B) the *SUPT3H* locus in the combined GWA analysis.
Linkage disequilibrium of variants is shown with A) rs41314453, B) rs10456544.

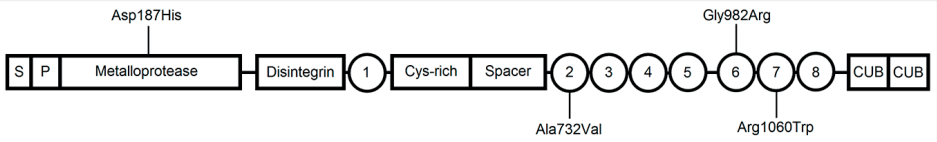


Figure 2. Location of the independently associated nonsynomymous variants across the functional domains of ADAMTS13.
Asp187His is rs148312697, Ala732Val is rs41314453, Gly982Arg is rs36222275, and Arg1060Trp is rs142572218. Thrombospondin type 1 repeats 1-8 are shown as circles. Cys-rich indicates the cysteine rich domain, and CUB indicates the C1r-C1s, urinary epidermal growth factor, and bone morphogenetic protein domains.

of the variance. Through the combined analysis of our discovery and replication samples, we present preliminary evidence of independent associations with two further SNPs in *ADAMTS13* (rs3118667 and rs139911703), and with a SNP in the *SUPT3H* gene (rs10456544). Furthermore, in a subset of our discovery sample, we found 3 independently associated rare variants in *ADAMTS13* (rs148312697, rs142572218, and rs36222275). Finally, we established a lower bound for the heritability of ADAMTS13 activity at 35%.

The most significant SNP, rs41314453, is a nonsynonymous exonic variant in the thrombospondin type 1 repeat 2 domain that is also known as Ala732Val. It is in linkage disequilibrium with several intronic SNPs in *ADAMTS13*, as well as to SNPs in regulatory regions of neighbouring genes. However, rs41314453 remains the most promising candidate causal SNP, because it has previously been shown, *in vitro*, to reduce ADAMTS13 levels by 40% and ADAMTS13 activity by 29%.[29] The decrease in activity appeared to be mediated completely by the decrease in protein concentration rather than a decrease in the specific activity (activity per milligram of ADAMTS13), and the decrease in levels was not linked to decreased synthesis.[29] This suggests that the underlying mechanism is a decreased secretion of ADAMTS13.

The secondary signal at the *ADAMTS13* locus, rs3118667, is a synonymous SNP that has not previously been reported to be associated with ADAMTS13. It is not in strong linkage disequilibrium with other SNPs. Thus, the mechanism behind this signal is unclear. The third signal at the *ADAMTS13* locus, rs139911703, is an intronic SNP in *OBP2B*. It is not strongly correlated to any variant in the *ADAMTS13* gene, but it is in perfect linkage disequilibrium with rs36218903, an intronic variant in *ABO*. The underlying mechanism may thus involve the *ABO* gene although we cannot exclude an effect on the regulation of the *ADAMTS13* gene, or correlation with an unknown coding variant. It is unclear how *ABO* could regulate ADAMTS13 activity. Variation in the glycan structures attached to VWF that are encoded by the *ABO* gene has been linked to the cleavage rate: cleavage was faster with VWF originating from individuals with blood group O than with VWF originating from individuals with non-O

blood groups.[30,31] However, this effect on the cleavage rate is not reflected in the ADAMTS13 activity measurements in this study, as the measurements are based on an introduced peptide spanning the VWF cleavage site.

Only one SNP outside of the *ADAMTS13* locus was associated with ADAMTS13 activity: rs10456544 in *SUPT3H*, which encodes the protein Spt3.[32] As a part of the SPT3-TAF9-GCN5 acetyltransferase (STAGA) complex, Spt3 is involved in transcription activation.[33] The STAGA complex acetylates histones, reconfiguring the DNA around the histones into a more accessible structure, allowing for increased transcription.[34] In yeast, around 3% of the genome is dependent on Spt3 for expression.[35] The main role of the Spt3 subunit in STAGA is to recruit the transcription factor II D complex (TFIID), which then binds to TATA box motifs in promoters, enabling RNA polymerase II to position itself appropriately for transcription.[36] The *ADAMTS13* promotor does not have a known TATA box motif, but it does have an Sp1 binding site, which can allow TFIID to bind to TATA-less promoters.[37,38] We thus hypothesize that rs10456544 is associated to ADAMTS13 activity through a disturbance to these basal transcription activation processes. As ADAMTS13 does not appear to be heavily regulated by transcription factors, the sensitivity to these processes might be increased.[37] The possible relationship between Spt3 and ADAMTS13 activity should be confirmed through replication of the association and functional work.

Of the three associated rare SNPs, rs148312697 (Asp187His), located in the metalloprotease domain, has been shown in mice to reduce ADAMTS13 activity and secretion and to cause TTP.[39] Another variant at the same position (Asp187Ala) has also been shown to reduce proteolytic function.[40] rs142572218 (Arg1060Trp) has been identified as a causal mutation for late-onset adult TTP, and has been shown to profoundly decrease secretion, but not the specific activity.[41] rs36222275 (Gly982Arg) has not previously been associated to ADAMTS13 activity. The effect size is smaller than that of the other two rare variants and rs41314453, the lead common variant. We were able to identify this rare variant with an intermediate effect size because of our hypothesis driven approach, but it will need to be confirmed either in vitro or through replication in other association studies.

Nonsynonymous variant rs28647808, or Pro618Ala, has previously been used as a genetic proxy of ADAMTS13 activity.[42] Indeed, several lines of experimental evidence support a causal role for Pro618Ala.[29,43] In the combined analysis of our discovery and replication samples, Pro618Ala was well-imputed (imputation quality > 0.9), and was associated with ADAMTS13 activity (Beta = -4.5, *P*-value = 7.3×10^{-16}, Frequency = 9.8%). However, this association disappeared after adjusting for the lead variant, rs41314453, with which it is in modest linkage disequilibrium (R^2 = 0.18). In line with our results, studies by Miyata et al and Kokame et al have found no association between Pro618Ala and ADAMTS13 activity in the Japanese general population.[43,44]

Our results therefore do not support a causal role of rs28647808 in the regulation of ADAMTS13 activity, and suggest that rs41314453 may be a more suitable genetic proxy for future studies.

Similarly, another polymorphism that has been associated to ADAMTS13 activity in the literature,[44] rs2301612 or Gln448Glu, was not strongly associated with ADAMTS13 activity in our study (Beta = 1.6, *P*-value = 1.4×10^{-6}, Frequency = 43.6%). The effect direction was consistent with the literature. Interestingly, the association became stronger upon adjustment for *ADAMTS13* lead variant rs41314453 (Beta = 2.6, *P*-value = 1.1×10^{-15}), but was again attenuated when further adjusted for secondary variant rs3118667 (Beta = 1.3, *P*-value = 1.4×10^{-3}).

In the discovery GWA analysis, we only found associations with SNPs within the *ADAMTS13* gene itself. In the combined analysis of the discovery and replication samples only one SNP at another locus was genome-wide significant. While this is likely related to the small sample size, the unbalanced genetic architecture is not surprising. ADAMTS13 is constantly synthesized and secreted in its active form. Previous work suggests that ADAMTS13 transcription is stable and not significantly regulated by transcription factors.[37] This leaves little room for strong regulators. Furthermore, while several factors are known to influence the rate at which ADAMTS13 cleaves VWF, these are not captured by the measurement of ADAMTS13 activity. The measurement is based on the rate at which an introduced peptide similar to VWF is cleaved. However, *in vivo*, alterations to VWF that disrupt its interactions with ADAMTS13 may also affect the cleavage rate. For example, mutations involved in type 2A von Willebrand disease have been shown to increase the cleavage rate.[45]

Apart from synthesis and secretion, ADAMTS13 activity is further determined by degradation, and the specific activity. ADAMTS13 degradation is known to occur in the presence of thrombin and plasmin.[46] However, the level of ADAMTS13 degradation is minimal, since coagulation and fibrinolysis are normally only occurring at a very low level. We therefore expect the regulation of ADAMTS13 degradation to explain a very small part of the genetic associations with ADAMTS13 activity.

In patients with congenital ADAMTS13 deficiency, who often suffer from TTP, the main underlying mechanisms are changes in secretion and specific activity.[11] This is in line with our results in this population based study. Functional work has previously been done for three of the variants associated with ADAMTS13 activity in our study, and two of these reduce secretion, while one reduces the specific activity.[29,40,41]

The strengths of this study include our genome-wide hypothesis-free approach, which, in contrast to the targeted genotyping of a few candidate SNPs, allowed us to systematically examine the ADAMTS13 locus. Secondly, the use of GoNL as a reference panel for the imputation of unmeasured SNPs was particularly appropriate, as this reference panel is based specifically on the Dutch population. Thirdly, we

were able to replicate our common variant results in a non-overlapping sample that was ethnically similar to the discovery sample and used the same assay to measure ADAMTS13 activity. Finally, the rare variant and conditional analyses we performed allowed us to gain a detailed view of the *ADAMTS13* locus.

However, while two of the rare variant associations were backed up by previous functional work, we were not able to replicate our rare variant associations because participants in RS-II were not genotyped using the exome chip. Neither were we able to replicate the associations with rs3118667 and rs139911703 in *ADAMTS13* nor rs10456544 in *SUPT3H*, as these associations were identified by combining our discovery and replication samples. Additionally, the limited sample size allowed us to detect only the strongest associations with ADAMTS13 activity. This will be improved as more studies with genome-wide SNP array data measure ADAMTS13 activity or levels. Although we replicated our results in a non-overlapping sample, both samples were from the Rotterdam Study and were measured together. Thus, the samples were not completely independent from one another. Finally, our estimate of the heritability should be interpreted as the lower bound of the heritability for two reasons. First, it is based on imperfectly imputed SNPs that may in turn be only partially correlated to the underlying causal variants. Second, it is only based on common SNPs, while a portion of the heritability is likely to stem from rare variants. Estimates from twin and family studies are required for further precision.

In conclusion, in our study we robustly identified a strong association between rs41314453 in the *ADAMTS13* gene and ADAMTS13 activity, and we present preliminary evidence of association with another five genetic variants in *ADAMTS13* and one variant in the *SUPT3H* gene. Explaining between 3.6 and 6.5 percent of the variance, rs41314453 appears to be the main genetic determinant of ADAMTS13 activity.

Supplement available online at:
http://www.bloodjournal.org/content/125/25/3949

REFERENCES

1. Fujikawa K, Suzuki H, McMullen B, Chung D. Purification of human von Willebrand factor-cleaving protease and its identification as a new member of the metalloproteinase family. *Blood*. 2001;98(6):1662-1666.
2. Gerritsen HE, Robles R, Lammle B, Furlan M. Partial amino acid sequence of purified von Willebrand factor-cleaving protease. *Blood*. 2001;98(6):1654-1661.
3. Zheng X, Chung D, Takayama TK, et al. Structure of von Willebrand factor-cleaving protease (ADAMTS13), a metalloprotease involved in thrombotic thrombocytopenic purpura. *J Biol Chem*. 2001;276(44):41059-41063.
4. Tsai HM. Thrombotic thrombocytopenic purpura: a thrombotic disorder caused by ADAMTS13 deficiency. *Hematol Oncol Clin North Am*. 2007;21(4):609-632, v.
5. Bongers TN, de Bruijne EL, Dippel DW, et al. Lower levels of ADAMTS13 are associated with cardiovascular disease in young patients. *Atherosclerosis*. 2009;207(1):250-254.
6. Sonneveld MA, de Maat MP, Leebeek FW. Von Willebrand factor and ADAMTS13 in arterial thrombosis: a systematic review and meta-analysis. *Blood Rev*. 2014;28(4):167-178.
7. Andersson HM, Siegerink B, Luken BM, et al. High VWF, low ADAMTS13, and oral contraceptives increase the risk of ischemic stroke and myocardial infarction in young women. *Blood*. 2012;119(6):1555-1560.
8. Chion CK, Doggen CJ, Crawley JT, Lane DA, Rosendaal FR. ADAMTS13 and von Willebrand factor and the risk of myocardial infarction in men. *Blood*. 2007;109(5):1998-2000.
9. Crawley JT, Lane DA, Woodward M, Rumley A, Lowe GD. Evidence that high von Willebrand factor and low ADAMTS-13 levels independently increase the risk of a non-fatal heart attack. *J Thromb Haemost*. 2008;6(4):583-588.
10. Tseng SC, Kimchi-Sarfaty C. SNPs in ADAMTS13. *Pharmacogenomics*. 2011;12(8):1147-1160.
11. Lotta LA, Garagiola I, Palla R, Cairo A, Peyvandi F. ADAMTS13 mutations and polymorphisms in congenital thrombotic thrombocytopenic purpura. *Hum Mutat*. 2010;31(1):11-19.
12. Hofman A, Darwish Murad S, van Duijn CM, et al. The Rotterdam Study: 2014 objectives and design update. *Eur J Epidemiol*. 2013;28(11):889-926.
13. Hofman A, Grobbee DE, de Jong PT, van den Ouweland FA. Determinants of disease and disability in the elderly: the Rotterdam Elderly Study. *Eur J Epidemiol*. 1991;7(4):403-422.
14. Kokame K, Nobe Y, Kokubo Y, Okayama A, Miyata T. FRETS-VWF73, a first fluorogenic substrate for ADAMTS13 assay. *Br J Haematol*. 2005;129(1):93-100.
15. Boomsma DI, Wijmenga C, Slagboom EP, et al. The Genome of the Netherlands: design, and project goals. *Eur J Hum Genet*. 2014;22(2):221-227.
16. Deelen P, Menelaou A, van Leeuwen EM, et al. Improved imputation quality of low-frequency and rare variants in European samples using the 'Genome of The Netherlands'. *Eur J Hum Genet*. 2014;22(11):1321-1326.
17. Genome of the Netherlands Consortium. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat Genet*. 2014;46(8):818-825.
18. Grove ML, Yu B, Cochran BJ, et al. Best practices and joint calling of the HumanExome BeadChip: the CHARGE Consortium. *PLoS One*. 2013;8(7):e68095.
19. Aulchenko YS, Struchalin MV, van Duijn CM. ProbABEL package for genome-wide association analysis of imputed data. *BMC Bioinformatics*. 2010;11:134.
20. Pruim RJ, Welch RP, Sanna S, et al. LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics*. 2010;26(18):2336-2337.

21. Ward LD, Kellis M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res*. 2012;40(Database issue):D930-934.
22. Encode Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012;489(7414):57-74.
23. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics*. 2010;26(17):2190-2191.
24. Liu X, Jian X, Boerwinkle E. dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. *Hum Mutat*. 2011;32(8):894-899.
25. Peloso GM, Auer PL, Bis JC, 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*. 2014;94(2):223-232.
26. Wu MC, Lee S, Cai T, et al. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet*. 2011;89(1):82-93.
27. Mathieson I, McVean G. Differential confounding of rare and common variants in spatially structured populations. *Nat Genet*. 2012;44(3):243-246.
28. Yang J, Benyamin B, McEvoy BP, et al. Common SNPs explain a large proportion of the heritability for human height. *Nat Genet*. 2010;42(7):565-569.
29. Plaimauer B, Fuhrmann J, Mohr G, et al. Modulation of ADAMTS13 secretion and specific activity by a combination of common amino acid polymorphisms and a missense mutation. *Blood*. 2006;107(1):118-125.
30. Jenkins PV, O'Donnell JS. ABO blood group determines plasma von Willebrand factor levels: a biologic function after all? *Transfusion*. 2006;46(10):1836-1844.
31. Bowen DJ. An influence of ABO blood group on the rate of proteolysis of von Willebrand factor by ADAMTS13. *J Thromb Haemost*. 2003;1(1):33-40.
32. Yu J, Madison JM, Mundlos S, Winston F, Olsen BR. Characterization of a human homologue of the Saccharomyces cerevisiae transcription factor spt3 (SUPT3H). *Genomics*. 1998;53(1):90-96.
33. Martinez E, Kundu TK, Fu J, Roeder RG. A human SPT3-TAFII31-GCN5-L acetylase complex distinct from transcription factor IID. *J Biol Chem*. 1998;273(37):23781-23785.
34. Verdone L, Caserta M, Di Mauro E. Role of histone acetylation in the control of gene expression. *Biochem Cell Biol*. 2005;83(3):344-353.
35. Lee TI, Causton HC, Holstege FC, et al. Redundant roles for the TFIID and SAGA complexes in global transcription. *Nature*. 2000;405(6787):701-704.
36. Eisenmann DM, Arndt KM, Ricupero SL, Rooney JW, Winston F. SPT3 interacts with TFIID to allow normal transcription in Saccharomyces cerevisiae. *Genes Dev*. 1992;6(7):1319-1331.
37. Claus RA, Bockmeyer CL, Kentouche K, et al. Transcriptional regulation of ADAMTS13. *Thromb Haemost*. 2005;94(1):41-45.
38. Gill G, Pascal E, Tseng ZH, Tjian R. A glutamine-rich hydrophobic patch in transcription factor Sp1 contacts the dTAFII110 component of the Drosophila TFIID complex and mediates transcriptional activation. *Proc Natl Acad Sci U S A*. 1994;91(1):192-196.
39. De Cock E, Hermans C, De Raeymaecker J, et al. The novel ADAMTS13-p.D187H mutation impairs ADAMTS13 activity and secretion and contributes to thrombotic thrombocytopenic purpura in mice. *J Thromb Haemost*. 2015;13(2):283-292.
40. de Groot R, Lane DA, Crawley JT. The ADAMTS13 metalloprotease domain: roles of subsites in enzyme activity and specificity. *Blood*. 2010;116(16):3064-3072.

41. Camilleri RS, Cohen H, Mackie IJ, et al. Prevalence of the ADAMTS-13 missense mutation R1060W in late onset adult thrombotic thrombocytopenic purpura. *J Thromb Haemost*. 2008;6(2):331-338.
42. Rurali E, Noris M, Chianca A, et al. ADAMTS13 predicts renal and cardiovascular events in type 2 diabetic patients and response to therapy. *Diabetes*. 2013;62(10):3599-3609.
43. Miyata T, Kokame K, Matsumoto M, Fujimura Y. ADAMTS13 activity and genetic mutations in Japan. *Hamostaseologie*. 2013;33(2):131-137.
44. Kokame K, Kokubo Y, Miyata T. Polymorphisms and mutations of ADAMTS13 in the Japanese population and estimation of the number of patients with Upshaw-Schulman syndrome. *J Thromb Haemost*. 2011;9(8):1654-1656.
45. Tsai HM, Sussman, II, Ginsburg D, et al. Proteolytic cleavage of recombinant type 2A von Willebrand factor mutants R834W and R834Q: inhibition by doxycycline and by monoclonal antibody VP-1. *Blood*. 1997;89(6):1954-1962.
46. Crawley JT, Lam JK, Rance JB, et al. Proteolytic inactivation of ADAMTS13 by thrombin and plasmin. *Blood*. 2005;105(3):1085-1093.

Chapter 3

ADAMTS13: association with cardiovascular risk factors

Chapter 3.1

ADAMTS13 activity and decline in kidney function

Manuscript based on this chapter

Sanaz Sedaghat*, Paul S. de Vries*, Johan Boender, Michelle A.H. Sonneveld, Ewout J. Hoorn, Albert Hofman, Moniek P.M. de Maat, Oscar H. Franco, M. Arfan Ikram, Frank W.G. Leebeek, and Abbas Dehghan.
*Contributed equally to this manuscript.

Von Willebrand factor, ADAMTS13 activity and decline in kidney function: a cohort study.

Submitted.

ABSTRACT

Background: Altered levels of von Willebrand factor and ADAMTS13 can promote thrombosis and disturb blood flow in kidney microcirculations.

Methods: In this study, we investigated the association of serum von Willebrand factor antigen, ADAMTS13 activity, and the von Willebrand factor-to-ADAMTS13 ratio in relation to decline in kidney function. The annual decline in estimated GFR, doubling of creatinine, halving of estimated GFR, and new onset chronic kidney disease (estimated GFR < 60 ml/min/1.73m^2) were assessed during a median follow up of 11 years.

Results: Higher von Willebrand factor-to-ADAMTS13 ratio was associated with steeper annual decline in estimated GFR (0.05 ml/min; 95% confidence interval: 0.01, 0.09) and higher risk of new onset chronic kidney disease (odds ratio: 1.14; 95% confidence interval: 1.01, 1.29). Likewise, higher von Willebrand factor-to-ADAMTS13 ratio was associated with higher risk of doubling of creatinine (odds ratio: 2.16; 95% confidence interval: 1.24, 3.76) and halving of estimated GFR (odds ratio: 1.44; 95% confidence interval: 1.01, 2.04). All these associations were independent of age, sex, cardiovascular risk factors and blood group.

Conclusions: In this population-based study, we observed that higher von Willebrand factor-to-ADAMTS13 ratio is associated with decline in kidney function over time. This finding suggests a role of elevated prothrombotic factors in the development and progression of kidney disease.

INTRODUCTION

Von Willebrand factor (VWF) is a multimeric glycoprotein which mediates platelet adhesion and aggregation.[1] VWF function is partly regulated by the VWF protease, ADAMTS13.[1] ADAMTS13 cleaves ultra-large VWF multimers into smaller multimers that are less procoagulant.[1,2] Therefore, the imbalance between VWF and ADAMTS13 is an important indicator of a prothrombotic state.[3] The significance of deficiency in ADAMTS13 is most apparent in thrombotic thrombocytopenic purpura (TTP) patients. Due to severe ADAMTS13 deficiency, TTP patients have higher loads of ultralarge VWF multimers, resulting in microthrombi formation and subsequent circulation disturbances. Given the dependency of kidney function on the adequate blood flow to the glomerulus, the kidney is one of the most susceptible organs to thrombotic events in its microcirculation.[4] The imbalance between VWF and ADAMTS13 may promote thrombosis in kidney vessels, leading to disturbances in kidney circulation and thereby contributing to the decline in kidney function and to the development of chronic kidney disease (CKD).[5] In fact, renal insufficiency is one of the hallmark clinical characteristics of TTP patients.[6] While previous animal studies[4] and studies in patient groups[3,5] suggest a link between VWF and ADAMTS13 with kidney function, whether this link extends to individuals from general populations remains to be elucidated. We investigated the association of VWF-to-ADAMTS13 ratio, VWF, and ADAMTS13 activity with decline in kidney function in the population-based Rotterdam study.

METHODS

Study population

The present study is embedded within the framework of the population-based Rotterdam Study. The design of the Rotterdam study has been described previously.[7] In brief, the cohort started in 1990, consisting of 7,983 participants aged 55 years or older living in Ommoord, a district of Rotterdam in the Netherlands (RS-I). In 2000, the first extension of the Rotterdam Study (RS-II) started, adding 3,011 new participants. VWF:Ag and ADAMTS13 activity were evaluated at the third visit of RS-I (1997-1999) and the first visit of RS-II (2000-2001). Among individuals with both VWF:Ag and ADAMTS13 activity measurements, 2,479 participants had repeated measurements of creatinine for the evaluation of longitudinal kidney function. The median time elapsed between the two creatinine measurements was 11 years (range: 7.8-13.6).

The Rotterdam Study has been approved by the medical ethics committee according to the Population Study Act Rotterdam Study, executed by the Ministry of Health,

Welfare and Sports of the Netherlands. A written informed consent was obtained from all participants.

Measurement of VWF:Ag and ADAMTS13 activity

Fasting venous blood samples were taken at the research center and collected in citrated tubes. Samples were stored at -80°C. VWF:Ag was determined with an in-house ELISA with polyclonal rabbit antihuman VWF antibodies (DakoCytomation, Glostrop, Denmark) for catching and tagging.[8] The intra-assay coefficient of variation was 5.8% and the interassay coefficient of variation was 7.8%.[8] ADAMTS13 activity was measured using the Fluorescence Resonance Energy Transfer Substrate VWF 73 kinetic assay (FRETS-VWF73).[9] Samples of VWF and ADAMTS13 were measured against a reference curve of serial dilutions of normal human plasma, calibrated against the international standard (Siemens, Germany).[9]

Measurement of estimated glomerular filtration rate (eGFR)

Serum creatinine was determined using an enzymatic assay method. Creatinine values were standardized to isotope-dilution mass spectrometry–traceable (IDMS) measurements. In order to calibrate, we aligned the mean values of serum creatinine with serum creatinine values of the participants of the Third National Health and Nutrition Examination Survey (NHANES III) in different gender and age groups (<60, 60-69, ≥70).[10] eGFR was calculated according to the Chronic Kidney Disease Epidemiology Collaboration (CKD-EPI) formula.[11] To calculate the annual eGFR decline, we first subtracted the eGFR values of the follow-up examination from the eGFR values at baseline and then divided by the time, in years, between the two visits. New onset CKD cases were defined among the individuals with eGFR >60 ml/min/1.73 m^2 at baseline, who had a decline in eGFR to less than 60 ml/min/1.73 m^2 between the two periodical examinations.[11] Doubling of creatinine and halving of eGFR between the two periodical visits were also defined to assess the kidney function over time.[12]

Covariates

Body mass index was calculated by dividing weight in kilograms by height in meters squared. Information on smoking and alcohol consumption was acquired from questionnaires. Participants were asked for the average daily consumption of alcohol and data is presented as grams per day. Smoking was categorized in never, former and current smoking. Blood pressure was measured twice by an oscillometric device after five minutes of rest and the mean was taken as the subject's reading. Information on medication use was based on home interview. Serum total cholesterol and high-density lipoprotein cholesterol levels were determined using an automated enzymatic method. Coronary heart disease was considered as experiencing myocardial

infarction or coronary revascularization procedures. Diabetes mellitus was defined by the use of blood glucose lowering drugs and/or a fasting serum glucose level greater than or equal to 7.0 mmol/l at baseline or a non-fasting serum glucose level greater than or equal to 11.1 mmol/l. Blood group was defined based on rs687289 variant, which discriminates blood group O from non-O status.[13]

Statistical analysis

The association of VWF-to-ADAMTS13 ratio, VWF:Ag, and ADAMTS13 activity with annual decline in eGFR was evaluated using linear regression models. Logistic regressions were used to estimate the odds ratios for the association of VWF-to-ADAMTS13 ratio, VWF:Ag, and ADAMTS13 activity with new onset CKD, doubling of creatinine and halving of eGFR. Betas were estimated per SD increase for VWF:Ag, ADAMTS13 activity and VWF-to-ADAMTS13 ratio. Since measures of VWF-to-ADAMTS13 ratio and VWF:Ag were not normally distributed, they were natural-log transformed. We performed analyses using two models. In the first model analyses were adjusted for age, sex, cohort (Rotterdam Study 1 or Rotterdam Study 2), and baseline eGFR (only for longitudinal analyses). In the second model, we further adjusted the analyses for systolic and diastolic blood pressure, body mass index, alcohol consumption, smoking, high-density lipoprotein cholesterol, total cholesterol, history of diabetes mellitus and coronary heart disease, blood group (O or non-O), and antihypertensive and antithrombotic medications. All analyses with new onset CKD, doubling of creatinine and halving of eGFR as an outcome were adjusted for the follow-up time elapsed between the two measurements of creatinine. We divided participants into tertiles of VWF-to-ADAMTS13 ratio and compared participants from the second and third tertile with participants from the first tertile (reference category). To investigate whether the association of prothrombotic factors and decline in kidney function differs based on gender, age, and blood group, we assessed the interaction of the prothrombotic factors and the aforementioned characteristics by adding an interaction term in the model. The interaction term was the product of the interacting factor and prothrombotic factors. In addition, we performed a series of stratified analyses by separately studying the association of prothrombotic factors and decline in kidney function in participants with blood group O and non-O, in men and women and in participants younger and older than 65 years. Evaluating linearity assumption, there was neither departure from linearity for the linear regression models and nor for logistic regressions, using fractional polynomials. We performed multiple imputation for missing data in the covariates (< 8% for all covariates), using a Markov Chain Monte Carlo method. The calibration of GFR measurements and the evaluation of linearity assumptions were done using R version 2.15.0. All other analyses were carried out using SPSS 20.0.2 for windows.

Table 1. Baseline characteristics of study participants

Characteristics	n= 2479
Age, years	65.1 (5.8)
Men	1056 (42.6)
Systolic blood pressure, mmHg	139.8 (19.9)
Diastolic blood pressure, mmHg	77.1 (10.5)
Body mass index, kg/m^2	26.9 (3.7)
Alcohol, g/day	5.7 (0.7-20.0)
Current smoker	433 (17.5)
Total cholesterol, mmol/l	5.8 (0.9)
HDL cholesterol, mmol/l	1.4 (0.3)
Blood group O	1185 (43.8)
History of diabetes mellitus	210 (8.5)
History of coronary heart disease	145 (5.8)
Antithrombotic agents	311 (12.5)
Antihypertensive medication	645 (26.0)
Estimated glomerular filtration rate (creatinine), mL/min/1.73 m^2	78.5 (13.1)
Von Willebrand factor antigen, %	112 (88-146)
ADAMTS13 activity, %	94.3 (16.8)
vWF-to-ADAMTS13 ratio	1.2 (0.9-1.6)

Categorical variables are presented as numbers (percentages), continuous variables as means (standard deviations) and von Willebrand factor antigen, vWF-to-ADAMTS13 ratio and alcohol intake are presented as medians (interquartile ranges).

RESULTS

The characteristics of 2479 study participants are presented in **Table 1**. The average age of the participants was 65±6 years and 43% were male. The mean eGFR based on creatinine measurements was 78.5±13 mL/min/1.73 m^2. Participants had average VWF antigen (VWF:Ag) level of 112% and average ADAMTS13 activity of 94.3 %. The correlation between VWF:Ag and ADAMTS13 activity was minimal (r= -0.08, p <0.01).

The median time elapsed between the two eGFR estimates was 11 years (range: 7.8-13.6). The association of VWF-to-ADAMTS13 ratio, VWF:Ag, and ADAMTS13 activity with annual decline in kidney function is presented in **Table 2**. Higher VWF-to-ADAMTS13 ratio, in model I was associated with steeper annual decline in eGFR (0.06 mL/min/year; 95%CI: 0.02, 0.09) and a higher risk of developing CKD (1.13; 95%CI: 1.01, 1.27). Similarly, higher VWF-to-ADAMTS13 ratio, in model I was associated with higher risk of doubling of creatinine (1.90; 95%CI: 1.15, 3.13) and halving of eGFR (1.40; 95%CI: 1.02, 1.93). Adjustment for potential confounders did not change the associations. Each SD higher VWF:Ag, in model I, was associated with 0.05 mL/

Table 2. Association of von Willebrand factor antigen, ADAMTS13 activity, and VWF- to-ADAMTS13 ratio with decline in kidney function.

	Annual eGFR decline N= 2479			New onset CKD N (case) =2272 (500)			Doubling of creatinine N (case) =2479 (18)			Halving of eGFR N (case) =2479 (43)		
	Beta	95% CI	*P*-value	OR	95% CI	*P*-value	OR	95% CI	*P*-value	OR	95% CI	*P*-value
VWF-to-ADAMTS13 ratio												
Model I	0.06	0.02, 0.09	<0.01	1.13	1.01, 1.27	0.03	1.90	1.15, 3.13	0.01	1.40	1.02, 1.93	0.04
Model II	0.05	0.01, 0.09	<0.01	1.14	1.01, 1.29	0.04	2.16	1.24, 3.76	<0.01	1.44	1.01, 2.04	0.04
VWF:Ag												
Model I	0.05	0.01, 0.08	0.01	1.14	1.01,1.28	0.02	1.54	0.95, 2.51	0.08	1.32	0.96, 1.81	0.09
Model II	0.03	-0.01, 0.07	0.07	1.12	0.98, 1.27	0.08	1.62	0.94, 2.80	0.09	1.30	0.91, 1.84	0.15
ADAMTS13 activity												
Model I	-0.04	-0.07, 0.00	0.08	0.98	0.87, 1.11	0.78	0.60	0.36, 1.02	0.06	0.82	0.58, 1.16	0.26
Model II	-0.05	-0.09, -0.01	0.01	0.92	0.81, 1.04	0.19	0.49	0.28, 0.84	<0.01	0.74	0.52, 1.05	0.09

Betas/odds ratios and 95% CI are calculated per standard deviation measures of log transformed VWF:Ag, ADAMTS13 activity and log transformed VWF-to-ADAMS13 ratio.
Model I: Adjusted for age, sex, cohort, and baseline eGFR.
Model II: Additionally adjusted for systolic blood pressure, diastolic blood pressure, antihypertensive medication, antithrombotic agents, alcohol intake, smoking, total cholesterol, high density lipoprotein cholesterol, lipid-lowering medication, diabetes mellitus, history of coronary heart disease, and body mass index, blood group (O and non-O), and follow up time (for analyses with new onset CKD, doubling of creatinine, and halving of eGFR).
Abbreviations: CI: confidence interval, eGFR: creatinine based estimated glomerular filtration rate, VWF:Ag: von Willebrand factor antigen, CKD: chronic kidney disease, OR: odds ratio.

min/year (95%: 0.01, 0.08) unit steeper annual decline in eGFR and 14% (95%CI: 1.01, 1.28) higher risk of new onset CKD. The associations were not present after adjustments for potential confounders. There was no association between VWF:Ag and risk of doubling of creatinine or halving of eGFR (all p>0.05).

Each SD lower ADAMTS13 activity was associated with 0.05 ml/min unit steeper annual decline in eGFR (95% CI: 0.01, 0.09), after adjusting for potential confounders in model II. There was no association between ADAMTS13 and risk of new onset CKD, or halving of eGFR.

Analyses of the tertiles of the VWF-to-ADAMTS13 ratio and measures of decline in kidney function are presented in **Figure 1**. Participants in the third tertile of the VWF-to-ADAMTS13 ratio compared to participants in the first tertile had steeper decline in eGFR and higher risk of developing new onset CKD and doubling of creatinine.

In the stratified analyses, there was no statistically significant difference in the strength of the association of VWF-to-ADAMTS13 ratio, VWF:Ag, and ADAMTS13 activity with decline in kidney function in subgroups of participants based on their blood group, gender, and age (**Supplemental Figure 1**).

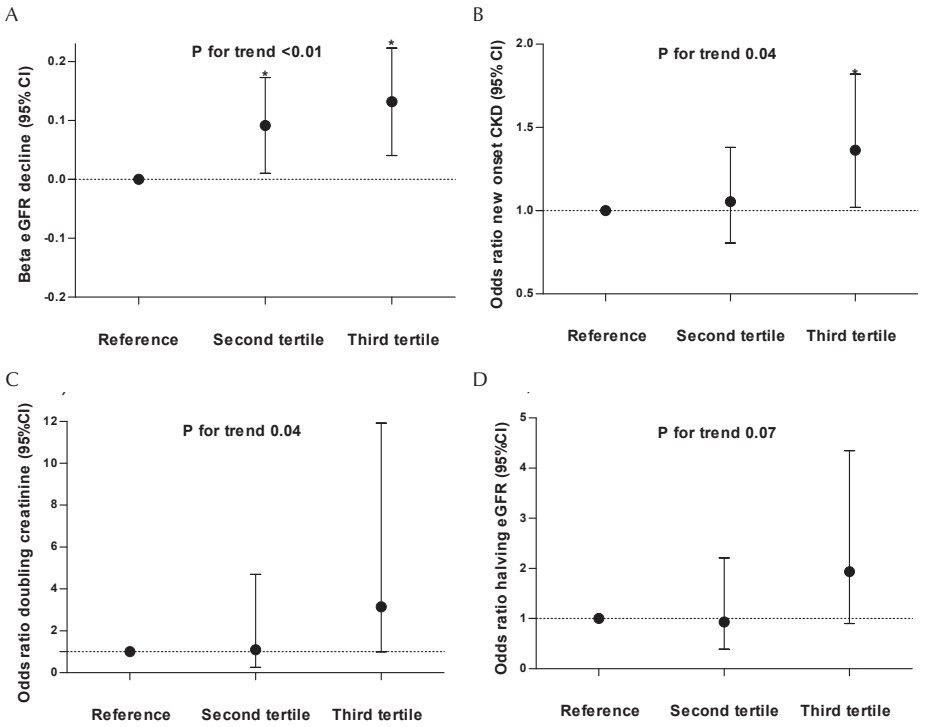


Figure 1. Association of VWF- to-ADAMTS13 ratio tertiles with A) annual decline in eGFR, B) new onset CKD, C) halving of eGFR and D) doubling of creatinine.
VWF-to-ADAMTS13 ratio tertiles (reference: < 0.1, second: 0.01-0.02, third: ≥0.02).
*represents a *P*-value<0.05 when a tertile was compared to the reference category (first tertile).

DISCUSSION

In this population-based study, we found that higher VWF-to-ADAMTS13 ratio, higher VWF:Ag, and lower ADAMTS13 activity are associated with steeper decline in kidney function independent of potential confounders.

A limited number of studies investigated a potential role for VWF and ADAMTS13 in relation to kidney function.[14-17] Previous cross-sectional studies reported higher levels of VWF and lower ADAMTS13 activity in patients with chronic kidney disease and end stage renal disease.[14,16,17] Apart from the cross-sectional observations, few studies reported an association between higher levels of VWF:Ag and progression of CKD.[18-22] Regarding the role of ADAMTS13, the link between its activity and CKD development has been investigated only in small groups of patients.[3,5,23,24] Ono et al., found that lower ADAMTS13 activity was associated with higher serum creatinine levels and future risk of kidney injury.[5] This study was performed in patients with sepsis and severe deficiency in ADAMTS13 activity. In the current large population-

based study we observed a clear association between VWF-to-ADAMTS13 ratio, VWF:Ag, and ADAMTS13 activity and decline in kidney function. Of note, although the effect estimates indicate a slight increase in kidney disease risk, previous studies showed that even trivial declines in eGFR are associated with considerable risk of future end stage renal disease.[25]

The plasma concentration of VWF and ADAMTS13 has been shown to be influenced by cardiovascular risk factors and differ based on certain characteristics.[3,8,26-29] For example, individuals with type O blood group have 25 percent lower VWF than those with non O blood group.[26] It is reported that VWF level and ADAMTS13 activity differs between men and women,[28] and in different age ranges.[30] It is also well-known that cardiovascular risk factors can influence the kidney function.[31] Therefore, the association of VWF:Ag, ADAMTS13 activity and their ratio with decline in kidney function may be confounded or mediated by these factors. In this study, adjustments for cardiovascular risk factors, medications and blood group did not change our findings. In addition, we did not observe any differences in the association of prothrombotic factors and decline in kidney function in different subgroups of participants, indicating that the associations of VWF-to-ADAMTS13 ratio, VWF:Ag, and ADAMTS13 activity with decline in kidney function are independent of cardiovascular risk factors and blood group.

VWF is known as an endothelial function marker.[1] Patients with CKD are more prone to endothelial damage and hence higher levels of VWF.[25] Therefore, it could be speculated that the steeper kidney function decline is a reflection of existing endothelial dysfunction at baseline. However, the prospective nature of our findings, adjustment of longitudinal analyses for baseline eGFR, as well as excluding participants with baseline eGFR less than 60 mL/min/1.73 m^2 rule out this conjecture.

Further evidence to support the etiologic role of ADAMTS13 on progression of kidney function can be provided by genetic variants in the *ADAMTS13* gene. Severe deficiency in ADAMTS13 caused by auto-antibodies or defects in the *ADAMTS13* gene is the cause of TTP and, in fact, acute kidney injury occurs in over 50% of TTP patients.[6,23,32] Furthermore, a Pro618Ala polymorphism in *ADAMTS13* is shown to be predictive of renal events in normoalbuminuric type 2 diabetic patients.[24] In addition, in a porcine model of *Escherichia coli* sepsis, decreased ADAMTS13 activity and increased large VWF multimers, was reported along with glomerular microthrombi enriched with platelets and VWF, and acute kidney injury.[4] Taken together, this suggests a potential causal role for VWF, ADAMTS13 and particularly the imbalance between them in relation to decline in kidney function.

We observed a stronger association between VWF-to-ADAMTS13 ratio and decline in kidney function compared to levels of VWF:Ag or ADAMTS13 activity, separately. It is known that ultra-large VWF multimers are more procoagulant; however, measur-

ing ultra-large VWF is technically difficult and laborious.[3] In line with our observation, several studies have indicated that the imbalance between VWF concentration and ADAMTS13 activity, rather than levels of VWF:Ag or ADAMTS13 activity, may allow a better evaluation of the prothrombotic state.[3,33,34]

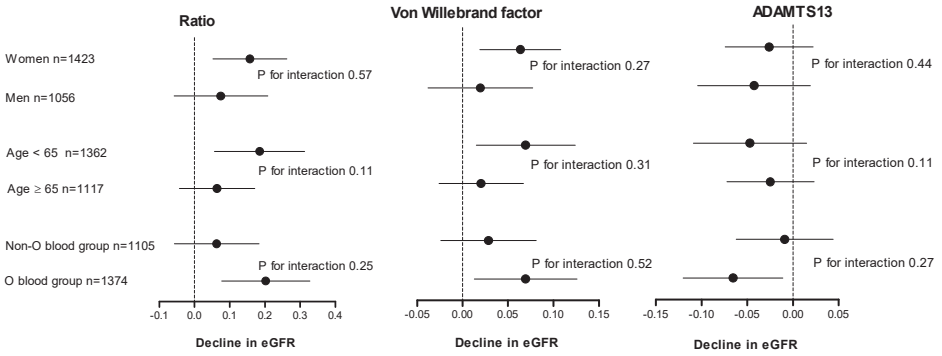
The population-based design of this study, the large sample size, prospective setting, and the availability of extensive data on various socio-demographic and cardiovascular risk factors that enabled us to control for several potential confounders, can be marked as the main strengths of this study. Limitations of this study should also be acknowledged. No data on albuminuria were available, which is an important element in defining CKD. In addition, although the definition of CKD based on KDIGO criteria requires two values of eGFR less than 60 ml/min/1.73m^2 at least 90 days apart, we only had a single measurement of eGFR. However, CKD definition based on eGFR < 60 ml/min /1.73 m^2 has been used previously in the population-based research setting.[35]

In conclusion, we observed that VWF-to-ADAMTS13 ratio, VWF:Ag, and ADAMTS13 activity are independently associated with decline in kidney function in the general population setting. Future studies are needed to explore whether the prediction of kidney function decline could be improved by monitoring VWF, ADAMTS13 and more specifically the imbalance between them.

REFERENCES

1. Sonneveld MA, de Maat MP, Leebeek FW. Von Willebrand factor and ADAMTS13 in arterial thrombosis: a systematic review and meta-analysis. Blood Rev. 2014;28:167-178.
2. Ruggeri ZM. The role of von Willebrand factor in thrombus formation. Thromb Res. 2007;120 Suppl 1:S5-9.
3. Fukushima H, Nishio K, Asai H, et al. Ratio of von Willebrand factor propeptide to ADAMTS13 is associated with severity of sepsis. Shock. 2013;39:409-414.
4. Bockmeyer CL, Reuken PA, Simon TP, et al. ADAMTS13 activity is decreased in a septic porcine model. Significance for glomerular thrombus deposition. Thromb Haemost. 2011;105:145-153.
5. Ono T, Mimuro J, Madoiwa S, et al. Severe secondary deficiency of von Willebrand factor-cleaving protease (ADAMTS13) in patients with sepsis-induced disseminated intravascular coagulation: its correlation with development of renal failure. Blood. 2006;107:528-534.
6. Zafrani L, Mariotte E, Darmon M, et al. Acute renal failure is prevalent in patients with thrombotic thrombocytopenic purpura associated with low plasma ADAMTS13 activity. J Thromb Haemost. 2015;13:380-389.
7. Hofman A, Brusselle GG, Darwish Murad S, et al. The Rotterdam Study: 2016 objectives and design update. Eur J Epidemiol. 2015;30:661-708.
8. Wieberdink RG, van Schie MC, Koudstaal PJ, et al. High von Willebrand factor levels increase the risk of stroke: the Rotterdam study. Stroke. 2010;41:2151-2156.
9. Kokame K, Nobe Y, Kokubo Y, Okayama A, Miyata T. FRETS-VWF73, a first fluorogenic substrate for ADAMTS13 assay. Br J Haematol. 2005;129:93-100.
10. Coresh J, Astor BC, McQuillan G, et al. Calibration and random variation of the serum creatinine assay as critical elements of using equations to estimate glomerular filtration rate. Am J Kidney Dis. 2002;39:920-929.
11. Inker LA, Schmid CH, Tighiouart H, et al. Estimating glomerular filtration rate from serum creatinine and cystatin C. N Engl J Med. 2012;367:20-29.
12. Sedaghat S, Cremers LG, de Groot M, et al. Kidney function and microstructural integrity of brain white matter. Neurology. 2015;85:154-161.
13. Sonneveld MA, van Dijk AC, van den Herik EG, et al. Relationship of Von Willebrand Factor with carotid artery and aortic arch calcification in ischemic stroke patients. Atherosclerosis. 2013;230:210-215.
14. Shen L, Lu G, Dong N, Jiang L, Ma Z, Ruan C. Von Willebrand factor, ADAMTS13 activity, TNF-alpha and their relationships in patients with chronic kidney disease. Exp Ther Med. 2012;3:530-534.
15. Warrell RP, Jr., Hultin MB, Coller BS. Increased factor VIII/von Willebrand factor antigen and von Willebrand factor activity in renal failure. Am J Med. 1979;66:226-228.
16. Rios DR, Carvalho MG, Figueiredo RC, et al. ADAMTS13 and Von Willebrand factor in patients undergoing hemodialysis. J Thromb Thrombolysis. 2012;34:73-78.
17. Chen J, Hamm LL, Mohler ER, et al. Interrelationship of Multiple Endothelial Dysfunction Biomarkers with Chronic Kidney Disease. PLoS One. 2015;10:e0132047.
18. Clausen P, Feldt-Rasmussen B, Jensen G, Jensen JS. Endothelial haemostatic factors are associated with progression of urinary albumin excretion in clinically healthy subjects: a 4-year prospective study. Clin Sci (Lond). 1999;97:37-43.

19. Stehouwer CD, Gall MA, Twisk JW, Knudsen E, Emeis JJ, Parving HH. Increased urinary albumin excretion, endothelial dysfunction, and chronic low-grade inflammation in type 2 diabetes: progressive, interrelated, and independently associated with risk of death. Diabetes. 2002;51:1157-1165.
20. Stuveling EM, Bakker SJ, Hillege HL, de Jong PE, Gans RO, de Zeeuw D. Biochemical risk markers: a novel area for better prediction of renal risk? Nephrol Dial Transplant. 2005;20:497-508.
21. Bash LD, Erlinger TP, Coresh J, Marsh-Manzi J, Folsom AR, Astor BC. Inflammation, hemostasis, and the risk of kidney function decline in the Atherosclerosis Risk in Communities (ARIC) Study. Am J Kidney Dis. 2009;53:596-605.
22. Tin A, Grams ME, Maruthur NM, et al. Hemostatic Factors, APOL1 Risk Variants, and the Risk of ESRD in the Atherosclerosis Risk in Communities Study. Clin J Am Soc Nephrol. 2015;10:784-790.
23. Bramham K, Hilton R, Horsfield C, McDonald V, Camilleri R, Hunt BJ. ADAMTS-13 deficiency: can it cause chronic renal failure? Nephrol Dial Transplant. 2011;26:742-744.
24. Rurali E, Noris M, Chianca A, et al. ADAMTS13 predicts renal and cardiovascular events in type 2 diabetic patients and response to therapy. Diabetes. 2013;62:3599-3609.
25. Coresh J, Turin TC, Matsushita K, et al. Decline in estimated glomerular filtration rate and subsequent risk of end-stage renal disease and mortality. JAMA. 2014;311:2518-2531.
26. Franchini M, Capra F, Targher G, Montagnana M, Lippi G. Relationship between ABO blood group and von Willebrand factor levels: from biology to clinical implications. Thromb J. 2007;5:14.
27. Skeppholm M, Kallner A, Kalani M, Jorneskog G, Blomback M, Wallen HN. ADAMTS13 and von Willebrand factor concentrations in patients with diabetes mellitus. Blood Coagul Fibrinolysis. 2009;20:619-626.
28. Terrell DR, Vesely SK, Kremer Hovinga JA, Lammle B, George JN. Different disparities of gender and race among the thrombotic thrombocytopenic purpura and hemolytic-uremic syndromes. Am J Hematol. 2010;85:844-847.
29. de Vries PS, Boender J, Sonneveld MA, et al. Genetic variants in the ADAMTS13 and SUPT3H genes are associated with ADAMTS13 activity. Blood. 2015;125:3949-3955.
30. Kokame K, Sakata T, Kokubo Y, Miyata T. von Willebrand factor-to-ADAMTS13 ratio increases with age in a Japanese population. J Thromb Haemost. 2011;9:1426-1428.
31. Fassett RG, Venuthurupalli SK, Gobe GC, Coombes JS, Cooper MA, Hoy WE. Biomarkers in chronic kidney disease: a review. Kidney Int. 2011;80:806-821.
32. Sarode R, Gottschall JL, Aster RH, McFarland JG. Thrombotic thrombocytopenic purpura: early and late responders. Am J Hematol. 1997;54:102-107.
33. Matsukawa M, Kaikita K, Soejima K, et al. Serial changes in von Willebrand factor-cleaving protease (ADAMTS13) and prognosis after acute myocardial infarction. Am J Cardiol. 2007;100:758-763.
34. Claus RA, Bockmeyer CL, Budde U, et al. Variations in the ratio between von Willebrand factor and its cleaving protease during systemic inflammation and association with severity and prognosis of organ failure. Thromb Haemost. 2009;101:239-247.
35. Bash LD, Coresh J, Kottgen A, et al. Defining incident chronic kidney disease in the research setting: The ARIC Study. Am J Epidemiol. 2009;170:414-424.



Supplemental Figure 1. Association of von Willebrand factor antigen, ADAMTS13 activity, and VWF-to-ADAMTS13 ratio with decline in eGFR, stratified based on blood group, gender, and age.

Chapter 3.2

ADAMTS13 activity and incident type 2 diabetes

Manuscript based on this chapter

Paul S. de Vries*, Thijs T.W. van Herpt*, Symen Ligthart, Albert Hofman, M. Arfan Ikram, Mandy van Hoek, Eric J.G. Sijbrands, Oscar H. Franco, Moniek P.M. de Maat, Frank W.G. Leebeek, and Abbas Dehghan.

*Contributed equally to this manuscript

ADAMTS13 activity as a novel risk factor for incident diabetes: a population-based cohort study.

Submitted.

ABSTRACT

Background: ADAMTS13 is a protease that breaks down von Willebrand factor (VWF) multimers into smaller, less active particles. Because of VWF's previously reported association with an increased risk of incident type 2 diabetes, we aimed to examine the association of ADAMTS13 activity and VWF antigen with incident diabetes.

Methods: The study included 5,176 participants of the Rotterdam Study, a prospective population-based cohort study. All participants were free of diabetes at baseline. The median follow up time was 11.2 years. Cox proportional hazard models were used to examine the association of ADAMTS13 activity and VWF antigen with incident diabetes.

Results: ADAMTS13 activity was associated with an increased risk of incident diabetes (HR: 1.17; 95%CI: 1.08 to 1.27) after adjustment for known risk factors and VWF antigen. Although ADAMTS13 activity was positively associated with fasting glucose and insulin, the association with incident diabetes did not change when we adjusted for these covariates. VWF antigen was associated with incident diabetes, but this association was attenuated when adjusted for known risk factors. ADAMTS13 activity was also associated with incident prediabetes after adjustment for known risk factors (HR: 1.11; 95%CI: 1.03, 1.20), while VWF antigen was not.

Conclusions: ADAMTS13 activity is thus an independent risk factor for incident type 2 diabetes and this association is unlikely to be the consequence of reverse causation. As the association between ADAMTS13 and diabetes did not appear to be explained by its cleavage of VWF, ADAMTS13 may have an independent role in the development of diabetes.

INTRODUCTION

ADAMTS13 is a protease that reduces the activity of von Willebrand factor (VWF) in platelet adhesion and aggregation by cleaving prothrombotic VWF multimers into smaller particles.[1,2] This is ADAMTS13's only known function. Low ADAMTS13 levels and activity are associated with an increased risk of various thrombotic diseases, including ischemic stroke and myocardial infarction.[3-8] Additionally, low ADAMTS13 activity may contribute to renal and cardiovascular complications of diabetes,[9-11] but the association of ADAMTS13 with diabetes itself remains unexplored. Elevated levels of VWF have been associated with an increased risk of type 2 diabetes.[12-16] This association has been attributed primarily to VWFs role as a marker of endothelial dysfunction rather than its role in thrombosis.[17]

VWF may also be associated with diabetes through its prothrombotic effect. This would be in line with emerging evidence that vascular disease may contribute to the development of diabetes.[18] Low ADAMTS13 activity and high VWF levels may exacerbate small vessel disease, which in turn may contribute to the development of diabetes.[19-21] If VWF is associated with diabetes through its prothrombotic function, then we expect ADAMTS13, with its antithrombotic function, to be inversely associated with the risk of diabetes. On the other hand, still little is known about the regulation of ADAMTS13 and its role as a marker of other physiological processes.[22]

Further investigation of the association of ADAMTS13 and VWF with diabetes may therefore clarify the role of both factors in the development of diabetes. In this study, we thus aimed to examine the association between ADAMTS13 activity and VWF antigen with incident diabetes in a large prospective population-based cohort study.

METHODS

Study description and population

The Rotterdam Study is a prospective population-based cohort study initiated in 1990 to study the determinants of several chronic diseases in older adults.[23] The first cohort (RS-I) includes 7,983 inhabitants of Ommoord, a district of Rotterdam in the Netherlands, who were 55 years or older. The first examination took place between 1990 and 1993. The third visit, including 4,797 participants, took place between March 1997 and December 1999, and was used as the baseline in this study. The second cohort (RS-II), established between February 2000 and December 2001, includes another 3,011 inhabitants of Ommoord who either reached the age of 55 years after the recruitment phase of RS-I or who had migrated into the research area. Thus, there is no overlap in participants across the two cohorts. There were no eli-

gibility criteria to enter the Rotterdam Study except age and residential area (postal code). The Rotterdam Study has been approved by the Medical Ethics Committee of the Erasmus MC and by the Ministry of Health, Welfare and Sport of the Netherlands, implementing the Wet Bevolkingsonderzoek: ERGO (Population Studies Act: Rotterdam Study). All included participants provided written informed consent to participate in the study and to obtain information from their treating physicians.

Ascertainment of prediabetes and diabetes

Diabetes and prediabetes at baseline and during follow-up was ascertained using records kept by general practitioners, hospital discharge letters, and glucose measurements from Rotterdam Study visits, which take place approximately every 4 years.[24] Diabetes, prediabetes and normoglycemia were defined according to the most recent World Health Organization guidelines.[25] Prediabetes was defined as a fasting blood glucose between 6.0 mmol/L and 7.0 mmol/L or a non-fasting blood glucose between 7.7 mmol/L and 11.1 mmol/L (when fasting samples were absent); diabetes was defined as a fasting blood glucose higher than 7.0 mmol/L, a non-fasting blood glucose ^{3} 11.1 mmol/L (when fasting samples were absent), or the use of blood glucose lowering medication. Information regarding the use of blood glucose lowering medication was derived from both home interviews and pharmacy records.[24] At baseline, more than 99% of the Rotterdam Study population was covered by the pharmacies in the study area. All potential events of prediabetes and diabetes were independently adjudicated by two study physicians, and in the case of disagreement consensus was sought with the help of an endocrinologist. We used follow-up data until January 1st 2012.

ADAMTS13 activity and VWF antigen measurements

Citrated plasma samples were collected at the third visit of RS-I and the baseline examination of RS-II, and stored at -80°C. Between June and October 2013, we measured ADAMTS13 activity using a kinetic assay based on the Fluorescence Resonance Energy Transfer Substrate VWF 73 (FRETS-VWF73) assay.[26] Plasma samples were measured against a reference curve of serial dilutions of normal human plasma defined to have an ADAMTS13 activity of 1 IU/ml, and we express ADAMTS13 activity as a percentage of this. The ADAMTS13 activity of 6,258 participants was measured: 3,791 from RS-I, and 2,467 from RS-II.

Between July and October of 2008, VWF antigen was determined in IU/ml with an in-house ELISA with polyclonal rabbit antihuman VWF antibodies (DakoCytomation, Glostrop, Denmark) for catching and tagging.[27] The intra-assay coefficient of variation was 5.8% and the inter-assay coefficient of variation was 7.8%. VWF antigen was measured in 3,968 individuals from RS-I, and 2,561 individuals from RS-II.

In total, 5,176 participants with VWF and ADAMTS13 measurements also had a fasting glucose measurement and were free of diabetes at baseline.

Covariates

Body mass index (BMI) was calculated by dividing weight in kilograms by height in meters squared. Information on current smoking was acquired from questionnaires. Lipid-lowering (statins, fibrates, and other lipid modifying agents), antihypertensive (diuretics, beta-blocking agents, ACE-inhibitors, calcium channel blockers), and antithrombotic medication use was assessed during a structured interview. Blood pressure was measured twice by an oscillometric device after five minutes of rest and the mean was taken as the subject's reading. Serum total cholesterol and high- density lipoprotein (HDL) cholesterol levels were determined using an automated enzymatic method. Blood glucose and insulin levels were quantified using standard laboratory techniques. Serum alanine-aminotransferase (ALAT) levels were measured using a Merck Diagnostica kit on an Elan Autoanalyzer (Merck, Whitehouse Station, NJ, USA). White blood cell count was assessed in citrate plasma with a Coulter Counter T540 (Coulter Electronics, Hialeah, Florida, USA). C-reactive protein (CRP) was measured using CRPL3, an immunoturbidometric assay (Roche Diagnostics, Indianapolis, IN, USA). Prevalent coronary heart disease (CHD) was defined as having a history of myocardial infarction or coronary revascularization procedures, as previously described.[24]

Statistical analysis

Statistical analyses were performed in SPSS version 21 (IBM Corp, Armonk, NY, USA) and R version 3.1.3 (R Foundation for Statistical Computing, Vienna, Austria). Missing values forcovariates were imputed in SPSS using single imputation based on expectation maximization. Each of the covariates had missing values for less than 5% of the participants. VWF antigen, HDL cholesterol, CRP, ALAT, and fasting insulin were natural-log transformed. We used linear regression models to test the association of ADAMTS13 activity and VWF antigen with baseline fasting glucose and fasting insulin. Individuals with prevalent diabetes were excluded in all analyses.

The association of ADAMTS13 activity and VWF antigen with incident diabetes was examined using Cox proportional hazards models. The assumption of proportional hazards was met. Three adjustment models were used. Model 1 was adjusted for age, sex, and cohort. Model 2 was additionally adjusted for HDL and total cholesterol, lipid-lowering medication, BMI, CRP, current smoking, antithrombotic medication, ALAT, white blood cell count, systolic blood pressure, antihypertensive medication, and prevalent CHD. Model 3 was additionally adjusted for fasting glucose and insulin. In Model 1 ADAMTS13 activity and VWF antigen were tested separately, whereas in

Table 1. Baseline characteristics of the study population.

	Mean (SD) or Percentage N = 5,176
Age (years)	69.0 (8.1)
Sex (female)	57.7
Body mass index (kg/m^2)	26.7 (3.8)
High-density lipoprotein cholesterol (mmol/L)	1.4 (0.4)
Total cholesterol (mmol/L)	5.9 (1.0)
Lipid-lowering medication use	11.4
Systolic blood pressure (mmHg)	142.1 (21.0)
Antihypertensive medication use	20.8
Alanine aminotransferase (U/L)	22.6 (13.0)
Current smoking	12.5
C-reactive protein (mg/L)	3.1 (5.6)
White blood cell count (10^9 cells/L)	6.7 (1.9)
Prevalent coronary heart disease	7.3
Prevalent prediabetes	18.2
Fasting glucose (mmol/L)	5.5 (0.5)
Fasting insulin (pmol/L)	74.3 (42.3)
Antithrombotic medication use	17.4
ADAMTS13 activity (%)	91.0 (17.2)
VWF antigen (IU/ml)	1.3 (0.6)

Both ADAMTS13 activity and VWF antigen were positively associated with baseline fasting insulin, and ADAMTS13 activity was positively associated with baseline fasting glucose (**Supplemental Table 1)**. Nevertheless, when adjusting for fasting glucose and insulin in Model 3, the effect sizes did not change. These associations were robust to the exclusion of participants with prevalent CHD and baseline, and the exclusion of users of lipid-lowering, antihypertensive, and antithrombotic medication **(Supplemental Table 2**).

There was a significant interaction between ADAMTS13 activity and VWF antigen with incident diabetes (*P*-value: 0.01). As shown in Figure 1, the association of ADAMTS13 activity with incident diabetes was strongest in the fourth quartile of VWF antigen (HR: 1.49: 95%CI: 1.27 to 1.75).

Furthermore, ADAMTS13 activity was also associated with an 11% (HR: 1.11; 95%CI: 1.03 to 1.19) increased risk of prediabetes per SD in Model 1, and this association was similar in Model 2 and 3 (Figure 2). In contrast, VWF antigen was not associated with incident prediabetes.

Models 2 and 3 the analysis of ADAMTS13 activity was adjusted for VWF antigen and vice versa. We examined the interaction between ADAMTS13 activity and VWF antigen on incident diabetes using a multiplicative interaction term, and adjusting for age, sex, and cohort. Results are shown per SD of VWF antigen and ADAMTS13 activity.

To test whether associations with incident diabetes were driven by participants with prevalent CHD, or users of lipid-lowering, antihypertensive, and antithrombotic medication, we excluded participants in each of these subgroups in a sensitivity analysis.

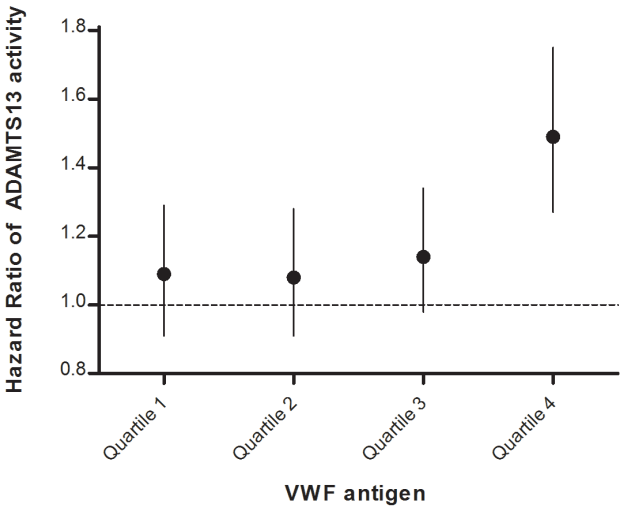


Figure 1. Hazard ratios of ADAMTS13 activity (per SD) for incident diabetes across quartiles of VWF antigen: interaction between ADAMTS13 and VWF.

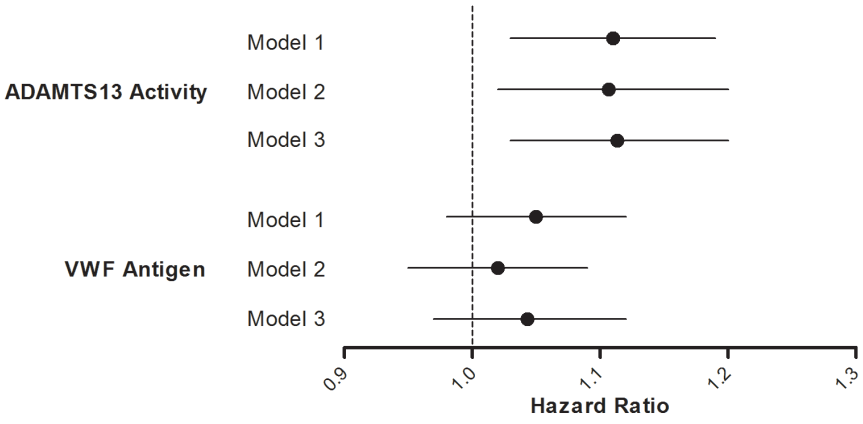


Figure 2. Hazard ratios of ADAMTS13 activity and log transformed VWF antigen (per SD) for incident prediabetes excluding participants with prediabetes at baseline (862 events in 4,232 participants).

Finally, to explore the association of ADAMTS13 activity and VWF antigen with the early stages of dysglycemia, we examined incident prediabetes using the same models as for incident diabetes, but additionally excluding participants with prevalent prediabetes.

RESULTS

Baseline characteristics are shown in **Table 1**. Among the 5,176 participants without prevalent diabetes at baseline, the mean (SD) age was 69.0 (8.1), and 57.7% were women. In a median follow-up time of 11.2 years (IQR: 9.8, 12.6), 638 participants out of 5,176 participants developed diabetes.

Associations of ADAMTS13 activity and VWF antigen with incident diabetes are shown in **Table 2**. ADAMTS13 activity was associated with a 19% increased risk of incident diabetes per SD in Model 1 (Hazard ratio [HR]: 1.19; 95% confidence intervals [95%CI]: 1.10 to 1.30), and this association remained unchanged in Model 2. As

Table 2. Hazard ratios of ADAMTS13 activity and log transformed VWF antigen (per SD) on incident diabetes (638 events in 5176 participants).

	ADAMTS13 activity		VWF antigen	
	Hazard Ratio (95%CI)	***P*-value**	**Hazard Ratio (95%CI)**	***P*-value**
Model 1	1.19 (1.10, 1.30)	0.00003	1.12 (1.03, 1.21)	0.008
Model 2	1.17 (1.08, 1.27)	0.0001	1.06 (1.00, 1.15)	0.2
Model 3	1.17 (1.08, 1.27)	0.0001	1.07 (0.99, 1.17)	0.1

Adjustments: Model 1: Adjusted for age, sex, and cohort. Model 2: Additionally adjusted for HDL and total cholesterol, lipid-lowering medication, body-mass index, CRP, current smoking, antithrombotic medication, ALAT, white blood cell count, systolic blood pressure, antihypertensive medication, and prevalent CHD. The analysis of VWF antigen was adjusted for ADAMTS13 activity and vice versa. Model 3: Additionally adjusted for glucose and insulin levels. HDL cholesterol, CRP, ALAT, and insulin were natural-log transformed when used.

Table 3. Hazard ratios of ADAMTS13 activity quartiles on incident diabetes.

	Model 1		Model 2		Model 3	
	Hazard Ratio (95%CI)	***P*-value**	**Hazard Ratio (95%CI)**	***P*-value**	**Hazard Ratio (95%CI)**	***P*-value**
Quartile 1	Reference		Reference		Reference	
Quartile 2	1.12 (0.89, 1.42)	0.3	1.10 (0.87, 1.40)	0.4	1.12 (0.88, 1.42)	0.4
Quartile 3	1.26 (1.00, 1.60)	0.05	1.31 (1.04, 1.65)	0.02	1.36 (1.08, 1.72)	0.01
Quartile 4	1.47 (1.16, 1.86)	0.001	1.46 (1.15, 1.85)	0.002	1.48 (1.17, 1.87)	0.001

Adjustments: Model 1: Adjusted for age, sex, and cohort. Model 2: Additionally adjusted for VWF antigen, HDL and total cholesterol, lipid-lowering medication, body-mass index, CRP, current smoking, antithrombotic medication, ALAT, white blood cell count, systolic blood pressure, antihypertensive medication, and prevalent CHD. Model 3: Additionally adjusted for glucose and insulin levels. HDL cholesterol, CRP, ALAT, and insulin were natural-log transformed when used.

shown in **Table 3**, participants in the highest quartile of ADAMTS13 activity had a 46% increased risk compared to participants in the lowest quartile (HR: 1.46; 95%CI: 1.15, 1.85). VWF antigen was associated with a 12% (HR: 1.12; 95%CI: 1.03 to 1.21) increased risk of incident diabetes per SD in Model 1. However, the association was attenuated to 6% (HR: 1.06; 95%CI: 1.00 to 1.15) increased risk per SD after adjustment for additional covariates in Model 2.

DISCUSSION

In our study, ADAMTS13 activity was associated with an increased risk of incident diabetes, even after adjustment for other known risk factors, including VWF antigen, fasting glucose, and fasting insulin. Furthermore, ADAMTS13 activity was also associated with the incidence of prediabetes among participants with normoglycemia at baseline. VWF antigen was also associated with an increased risk of diabetes, but this association was attenuated after adjustment for known risk factors.

To our knowledge, the association of ADAMTS13 with diabetes has not previously been studied with diabetes as the primary outcome, and we are the first to examine this association in a large prospective population-based cohort study. One cross-sectional study reported the association between ADAMTS13 and prevalent diabetes.[11] The researchers did not observe a statistically significant difference in ADAMTS13 levels between 86 cases of diabetes and 26 healthy controls.[11] Our results for VWF are consistent with previous studies. VWF has been associated with incident diabetes in a range of studies,[12-16] but in general the association weakened after adjustment for confounder and became non-significant. In the Framingham Heart Study, however, VWF remained significantly associated after adjustment for a wide range of potential confounders, including insulin resistance.[13] VWF is a marker of endothelial dysfunction, and this is thought to explain the association between VWF and diabetes.[17] We report an interaction between ADAMTS13 activity and VWF, with the largest effect of ADAMTS13 activity among participants in the highest quartile of VWF. This interaction suggests that the effect of ADAMTS13 is mainly present in individuals with advanced endothelial dysfunction.

The mechanism underlying the association of ADAMTS13 activity with diabetes remains unclear. Because the association was robust to the adjustment for baseline fasting glucose and insulin, and because ADAMTS13 activity was also associated with incident prediabetes, the possibility of reverse causation is limited. However, the association between ADAMTS13 activity and diabetes is unlikely to be explained by its only known function as a cleaving protease of VWF, because in that case we would expect VWF (prothrombotic) and ADAMTS13 activity (antithrombotic) to be associ-

ated with diabetes in opposite directions. An alternative hypothesis is an additional functionality of ADAMTS13 beyond VWF cleavage. ADAMTS13 is part of the ADAMTS family of enzymes, which are metalloendopeptidades with a diversity of functions in vascular biology.[28] Finally, the association could be explained by pathways that respond to ADAMTS13. For example, there is preliminary evidence that ADAMTS13 regulates the expression and phosphorylation of vascular endothelial growth factor, which is known to contribute to microvascular complications of diabetes.[29,30] However, ADAMTS13 was only discovered in 2001, and since then most research has focused on its interactions with VWF and its role in TTP.[1,2] Therefore, we believe that further research is required to elucidate such pathways.

We measured ADAMTS13 activity using the FRETS assay, which is based on an introduced peptide spanning the VWF cleavage site.[26] ADAMTS13 antigen is an alternative measurement, which corresponds to the abundance of ADAMTS13. Future studies should investigate whether ADAMTS13 activity or antigen is most strongly associated to diabetes. If the association with diabetes is strongest with ADAMTS13 antigen, then the association of markers of ADAMTS13 gene expression, synthesis, secretion, and degradation with diabetes should be explored. Alternatively, a stronger association with ADAMTS13 activity points towards a downstream implication of VWF cleavage, albeit not the decreased activity of VWF itself.

The strengths of our study include the comprehensive assessment of incident diabetes and prediabetes, using medical records, linkage with pharmacies in the study area, and standardized blood glucose measurements at each of the follow up visits. Additionally, we used data from a well-characterized prospective population-based cohort study, which allowed us to correct for a wide range of covariates. We used a long follow up period, and adjusted for baseline fasting glucose and insulin to reduce the possibility of reverse causation. By also examining associations with incident prediabetes, we provide insight into the early development of subclinical disease.

The main limitation of our study is that, as in all epidemiological studies, we cannot rule out residual confounding. Furthermore, our results were found in individuals of European ancestry, and may not be generalizable to other populations. In addition, we included individuals aged 55 years and older and effect estimates might not be generalizable to younger ages.

In conclusion, we identified ADAMTS13 activity as a novel independent marker of incident diabetes, associated with both diabetes and prediabetes. Future research is necessary to confirm this association and to elucidate the biology underlying this association. Exploration of alternative mechanisms of ADAMTS13 beyond VWF cleavage is warranted as the association may not be explained by its antithrombotic function.

REFERENCES

1. Fujikawa K, Suzuki H, McMullen B, Chung D. Purification of human von Willebrand factor-cleaving protease and its identification as a new member of the metalloproteinase family. Blood. 2001;98:1662-1666.
2. Gerritsen HE, Robles R, Lammle B, Furlan M. Partial amino acid sequence of purified von Willebrand factor-cleaving protease. Blood. 2001;98:1654-1661.
3. Andersson HM, Siegerink B, Luken BM, et al. High VWF, low ADAMTS13, and oral contraceptives increase the risk of ischemic stroke and myocardial infarction in young women. Blood. 2012;119:1555-1560.
4. Bongers TN, de Bruijne EL, Dippel DW, et al. Lower levels of ADAMTS13 are associated with cardiovascular disease in young patients. Atherosclerosis. 2009;207:250-254.
5. Chion CK, Doggen CJ, Crawley JT, Lane DA, Rosendaal FR. ADAMTS13 and von Willebrand factor and the risk of myocardial infarction in men. Blood. 2007;109:1998-2000.
6. Crawley JT, Lane DA, Woodward M, Rumley A, Lowe GD. Evidence that high von Willebrand factor and low ADAMTS-13 levels independently increase the risk of a non-fatal heart attack. J Thromb Haemost. 2008;6:583-588.
7. Sonneveld MA, de Maat MP, Leebeek FW. Von Willebrand factor and ADAMTS13 in arterial thrombosis: a systematic review and meta-analysis. Blood Rev. 2014;28:167-178.
8. Sonneveld MAH, de Maat MPM, Portegies MLP, et al. Low ADAMTS13 activity is a strong risk factor for ischemic stroke: a prospective cohort study - the Rotterdam Study. ASH Annual Meeting and Exposition. San Francisco, CA, USA; 2014.
9. Rossing P, Lajer M. Can ADAMTS13 lead us to the paradise of personalized medicine? Diabetes. 2013;62:3331-3332.
10. Rurali E, Noris M, Chianca A, et al. ADAMTS13 predicts renal and cardiovascular events in type 2 diabetic patients and response to therapy. Diabetes. 2013;62:3599-3609.
11. Taniguchi S, Hashiguchi T, Ono T, et al. Association between reduced ADAMTS13 and diabetic nephropathy. Thromb Res. 2010;125:e310-316.
12. Duncan BB, Schmidt MI, Offenbacher S, Wu KK, Savage PJ, Heiss G. Factor VIII and other hemostasis variables are related to incident diabetes in adults. The Atherosclerosis Risk in Communities (ARIC) Study. Diabetes Care. 1999;22:767-772.
13. Meigs JB, O'Donnell C J, Tofler GH, et al. Hemostatic markers of endothelial dysfunction and risk of incident type 2 diabetes: the Framingham Offspring Study. Diabetes. 2006;55:530-537.
14. Muris DM, Houben AJ, Schram MT, Stehouwer CD. Microvascular dysfunction is associated with a higher incidence of type 2 diabetes mellitus: a systematic review and meta-analysis. Arterioscler Thromb Vasc Biol. 2012;32:3082-3094.
15. Thorand B, Baumert J, Chambless L, et al. Elevated markers of endothelial dysfunction predict type 2 diabetes mellitus in middle-aged men and women from the general population. Arterioscler Thromb Vasc Biol. 2006;26:398-405.
16. Wannamethee SG, Sattar N, Rumley A, Whincup PH, Lennon L, Lowe GD. Tissue plasminogen activator, von Willebrand factor, and risk of type 2 diabetes in older men. Diabetes Care. 2008;31:995-1000.
17. Mannucci PM. von Willebrand factor: a marker of endothelial damage? Arterioscler Thromb Vasc Biol. 1998;18:1359-1362.

18. Izzo R, de Simone G, Trimarco V, et al. Hypertensive target organ damage predicts incident diabetes mellitus. Eur Heart J. 2013;34:3419-3426.
19. Jaap AJ, Shore AC, Tooke JE. Relationship of insulin resistance to microvascular dysfunction in subjects with fasting hyperglycaemia. Diabetologia. 1997;40:238-243.
20. Tal MG. Type 2 diabetes: Microvascular ischemia of pancreatic islets? Med Hypotheses. 2009;73:357-358.
21. Tooke JE. Microvascular function in human diabetes. A physiological perspective. Diabetes. 1995;44:721-726.
22. de Vries PS, Boender J, Sonneveld MA, et al. Genetic variants in the ADAMTS13 and SUPT3H genes are associated with ADAMTS13 activity. Blood. 2015.
23. Hofman A, Darwish Murad S, van Duijn CM, et al. The Rotterdam Study: 2014 objectives and design update. Eur J Epidemiol. 2013;28:889-926.
24. Leening MJ, Kavousi M, Heeringa J, et al. Methods of data collection and definitions of cardiac outcomes in the Rotterdam Study. Eur J Epidemiol. 2012;27:173-185.
25. World Health Organization. Definition and diagnosis of diabetes mellitus and intermediate hyperglycemia: report of a WHO/IDF consultation. Geneva: World Health Organization; 2006:1:50.
26. Kokame K, Nobe Y, Kokubo Y, Okayama A, Miyata T. FRETS-VWF73, a first fluorogenic substrate for ADAMTS13 assay. Br J Haematol. 2005;129:93-100.
27. Wieberdink RG, van Schie MC, Koudstaal PJ, et al. High von Willebrand factor levels increase the risk of stroke: the Rotterdam study. Stroke. 2010;41:2151-2156.
28. Kelwick R, Desanlis I, Wheeler GN, Edwards DR. The ADAMTS (A Disintegrin and Metalloproteinase with Thrombospondin motifs) family. Genome Biol. 2015;16:113.
29. Lee M, Keener J, Xiao J, Long Zheng X, Rodgers GM. ADAMTS13 and its variants promote angiogenesis via upregulation of VEGF and VEGFR2. Cell Mol Life Sci. 2015;72:349-356.
30. Wirostko B, Wong TY, Simo R. Vascular endothelial growth factor and diabetic complications. Prog Retin Eye Res. 2008;27:608-621.

Supplemental Table 1. Cross-sectional association of ADAMTS13 activity and VWF antigen (per SD) with fasting glucose and natural-log transformed fasting insulin.

	ADAMTS13 activity		VWF antigen	
	β coefficient (95%CI)	***P*-value**	**β coefficient (95%CI)**	***P*-value**
Glucose				
Model 1	0.03 (0.01, 0.04)	0.001	0.01 (-0.00, 0.03)	0.08
Model 2	0.02 (0.01, 0.04)	0.003	-0.01 (-0.02, 0.01)	0.4
Insulin				
Model 1	0.06 (0.04, 0.07)	6×10^{-15}	0.07 (0.05, 0.08)	1×10^{-19}
Model 2	0.05 (0.04, 0.06)	3×10^{-15}	0.03 (0.02, 0.05)	2×10^{-8}

β coefficient refers to the 1 unit increase in fasting glucose (mmol/L) or insulin (natural-log transformed pmol/L) per 1 standard deviation increase in VWF antigen or ADAMTS13 activity.
Adjustments: Model 1: Adjusted for age, sex, and cohort. Model 2: Additionally adjusted for HDL and total cholesterol, lipid-lowering medication, body-mass index, CRP, current smoking, antithrombotic medication, ALAT, white blood cell count, systolic blood pressure, antihypertensive medication, and prevalent CHD. The analysis of VWF antigen was adjusted for ADAMTS13 activity and vice versa. HDL cholesterol, CRP, and ALAT were natural-log transformed.

Supplemental Table 2. Association of ADAMTS13 activity and VWF antigen (per SD) with incident diabetes after exclusions based on disease and medication use at baseline*.

	ADAMTS13 Activity		VWF antigen	
	Hazard Ratio (95%CI)	***P*-value**	**Hazard Ratio (95%CI)**	***P*-value**
Excluding cases of prevalent CHD: 565 events in 4,674 participants				
Model 1	1.20 (1.10, 1.30)	0.00007	1.14 (1.05, 1.25)	0.002
Model 2	1.17 (1.07, 1.27)	0.0004	1.09 (1.00, 1.19)	0.06
Model 3	1.18 (1.08, 1.28)	0.0002	1.09 (1.00, 1.20)	0.05
Excluding antithrombotic medication users: 490 events in 4,062 participants				
Model 1	1.20 (1.09, 1.32)	0.0001	1.14 (1.04, 1.25)	0.006
Model 2	1.15 (1.05, 1.27)	0.002	1.07 (0.98, 1.18)	0.1
Model 3	1.16 (1.06, 1.27)	0.001	1.09 (0.99, 1.20)	0.08
Excluding lipid-lowering medication users: 526 events in 4,372 participants				
Model 1	1.18 (1.08, 1.29)	0.0005	1.12 (1.03, 1.23)	0.01
Model 2	1.15 (1.05, 1.26)	0.002	1.05 (0.96, 1.15)	0.3
Model 3	1.15 (1.05, 1.25)	0.002	1.07 (0.98, 1.17)	0.2
Excluding antihypertensive medication users: 415 events in 3,837 participants				
Model 1	1.19 (1.07, 1.32)	0.001	1.12 (1.01, 1.24)	0.03
Model 2	1.20 (1.08, 1.33)	0.0005	1.06 (0.96, 1.18)	0.2
Model 3	1.20 (1.08, 1.33)	0.0005	1.11 (1.00, 1.23)	0.05

*Exclusions were based on non-imputed variables.
Adjustments: Model 1: Adjusted for age, sex, and cohort. Model 2: Additionally adjusted for HDL and total cholesterol, lipid-lowering medication, body-mass index, CRP, current smoking, antithrombotic medication, ALAT, white blood cell count, systolic blood pressure, antihypertensive medication, and prevalent CHD. The analysis of VWF antigen was adjusted for ADAMTS13 activity and vice versa. Model 3: Additionally adjusted for glucose and insulin levels. HDL cholesterol, CRP, ALAT, and insulin were natural-log transformed when used.

Chapter 4

Genetic risk of coronary heart disease

Chapter 4.1

Genetic risk prediction of coronary heart disease

Manuscript based on this chapter

Paul S. de Vries, Maryam Kavousi, Symen Ligthart, André G. Uitterlinden, Albert Hofman, Oscar H. Franco, and Abbas Dehghan.

Incremental predictive value of 152 single nucleotide polymorphisms in the 10-year risk prediction of incident coronary heart disease: the Rotterdam Study.

International Journal of Epidemiology. 2015; 44(2): 682-8.

ABSTRACT

Background: The aim of this study was to examine the incremental predictive value of genetic risk scores of coronary heart disease (CHD) in the 10-year risk prediction of incident CHD.

Methods: In 5,899 subjects, we used 152 single nucleotide polymorphisms (SNPs) associated with coronary artery disease by the CARDIoGRAMplusC4D consortium to construct three weighted genetic risk scores: 1) GRS_{gws} based on 49 genome-wide significant SNPs, 2) GRS_{fdr} based on 103 suggestively associated SNPs, and 3) GRS_{all} based on all 152 SNPs. We examined the changes in discrimination and reclassification of incident CHD when adding the genetic risk scores to models including traditional risk factors. We repeated the analysis for prevalent CHD.

Results: The genetic risk scores were associated with incident CHD despite adjustment for traditional risk factors and family history: participants had a 13% higher rate of CHD per standard deviation increase in GRS_{all}. GRS_{all} improved the C-statistic by 0.006 (CI95%: 0.000, 0.013) beyond age and sex, 0.003 (CI95%: -0.001, 0.008) beyond traditional risk factors and 0.003 (CI95%: -0.001, 0.007 beyond traditional risk factors and family history. The genetic risk scores did not improve reclassification. GRS_{all} strongly improved both discrimination and reclassification of prevalent CHD, even beyond traditional risk factors and family history, with a C-statistic improvement of 0.009 (0.003, 0.015).

Conclusions: Although the genetic risk scores based on 152 SNPs were associated with incident CHD, they did not improve risk prediction. This discrepancy may be the result of SNP discovery for prevalent rather than incident CHD, since the SNPs do improve prediction for prevalent disease.

INTRODUCTION

Primary and secondary prevention programs are widely performed using risk prediction models based on traditional risk factors to identify individuals at high risk for coronary heart disease (CHD). Optimizing these risk prediction models could therefore directly translate into improved prevention and management of CHD-related morbidity and mortality. As CHD has a strong heritable component,[1,2] adding genetic markers to prediction models could improve risk prediction. This assumption has been tested in studies using genetic risk scores based on single nucleotide polymorphisms (SNPs).[3-13] Overall, the studies show that prediction is not meaningfully improved by currently validated CHD SNPs.[3-13] Nevertheless, the set of CHD SNPs is growing through the efforts of international consortia, and a recent genome-wide association study (GWAS) by the CARDIoGRAMplusC4D consortium raised the number of independent CHD SNPs from 31 to 153.[14] Collectively these SNPs explain around 10% of the genetic variance,[14] which suggests that we are now in a better position to implement SNPs in risk prediction of CHD.

These SNPs, however, were identified using case-control and cross-sectional designs. In these study designs SNPs associated with a favorable prognosis after CHD events may be overrepresented in cases. As a consequence, the association of these SNPs may not fully translate to incident CHD, leading to markers that are spuriously associated with CHD.

We hypothesized that adding genetic risk scores based on CHD SNPs would improve 10 year CHD risk prediction when added to traditional risk factors. To evaluate our hypothesis we constructed three genetic risk scores based on CHD SNPs found by the CARDIoGRAMplusC4D consortium. We then examined whether risk prediction improved when we added the genetic risk scores to three models including: 1) age and sex, 2) age, sex and traditional risk factors, 3) age, sex, traditional risk factors and family history. To examine differences between incident and prevalent CHD, we repeated the analysis for prevalent CHD.

METHODS

Study population

This study was conducted within the Rotterdam Study, an ongoing prospective population-based cohort study of inhabitants of Ommoord, a district of Rotterdam in the Netherlands. The Rotterdam Study has been described in detail elsewhere.[15,16] In the year 1990, inhabitants of Ommoord who were 55 years old or over were invited to participate. Baseline examination lasted from 1990 to 1993 and included

7,983 participants, of whom 7,758 gave their informed consent for follow-up data collection. Follow-up examinations were carried out every 3 to 5 years. The study was approved by the Medical Ethics Committee of Erasmus University, Rotterdam, the Netherlands, and all included participants gave their written informed consent.

Genotyping and imputation

Genotyping was successfully conducted in 5,899 participants who agreed to be followed-up using the Illumina 550K. Imputation was done with reference to HapMap release 22 CEU using the maximum likelihood method implemented in MACH.[17-19] The imputation quality of the SNPs is presented in **Supplemental Table 1**.

Genetic risk scores

To construct genetic risk scores we used 153 uncorrelated SNPs associated with CHD by the CARDIoGRAMplusC4D consortium, of which 49 attained genome-wide significance and the remaining 104 had a false discovery rate of less (FDR) than 10% in an FDR analysis.[14] Out of the 153 SNPs, 152 were either genotyped or imputed in the Rotterdam Study. We calculated weighted dosages by multiplying the risk allele (the allele previously reported to increase the risk of CHD) dosage of each SNP with its previously reported effect size (lnOR)[14]. GRS_{gws} was constructed using the 49 genome-wide significant SNPs, GRS_{fdr} using the 103 additional SNPs that were found in the FDR analysis, and GRS_{all} using all 152 SNPs. Genetic risk scores were computed using the PredictABEL package in R version 2.15.1 (R Foundation for Statistical Computing, Vienna, Austria).[20]

Coronary heart disease

CHD events included myocardial infarctions, all CHD mortality, and revascularization. Cardiovascular outcome definitions as well as data collection methods are presented in detail elsewhere.[21] In brief, participants with general practitioners in the district of Ommoord were continuously monitored for fatal and nonfatal cardiovascular events through automated linkage with files from general practitioners and hospitals. Participants with general practitioners outside of Ommoord were monitored through annual checks of their medical records. All reported events were independently reviewed and coded by two research physicians. Codes on which the research physicians disagreed were discussed to reach consensus, and a medical expert in cardiovascular disease subsequently reviewed all events.

Traditional risk factors and family history

Serum total and high density lipoprotein (HDL) cholesterol concentrations were determined at baseline within 2 weeks after sampling by an automated enzymatic

procedure in non-fasting blood samples (Kone Specific Analyzer, Kone Instruments). Blood pressure was measured while seated using a random-zero sphygmomanometer at the right brachial artery. The average of two consecutive measurements was used. Diabetes was defined as fasting plasma glucose levels ≥ 7 mmol/L or non-fasting plasma glucose ≥ 11.1 mmol/L, or use of medications indicated for the treatment of diabetes. Current smoking status (yes/no), family history of myocardial infarction in first degree relatives (yes/no), lipid-lowering medication use (yes/no), and antihypertensive medication use (yes/no) were assessed during a structured interview at baseline by trained research assistants.

Statistical analyses

Statistical analyses were done using SPSS version 20 (IBM Corp., Armonk, NY) and R version 2.15.1. Missing values for all covariates were imputed using expectation maximization in SPSS. Participants with prevalent CHD at baseline were excluded, and hazard rates were computed using cox proportional hazards models. Three adjustment models were used. Model 1 was adjusted only for age and sex. Model 2 was further adjusted for total and HDL cholesterol, systolic blood pressure, prevalent type 2 diabetes, antihypertensive medication, lipid-lowering medication, and current smoking. Model 3 was additionally adjusted for family history of myocardial infarction. In addition to standard *P*-values, we computed Bonferroni corrected *P*-values for the associations of the genetic risk scores with incident CHD using the p.adjust function in R. We applied a correction for 9 statistical tests (the 3 genetic risk scores were each tested in 3 models). All models met the assumption of proportional hazards, which was tested for each model using the "cox.zph" function in R. Absolute 10-year risk was estimated as explained by Wilson *et al*.[22] These predicted risks were used to classify participants into low (< 5%), intermediate-low (5-10%), intermediate-high (10-20%), and high (> 20%) risk categories. Changes in C-statistic were used to assess improvements in discrimination, and the categorical net reclassification improvement (NRI) was used to assess improvements in reclassification.[23] NRIs were calculated using the prospective form applicable to survival data as introduced by Pencina *et al*.[24] We used 10,000 bootstrap resamples to generate 95% confidence intervals for changes in C-statistic and prospective NRI. We performed several additional analyses. First, improvements in prediction were also calculated in the subgroup of 2082 participants who were under 65 years old at baseline. Secondly, we examined the association of the genetic risk scores with myocardial infarction and estimated the corresponding improvements in prediction. Furthermore, we used cox proportional hazard models to examine the association between family history and incident CHD using different adjustments: age and sex adjusted, further

adjusted for traditional risk factors, and further adjusted for each of the genetic risk scores.

The genetic risk scores used SNPs that were identified for prevalent rather than incident CHD. To examine whether this affects their predictive value, we repeated the analysis separately for prevalent cases. For prevalent CHD, odds ratios were computed using logistic regression, and both the predicted risks and NRIs were calculated using PredictABEL.[20] Nagelkerke's R^2 was used to estimate the variance in incident and prevalent CHD explained by different combinations of risk factors.[25]

RESULTS

Out of 5,899 participants, 485 participants had prevalent CHD at baseline. During a mean follow up period of 12.8 years, 964 CHD events (460 myocardial infarctions) occurred among the remaining 5,414 individuals. Of these events, 571 (270 myocardial infarctions) occurred within 10 years. Baseline characteristics of the study population are shown in **Table 1**, and baseline characteristics by CHD status are shown in **Supplemental Table 2**).

All three genetic risk scores were associated with incident CHD. The associations were attenuated when adjusting for traditional risk factors, and further attenuated when additionally adjusted for family history. These associations are shown in **Table 2**. The association between family history and incident CHD largely remained stable when the genetic risk scores were added to the model (**Supplemental Table 3**).

Improvements in discrimination and reclassification of incident CHD are shown in **Table 3**. The largest improvement in risk prediction was achieved by GRS_{all} beyond age and sex ($\Delta C = 0.006$, 95%CI: 0.000, 0.013); however, it did not improve reclassification. Furthermore, improvements in discrimination or reclassification beyond traditional risk factors or traditional risk factors + family history were very modest. In participants under the age of 65 the genetic risk scores lead to greater improvements in prediction than in the entire sample, although these were accompanied by larger confidence intervals (**Supplemental Table 4**). The associations and improvements in prediction were considerably weaker for incident MI than for prevalent CHD (**Supplemental Tables 5 and 6**).

All three genetic risk scores were associated with prevalent CHD (**Supplemental Table 7**), and these associations were stronger than the associations with incident CHD. Improvements in the prediction of prevalent CHD were almost always markedly higher than improvements in prediction of incident CHD events (**Supplemental Table 8**). All three genetic risk scores improved discrimination beyond the three models. GRS_{all} improved discrimination the most (ΔC 0.009

beyond traditional risk factors and family history, 95%CI: 0.003, 0.015). GRS_{all} also improved reclassification beyond the three models, while GRS_{gws} only improved reclassification beyond age + sex and traditional risk factors. GRS_{fdr} did not improve reclassification beyond any of the models.

Table 1. Baseline characteristics of the 5,899 participants included in this study.

	Mean (SD) or percentage
Age (years)	69.3 (9.0)
Sex (% males)	40.9
Total cholesterol (mmol/L)	6.6 (1.2)
HDL cholesterol (mmol/L)	1.34 (0.4)
Lipid lowering medication use	2.5
Antihypertensive medication use	13.3
Systolic blood pressure (mmHg)	139.2 (22.3)
Diastolic blood pressure (mmHg)	73.7 (11.5)
Prevalent Type 2 Diabetes	10.6
Current smoking	23.1

Abbreviations: **BMI:** Body mass index; **HDL:** High-density lipoprotein

The percentage of variance in incident and prevalent CHD explained by the genetic risk scores, risk factors, and their combinations are shown in **Supplemental Table 9**. Genetic risk scores consistently explained a larger proportion of the variance of prevalent CHD than of incident CHD: GRS_{all} explained 1.5% of the variance of prevalent CHD, but only 0.7% of the variance of incident CHD. In both cases, only 0.1% of the variance was also explained by family history. GRS_{all} explained a larger proportion of the variance of both incident and prevalent CHD than family history, age, total cholesterol, systolic blood pressure, smoking, and lipid lowering medication use.

DISCUSSION

In this study we showed that genetic risk scores based on up to 152 SNPs so far identified for prevalent CHD are associated with incident CHD, though they do not lead to clinically relevant improvements in 10-year risk prediction of CHD.

SNPs could be used in CHD risk prediction in two different settings. The first is to use genetic data in adults and elderly subjects to improve risk prediction beyond current CHD risk prediction models. Our results show that currently available SNPs are not sufficient for this application. A second use of SNPs is to estimate the future risk of CHD earlier in life. This could be in the form of lifetime risk, or in the form of 10 year risk at different ages. In this setting SNPs are already useful if they improve

Table 2. Hazard ratios (95% confidence intervals) per SD change of genetic risk scores for incident CHD.

	Model 1	*P*-value	Bonferroni Corrected *P*-value*	Model 2	*P*-value	Bonferroni Corrected *P*-value*	Model 3	*P*-value	Bonferroni Corrected *P*-value*
GRS_{gws}	1.13 (1.06, 1.20)	0.00014	0.0013	1.12 (1.05, 1.19)	0.00054	0.0049	1.11 (1.05, 1.19)	0.00076	0.0068
GRS_{fdr}	1.09 (1.03, 1.17)	0.0051	0.046	1.08 (1.01, 1.15)	0.02	0.18	1.07 (1.01, 1.14)	0.032	0.29
GRS_{all}	1.15 (1.08, 1.23)	1.1×10^{-5}	9.9×10^{-5}	1.13 (1.06, 1.21)	0.00012	0.0011	1.13 (1.06, 1.20)	0.00022	0.0020

*Bonferroni-corrected *P*-values are corrected for 9 statistical tests.

Abbreviations: **CHD:** Coronary heart disease; **GRS_{gws}**: Genetic risk score including only CHD SNPs significant according to genome-wide significance; **GRS_{fdr}**: Genetic risk score including only CHD SNPs significant according to false discovery rate analysis; **GRS_{all}**: Genetic risk score including all significant CHD SNPs.

Adjustments: **Model 1:** age and sex adjusted; **Model 2**: Further adjusted for total and HDL cholesterol, systolic blood pressure, prevalent type 2 diabetes, antihypertensive medication, lipid-lowering medication, and current smoking; **Model 3:** Further adjusted for family history of myocardial infarction.

Table 3. Improvements in discrimination and reclassification of incident CHD when adding genetic risk scores to 10 year risk prediction models.

	C	ΔC	NRI
Model 1	0.684		
GRS_{gws}		0.004 (-0.001, 0.009)	0.023 (-0.021, 0.067)
GRS_{fdr}		0.004 (-0.001, 0.008)	0.003 (-0.04, 0.046)
GRS_{all}		0.006 (0.000, 0.013)	0.034 (-0.014, 0.081)
Model 2	0.716		
GRS_{gws}		0.002 (-0.001, 0.006)	0.014 (-0.019, 0.047)
GRS_{fdr}		0.002 (-0.001, 0.005)	0.01 (-0.024, 0.044)
GRS_{all}		0.003 (-0.001, 0.008)	0.022 (-0.018, 0.061)
Model 3	0.716		
GRS_{gws}		0.002 (-0.001, 0.006)	0.016 (-0.019, 0.051)
GRS_{fdr}		0.002 (-0.001, 0.004)	0.007 (-0.026, 0.04)
GRS_{all}		0.003 (-0.001, 0.007)	0.017 (-0.025, 0.058)

Abbreviations: **CHD:** Coronary heart disease; **C:** C-statistic before adding genetic risk scores to the model; **Δ*C:*** Improvement in C-statistic when adding the genetic risk score to base models; **NRI:** Net reclassification improvement when adding the genetic risk score to base models; **GRS_{gws}**: Genetic risk score including only CHD SNPs significant according to genome-wide significance; **GRS_{fdr}**: Genetic risk score including only CHD SNPs significant according to false discovery rate analysis; **GRS_{all}**: Genetic risk score including all significant CHD SNPs.

Adjustments: **Model 1:** includes age and sex; **Model 2:** Further includes total and HDL cholesterol, systolic blood pressure, prevalent type 2 diabetes, antihypertensive medication, lipid-lowering medication, and current smoking; **Model 3:** Further includes for family history of myocardial infarction.

prediction over age and sex. Our study suggests that current GWAS findings may be more useful for this setting.

Several studies have shown that genetic risk scores based on SNPs for prevalent CHD are associated with incident CHD though improvements in prediction are generally very small.[3-7] Ganna et al have previously tested a genetic risk score similar to GRS_{gws},[7] and they found slightly larger improvements in discrimination and reclassification. In contrast to our study, they recalculated the weight of each included SNP in an independent prospective cohort. This step may partly explain the differences between our studies. Another study suggested that SNPs might be especially useful in specific subgroups such as middle aged men.[4] Our study was not sufficiently powered to examine predictive improvements in this subgroup, but we did find greater improvements in prediction when we limited our analysis to participants under 65 years old.

Our genetic risk scores were based on GWA studies. Given that collecting the large number of cases needed for adequate statistical power is easier in a case-control setting with prevalent cases, a large proportion of studies included in these GWA studies are composed of case-control studies. Such a design, though statistically more powerful, may lead to the identification of SNPs that are related to improved survival after events rather than SNPs that increase the risk of event. This is known as Neyman's bias or incidence-prevalence bias.[26] If so, the identified SNPs for CHD, and hence the genetic risk score herewith evaluated, might represent a mixture of SNPs associated with CHD risk and SNPs associated with an improved survival after a CHD event. Indeed, we found a striking rise in the incremental value of the genetic risk scores when we used prevalent CHD as the outcome instead of incident CHD. Furthermore, a previous study of prevalent CHD also found a large C-statistic improvement beyond traditional risk factors (0.008) in contrast to the small improvements found by studies of incident CHD.[8] This difference suggests that the inability of SNPs to contribute to risk prediction is in part explained by the cross-sectional discovery panel. This is also supported by our findings as percentage of variance explained. For instance the variance explained by GRS_{all} in prevalent CHD was twice as large as in incident CHD. This bias may hamper the ability of genetic risk scores to improve prediction of first CHD events in populations free of CHD.[27] We present only preliminary evidence that this is influencing risk prediction: prevalent events occurred earlier in life than incident events, and this may partly explain the observed differences in risk prediction. Individuals experiencing CHD events at a younger age may be genetically enriched for CHD SNPs. In line with this, the percentage of individuals with a family history of myocardial infarction is slightly higher in prevalent cases than in incident cases.

A potential solution may be to recalculate the weight of each included SNP in an independent prospective cohort as done by Ganna et al.[7] Nevertheless, this approach still assumes that important SNPs for prevalent CHD are also important for incident CHD, and did not lead to substantially higher indices of discrimination and reclassification. Instead, it may be necessary to conduct a GWAS on incident CHD restricted to prospective cohort studies.

Conducting large-scale genetic studies in prospective cohort studies is likely to lead to more clinically relevant SNPs for prediction, but there are further developments that may also achieve this goal. First, increasing the discovery GWAS sample size will continue to lead to more effective genetic risk scores, by identifying new SNPs and by refining the effect estimates of known SNPs. Chatterjee *et al.* projected that the predictive performance of genetic risk scores for CHD may keep improving as GWAS samples increase to as much as ten times their current size.[28] Our study also supports intensifying the discovery effort: the most effective risk score not only included SNPs robustly associated with CHD, but also 103 further SNPs suggestively associated with CHD. Second, denser genotyping arrays, denser imputation panels, exome and whole-genome sequencing studies may yield low-frequency and rare variants for CHD that were hidden from GWAS. While common variants usually have small effect sizes due to evolutionary constraints, rarer variants may also have intermediate to large effect sizes. Therefore, while a single rare variant only explains a small proportion of variance in the general population, it can explain a large proportion of variance in families where it is present.

Family history only overlapped slightly with the genetic risk scores in the variance of CHD explained, providing largely independent information. Our results suggest that family history largely tags genetic variants that are not well covered by GWAS, or aspects of the shared environment that are independent of traditional risk factors. These hidden risk factors appear to affect CHD risk by increasing the burden of subclinical atherosclerosis.[29]

This study has certain strengths and limitations. Firstly, we examined the association between the genetic risk scores and both incident and prevalent CHD in the same population, allowing us to compare these associations. Since associated SNPs were identified using the largest available GWAS of CHD, a relatively large set of CHD SNPs with well-estimated weights was used, including multiple independently associated SNPs per locus when known. Previous studies have focused on genome-wide significant SNPs to include only the most robustly associated SNPs. This was also our approach for GRS_{gws}, but by including both genome-wide significant SNPs and suggestively associated SNPs in GRS_{all}, we were able to create a stronger genetic instrument than GRS_{gws}. In addition, this study included individuals of 55 years and older, which corresponds well with the target population for prediction. On the

other hand, our population consisted entirely of Caucasians, and our results may not be generalizable to other populations. Furthermore, we used a crude measure of family history. First, family history was only available for myocardial infarction and not for CHD in general. Second, family history was obtained during an interview, and may not always be complete. Third, participants were only asked about first degree relatives. However, these limitations reflect difficulties in measuring family history that also arise in clinical practice.

While our results do not support a role for currently available common SNPs in CHD risk prediction in the traditional setting, they do suggest that it could already improve prediction of future CHD earlier in life, when other variables used in prediction are not yet available. Our results also suggest that SNPs identified through GWAS of prevalent disease may not be optimally suited for the prediction of incident disease. This mismatch may extend to other diseases with high mortality rates.

Supplement available online at:
http://ije.oxfordjournals.org/content/44/2/682/suppl/DC1

REFERENCES

1. Zdravkovic S, Wienke A, Pedersen NL, de Faire U. Genetic influences on angina pectoris and its impact on coronary heart disease. Eur J Hum Genet. 2007;15:872-877.
2. Zdravkovic S, Wienke A, Pedersen NL, Marenberg ME, Yashin AI, De Faire U. Heritability of death from coronary heart disease: a 36-year follow-up of 20 966 Swedish twins. J Intern Med. 2002;252:247-254.
3. Brautbar A, Pompeii LA, Dehghan A, et al. A genetic risk score based on direct associations with coronary heart disease improves coronary heart disease risk prediction in the Atherosclerosis Risk in Communities (ARIC), but not in the Rotterdam and Framingham Offspring, Studies. Atherosclerosis. 2012;223:421-426.
4. Hughes MF, Saarela O, Stritzke J, et al. Genetic markers enhance coronary risk prediction in men: the MORGAM prospective cohorts. PLoS One. 2012;7:e40922.
5. Morrison AC, Bare LA, Chambless LE, et al. Prediction of coronary heart disease risk using a genetic risk score: the Atherosclerosis Risk in Communities Study. Am J Epidemiol. 2007;166:28-35.
6. Thanassoulis G, Peloso GM, Pencina MJ, et al. A genetic risk score is associated with incident cardiovascular disease and coronary artery calcium: the Framingham Heart Study. Circ Cardiovasc Genet. 2012;5:113-121.
7. Ganna A, Magnusson PK, Pedersen NL, et al. Multilocus genetic risk scores for coronary heart disease prediction. Arterioscler Thromb Vasc Biol. 2013;33:2267-2272.
8. Davies RW, Dandona S, Stewart AF, et al. Improved prediction of cardiovascular disease based on a panel of single nucleotide polymorphisms identified through genome-wide association studies. Circ Cardiovasc Genet. 2010;3:468-474.
9. Lluis-Ganella C, Lucas G, Subirana I, et al. Additive effect of multiple genetic variants on the risk of coronary artery disease. Rev Esp Cardiol. 2010;63:925-933.
10. Lluis-Ganella C, Subirana I, Lucas G, et al. Assessment of the value of a genetic risk score in improving the estimation of coronary risk. Atherosclerosis. 2012;222:456-463.
11. Paynter NP, Chasman DI, Pare G, et al. Association between a literature-based genetic risk score and cardiovascular events in women. JAMA. 2010;303:631-637.
12. Ripatti S, Tikkanen E, Orho-Melander M, et al. A multilocus genetic risk score for coronary heart disease: case-control and prospective cohort analyses. Lancet. 2010;376:1393-1400.
13. Vaarhorst AA, Lu Y, Heijmans BT, et al. Literature-based genetic risk scores for coronary heart disease: the Cardiovascular Registry Maastricht (CAREMA) prospective cohort study. Circ Cardiovasc Genet. 2012;5:202-209.
14. Consortium CAD, Deloukas P, Kanoni S, et al. Large-scale association analysis identifies new risk loci for coronary artery disease. Nat Genet. 2013;45:25-33.
15. Hofman A, Grobbee DE, de Jong PT, van den Ouweland FA. Determinants of disease and disability in the elderly: the Rotterdam Elderly Study. Eur J Epidemiol. 1991;7:403-422.
16. Hofman A, Darwish Murad S, van Duijn CM, et al. The Rotterdam Study: 2014 objectives and design update. Eur J Epidemiol. 2013;28:889-926.
17. Erdmann J, Grosshennig A, Braund PS, et al. New susceptibility locus for coronary artery disease on chromosome 3q22.3. Nat Genet. 2009;41:280-282.
18. Waterworth DM, Ricketts SL, Song K, et al. Genetic variants influencing circulating lipid levels and risk of coronary artery disease. Arterioscler Thromb Vasc Biol. 2010;30:2264-2276.
19. International HapMap C. The International HapMap Project. Nature. 2003;426:789-796.

20. Kundu S, Aulchenko YS, van Duijn CM, Janssens AC. PredictABEL: an R package for the assessment of risk prediction models. Eur J Epidemiol. 2011;26:261-264.
21. Leening MJ, Kavousi M, Heeringa J, et al. Methods of data collection and definitions of cardiac outcomes in the Rotterdam Study. Eur J Epidemiol. 2012;27:173-185.
22. Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. Circulation. 1998;97:1837-1847.
23. Pencina MJ, D'Agostino RB, Sr., D'Agostino RB, Jr., Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. Stat Med. 2008;27:157-172; discussion 207-112.
24. Pencina MJ, D'Agostino RB, Sr., Steyerberg EW. Extensions of net reclassification improvement calculations to measure usefulness of new biomarkers. Stat Med. 2011;30:11-21.
25. Nagelkerke NJD. A note on a general definition of the coefficient of determination. Biometrika. 1991;78:691-692.
26. Hill G, Connelly J, Hebert R, Lindsay J, Millar W. Neyman's bias re-visited. J Clin Epidemiol. 2003;56:293-296.
27. Janssens AC, van Duijn CM. Genome-based prediction of common diseases: methodological considerations for future research. Genome Med. 2009;1:20.
28. Chatterjee N, Wheeler B, Sampson J, Hartge P, Chanock SJ, Park JH. Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. Nat Genet. 2013;45:400-405.
29. Nasir K, Budoff MJ, Wong ND, et al. Family history of premature coronary heart disease and coronary artery calcification: Multi-Ethnic Study of Atherosclerosis (MESA). Circulation. 2007;116:619-626.

Chapter 4.2

Association of miR-4513 with cardiovascular disease and its risk factors

Manuscript based on this chapter

Mohsen Ghanbari, Paul S. de Vries, Hans de Looper, Marjolein J. Peters, Claudia Schurmann, Hanieh Yaghootkar, Marcus Dörr, Timothy M. Frayling, Andre G. Uitterlinden, Albert Hofman, Joyce B.J. van Meurs, Stefan J. Erkeland, Oscar H. Franco, and Abbas Dehghan.

A genetic variant in the seed region of miR-4513 shows pleiotropic effects on lipid and glucose homeostasis, blood pressure and coronary artery disease.

Human Mutation. 2014; 35(12): 1524-31.

ABSTRACT

Background: MicroRNAs (miRNA) play a crucial role in the regulation of diverse biological processes by post-transcriptional modulation of gene expression. Genetic polymorphisms in miRNA-related genes can potentially contribute to a wide range of phenotypes. The effect of such variants on cardiometabolic diseases has not yet been defined.

Methods: We systematically investigated the association of genetic variants in the seed regions of miRNAs with cardiometabolic phenotypes, using the thus far largest genome wide association studies on 17 cardiometabolic traits/diseases.

Results: We found that rs2168518:G>A, a seed region variant of miR-4513, associates with fasting glucose, LDL-cholesterol and total cholesterol, systolic and diastolic blood pressure and risk of coronary artery disease. We experimentally showed that miR-4513 expression is significantly reduced in presence of the rs2168518 mutant allele. We sought to identify miR-4513 target genes that may mediate these associations and revealed five genes (*PCSK1*, *BNC2*, *MTMR3*, *ANK3* and *GOSR2*) through which these effects might be taking place. Using luciferase reporter assays we validated *GOSR2* as a target of miR-4513 and further demonstrated that the miRNA mediated regulation of this gene is changed by rs2168518.

Conclusions: Our findings indicate a pleiotropic effect of miR-4513 on cardiometabolic phenotypes and may improve our understanding of the pathophysiology of cardiometabolic diseases.

INTRODUCTION

MicroRNAs (miRNAs) are a class of small non-coding RNAs spanning 20-24 nucleotides that function as crucial regulators in a broad range of biological processes.[1] Since the first miRNA was discovered in the early 1990s, over 1500 miRNAs have been identified with confidence in humans.[2,3] These miRNAs together can regulate expression levels of approximately 60% of all human protein-coding genes.[4] In recent years miRNAs have been widely studied as potential diagnostic biomarkers and therapeutic targets in complex disorders.[5] Furthermore, miRNAs have gained attention as important modulators of cardiovascular diseases such as myocardial infarction,[6,7] cardiac hypertrophy,[8] and heart failure,[9] as well as various metabolic processes such as insulin production,[10] glucose homeostasis,[11] lipid metabolism,[12] and obesity.[13]

MiRNAs are post-transcriptional regulators of gene expression by interacting with the 3′ untranslated region (3′UTR) of the target mRNAs.[1] Thereby they repress translation and to a lesser extent accelerate the decay of target transcripts.[14] Given the central role of miRNAs in gene expression, genetic polymorphisms in the corresponding sequences of a miRNA may contribute to a wide range of phenotypic variation and disease susceptibility.[15,16] The core of a mature miRNA, called the "seed region", includes nucleotides 2-8 from the 5′ end, and plays a critical role in target gene recognition and interaction.[17] Genetic variation within this critical region of miRNA may both disrupt the interaction of a miRNA with target transcripts and create illegitimate miRNA targets.[18,19] Therefore, miRNA seed polymorphisms are expected to alter the expression profile of target genes and subsequently affect corresponding phenotypes; however, so far only very few pathogenic variants have been evidenced in cardiovascular disease and metabolic syndrome.

In the present study, we aimed to systematically investigate the association between miRNA seed polymorphisms and a number of cardiometabolic traits and diseases. In addition, we sought to determine whether any of the target genes of the identified miRNAs may mediate their effects on cardiometabolic phenotypes.

METHODS

Identification of miRNA seed polymorphisms

A flow chart of our approach to retrieve single nucleotide polymorphisms (SNPs) in miRNA seed regions is shown in **Figure 1**. We systematically screened all known human miRNAs to identify variants in their seed regions, by reviewing the literature and searching the following online databases: microSNiPer,[20] PolymiRTS,[21] Patrocles,[22]

and miRvar.[23] We included variants with minor allele frequency (MAF) > 0.01. Since previous genome-wide association (GWA) meta-analysis on cardiometabolic traits/diseases have been performed using HapMap imputed data, we focused on SNPs that were present in the international HapMap project (release 22) (http://www.hapmap.org/).[24] For SNPs that were not present in the HapMap, we used the SNAP web tool to find proxy SNPs in high Linkage Disequilibrium (LD) (R^2 > 0.8 and distance <200 kb) (http://www.broadinstitute.org/ mpg/snap/id).[25]

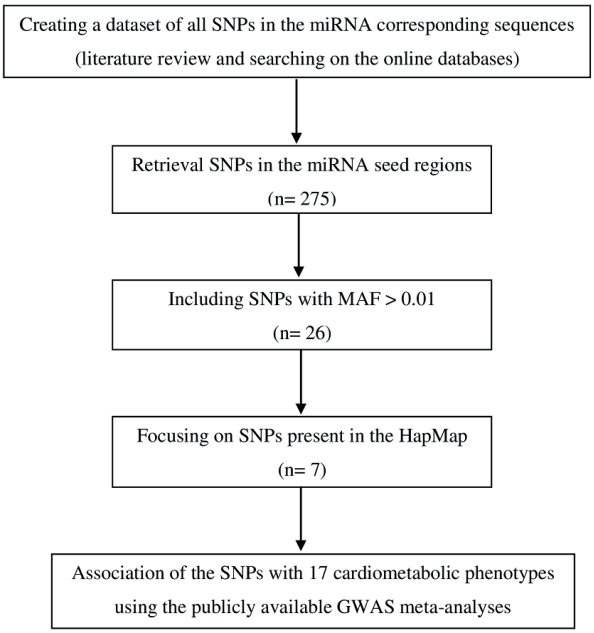


Figure 1. Identification of polymorphisms within the miRNA seed regions.
The flow chart describes the selection process to retrieve SNPs in the seed regions of miRNAs.

Association of miRNA seed polymorphisms with cardiometabolic phenotypes

We examined the association of miRNA seed SNPs with cardiometabolic phenotypes using the thus far largest available GWA meta-analyses of 17 cardiometabolic traits and diseases. **Table 1** shows a description of cardiometabolic phenotypes and consortia that we used in this study.

Data on glycemic traits have been contributed by the Meta-Analyses of Glucose and Insulin-related traits Consortium (MAGIC) investigators, including fasting glucose, serum glucose after 2hr, fasting insulin, fasting pro-insulin, HbA1c, HOMA-B, and HOMA-IR from up to 133,000 individuals (http://www.magicinvestigators.org).[26-30] The DIAbetes Genetics Replication and Meta-analysis (DIAGRAMv3) con-

sortium has done a GWA meta-analysis in 12,171 T2D cases and 56,862 controls.[31] The Global Lipid Genetics Consortium (GLGC) has carried out GWA studies of plasma concentrations of total cholesterol, low-density lipoprotein cholesterol (LDL), high-density lipoprotein cholesterol (HDL) and triglycerides for approximately 100,000 individuals.[32] The Genetic Investigation of ANthropometric Traits (GIANT) consortium has performed GWA studies on anthropometric traits including body mass index (BMI) of over 120,000 and waist/hip ratio (WHR) adjusted for BMI of 77,000 individuals.[33,34] The Global BPgen consortium has done GWA studies on systolic and diastolic blood pressure in over 71,000 individuals. Individuals under treatment for hypertension were imputed to have 15 mm Hg higher systolic blood pressure and 10 mm Hg higher diastolic blood pressure than the observed measurements.[35] The CARDIoGRAMplusC4D consortium conducted a GWA study in 63,746 CAD cases and 130,681 controls. In this study, they have assessed 79,138 important related SNPs with CAD on the Metabochip.[36]

Table 1. Description of publicly available GWA meta-analysis on cardiometabolic phenotype**s.**

Phenotype	Consortium	Sample Size	Reference
Glycemic indices			
Fasting glucose	MAGIC	133,010	26
Fasting insulin	MAGIC	108,557	26
Glucose after2h	MAGIC	42,854	27
Pro-insulin	MAGIC	10,701	28
HbA1c	MAGIC	46,368	29
HOMA-B	MAGIC	46,186	30
HOMA-IR	MAGIC	46,186	30
Type 2 diabetes	DIAGRAM	12,171 cases/56,862 controls	31
Lipid traits			
Total Cholesterol	GLGC	100,184	32
Triglycerides	GLGC	96,598	32
HDL Cholesterol	GLGC	99,900	32
LDL Cholesterol	GLGC	95,454	32
Anthropometric measures			
BMI	GIANT	123,764	33
WHR	GIANT	77,105	34
Blood pressure			
Systolic BP	Global BPGen	71,225	35
Diastolic BP	Global BPGen	71,225	35
Coronary artery disease	CARDIoGRM	63,746 case/130,681 control	36

Shown are 17 cardiometabolic phenotypes and publicly available GWAS meta-analysis on these traits/disease that we used for the association studies.

Effect of miRNA seed variants on miRNA processing and expression

When a miRNA seed variant was associated with cardiometabolic phenotypes, we used the Vienna RNAfold algorithm to predict the effect of variant on the secondary structure and processing of the primary miRNA sequence.[37] Furthermore, we examined whether the SNP affects mature miRNA expression. We cloned the pre-miRNA sequence containing wild type or mutant alleles behind the gene encoding green fluorescent protein (GFP) in the expression plasmid MSCV-BC,[38] resulting in GFP-miRNA fusion transcripts. HEK293 cell transfection, total RNA isolation and quantitative PCRs were performed as previously described.[38]

Association of miRNA target genes with cardiometabolic phenotypes

To explore the putative mediatory role of the target genes of miRNA associated with cardiometabolic phenotypes, we investigated the association of genetic variants in the target genes with the associated phenotypes. The significance threshold for this analysis was set by using a Bonferroni correction based on the number of independent SNPs. We calculated the number of independent SNPs using of Linkage disequilibrium based SNP pruning in PLINK with $R^2 > 0.5$ (http://pngu.mgh.harvard.edu/~purcell /plink/). The TargetScan database was used to identify target gene information, including their context score, and evolutionary conserved sites of miRNAs (release 6.2) (http://www.TargetScan.org/).

Expression quantitative trait loci (eQTL)

We examined the effect of miRNA seed SNPs on the expression levels of miRNA target genes using whole blood *trans*-eQTL and on their host genes expression using *cis*-eQTL data from the Rotterdam Study (n=762). We further replicated the eQTL analyses in two other cohorts; SHIP-TREND (n=963) and InCHIANTI (n=611). The designs of these cohorts have been described in detail elsewhere.[39-41]

Association of SNPs or their proxies, based on an $R^2 > 0.7$, were assessed with gene expression levels in whole blood cells. Whole-blood cells were collected in PAXgene-tubes (Becton Dickinson). Total RNA was isolated using PAXgene Blood RNA kits (Qiagen), and to ensure a constant high quality of the RNA preparations, all RNA samples were analyzed using the Labchip GX (Calliper) according to the manufacturer's instructions. Samples with an RNA Quality Score>=7 were amplified and labelled (AmbionTotalPrep RNA), and hybridized to the Illumina Whole-Genome Expression Beadchips (HumanHT-12 v4). Processing of the samples was performed at the Genetic Laboratory of Internal Medicine, Erasmus University Medical Center Rotterdam. The RS-III expression dataset is available at GEO (Gene Expression Omnibus) public repository under the accession GSE 33828. For normalization, raw intensity data generated with the expression arrays were exported from Illu-

mina's GenomeStudio V 2010.1 Gene Expression Module to the R environment and quantile normalized and log2-transformed, as well as probe-centered, and sample-standardized.

We used the eQTL mapping pipeline called MegaQTL. eQTLs were deemed cis when the distance between the SNP chromosomal position and the probe midpoint was less than 250 kb; eQTLs were deemed trans when the distance between the SNP chromosomal position and the probe midpoint position was larger than 5 Mbp. eQTLs were mapped using Spearman's rank correlation, using the imputation dosage values as genotypes. Resultant correlations were then converted to *P*-values and their respective z-scores weighted with the square root of the sample size. The model was adjusted for 40 principal components, of which 18 components capture different blood count parameters.[42]

Luciferase reporter assay

We used luciferase reporter assay system to validate the predicted interaction of a miRNA with its identified target genes and also to determine the functional consequence of the miRNA seed SNP on the binding of miRNA to the target genes. To amplify the mature miR-4513 sequence, we used a forward primer containing *XhoI* restriction site (AACTCGAGAGGATGTGGTCTTTGCATCT TC) and a reverse primer containing *EcoRI* restriction site (AAGAATTCCCTCCAGTCTCCCCACCTAG). The miRNA sequences with major or minor alleles were cloned in the MSCV-BC vector. In addition, the 3'UTR sequence of *GOSR2* was amplified with the forward primer (AATCTAGAGTGATCCCAGCGACTCTTCA) containing the restriction enzyme site *XbaI* and the reverse primer (AAGGGCCCCCGTAGAGATGGCAGGGACT), containing an *ApaI* restriction site. The 3'UTR fragment of *GOSR2*, containing the putative target site of miR-4513, was cloned in the pGL3 Luciferase reporter vector.[38] All constructs were confirmed by Sanger sequencing. HEK293 cells were plated into 12-well plates and co-transfected with MSCV-wild type miR-4513 (contain major allele) or MSCV-mutant miR-4513 (contain minor allele) and pGL3 containing the 3'UTR fragment of *GOSR2*. Luciferase activity was measured with the Dual-Glo Luciferase Assay System according to manufacturer's protocol (Promega). Renilla luciferase activity was normalize to the corresponding firefly luciferase activity and plotted as a percentage of the control. The experiments were performed in triplicate.

Potential functional roles and pathway analysis for the identified miRNA target genes

To explore the pathways and networks in which the identified miRNA's target genes play a role, we performed Ingenuity Pathway Analysis (IPA). IPA is a knowledge database generated from peer-reviewed scientific publications that enables the

discovery of highly represented biological mechanisms, pathways or functions most relevant to the genes of interest from large, quantitative datasets (http://www.ingenuity.com/products/ipa/). We uploaded lists of target genes of miRNAs found to be associated with cardiometabolic phenotypes, and performed a core analysis with the default settings in IPA. We mapped these target genes to biological functions or canonical pathways. We looked at each gene separately to identify the associated pathways and biological networks. We further sought to determine whether the highlighted target genes of a miRNA that were found to be associated with cardiometabolic phenotypes are correlated together.

RESULTS

Genetic variants in the miRNA seed regions

We retrieved all possible SNPs in the miRNA corresponding sequences, of which a total number of 275 SNPs selected in the miRNA seed regions (**Supplemental Table 1**). We included SNPs with MAF > 0.01 (n=26) and focused on the SNPs present in the HapMap project (n=5). Using the SNAP web tool, we found 2 proxy SNPs in high LD (R^2 > 0.8 and distance < 200 kb) with 2 further miRNA seed variants that were not present in HapMap (**Figure 1**). Thus, we examined 7 SNPs pertaining to 7 different miRNAs, including miR-146a-3p, miR-548a, miR-1178-5p, miR-1269b, miR-4513, miR-4741, and miR-6499-5p (**Table 2**).

Table 2. MiRNA seed variants with MAF > 0.01 and present in the HapMap project.

SNP ID	Chr.	Coded allele	Non-coded allele	MAF (Coded allele)	SNP proxy	miRNA ID	miRNA location
rs2910164	5	C	G	0.24	-	miR-146a-3p	Intergenic
rs3734050	5	T	C	0.098	-	miR-6499-5p	FAT2
rs7210937	4	C	G	0.074	-	miR-1269b	ARHGAP44
rs7311975	12	C	T	0.028	-	miR-1178-5p	CIT
rs515924	6	G	A	0.15	rs676103*	miR-548a	Intergenic
rs2168518	15	A	G	0.31	rs1378942**	miR-4513	CSK
rs7227168	18	T	C	0.12	rs7239066***	miR-4741	RBBP8

Shown are 7 miRNA seed SNPs with minor allele frequency (MAF) > 0.01 which are present in the HapMap project (release 22). For those SNPs that were not present in HapMap imputed data, we used their proxies in high linkage disequilibrium (LD), marked by star.
* R^2=1.0 and distance =189bp (A/G)
** R^2=1.0 and distance =928bp (A/C)
*** R^2=1.0 and distance = 1351bp (A/G)

A miR-4513 seed variant associates with multiple cardiometabolic phenotypes

The genetic association analysis of 7 miRNA seed SNPs with 17 cardiometabolic traits/ diseases are shown in **Supplemental Table 2**. We used a Bonferroni correction to compensate for 119 tests (7×17=119), resulting in a *P*-value of 4.2×10^{-4} as a threshold of study-wide significance. We found rs1378942:G>T, a proxy in full LD (R^2= 1.0) with rs2168518:C>T in the seed region of miR-4513 **(Supplemental Figure 1)**, to be significantly associated with multiple cardiometabolic phenotypes. Among glycemic traits, rs1378942 was significantly associated with increased levels of fasting glucose (effective allele: A, *P*-value=2.5×10^{-4}, β= 1.2×10^{-2}). For lipid traits, the A allele of rs1378942 was significantly associated with higher LDL (*P*-value = 5.6×10^{-5}, z-score=4.03) and total cholesterol (*P*-value= 5.7×10^{-5}, z-score= 4.02). This allele was also significantly associated with higher systolic (*P*-value=3.4×10^{-10}) and diastolic blood pressure (*P*-value=3.5×10^{-12}). Moreover, the A allele of rs1378942 showed a suggestive association with increased risk of CAD (*P*-values=9.2×10^{-4}). Additionally, we generated regional association plots of the related genomic region of this SNP for the identified traits using LocusZoom web tool (Version 1.1.).[15] **Supplemental Figure 2** illustrates the association of rs1378942 with these traits in regional association plots, showing that this SNP either has the strongest association with the trait in the given genomic region or is one of the strongest ones.

rs2168518 affects the miR-4513 processing and expression

We observed 0.49 kcal/mol difference in the free energy of the thermodynamic ensemble of the mutant versus the wild type primary miR-4513 sequence, which may affect the processing of the primary miRNA (**Supplemental Figure 3**). We cloned the pre-miR-4513 sequence (containing the wild type or mutant alleles) behind the GFP in the expression plasmid to examine the effect of rs2168518 on the level of mature miR-4513 expression. Transient transfection experiments in HEK293 cells showed a significant reduced level of miR-4513 from the mutant allele relative to GFP compared to the wild type allele (*P*-value =0.0048) (**Figure 2**).

miR-4513 target genes are associated with cardiometabolic phenotypes

We examined the association of all 109 predicted target genes of miR-4513 with the cardiometabolic traits to identify their putative mediatory roles in our findings **(Supplemental Table 3)**. After applying a Bonferroni correction to compensate for the multiple testing, we found five target genes to be significantly associated with the identified traits, including *PSCK1* with fasting glucose (*P*-value = 8.1×10^{-6}), *BNC2* with LDL (*P*-value=7.6×10^{-6}) and total cholesterol (*P*-value=6.6×10^{-6}) , *MTMR3* with total cholesterol (*P*-value =3.6×10^{-5}), *GOSR2* with systolic blood pressure (*P*-value=7.3×10^{-7}) and CAD (*P*-value= 1.5×10^{-6}), and *ANK3* with systolic blood pressure (*P*-value=3.9×10^{-5}) **(Figure 3)**.

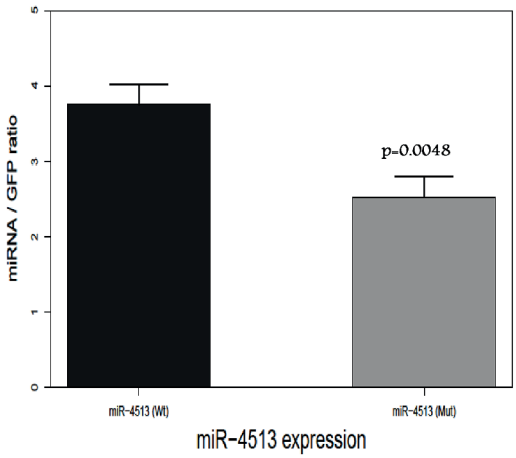


Figure 2. The effect of rs2168518 on miR-4513 expression containing the wild type or mutant alleles. This figure illustrates a significant reduced level of mature miR-4513 from the mutant allele relative to GFP compared to the wild type allele.

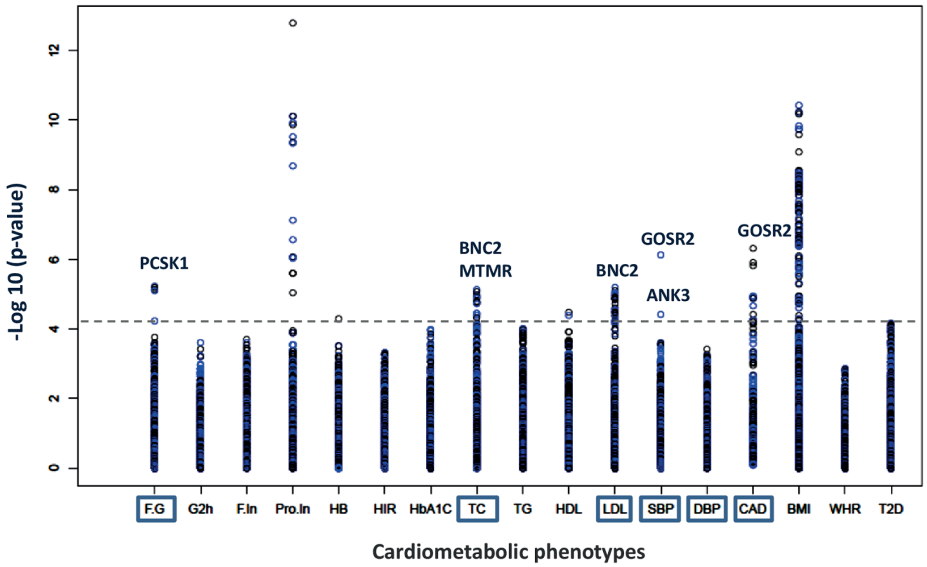


Figure 3. Association of miR-4513 target genes with the identified cardiometabolic traits.
This figure shows the association of 2,261 SNPs in the 109 predicted target genes of miR-4513 with 17 cardiometabolic traits. Dashed line indicates the significance threshold set at *P*-value < 2.2×10^{-5} (Bonferroni adjusted for 2,261tests). We highlighted the target genes which are most suspected to be influenced by the significantly associated SNPs. F.G, Fasting glucose; G2H, Glucose after 2hours; F.In, Fasting insulin; Pro.In, Pro-Insulin; HB, Homa-B; HIR, Homa-IR; HbA1c; TC, Total cholesterol; TG, Triglycerides; HDL, High density lipoprotein; LDL, Low density lipoprotein; SBP, Systolic blood pressure; DBP, Diastolic blood pressure; T2D, Type 2 diabetes; CAD, Coronary artery disease; BMI, Body mass index; WHR, Waist to hip ratio.

Association of rs2168518 with miR-4513 target genes and CSK expression

We examined the effect of SNP rs2168518 in miR-4513 on the expression levels of five identified genes using blood *trans*-eQTL data in 2,336 individuals. We did not find a statistically significant difference in the expression levels of target genes *PSCK1*, *BNC2*, *MTMR3*, *GOSR2*, and *ANK3* across different alleles of rs2168518. However, there was a positive trend in the mean RNA-expression levels of *GOSR2* in individuals carrying the risk allele of rs2168518 **(Supplemental Table 4)**. Our *cis*-eQTL analysis showed a significant association between rs2168518 and expression of miR-4513's host gene *CSK* (z-score=16.2, *P*-value=5.1×10^{-59}).

rs2168518 affects miR-4513 controlled expression of GOSR2

Next, we investigated whether the rs2168518 in miR-4513 effects on the expression level of *GOSR2* in-vitro. Therefore, we generated expression vectors with either the wild type (containing the major allele) or mutant miR-4513 (containing the minor allele) and co-transfected these constructs with Luciferase reporters containing the 3′ UTR of *GOSR2*. Overexpression of miR-4513 significantly decreased the Luciferase activity of *GOSR2* 3′UTR fragment by 45% (*P*-value=0.04), indicating *GOSR2* is a direct miR-4513 target (**Figure 4**). In addition, the rs2168518 SNP caused a reduced miR-4513 activity compared to the wild type miRNA, when the miRNA was overexpressed at lower levels, suggesting that the target repression efficiency, but not the specificity is changed by this SNP (**Figure 4**).

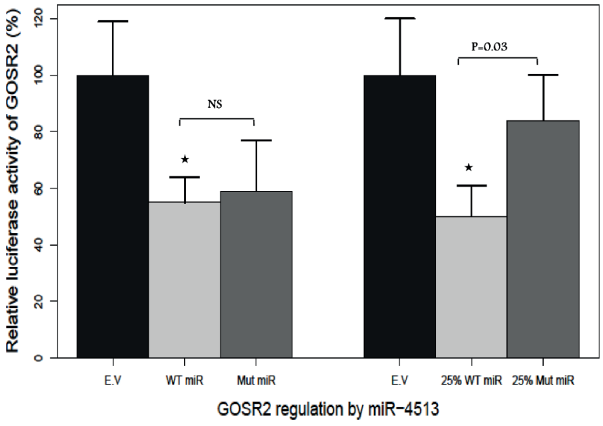


Figure 4. Luciferase reporter assays of the *GOSR2* 3′UTR in presence of miR-4513 containing the wild type or mutant alleles of rs218518.
This figure illustrates Luciferase reporter assays of cells transfected with pGL3 vector coupled to the miR-4513 (wild type and mutant) and 3′UTR regions of *GOSR2*. A significance differences of the mean relative luciferase activity between cells transfected with pGL3 vector coupled to 3′UTR region of *GOSR2* with or without miR-4513 (wild types) marked by * sign. This figure further shows rs2168518 mutant allele in miR-4513 affects the repression of *GOSR2* when overexpressed at lower levels (25% of normal).

Potential roles of the identified miR-4513 target genes in cardiometabolic phenotypes

The IPA core analysis was performed to determine the canonical pathways and networks that link the five identified miR-4513 target genes with the associated phenotypes. In agreement with our findings in the association study, there was a link between *PCSK1* and the insulin biosynthesis pathway and hyperglycemia. In addition, *MTMR3* and *BNC2* were correlated with lipid metabolism, *GOSR2* was associated with CAD and myocardial infarction and *ANK3* was linked with pulmonary and renal hypertension (**Supplemental Figure 4**). We further generated interaction networks between these five target genes of miR-4513 and their associated phenotypes. **Supplemental Figure 5** illustrates a potential pleiotropic effect of miR-4513 on cardiometabolic traits and diseases.

DISCUSSION

We found that rs2168518, a variant in the seed region of miR-4513, associates with fasting glucose, LDL and total cholesterol, and systolic and diastolic blood pressure. We identified five miR-4513 target genes, *GOSR2*, *ANK3*, *PCSK1*, *BNC2*, and *MTMR3*, as potential mediators of these associations. We then experimentally showed two mechanisms through which rs2168518 affects miR-4513 function. First, the rs2168518 mutant allele decreases miR-4513 expression. Second, rs2168518 reduces the ability of miR-4513 to repress the target genes (*GOSR2*) expression compared to the wild type in a concentration dependent manner.

In recent years numerous studies have provided strong evidence showing miRNAs as major players in complex disorders.[43,44] In addition, large advances have been made to identify the regulatory role of miRNAs in the pathophysiology of cardiometabolic diseases.[4,45,46] Since each miRNA regulates the expression of a large number of genes, genetic polymorphisms in miRNA corresponding sequences are expected to contribute to phenotypic variation and subsequently disease susceptibility.[47,48] Previous studies have reported an appreciable level of variation at miRNA binding sites and associated some of them with complex disorders.[49] However, since genetic variation in miRNA seed regions has important phenotypic consequences, they are not expected to be common. Polymorphisms in the seed of miRNAs have a strong effect on miRNA interaction with its target genes. For instance, a variant in the miR-96 seed region results in non-syndromic progressive hearing loss, and variants in the seed regions of miR-146a-3p and miR-499a-3p are associated with an increased risk of cancer.[19,50,51] Although variants on the miRNA target sites have previously linked with metabolic disorders,[52] the association of miRNA seed polymorphisms with

cardiometabolic phenotypes were not defined yet. Here we applied a systematic approach to investigate the association of miRNA seed SNPs with different cardiometabolic phenotypes. In agreement with previous studies, we show that common variants do not frequently occur within the seed region of miRNAs, and because of that many of the SNPs are not present in HapMap imputed data and are of negligible population genetic importance.[49]

However, we found that the SNP rs2168518 in miR-4513 is associated with fasting glucose, LDL and total cholesterol, blood pressure, and CAD. This is the first finding concerning the role of miR-4513 in disease since its discovery by deep sequencing in 2010.[53] We showed that the mature miR-4513 expression from the minor allele of rs2168518 is significantly reduced. The lower miR-4513 levels may be explained two possible mechanisms, which are not mutually exclusive. First, this variant could affect the expression of mature miRNA by interfering with miRNA processing efficiency and components such as the RNA-induced silencing complex (RISC) assembly and Dicer cleavage.[54,55] Second, the stability of rs2168518 containing miR-4513 may be reduced due to aberrant RISC loading and RNA degradation mechanisms.[56]

We highlighted five predicted target genes of miR-4513, *PCSK1*, *BNC2*, *MTMR3*, *ANK3* and *GOSR2*, as potential mediators of this effect on cardiometabolic phenotypes. We revealed a significant association between *PCSK1* and fasting glucose. This gene has previously been associated with obesity,[57] glucose metabolism, insulin secretion and risk of T2D.[58] *BNC2* is associated with HbA1c and glucose in type 1 diabetes.[58,59] Our results here indicate that this gene is also a regulator of cholesterol metabolism. *MTMR3* is an inositol lipid 3-phosphatase which is involved in lipid metabolism.[60] In agreement with our study, a recent large-scale meta-analysis of GWA studies of lipid traits has reported *MTMR3* to be associated with LDL cholesterol.[61] Our findings further showed an association between *ANK3* and higher systolic blood pressure. This gene has been previously highlighted to be involved in cardiac arrhythmia[62] and psychological disorders like bipolar disorders.[63] In addition, our pathway analysis using Ingenuity showed *ANK3* to be linked to pulmonary and renal hypertension. Finally, we report *GOSR2* to be associated with blood pressure and CAD by use of GWA study data. Previous studies of other investigators have also shown it to be associated with increased hypertension[64] and pulse pressure.[65] These findings indicate that our approach is valid to identify miRNA target genes that may mediate the effect of a miRNA on the studied traits. Since each miRNA regulate a large number of target genes, miRNAs have the potential to play a pleiotropic role in biological pathways. We demonstrate the pleiotropic effect of miR-4513 on cardiometabolic traits may be through its highlighted target genes.

Gene expression patterns are highly variable across tissues. Therefore, although we did not find an association between rs2168518 and blood expression levels of

the highlighted target genes, this does not rule out an effect in other tissues. Accordingly, previous studies have shown that *trans*-regulatory effects of gene expression are highly complex and with small effect size.[66] However, we identified a positive trend in the RNA-expression levels of *GOSR2* in individuals carrying the risk allele of rs2168518 in blood. Therefore, to have higher priority about the functional effect of rs2168518 on the expression of *GOSR2,* we employed the luciferase reporter assay system. We experimentally validated *GOSR2* as target genes of miR-4513, which is the first report of a validated target gene for this miRNA. We then showed that miR-4513 mediated regulation of *GOSR2* was only significantly affected by SNP rs2168518 at lower concentration. This dose-dependent effect of the miRNA concentrations can be explained by the minimal concentration that is necessary for a miRNA to regulate the target gene.[67] Alternatively, this may further indicates that rs2168518 changes the expression levels of mature miR-4513 rather than impairing the targeting. We found an association between rs2168518 and the expression of its host gene *CSK* in blood. Several reports demonstrate that the expression profiles of intragenic miRNAs are highly correlated with their corresponding host genes.[68-70] Therefore, it is possible to use the miRNA host gene expression as a proxy to monitor the expression of its embedded miRNA.[71] The identified association of rs2168518 with expression levels of *CSK* may subsequently indicate an altered expression of miR-4513 in individual carrying the mutant allele.

Previous GWA studies reported rs1378942, the SNP we used as a proxy for rs2168518, to be significantly associated with systolic and diastolic blood pressure and annotated that to *CSK*.[72] However, our results indicate that rs1378942 is tagging the altered function of miR-4513 caused by rs216518, and the resulting up-regulation of *GOSR2*. Furthermore, *GOSR2* has been robustly associated with blood pressure traits: in our study with systolic blood pressure, and previously with hypertension and pulse pressure.[64,65] Our findings further indicate that *GOSR2* is significantly associated with CAD. This may suggests miR-4513 as a candidate miRNA for blood pressure and CAD. Thus, it would be interesting to do further research on miR-4513 including expression levels of this miRNA in hypertensive and CAD patients.

To our knowledge, this is the first study to systematically investigate the association of genetic variations in the seed regions of miRNAs with cardiometabolic phenotypes. We demonstrate that a cardiometabolic-associated variant in the miR-4513 region seed affects the miRNA expression and activity. We provide data supporting a pleiotropic role for miR-4513 in cardiometabolic traits and highlight a number of its target genes including *GOSR2* as potential mediators. This may improve our understanding of the pathophysiology of cardiometabolic disorders. Moreover, our work introduces the investigation of miRNA variants as a novel approach to study the putative role of miRNAs in complex disorders. Given that the first phase of GWA

studies is complete, and information on the association of millions of SNPs with complex disorders is available, the time is ripe to apply this kind of approach to a wide range of traits and diseases to detect miRNA involved in complex disorders.

Supplement available online at:
http://onlinelibrary.wiley.com/doi/10.1002/humu.22706/suppinfo

REFERENCES

1. He L, Hannon GJ. MicroRNAs: small RNAs with a big role in gene regulation. Nat Rev Genet. 2004;5:522-531.
2. Lee RC, Feinbaum RL, Ambros V. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. Cell. 1993;75:843-854.
3. Kozomara A, Griffiths-Jones S. miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Res. 2014;42:D68-73.
4. Friedman RC, Farh KK, Burge CB, Bartel DP. Most mammalian mRNAs are conserved targets of microRNAs. Genome Res. 2009;19:92-105.
5. Kong YW, Ferland-McCollough D, Jackson TJ, Bushell M. microRNAs in cancer management. Lancet Oncol. 2012;13:e249-258.
6. Corsten MF, Dennert R, Jochems S, et al. Circulating MicroRNA-208b and MicroRNA-499 reflect myocardial damage in cardiovascular disease. Circ Cardiovasc Genet. 2010;3:499-506.
7. Eulalio A, Mano M, Dal Ferro M, et al. Functional screening identifies miRNAs inducing cardiac regeneration. Nature. 2012;492:376-381.
8. Song DW, Ryu JY, Kim JO, Kwon EJ, Kim DH. MicroRNA-19a/b family positively regulates cardiomyocyte hypertrophy by targeting atrogin-1 and MuRF-1. Biochem J. 2013.
9. Stather PW, Sylvius N, Wild JB, Choke E, Sayers RD, Bown MJ. Differential MicroRNA Expression Profiles in Peripheral Arterial Disease. Circ Cardiovasc Genet. 2013;6:490-497.
10. Trajkovski M, Hausser J, Soutschek J, et al. MicroRNAs 103 and 107 regulate insulin sensitivity. Nature. 2011;474:649-653.
11. Gauthier BR, Wollheim CB. MicroRNAs: 'ribo-regulators' of glucose homeostasis. Nat Med. 2006;12:36-38.
12. Chen L, Song J, Cui J, et al. microRNAs regulate adipocyte differentiation. Cell Biol Int. 2013.
13. Kornfeld JW, Baitzel C, Konner AC, et al. Obesity-induced overexpression of miR-802 impairs glucose metabolism through silencing of Hnf1b. Nature. 2013;494:111-115.
14. Bartel DP. MicroRNAs: target recognition and regulatory functions. Cell. 2009;136:215-233.
15. Pruim RJ, Welch RP, Sanna S, et al. LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics. 2010;26:2336-2337.
16. Lin E, Pei D, Huang YJ, Hsieh CH, Wu LS. Gene-gene interactions among genetic variants from obesity candidate genes for nonobese and obese populations in type 2 diabetes. Genet Test Mol Biomarkers. 2009;13:485-493.
17. Petersen CP, Bordeleau ME, Pelletier J, Sharp PA. Short RNAs repress translation after initiation in mammalian cells. Mol Cell. 2006;21:533-542.
18. Richardson K, Nettleton JA, Rotllan N, et al. Gain-of-function lipoprotein lipase variant rs13702 modulates lipid traits through disruption of a microRNA-410 seed site. Am J Hum Genet. 2013;92:5-14.
19. Mencia A, Modamio-Hoybjor S, Redshaw N, et al. Mutations in the seed region of human miR-96 are responsible for nonsyndromic progressive hearing loss. Nat Genet. 2009;41:609-613.
20. Barenboim M, Zoltick BJ, Guo Y, Weinberger DR. MicroSNiPer: a web tool for prediction of SNP effects on putative microRNA targets. Hum Mutat. 2010;31:1223-1232.
21. Bao L, Zhou M, Wu L, et al. PolymiRTS Database: linking polymorphisms in microRNA target sites with complex traits. Nucleic Acids Res. 2007;35:D51-54.

22. Hiard S, Charlier C, Coppieters W, Georges M, Baurain D. Patrocles: a database of polymorphic miRNA-mediated gene regulation in vertebrates. Nucleic Acids Res. 2010;38:D640-651.
23. Bhartiya D, Laddha SV, Mukhopadhyay A, Scaria V. miRvar: A comprehensive database for genomic variations in microRNAs. Hum Mutat. 2011;32:E2226-2245.
24. International HapMap C, Frazer KA, Ballinger DG, et al. A second generation human haplotype map of over 3.1 million SNPs. Nature. 2007;449:851-861.
25. Johnson AD, Handsaker RE, Pulit SL, Nizzari MM, O'Donnell CJ, de Bakker PI. SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. Bioinformatics. 2008;24:2938-2939.
26. Manning AK, Hivert MF, Scott RA, et al. A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. Nat Genet. 2012;44:659-669.
27. Saxena R, Hivert MF, Langenberg C, et al. Genetic variation in GIPR influences the glucose and insulin responses to an oral glucose challenge. Nat Genet. 2010;42:142-148.
28. Strawbridge RJ, Dupuis J, Prokopenko I, et al. Genome-wide association identifies nine common variants associated with fasting proinsulin levels and provides new insights into the pathophysiology of type 2 diabetes. Diabetes. 2011;60:2624-2634.
29. Soranzo N, Sanna S, Wheeler E, et al. Common variants at 10 genomic loci influence hemoglobin A(1)(C) levels via glycemic and nonglycemic pathways. Diabetes. 2010;59:3229-3239.
30. Dupuis J, Langenberg C, Prokopenko I, et al. New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. Nat Genet. 2010;42:105-116.
31. Scott RA, Lagou V, Welch RP, et al. Large-scale association analyses identify new loci influencing glycemic traits and provide insight into the underlying biological pathways. Nat Genet. 2012;44:991-1005.
32. Teslovich TM, Musunuru K, Smith AV, et al. Biological, clinical and population relevance of 95 loci for blood lipids. Nature. 2010;466:707-713.
33. Speliotes EK, Willer CJ, Berndt SI, et al. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. Nat Genet. 2010;42:937-948.
34. Heid IM, Jackson AU, Randall JC, et al. Meta-analysis identifies 13 new loci associated with waist-hip ratio and reveals sexual dimorphism in the genetic basis of fat distribution. Nat Genet. 2010;42:949-960.
35. International Consortium for Blood Pressure Genome-Wide Association Studies, Ehret GB, Munroe PB, et al. Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. Nature. 2011;478:103-109.
36. CARDIoGRAMplusC4D Consortium, Deloukas P, Kanoni S, et al. Large-scale association analysis identifies new risk loci for coronary artery disease. Nat Genet. 2013;45:25-33.
37. Lorenz R, Bernhart SH, Honer Zu Siederdissen C, et al. ViennaRNA Package 2.0. Algorithms Mol Biol. 2011;6:26.
38. Meenhuis A, van Veelen PA, de Looper H, et al. MiR-17/20/93/106 promote hematopoietic cell expansion by targeting sequestosome 1-regulated pathways in mice. Blood. 2011;118:916-925.
39. Ferrucci L, Bandinelli S, Benvenuti E, et al. Subsystems contributing to the decline in ability to walk: bridging the gap between epidemiology and geriatric practice in the InCHIANTI study. J Am Geriatr Soc. 2000;48:1618-1625.
40. Volzke H, Alte D, Schmidt CO, et al. Cohort profile: the study of health in Pomerania. Int J Epidemiol. 2011;40:294-307.

41. Hofman A, Darwish Murad S, van Duijn CM, et al. The Rotterdam Study: 2014 objectives and design update. Eur J Epidemiol. 2013;28:889-926.
42. Westra HJ, Peters MJ, Esko T, et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. Nat Genet. 2013.
43. Croce CM. Causes and consequences of microRNA dysregulation in cancer. Nat Rev Genet. 2009;10:704-714.
44. van Rooij E, Olson EN. MicroRNA therapeutics for cardiovascular disease: opportunities and obstacles. Nat Rev Drug Discov. 2012;11:860-872.
45. Fichtlscherer S, Zeiher AM, Dimmeler S. Circulating microRNAs: biomarkers or mediators of cardiovascular diseases? Arterioscler Thromb Vasc Biol. 2011;31:2383-2390.
46. Kumar M, Nath S, Prasad HK, Sharma GD, Li Y. MicroRNAs: a new ray of hope for diabetes mellitus. Protein Cell. 2012;3:726-738.
47. Gong J, Tong Y, Zhang HM, et al. Genome-wide identification of SNPs in microRNA genes and the SNP effects on microRNA target binding and biogenesis. Hum Mutat. 2012;33:254-263.
48. Abelson JF, Kwan KY, O'Roak BJ, et al. Sequence variants in SLITRK1 are associated with Tourette's syndrome. Science. 2005;310:317-320.
49. Saunders MA, Liang H, Li WH. Human polymorphism at microRNAs and microRNA target sites. Proc Natl Acad Sci U S A. 2007;104:3300-3305.
50. Dorn GW, 2nd, Matkovich SJ, Eschenbacher WH, Zhang Y. A human 3' miR-499 mutation alters cardiac mRNA targeting and function. Circ Res. 2012;110:958-967.
51. Wang J, Wang Q, Liu H, et al. The association of miR-146a rs2910164 and miR-196a2 rs11614913 polymorphisms with cancer risk: a meta-analysis of 32 studies. Mutagenesis. 2012;27:779-788.
52. Zhao X, Ye Q, Xu K, et al. Single-nucleotide polymorphisms inside microRNA target sites influence the susceptibility to type 2 diabetes. J Hum Genet. 2013;58:135-141.
53. Jima DD, Zhang J, Jacobs C, et al. Deep sequencing of the small RNA transcriptome of normal and malignant human B cells identifies hundreds of novel microRNAs. Blood. 2010;116:e118-127.
54. Khvorova A, Reynolds A, Jayasena SD. Functional siRNAs and miRNAs exhibit strand bias. Cell. 2003;115:209-216.
55. Krol J, Sobczak K, Wilczynska U, et al. Structural features of microRNA (miRNA) precursors and their relevance to miRNA biogenesis and small interfering RNA/short hairpin RNA design. J Biol Chem. 2004;279:42230-42239.
56. Kawamata T, Seitz H, Tomari Y. Structural determinants of miRNAs for RISC loading and slicer-independent unwinding. Nat Struct Mol Biol. 2009;16:953-960.
57. Heni M, Haupt A, Schafer SA, et al. Association of obesity risk SNPs in PCSK1 with insulin sensitivity and proinsulin conversion. BMC Med Genet. 2010;11:86.
58. Paterson AD, Waggott D, Boright AP, et al. A genome-wide association study identifies a novel major locus for glycemic control in type 1 diabetes, as measured by both A1C and glucose. Diabetes. 2010;59:539-549.
59. Hertel JK, Johansson S, Raeder H, et al. Evaluation of four novel genetic variants affecting hemoglobin A1c levels in a population-based type 2 diabetes cohort (the HUNT2 study). BMC Med Genet. 2011;12:20.
60. Walker DM, Urbe S, Dove SK, Tenza D, Raposo G, Clague MJ. Characterization of MTMR3. an inositol lipid 3-phosphatase with novel substrate specificity. Curr Biol. 2001;11:1600-1605.

61. Global Lipids Genetics C, Willer CJ, Schmidt EM, et al. Discovery and refinement of loci associated with lipid levels. Nat Genet. 2013.
62. Mohler PJ, Rivolta I, Napolitano C, et al. Nav1.5 E1053K mutation causing Brugada syndrome blocks binding to ankyrin-G and expression of Nav1.5 on the surface of cardiomyocytes. Proc Natl Acad Sci U S A. 2004;101:17533-17538.
63. Ferreira MA, O'Donovan MC, Meng YA, et al. Collaborative genome-wide association analysis supports a role for ANK3 and CACNA1C in bipolar disorder. Nat Genet. 2008;40:1056-1058.
64. International Consortium for Blood Pressure Genome-Wide Association S, Ehret GB, Munroe PB, et al. Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. Nature. 2011;478:103-109.
65. Wain LV, Verwoert GC, O'Reilly PF, et al. Genome-wide association study identifies six new loci influencing pulse pressure and mean arterial pressure. Nat Genet. 2011;43:1005-1011.
66. Grundberg E, Small KS, Hedman AK, et al. Mapping cis- and trans-regulatory effects across multiple tissues in twins. Nat Genet. 2012;44:1084-1089.
67. Shu J, Xia Z, Li L, et al. Dose-dependent differential mRNA target selection and regulation by let-7a-7f and miR-17-92 cluster microRNAs. RNA Biol. 2012;9:1275-1287.
68. Baskerville S, Bartel DP. Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes. RNA. 2005;11:241-247.
69. Karali M, Peluso I, Marigo V, Banfi S. Identification and characterization of microRNAs expressed in the mouse eye. Invest Ophthalmol Vis Sci. 2007;48:509-515.
70. Kim YK, Kim VN. Processing of intronic microRNAs. EMBO J. 2007;26:775-783.
71. Tsang J, Zhu J, van Oudenaarden A. MicroRNA-mediated feedback and feedforward loops are recurrent network motifs in mammals. Mol Cell. 2007;26:753-767.
72. Newton-Cheh C, Johnson T, Gateva V, et al. Genome-wide association study identifies eight loci associated with blood pressure. Nat Genet. 2009;41:666-676.

Chapter 4.3

Transcriptome-wide association study of carotid intima media thickness

Manuscript based on this chapter*
Paul S. de Vries, Markus Scholz, Stefan Weiss, Katharina Schramm, Klodian Dhana, Frank Beutner, Ulf Schminke, Carola S. Marzi, Marjolein J. Peters, Knut Krohn, Georg Homuth, Petra Wolf, Joyce B. van Meurs, Kerstin Wirkner, Marcus Dörr, Annette Peters, André G. Uitterlinden, Maryam Kavousi, Christian Herder, Melanie Waldenberger, Christa Meisinger, Wolfgang Rathmann, Joachim Thiery, Wolfgang Koenig, Jochen Seissler, Albert Hofman, Oscar H. Franco, Holger Prokisch, Henry Völzke, Markus Loeffler, and Abbas Dehghan.
*This author list may change according to contributions after the printing of this thesis.

Whole blood transcriptome-wide association study of atherosclerosis as measured by carotid intima media thickness.

Manuscript in preparation.

ABSTRACT

Background: Carotid intima media thickness (cIMT) is a marker of atherosclerosis and a predictor of cardiovascular disease. Whole blood gene expression levels may provide insights into the etiology and consequences of atherosclerosis.

Methods: We measured cIMT and genome-wide gene expression levels in whole blood of 5,647 individuals from four population-based cohort studies: KORA, LIFE-Adult, SHIP, and the Rotterdam Study. We examined the association of over 50,000 gene expression probes with cIMT adjusted for age, sex, batch effects, cell counts, RNA quality, fasting, and smoking status. In a sensitivity analysis, we further adjusted the model for traditional cardiovascular risk factors, and excluded participants with prevalent coronary heart disease. Finally, we explored whether probes mapping to genes identified for coronary heart disease were enriched for association with cIMT.

Results: After a Bonferroni correction (*P*-value < 9.2×10^{-7}), four probes mapping to three genes (*TNFAIP3*, *CEBPD*, and *METRNL*) were inversely associated with cIMT. Effect sizes and significance levels of the probes decreased after adjustment for traditional cardiovascular risk factors and exclusion of participants with prevalent coronary heart disease, but all remained nominally significant. Expression levels of genes that were previously implicated in coronary heart disease by genome-wide association studies were not enriched for association with cIMT.

Conclusions: Our results highlight the importance of inflammation in atherosclerosis as *TNFAIP3* and *METRNL* are anti-inflammatory genes, and *CEBPD* can be both pro and anti-inflammatory. Further research is needed to clarify whether the association between these genes and cIMT can indeed be explained through their anti-inflammatory properties.

INTRODUCTION

As a marker of atherosclerosis, carotid intima media thickness (cIMT) is a strong predictor of coronary heart disease (CHD) and stroke.[1,2] cIMT evaluates the full range of atherosclerosis: from early subclinical to full-blown clinical disease. Like CHD and stroke, cIMT has a moderate heritability,[3-7] and numerous loci have been identified through genetic association studies.[8-12] However, the genetic variants at these loci collectively explain only a small fraction of the heritability of cIMT. Furthermore, the ability of these genetic variants to predict incident cardiovascular disease remains limited.[13-16] Besides genetic association studies, alternative approaches harnessing genomic data may yield new loci associated with atherosclerosis.

One such approach is the transcriptome-wide association study, based on gene expression levels instead of genetic variants. Whole-blood is often used as it is feasible to measure on a large scale in a non-invasive manner, and also because it is a relevant tissue for atherosclerosis. Although several transcriptome-wide association studies have already identified genes whose expression is associated with cardiovascular disease, the overlap between the results of the different studies is very low.[17-23] No large-scale study has been performed on cIMT specifically.

Hence, within the framework of the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) consortium,[24] we aimed to robustly identify genes whose expression is associated with atherosclerosis. To this end, we profiled genome-wide gene expression levels in whole blood of 5,647 individuals with cIMT measurements available from four population-based cohort studies. We then replicated our findings in two further independent cohort studies.

METHODS

Study population

Individuals from four population-based cohort studies were included in the discovery analysis: 836 from KORA,[25,26] 2,973 from LIFE-Adult,[27] 856 from the Rotterdam Study,[28] and 982 from the Study of Health in Pomerania (SHIP).[29] The total sample size was 5,647. All studies were approved by appropriate research ethics committees and all participants signed informed consent prior to participation.

Measurement of carotid intima media thickness

cIMT of the common carotid artery was measured with high-resolution B-mode ultrasonography. cIMT was calculated as the mean of the maximum cIMT of the near

and far walls of the right and left common carotid arteries. When the intima media thickness of the near walls was unavailable, only the far walls were used.

Measurement of gene expression levels

Genome-wide gene expression levels in whole blood of up to 49,618 probes covering more than 25,000 genes were measured using the Ilumina HumanHT-12 Gene Expression BeadChip v3.0 or v4.0. In all four studies gene expression levels were measured based on blood that was drawn around the same time as cIMT was measured.

Statistical analysis

cIMT was natural-log transformed. We used a linear mixed model, adjusting for batch effects (examples: array ID and position on array) as random effects, and for further technical covariates (examples: RNA quality and storage time between sampling and RNA isolation), cell types (examples: granulocytes, lymphocytes, monocytes), age, sex, fasting state, and smoking status as fixed effects. We ran a separate model for each gene expression probe, using cIMT and the covariates as independent variables, and gene expression levels as the dependent variable. These analyses were done in R. Meta-analysis of the four studies was performed using inverse-variance fixed effects meta-analysis implemented in METAL.[30] We used a Bonferroni correction to adjust for multiple testing.

We performed additional analyses including further covariates relevant to atherosclerosis: total / high density lipoprotein (HDL) cholesterol ratio, systolic blood pressure, body mass index (BMI), prevalent type 2 diabetes, lipid-lowering medication and antihypertensive medication (Model 2). We also repeated the original model in only those individuals with data available on all of the additional covariates (Model 1). Finally, we reran the full model excluding individuals with prevalent CHD (Model 3).

We also examined whether the expression levels of genes related to CHD, as described by the CARDIoGRAMplusC4D consortium,[10] were enriched for associations with cIMT. For each genome-wide significant locus, we selected genes that the top variant or one of its proxies ($R^2 > 0.8$) were located in as exonic or intronic variants, and genes whose expression levels were associated with the top variant or one of its proxies. Associations between expression levels and genetics variants were examined using a publicly available dataset based on whole blood (http://genenetwork.nl/bloodeqtlbrowser/), and associations with a false discovery rate of less than 5% were considered significant.[31] A total of 48 genes were selected because they contained a top variant in-gene, and a total of 40 were selected because their expression levels were associated with one of the top variants, leading to a set of 74 unique CHD-related genes. We examined the association of expression levels of the

individually CHD-related genes with cIMT as described above, and we examined their collective enrichment for association with cIMT using Fisher's combined probability test.[32]

RESULTS

Clinical characteristics

Baseline characteristics of the studies included in the discovery analysis are shown in **Table 1**. The mean age of the participants across the four studies was 58.5 years, and 50.6% of participants were women. The mean BMI was 27.7 kg/m^2.

Transcriptome-wide association analysis

A total of 54,124 probes were included in the analysis, resulting in a Bonferroni corrected *P*-value threshold of 9.2×10^{-7}. There were 4 probes that were significantly associated with cIMT: ILMN_1780861 and ILMN_1688775 mapping to *METRNL*, ILMN_1702691 mapping to *TNFAIP3*, and ILMN_1782050 mapping to *CEBPD* (**Table 2**). All four probes were inversely associated with cIMT (**Figure 1**). The correlation between the 4 significant probes was low (**Figure 2**).

Table 1. Baseline characteristics of the four participating population-based cohort studies.

	KORA	LIFE-Adult	Rotterdam Study	SHIP
Sample size	836	2,973	856	982
Age	70.20 (5.34)	57.55 (12.48)	59.70 (8.02)	50.07 (13.74)
Sex (% women)	50.48	48.13	53.39	56.01
BMI (kg/m2)	28.99 (4.52)	27.43 (4.60)	27.71 (4.62)	27.28 (4.49)
HDL cholesterol (mmol/l)	1.44 (0.36)	1.58 (0.45)	1.40 (0.42)	1.48 (0.37)
Total cholesterol (mmol/l)	5.71 (1.03)	5.57 (1.07)	5.54 (1.08)	5.51 (1.07)
Lipid-lowering medication use (% yes)	24.28	15.2	27.0	7.33
Systolic blood pressure (mmHg)	128.48 (19.09)	128.97 (16.80)	134.53 (20.06)	124.33 (16.91)
Diastolic blood pressure mmHg)	73.93 (9.81)	75.46 (9.88)	82.92 (11.56)	76.50 (9.66)
Antihypertensive medication use (%yes)	56.82	44.50	27.27	29.33
Type 2 diabetes (% yes)	13.88	14.52	9.23	0.2
Current smoking (% yes)	6.22	20.92	27.10	18.43
Prevalent cardiovascular disease (% yes)	5.38	4.81	6.04	0.61
cIMT	0.97 (0.13)	0.75 (0.15)	0.96 (0.19)	0.73 (0.17)

Values are mean (SD) of percentages.
Abbreviations: BMI refers to body-mass index. HDL refers to high density lipoprotein. cIMT refers to carotid intima media thickness.

Table 2. Association of significant probes with cIMT in 5,647 individuals.

Probe ID	Locus	Gene	Effect Size	*P*-value
ILMN_1702691	6q23.3	*TNFAIP3*	-0.46	1.2×10^{-7}
ILMN_1782050	8q11.21	*CEBP*	-0.39	2.8×10^{-7}
ILMN_1688775	17q25.3	*METRNL*	-0.49	2.8×10^{-8}
ILMN_1780861	17q25.3	*METRNL*	-0.57	4.8×10^{-10}

Abbreviations: cIMT refers to carotid intima media thickness.

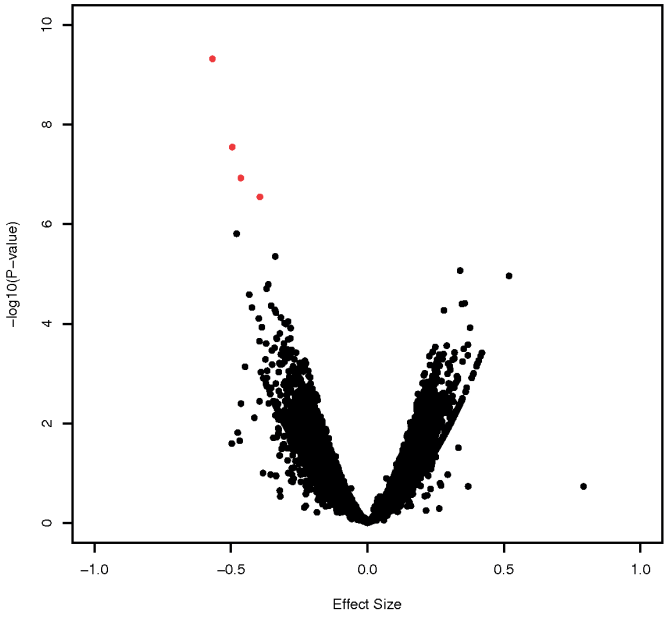


Figure 1. Volcano plot showing the –log10(*P*-value) of each probe plotted against the effect size, distinguishing between non-significant (black) and significant probes (red).

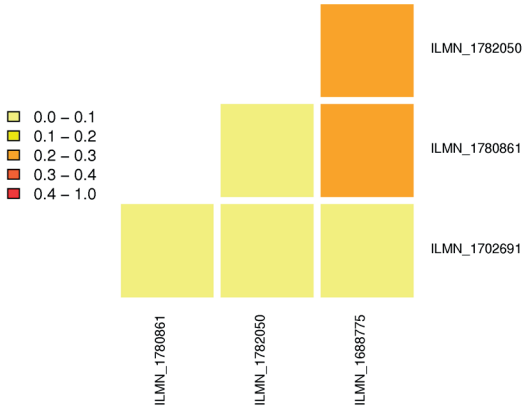


Figure 2. Correlation R^2 between the four probes that were significantly associated with cIMT.

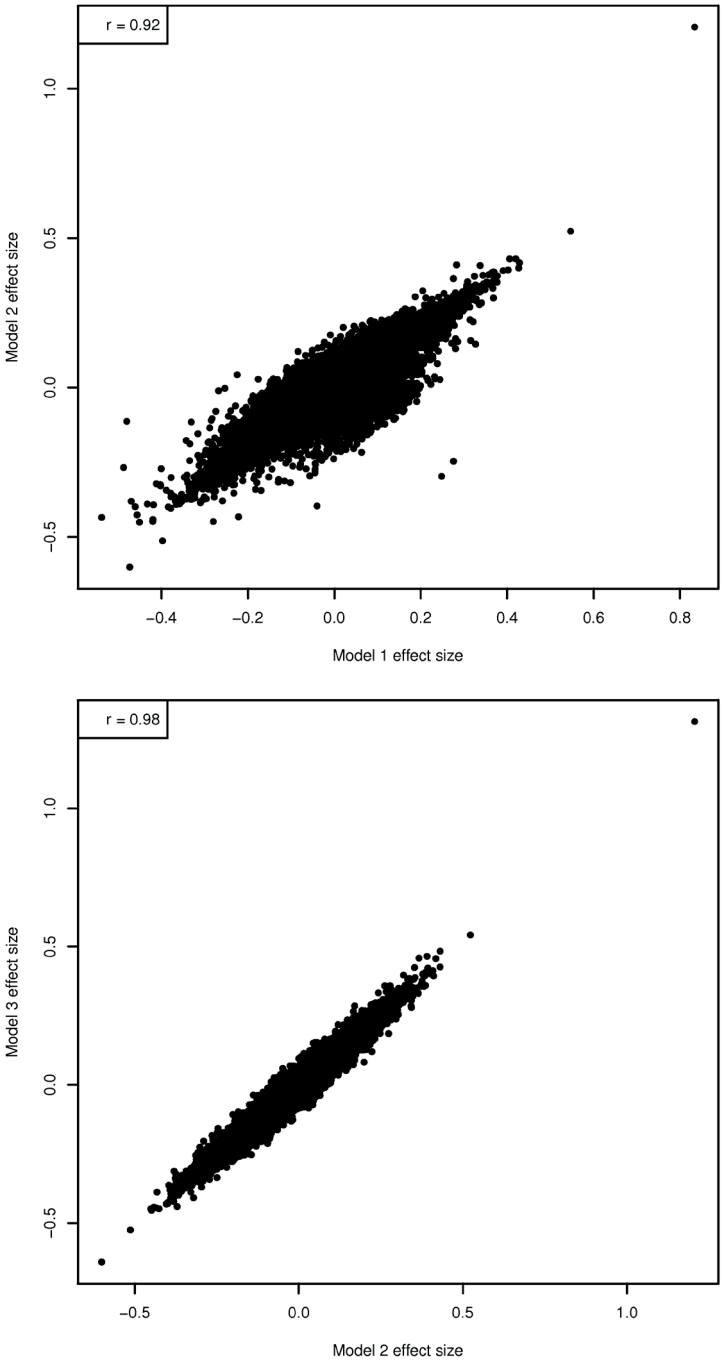


Figure 3A and 3B. Correlation of effect sizes between A) Model 1 and Model 2, and B) Model 2 and Model 3.

Additional Adjustments

As shown in **Figure 3A**, in general effect sizes and did not change substantially when we adjusted for total / high density lipoprotein (HDL) cholesterol ratio, systolic blood pressure, BMI, prevalent type 2 diabetes, lipid-lowering medication and antihypertensive medication (correlation $R^2 = 0.92$). As shown in **Figure 3B**, effect sizes also remained stable when we repeated the analysis excluding participants with prevalent CHD (correlation $R^2 = 0.98$). For the four significant probes in particular, effect sizes decreased when adjusted for additional covariates, though all probes remained nominally significant (**Table 3**). When participants with prevalent CHD were excluded, effect sizes remained stable or slightly increased. Of the four probes, the probe mapping to *TNFAIP3* was the most stable with effect estimates changing by less than 10% after adjustment.

CHD-related genes

68 of the 74 CHD-related genes had one or more probes that were included in the analysis. A total of 104 probes representing these genes were analysed. Collectively, the 104 probes of CHD-related genes were not enriched for association with cIMT (Fisher combined probability *P*-value = 0.75). None of the probes of CHD-related genes were associated with cIMT according to a less strict significance threshold corrected only for CHD genes ($0.05 / 104 = 4.8\times10^{-4}$).

Table 3. Additional adjustment analyses of significant probes.

		Model 1		Model 2		Model 3	
Probe ID	**Gene**	**Effect Size**	***P*-value**	**Effect Size**	***P*-value**	**Effect Size**	***P*-value**
ILMN_1702691	*TNFAIP3*	-0.43	1.1×10^{-6}	-0.39	1.1×10^{-5}	-0.40	1.6×10^{-5}
ILMN_1782050	*CEBP*	-0.41	1.4×10^{-7}	-0.32	2.8×10^{-5}	-0.36	7.6×10^{-6}
ILMN_1688775	*METRNL*	-0.47	2.4×10^{-7}	-0.38	2.2×10^{-5}	-0.39	2.2×10^{-5}
ILMN_1780861	*METRNL*	-0.54	6.6×10^{-9}	-0.43	1.6×10^{-6}	-0.45	1.9×10^{-6}

Adjustments: Model 1: batch effects, technical covariates, cell types, age, sex, fasting state, and smoking status. Only including individuals with data available on all of the additional Model 2 covariates. Model 2: Model 1 + total / HDL cholesterol ratio, systolic blood pressure, BMI, prevalent type 2 diabetes, lipid-lowering medication and antihypertensive medication. Model 3: Model 2 excluding participants with prevalent CHD.

DISCUSSION

We performed the first large-scale transcriptome-wide association study meta-analysis of cIMT including over 5,600 participants. We identified four gene expression probes mapping to three genes to be differentially expressed according to cIMT: *TNFAIP3*, *CEBPD*, and *METRNL*. The associations were robust to further adjustment for potential confounders, and excluding individuals with prevalent CHD did not

change the results. Probes at the three genes were not correlated to each other, suggesting that they represent separate mechanisms.

Expression levels of genes identified for CHD in the largest genome-wide association study were not associated with cIMT. Several possible explanations may explain the absence of associations. First, despite the predictive value of cIMT for CHD, cIMT and CHD may be too distinct as phenotypes to produce an overlap in associations with genes. In agreement, only one locus was found in genome-wide association studies of both cIMT and CHD.[8,33] Second, the genetic background of atherosclerosis and CHD may be differentially reflected through polymorphisms and gene expression levels. In a large-scale transcriptome-wide association study of blood pressure only two out of 34 genes were previously reported in relation to hypertension, and none were identified through genome-wide association studies.[34] Third, while blood is a relevant tissue for atherosclerosis, it may not be the tissue in which the genes identified by genome-wide association studies are primarily expressed.

TNFAIP3 encodes tumor necrosis factor α-induced protein-3, also known as A20, a protein involved in several inflammatory pathways. Most notably *TNFAIP3* is involved in the negative feedback regulation of NF-kappaB,[35] but it may also inhibit IFNγ/STATI signalling.[36] It is thus an anti-inflammatory protein, and low expression levels of *TNFAIP3* have been associated with inflammatory disorders such as rheumatoid arthritis.[37] In a small case-control study, genetic variants in *TNFAIP3* were associated both with increased odds of CHD and lower *TNFAIP3* expression in blood.[38] However, neither the association with CHD nor the association with expression levels was replicated in larger hypothesis-free studies.[10,31] The proposed anti-inflammatory properties of *TNFAIP3* are in line with our study, in which expression of *TNFAIP3* was inversely associated with cIMT.

CEBPD encodes CCAAT/Enhancer Binding Protein Delta (C/EBP-Delta), a transcription factor regulating several inflammatory genes.[39] Depending on the situation C/EBP-Delta can be both pro-inflammatory and anti-inflammatory: on the one hand, C/EBP-Delta may amplify the NF-kappaB response,[40,41] but on the other hand, C/EBP-Delta has been shown to have an anti-inflammatory role in pancreatic β-cells and brain pericytes,[42,43] while inhibiting the accumulation of amyloid plaques in Alzheimer's disease.[44] In our study, increased expression of *CEBPD* in blood is associated with less atherosclerosis as measured by cIMT.

The remaining two probes mapped to *METRNL*, which for meteorin-like protein (Metrnl). Metrnl increases thermogenesis in brown and beige adipocytes, and increases the expression of anti-inflammatory genes.[45] Brown and beige adipocytes may play a role in metabolic disease by inhibiting weight gain through thermogenesis.[46] Both the potential effects on adiposity and inflammation could explain the inverse association of *METRNL* expression with cIMT in our study.

All three genes identified in the transcriptome-wide association analyses thus appear to be related to inflammation. This is not surprising, given the importance of inflammation in atherosclerosis,[47,48] and the fact that expression levels were measured in whole blood, in which we expect most mRNA to originate from white blood cells. *TNFAIP3* and *METRNL* are both reported to have anti-inflammatory properties, which is consistent with the direction of the association in this study. *CEBPD*, on the other hand, is reported to have both inflammatory and anti-inflammatory properties. None of the three genes was reported to be significantly associated in a recent transcriptome-wide association study of interleukin-6 levels.[49] There has been no previous large-scale transcriptome-wide association study of cIMT, but several studies of CHD have been carried out. None of the three genes we report were significant in these previous studies.[17-23]

Strengths of this study include the large sample size, the hypothesis-free approach, and the strict correction for multiple testing. The main limitation of this study is the lack of replication. Although we consider whole blood to be a relevant tissue for the expression of genes associated with atherosclerosis, the use of only whole blood could be considered a limitation of this study. As gene expression is highly tissue specific, investigating other tissues, may yield important genes for atherosclerosis that remained hidden in this study.

Furthermore, the interpretation of the results is challenging because it is difficult to distinguish between genes whose expression influences atherosclerosis and genes whose expression is influenced by atherosclerosis. Although a longitudinal design could be used to focus on one of these two directions, reverse causation cannot be ruled out. Finally, the associations described in this study may be affected by residual confounding. We attempted to reduce the chance of confounding by correcting for batch effects, cell types, and, in an additional analysis, traditional cardiovascular risk factors. Nevertheless, other variables not covered in these models, as well as measurement error in the included variables may affect the results.

We identified novel three genes that were associated with atherosclerosis as measured by cIMT. All three genes are reported to be involved in inflammation, with *TNFAIP3* and *METRNL* having well described anti-inflammatory properties. Our results thus highlight the importance of inflammation in atherosclerosis, but further research is needed to clarify whether the association between these genes and cIMT can indeed be explained through their anti-inflammatory properties.

REFERENCES

1. Stein JH, Korcarz CE, Hurst RT, et al. Use of carotid ultrasound to identify subclinical vascular disease and evaluate cardiovascular disease risk: a consensus statement from the American Society of Echocardiography Carotid Intima-Media Thickness Task Force. Endorsed by the Society for Vascular Medicine. *J Am Soc Echocardiogr*. 2008;21(2):93-111; quiz 189-190.
2. Stein JH, Korcarz CE, Post WS. Use of carotid ultrasound to identify subclinical vascular disease and evaluate cardiovascular disease risk: summary and discussion of the American Society of Echocardiography consensus statement. *Prev Cardiol*. 2009;12(1):34-38.
3. Bak S, Gaist D, Sindrup SH, Skytthe A, Christensen K. Genetic liability in stroke: a long-term follow-up study of Danish twins. *Stroke*. 2002;33(3):769-774.
4. Rampersaud E, Bielak LF, Parsa A, et al. The association of coronary artery calcification and carotid artery intima-media thickness with distinct, traditional coronary artery disease risk factors in asymptomatic adults. *Am J Epidemiol*. 2008;168(9):1016-1023.
5. Sayed-Tabatabaei FA, van Rijn MJ, Schut AF, et al. Heritability of the function and structure of the arterial wall: findings of the Erasmus Rucphen Family (ERF) study. *Stroke*. 2005;36(11):2351-2356.
6. Swan L, Birnie DH, Inglis G, Connell JM, Hillis WS. The determination of carotid intima medial thickness in adults--a population-based twin study. *Atherosclerosis*. 2003;166(1):137-141.
7. Zdravkovic S, Wienke A, Pedersen NL, et al. Heritability of death from coronary heart disease: a 36-year follow-up of 20 966 Swedish twins. *J Intern Med*. 2002;252(3):247-254.
8. Bis JC, Kavousi M, Franceschini N, et al. Meta-analysis of genome-wide association studies from the CHARGE consortium identifies common variants associated with carotid intima media thickness and plaque. *Nat Genet*. 2011;43(10):940-947.
9. Bis JC, White CC, Franceschini N, et al. Sequencing of 2 subclinical atherosclerosis candidate regions in 3669 individuals: Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium Targeted Sequencing Study. *Circ Cardiovasc Genet*. 2014;7(3):359-364.
10. CARDIoGRAMplusC4D Consortium, Deloukas P, Kanoni S, et al. Large-scale association analysis identifies new risk loci for coronary artery disease. *Nat Genet*. 2013;45(1):25-33.
11. Ikram MA, Seshadri S, Bis JC, et al. Genomewide association studies of stroke. *N Engl J Med*. 2009;360(17):1718-1728.
12. Kilarski LL, Achterberg S, Devan WJ, et al. Meta-analysis in more than 17,900 cases of ischemic stroke reveals a novel association at 12q24.12. *Neurology*. 2014;83(8):678-685.
13. de Vries PS, Kavousi M, Ligthart S, et al. Incremental predictive value of 152 single nucleotide polymorphisms in the 10-year risk prediction of incident coronary heart disease: the Rotterdam Study. *Int J Epidemiol*. 2015;44(2):682-688.
14. Ibrahim-Verbaas CA, Fornage M, Bis JC, et al. Predicting stroke through genetic risk functions: the CHARGE Risk Score Project. *Stroke*. 2014;45(2):403-412.
15. Malik R, Bevan S, Nalls MA, et al. Multilocus genetic risk score associates with ischemic stroke in case-control and prospective cohort studies. *Stroke*. 2014;45(2):394-402.
16. Morrison AC, Bare LA, Chambless LE, et al. Prediction of coronary heart disease risk using a genetic risk score: the Atherosclerosis Risk in Communities Study. *Am J Epidemiol*. 2007;166(1):28-35.
17. Huan T, Zhang B, Wang Z, et al. A systems biology framework identifies molecular underpinnings of coronary heart disease. *Arterioscler Thromb Vasc Biol*. 2013;33(6):1427-1434.

18. Rosenberg S, Elashoff MR, Beineke P, et al. Multicenter validation of the diagnostic accuracy of a blood-based gene expression test for assessing obstructive coronary artery disease in nondiabetic patients. *Ann Intern Med*. 2010;153(7):425-434.
19. Sinnaeve PR, Donahue MP, Grass P, et al. Gene expression patterns in peripheral blood correlate with the extent of coronary artery disease. *PLoS One*. 2009;4(9):e7037.
20. Wingrove JA, Daniels SE, Sehnert AJ, et al. Correlation of peripheral-blood gene expression with the extent of coronary artery stenosis. *Circ Cardiovasc Genet*. 2008;1(1):31-38.
21. Joehanes R, Ying S, Huan T, et al. Gene expression signatures of coronary heart disease. *Arterioscler Thromb Vasc Biol*. 2013;33(6):1418-1426.
22. Nuhrenberg TG, Langwieser N, Binder H, et al. Transcriptome analysis in patients with progressive coronary artery disease: identification of differential gene expression in peripheral blood. *J Cardiovasc Transl Res*. 2013;6(1):81-93.
23. Taurino C, Miller WH, McBride MW, et al. Gene expression profiling in whole blood of patients with coronary artery disease. *Clin Sci (Lond)*. 2010;119(8):335-343.
24. Psaty BM, O'Donnell CJ, Gudnason V, 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*. 2009;2(1):73-80.
25. Holle R, Happich M, Lowel H, Wichmann HE, Group MKS. KORA--a research platform for population based health research. *Gesundheitswesen*. 2005;67 Suppl 1:S19-25.
26. Schurmann C, Heim K, Schillert A, et al. Analyzing illumina gene expression microarray data from different tissues: methodological aspects of data analysis in the metaxpress consortium. *PLoS One*. 2012;7(12):e50938.
27. Loeffler M, Engel C, Ahnert P, et al. The LIFE-Adult-Study: objectives and design of a population-based cohort study with 10,000 deeply phenotyped adults in Germany. *BMC Public Health*. 2015;15:691.
28. Hofman A, Brusselle GG, Darwish Murad S, et al. The Rotterdam Study: 2016 objectives and design update. *Eur J Epidemiol*. 2015;30(8):661-708.
29. Volzke H, Alte D, Schmidt CO, et al. Cohort profile: the study of health in Pomerania. *Int J Epidemiol*. 2011;40(2):294-307.
30. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics*. 2010;26(17):2190-2191.
31. Westra HJ, Peters MJ, Esko T, et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat Genet*. 2013;45(10):1238-1243.
32. Peng G, Luo L, Siu H, et al. Gene and pathway-based second-wave analysis of genome-wide association studies. *Eur J Hum Genet*. 2010;18(1):111-117.
33. CARDIoGRAMplusC4D Consortium. A comprehensive 1000 Genomes-based genome-wide association meta-analysis of coronary artery disease. *Nat Genet*. 2015;47(10):1121-1130.
34. Huan T, Esko T, Peters MJ, et al. A meta-analysis of gene expression signatures of blood pressure and hypertension. *PLoS Genet*. 2015;11(3):e1005035.
35. Coornaert B, Carpentier I, Beyaert R. A20: central gatekeeper in inflammation and immunity. *J Biol Chem*. 2009;284(13):8217-8221.
36. Moll HP, Lee A, Minussi DC, et al. A20 regulates atherogenic interferon (IFN)-gamma signaling in vascular cells by modulating basal IFNbeta levels. *J Biol Chem*. 2014;289(45):30912-30924.
37. Wang Z, Zhang Z, Yuan J, Li LI. Altered mRNA expression in peripheral blood mononuclear cells from patients with rheumatoid arthritis. *Biomed Rep*. 2015;3(5):675-680.

38. Boonyasrisawat W, Eberle D, Bacci S, et al. Tag polymorphisms at the A20 (TNFAIP3) locus are associated with lower gene expression and increased risk of coronary artery disease in type 2 diabetes. *Diabetes*. 2007;56(2):499-505.
39. Balamurugan K, Sterneck E. The many faces of C/EBPdelta and their relevance for inflammation and cancer. *Int J Biol Sci*. 2013;9(9):917-933.
40. Ko CY, Chang WC, Wang JM. Biological roles of CCAAT/Enhancer-binding protein delta during inflammation. *J Biomed Sci*. 2015;22:6.
41. Litvak V, Ramsey SA, Rust AG, et al. Function of C/EBPdelta in a regulatory circuit that discriminates between transient and persistent TLR4-induced signals. *Nat Immunol*. 2009;10(4):437-443.
42. Moore F, Santin I, Nogueira TC, et al. The transcription factor C/EBP delta has anti-apoptotic and anti-inflammatory roles in pancreatic beta cells. *PLoS One*. 2012;7(2):e31062.
43. Rustenhoven J, Scotter EL, Jansson D, et al. An anti-inflammatory role for C/EBPdelta in human brain pericytes. *Sci Rep*. 2015;5:12132.
44. Lutzenberger M, Burwinkel M, Riemer C, Bode V, Baier M. Ablation of CCAAT/Enhancer-Binding Protein Delta (C/EBPD): Increased Plaque Burden in a Murine Alzheimer's Disease Model. *PLoS One*. 2015;10(7):e0134228.
45. Rao RR, Long JZ, White JP, et al. Meteorin-like is a hormone that regulates immune-adipose interactions to increase beige fat thermogenesis. *Cell*. 2014;157(6):1279-1291.
46. Harms M, Seale P. Brown and beige fat: development, function and therapeutic potential. *Nat Med*. 2013;19(10):1252-1263.
47. Libby P, Ridker PM, Maseri A. Inflammation and atherosclerosis. *Circulation*. 2002;105(9):1135-1143.
48. Willeit P, Thompson SG, Agewall S, et al. Inflammatory markers and extent and progression of early atherosclerosis: Meta-analysis of individual-participant-data from 20 prospective studies of the PROG-IMT collaboration. *Eur J Prev Cardiol*. 2014.
49. Lin H, Joehanes R, Pilling LC, et al. Whole blood gene expression and interleukin-6 levels. *Genomics*. 2014;104(6 Pt B):490-495.

Chapter 5

General discussion

MAIN FINDINGS AND INTERPRETATION

Genetics of hemostatic factors

In **Chapter 2.1** we used the framework of the CHARGE consortium to identify 19 new loci for fibrinogen in a genome-wide association (GWA) study based on 1000 Genomes imputation. At the two most strongly associated loci we detected additional low-frequency (minor allele frequency [MAF] < 5%) and rare (MAF < 1%) variants independently associated with fibrinogen. In **Chapter 2.2** we also compared 1000 Genomes imputation to HapMap imputation in an identical sample, and found that 1000 Genomes imputation led to the discovery of roughly 20% more loci.

In **Chapter 2.3**, we used exome arrays to identify 2 low-frequency and 10 rare variants associated with fibrinogen, factor VII, factor VIII, and VWF that were independent of known associations.[1] In **Chapter 2.4** we used exome sequencing in a smaller sample to identify rare variants associated with fibrinogen, factor VII, factor VIII, and VWF. There was a large overlap between the findings of the exome array and exome sequencing studies, but both studies had unique findings. In the exome sequencing study we identified 3 new rare variants for factor VII and 2 new rare variants for factor VIII that were not discovered in the exome array study. For fibrinogen, there was also an overlap between the GWA study and the two exome studies.

Furthermore, in **Chapter 2.5**, we carried out a GWA study based on Genomes of the Netherlands imputation in the Rotterdam Study.[2] We identified 6 variants at the *ADAMTS13* locus and 1 variant at the *SUPT3H* locus that were independently associated with ADAMTS13 activity. Of the 6 variants at the *ADAMTS13* locus 1 was common, 2 were low-frequency, and 3 were rare variants.

ADAMTS13: association with cardiovascular risk factors

ADAMTS13 has so far primarily been investigated in relation to stroke and CHD. ADAMTS13 acts on VWF, and VWF has been associated with kidney function decline and type 2 diabetes.[3,4] In **Chapter 3.1** we found that VWF-to-ADAMTS13 ratio was related to kidney function decline, an important direct cause of morbidity and mortality, and a strong risk factor for cardiovascular disease. A higher ADAMTS13 activity was protective, as it was associated with a lower decline in kidney function. This finding was consistent with what we know about thrombotic thrombocytopenic purpura, a condition caused by a severe lack of ADAMTS13 that often results in kidney failure.

In contrast, in **Chapter 3.2** we found that ADAMTS13 activity was associated with a higher risk of incident type 2 diabetes. This association persisted despite adjustment for potential confounders, and for fasting glucose and insulin. ADAMTS13 activity was also associated with an increased risk of incident prediabetes. Thus, while ADAMTS13 may decrease the risk of cardiovascular disease through its antithrombotic

effects and its association with chronic kidney disease, it appears to increase the risk of cardiovascular disease through its association with diabetes.

Genetic risk of coronary heart disease

In **Chapter 4.1** we found that a genetic risk score using 152 genetic variants was not able to meaningfully improve risk prediction of incident coronary heart disease (CHD).[5] However, when we performed the analysis for prevalent CHD the improvements in prediction were considerably larger.

In **Chapter 4.2** we investigated the association of SNPs in the seed sequence of microRNAs with cardiovascular risk factors and disease.[6] The seed sequence consists of 5-6 nucleotides in every microRNA that determine to which target genes it can bind. We found that rs2168518, a variant in the seed sequence of miR-4513, was associated with fasting glucose, LDL-cholesterol and total cholesterol, systolic and diastolic blood pressure, and the risk of CHD. We experimentally showed that miR-4513 expression is significantly reduced in the presence of the rs2168518 mutant allele, and we highlighted five target genes that may mediate these associations. Using luciferase reporter assays we validated one of these genes, *GOSR2*, as a target of miR-4513. Additionally, we demonstrated that the microRNA mediated regulation of this gene is changed by rs2168518. This study highlights miR-4513 as a regulator of a range of cardiovascular risk factors and, ultimately, CHD. We were the first to implicate miR-4513 in human disease. In a second study Li *et al* investigated the association of the same variant, rs2168518, with clinical outcomes in CHD.[7] In 1,004 patients with angiographic CHD, they found that miR-4513 was associated with event-free survival and mortality, confirming the importance of this microRNA in cardiovascular disease.

In **Chapter 4.3**, we used a new type of omics, transcriptomics, to identify 3 genes (*TNFAIP3*, *CEBPD*, and *METRNL*) whose gene expression levels in blood were inversely associated with carotid intima media thickness, a measure of subclinical atherosclerosis. All three genes have previously been implicated in inflammation, with *TNFAIP3* and *METRNL* being described in the literature as anti-inflammatory genes, whereas *CEBPD* appears to have both pro and anti-inflammatory properties.[8-10]

METHODOLOGICAL CONSIDERATIONS

Genome-wide association studies

While traditional GWA studies are no longer novel, there are two key factors that ensure that they will keep delivering further results in the future. First, as more individuals are genotyped, the sample sizes available for GWA studies, and therefore

the statistical power, will keep increasing. This will lead to the discovery of further genetic associations that may be biologically informative or collectively useful in prediction.[11] Second, as more individuals are sequenced around the world, and the coverage those individuals are sequenced at increases, reference panels will keep improving. During the writing of this thesis, for example, both the HapMap and 1000 Genomes reference panels were updated,[12,13] and the Genomes of the Netherlands and UK10K reference panels were released.[14,15]

Whereas a significance threshold of 5×10^{-8}, correcting for one million independent tests, ensured a type I error rate of 5% for GWA studies based on HapMap imputation, the same might not be true for GWA studies based on 1000G imputation. As the imputation process is improved, further genetic variants are added to the analysis. Imputed variants are by definition correlated to directly genotyped variants; otherwise, the imputation process could not occur. Yet by combining information from multiple measured variants, an imputed variant can provide information that is independent from any one measured variant. This is also why GWA studies using HapMap imputation are corrected for one million tests even though genotyping arrays usually contain fewer variants than this. Several estimates for the number of tests being done using newer reference panels have been put forward,[16,17] but there is not yet a consensus. Thus, when using imputation based on new reference panels in GWA studies, extra care should be taken to limit the number of false positives. Deciding on a standard threshold for each reference panel is complicated by the large number of reference panels and the speed at which new versions of these reference panels are produced.

The associated variants found in future studies are likely to be either rarer or have smaller effect sizes, since most common variants with moderate to large effects have already been identified. Each of these variants individually will thus contribute less to heritability of the trait. However, the effect size of an associated variant discovered through a GWA study does not necessarily correspond to the importance of the gene underlying the association to the phenotype. Two relevant examples from the literature are *HMGCR* (coding for 3-hydroxy-3-methyl-glutaryl-CoA reductase) and *PCSK9* (coding for proprotein convertase subtilisin/kexin type 9).[18] Variants in both of these genes are associated with low-density lipoprotein (LDL) cholesterol with small effect sizes.[19] However, statins, drugs targeting *HMGCR*, are now the primary form of lipid-lowering medication. *PCSK9* was discovered more recently, but PCSK9 inhibitors have shown great promise in clinical trials as alternative or complementary lipid-lowering agents.[20] In this thesis, *STAT3* was among the new loci discovered in our GWA study of circulating fibrinogen. While the effect size of the most significant variant at the locus was small, this gene is thought to play a central role in regulating gene expression of fibrinogen genes as part of the acute phase response, and

many of the other associated loci appear to interact with it.[21] In the above examples the loci was already known to be related to the phenotype from previous research. There may, however, be other important genes remaining to be discovered with larger samples sizes that have not yet been highlighted using other research.

To identifying new important genes, however, an association from a GWA study is usually not enough. GWA studies do not directly identify genes but instead identify loci spanning hundreds of thousands of base pairs, and sometimes harboring many genes. Definitively identifying the gene underlying the association is usually not possible, and candidate genes are usually selected based on their distance to the lead variant. This approach is pragmatic but has severe limitations. Even if the true causal variant lies within a gene, the mechanism underlying the association may be completely independent of that gene. A high-profile example that recently came to light is the association between variants in the *FTO* gene and obesity. While the variants associated with obesity are located within the *FTO* gene, there is functional evidence that they regulate the expression of a gene called *IRX3*, and not the *FTO* gene itself.[22] Although *IRX3* and *FTO* are separated by over 500 million base pairs, the three dimensional structure of the DNA brings them closer together so that they can interact. While a causal role for FTO is not yet excluded,[23] this example illustrates the difficulty in using the location of associated variants to propose causal genes. In our GWA of fibrinogen, we also used associations with gene expression to provide information on the likely causal gene. For example, although we annotated the signal at 17q21.2 to *RAB5C* based on distance, we also found that the top variant was associated with expression levels of *STAT3* in blood. Even incorporating extra information such as gene expression may not always lead to a single plausible candidate. In some cases the top variant is associated with the expression of more than one gene, or none. Furthermore, blood is not always the relevant tissue to examine, and many databases of other tissues are limited by their small sample sizes.

Exome-wide association studies

The exome-wide association studies we performed, firstly using exon genotyping arrays and secondly using sequencing, also provide methodological insights. These new study designs were largely driven by the hypothesis that rare non-synonymous protein-coding variants are more likely to affect phenotypic variation. Thus, the designs reflect a balance between costs and anticipated benefits at a time when whole-genome sequencing was not yet affordable at a large scale. The major limitation of exome-based analyses is that noncoding regions are excluded, whilst they are also important for the genetic architecture of complex traits.[24] Although non-synonymous protein-coding variants are indeed enriched for associations with phenotypes, so are several other regulatory elements.[25,26] Furthermore, coding regions only comprise

a small percentage of the genome, so that despite their enrichment, most findings from GWA studies are still located in non-coding regions.[27]

As illustrated by the exome-based studies in this thesis, the bulk of the results from exome-based studies are rare variants in genes that were already known to be related to the phenotype. This still serves a purpose: in the case of hemostasis, for example, these rare variants may predispose individuals to bleeding disorders.[28,29] Nevertheless, many of these rare variants may also be identified using standard genotyping arrays and imputation. This is exemplified in our GWA study of circulating fibrinogen, in which we identify, among others, two rare variants with strong effects.

Above all, exome-based analyses in epidemiological studies should be seen as an intermediate step between traditional GWA studies and whole-genome sequencing studies. The scientific community has used these datasets as an opportunity to develop new analytical methods focused on rare variants that are now ready to be applied to whole-genome sequencing.

Genetic risk prediction

Genetic risk prediction studies of CHD, including our own, have been largely disappointing.[30-34] Nevertheless, this does not necessarily mean that genetic risk prediction of CHD will remain unfeasible in the future, as there are several ways how genetic risk prediction could still be improved.

The 152 genetic variants were identified in a large GWA study of CHD including a mix of incident and prevalent cases from cohort studies, case-control studies, and cross-sectional studies.[35] This GWA study may have been affected by a bias known as prevalence-incidence bias or Neyman's bias.[36] For example, in a cross-sectional study, certain factors can affect the chance of individuals with CHD being recruited: individuals with fatal CHD are not included, and individuals with severe CHD are less likely to participate. In such a cross-sectional study, the group of individuals with CHD will be enriched with individuals that suffered from non-fatal and mild CHD. Factors associated with a decreased severity of CHD may thus erroneously be associated with the risk of CHD itself. In a GWA study, this means that variants that reduce the severity of CHD are expected to be present at a higher frequency among cases than controls, and may be picked up as significant results. Additionally, variants associated with severe acute events may be biased towards the null. The susceptibility of different study designs to Neyman's bias is summarized in **Table 1**. In short, many of the study designs used in the GWA study of CHD are susceptible to Neyman's bias, and some of the proposed CHD variants may instead be variants that reduce the severity of CHD.

This could explain why the genetic risk score was more effective in predicting prevalent than incident CHD. If so, the implications for genetic risk prediction ex-

Table 1: Susceptibility of different study designs to Neyman's bias.

Study design	Susceptibility to Neyman's bias
Prospective cohort studies	
Incident cases	Not susceptible to Neyman's bias, because individuals with the disease are included regardless of survival.
Prevalent cases	Highly susceptible to Neyman's bias, because 1) individuals with fatal disease, whether sudden or not, are not included, and 2) individuals with non-fatal disease, especially when severe, are less likely to participate. The degree of Neyman's bias will depend on age-based inclusion criteria: a study of the elederly will be highly susceptible whereas a study of young adults will not.
Case-control studies	
Incidence-density sampling	Incident cases are included in the study as they occur. When nested in a cohort study the exposure and covariates have often been measured before the event occurs. Thus, even sudden fatal cases can be included. If not nested in a cohort study, they may still be susceptible to Neyman's bias for diseases that sometimes present themselves as sudden fatal events.
Cumulative incidence sampling	Prevalent cases available at the time of study initiation are included in the study. See explanation of prevalent cases in Prospective cohort studies.
Cross-sectional studies	Prevalent cases available at the time of study initiation are included in the study. These studies are highly susceptible to Neyman's bias. See explanation of prevalent cases in Prospective cohort studies.

tend beyond CHD to other diseases with a high mortality rate, such as cancer. Basing GWA studies on incident rather than prevalent cases is likely to be most beneficial for diseases involving acute events such as myocardial infarctions and strokes.

Thus, a first way how genetic risk prediction of CHD could be improved is by conducting large-scale GWA studies on incident CHD, rather than prevalent CHD, and using the variants and effect sizes from these studies to construct genetic risk scores. A second way to improve genetic risk prediction is to keep increasing the sample sizes of GWA studies. As sample size of GWA studies increase, the ability of the resulting genetic variants to predict disease will keep improving.[11] Although the new genetic variants will have smaller effect sizes, collectively they may still make a large contribution to the heritability, because as shown in **Figure 1**, variants with smaller effect sizes are much more numerous than variants with large effect sizes. Given that a limited number of studies have already found clinically relevant improvements in prediction using currently identified genetic variants,[37] it seems likely that further developments will lead to genetic risk scores that robustly improve prediction.

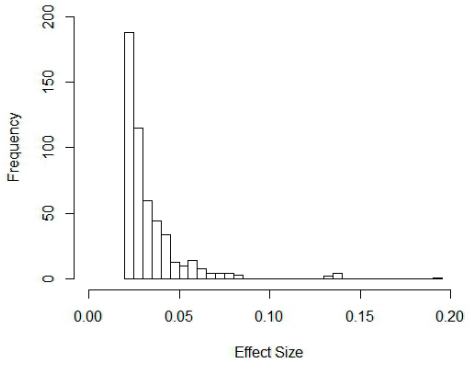


Figure 1: Absolute effect sizes of SNPs in the latest GWA study of height.[38]

Transcriptome-wide association studies

In the past transcriptomics has been applied primarily to small sample sizes. The resulting genes from these studies often did not replicate in independent studies. For example, there was not a single overlapping gene among the results of 3 independent transcriptome-wide association studies of CHD, despite the fact that each study identified more than 20 genes.[39-42] Yet it is unclear whether this heterogeneity is entirely attributable to the small sample sizes of previous studies. In this thesis, some of the findings were characterized by a high degree of heterogeneity. Gene expression levels are highly variable, with large changes occurring over small time spans. This variability may partially explain the heterogeneity and lack of robust, replicating findings. Lastly, confounding and effect modification may be an issue, as gene expression levels are highly dependent on environmental factors such as diet and lifestyle. Furthermore, gene expression levels are tissue specific, and measurements in the Rotterdam Study and other cohort studies are done on whole blood, including a variety of cell types. If the abundance of a specific cell type is associated with the phenotype of interest, then any probe associated with this cell type is likely to be associated with the phenotype through confounding. Although we adjust for counts of a selected number of cell types, this does not address the full range of cell types.

Besides introducing heterogeneity, these issues also make it difficult to interpret the results. Assuming there is a causal relationship between expression levels of a gene and the phenotype, the question remains what the direction of effect is: does the phenotype affect the expression levels or vice versa? In theory this can be addressed using a Mendelian randomization approach: if genetic variants associated with expression levels of the gene of interest are also associated with the phenotype of interest this suggests that gene expression levels influence the phenotype.[43] On the other hand, if genetic variants known to be associated with the phenotype are also associated with gene expression levels this suggests that the phenotype influences gene expression levels. Both directions can be explored, but there are two key limitations: 1) genetic variants may have pleiotropic effects and thereby influence the outcome through a pathway not involving the exposure and 2) the power needed to detect an association is much greater than in a normal association study,

and depends on the strength of the association between the genetic variants and the exposure. Applying Mendelian randomization to any trait thus requires careful consideration. While the approach can suggest causality or a lack thereof, it only rarely provides a definitive conclusion.

FUTURE RESEARCH

Molecular epidemiology

Despite the challenges associated with dynamic data such as transcriptomics, the field of molecular epidemiology is moving towards incorporating more of it. New dynamic omics approaches include microRNA profiling, epigenetics, metabolomics, proteomics, and microbiomics. The main features that these new approaches have in common with GWA studies is the use of large sample sizes, a hypothesis-free approach, and a strict Bonferroni-correct *P*-value threshold to define significant associations. Yet unlike GWA studies they suffer from many of the same issues as transcriptomics. The greatest challenge of the coming years will be to establish a set of guidelines for the conduct of these studies that will ensure that they produce robust, valid, and reliable results.

The other major change in the field will be the move from genotyping arrays and imputation to whole-genome sequencing.[44] While many epidemiological studies are now in the process of sequencing their participants, it is unclear how long it will take before new findings arising from whole-genome sequencing are widespread. The genotyping-imputation approach is estimated to capture 97% of the variation of common variants and 68% of the variation of rare variants.[45] One of the main advantages of whole-genome sequencing is thus likely to be the improved access to rare and population-specific variants, whereas the analysis of common variants will be improved to a smaller extent. The study of rare variants, however, requires large sample sizes that will initially be unavailable. Thus, as long as samples sizes using the genotyping-imputation approach are higher, the benefit of whole-genome sequencing is likely to be limited. For example, in our GWA study of circulating fibrinogen concentration we used 1000 genomes imputation, and we identified some of the same rare variants identified using whole-exome sequencing.

Although better access to rare and population specific variants is one of the objectives of whole-genome sequencing, the largest impact of whole-genome sequencing may be improvements in fine-mapping. In a traditional GWA study, the variant with the lowest *P*-value at a locus is selected as the lead variant and is reported in the results. All other things being equal this may be the optimal approach. However, all other things are often not equal: the imputation quality and sample size differ

among variants, and to make matter worse, some variants are not included at all. With whole-genome sequencing these issues can be avoided. All variants are directly measured and not imputed, so there are no differences in imputation quality. Sample sizes should be more consistent, since variants with poor imputation quality are no longer filtered out. Finally, although some QC filtering will still occur, more variants will be included.

Coming closer to the causal variant does not, by itself, guarantee the identification of the causal gene. However, the functional annotation of the genome is now rapidly evolving, spearheaded by large-scale efforts such as the ENCODE and Roadmap consortia.[24,26] These consortia have identified promotors, enhancers, DNAse hypersensitive regions, among other regulatory elements in a variety of cell types. Together, the identification of the correct causal variant and the availability of accurate functional annotation of the variant will increase the chance of selecting the correct causal gene. These developments may finally allow GWA studies to fully deliver on their aim of uncovering new biology.

Hemostasis and cardiovascular disease

We expect that the developments described above will continue to lead to new discoveries in the genetics of complex traits. For hemostasis factors and cardiovascular disease, these discoveries may help to define the association between the two. The Mendelian randomization approach described above, may in the future provide evidence for a causal relationship between hemostatic factors and cardiovascular disease, or a lack thereof. If there is a causal relationship, using a bi-directional Mendelian randomization approach may clarify the direction of the relationship. So far the use of genetic evidence to identify a causal relationship between hemostatic factors and cardiovascular disease has been only partially successful. Variants associated with VWF, including a variant in the *VWF* gene, are associated with venous thrombosis.[46]

On the contrary, there is evidence for a lack of a causal relationship between fibrinogen concentration and prevalent CHD and stroke. Variants found for fibrinogen concentration are not associated with these diseases.[47] A variant in one of the genes encoding fibrinogen, *FGG*, has been identified to be associated with venous thrombosis in GWA studies of venous thrombosis.[48] Interestingly, this is not one of the variants most associated with fibrinogen concentration, and also at the genome-wide level variants associated with fibrinogen level do not appear to be associated with venous thrombosis.[47] Instead of affecting fibrinogen concentration, the *FGG* variant might affect other aspects like fibrinogen activity or the proportion of different fibrinogen isoforms. Therefore, while fibrinogen concentration does not appear

to be causally related to venous thrombosis in the general population, fibrinogen might be.

Going beyond the hemostatic factors studied in this thesis, genes encoding several other hemostatic factors have been associated with CHD (plasminogen) and venous thrombosis (factor II, factor V, and factor XI).[35,48] Additionally, variants in the *ABO* gene, which are strongly associated with VWF, are associated with CHD and venous thrombosis.[35,48] The *ABO* gene codes for blood group, and thus its association with CHD and venous thrombosis might be explained by mechanisms not involving VWF.

One important limitation of the Mendelian randomization work done so far is the use of prevalent rather than incident CHD, stroke, and venous thrombosis. Genetic variants in hemostatic factors are likely to influence the severity of the thrombotic response to plaque rupture, rather than earlier stages of cardiovascular events. They thereby affect the risk of an event, but also the severity of the event, which can cause Neyman's bias (see **Table 1**). Associations of such variants with prevalent cardiovascular disease may be bias towards the null, and remain hidden. Large-scale Mendelian randomization studies using incident CHD, stroke, and venous thrombosis are thus needed to provide a conclusive answer regarding the causal role of hemostatic factors in cardiovascular disease.

CONCLUSIONS

In this thesis we identified many new genetic associations with hemostatic factors fibrinogen, factor VII, factor VIII, VWF, and ADAMTS13, providing new insight into their etiology. Additionally, we explored the association of ADAMTS13 with cardiovascular risk factors and uncovered a complex scenario where low ADAMTS13 activity is a risk factor for kidney function decline, but a protective factor for type 2 diabetes. We implicated miR-4513 in the etiology of several cardiovascular risk factors and CHD, and found expression levels of three genes to be associated with atherosclerosis.

REFERENCES

1. Huffman JE, de Vries PS, Morrison AC, et al. Rare and low-frequency variants and their association with plasma levels of fibrinogen, FVII, FVIII, and vWF. *Blood*. 2015.
2. de Vries PS, Boender J, Sonneveld MA, et al. Genetic variants in the ADAMTS13 and SUPT3H genes are associated with ADAMTS13 activity. *Blood*. 2015;125(25):3949-3955.

3. Bash LD, Erlinger TP, Coresh J, et al. Inflammation, hemostasis, and the risk of kidney function decline in the Atherosclerosis Risk in Communities (ARIC) Study. *Am J Kidney Dis*. 2009;53(4):596-605.
4. Meigs JB, O'Donnell C J, Tofler GH, et al. Hemostatic markers of endothelial dysfunction and risk of incident type 2 diabetes: the Framingham Offspring Study. *Diabetes*. 2006;55(2):530-537.
5. de Vries PS, Kavousi M, Ligthart S, et al. Incremental predictive value of 152 single nucleotide polymorphisms in the 10-year risk prediction of incident coronary heart disease: the Rotterdam Study. *Int J Epidemiol*. 2015;44(2):682-688.
6. Ghanbari M, de Vries PS, de Looper H, et al. A genetic variant in the seed region of miR-4513 shows pleiotropic effects on lipid and glucose homeostasis, blood pressure, and coronary artery disease. *Hum Mutat*. 2014;35(12):1524-1531.
7. Li Q, Chen L, Chen D, Wu X, Chen M. Influence of microRNA-related polymorphisms on clinical outcomes in coronary artery disease. *Am J Transl Res*. 2015;7(2):393-400.
8. Coornaert B, Carpentier I, Beyaert R. A20: central gatekeeper in inflammation and immunity. *J Biol Chem*. 2009;284(13):8217-8221.
9. Balamurugan K, Sterneck E. The many faces of C/EBPdelta and their relevance for inflammation and cancer. *Int J Biol Sci*. 2013;9(9):917-933.
10. Rao RR, Long JZ, White JP, et al. Meteorin-like is a hormone that regulates immune-adipose interactions to increase beige fat thermogenesis. *Cell*. 2014;157(6):1279-1291.
11. Chatterjee N, Wheeler B, Sampson J, et al. Projecting the performance of risk prediction based on polygenic analyses of genome-wide association studies. *Nat Genet*. 2013;45(4):400-405, 405e401-403.
12. International HapMap Consortium, Altshuler DM, Gibbs RA, et al. Integrating common and rare genetic variation in diverse human populations. *Nature*. 2010;467(7311):52-58.
13. 1000 Genomes Project Consortium, Abecasis GR, Auton A, et al. An integrated map of genetic variation from 1,092 human genomes. *Nature*. 2012;491(7422):56-65.
14. Genome of the Netherlands Consortium. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat Genet*. 2014;46(8):818-825.
15. UK10K Consortium. The UK10K project identifies rare variants in health and disease. *Nature*. 2015.
16. Huang J, Ellinghaus D, Franke A, Howie B, Li Y. 1000 Genomes-based imputation identifies novel and refined associations for the Wellcome Trust Case Control Consortium phase 1 Data. *Eur J Hum Genet*. 2012;20(7):801-805.
17. Li MX, Yeung JM, Cherny SS, Sham PC. Evaluating the effective numbers of independent tests and significant p-value thresholds in commercial genotyping arrays and public imputation reference datasets. *Hum Genet*. 2012;131(5):747-756.
18. Fox CS, Hall JL, Arnett DK, et al. Future translational applications from the contemporary genomics era: a scientific statement from the American Heart Association. *Circulation*. 2015;131(19):1715-1736.
19. Global Lipids Genetics Consortium, Willer CJ, Schmidt EM, et al. Discovery and refinement of loci associated with lipid levels. *Nat Genet*. 2013;45(11):1274-1283.
20. Roth EM, McKenney JM, Hanotin C, Asset G, Stein EA. Atorvastatin with or without an antibody to PCSK9 in primary hypercholesterolemia. *N Engl J Med*. 2012;367(20):1891-1900.
21. Fish RJ, Neerman-Arbez M. Fibrinogen gene regulation. *Thromb Haemost*. 2012;108(3):419-426.

22. Smemo S, Tena JJ, Kim KH, et al. Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nature*. 2014;507(7492):371-375.
23. Cedernaes J, Benedict C. Human obesity: FTO, IRX3, or both? *Mol Metab*. 2014;3(5):505-506.
24. Encode Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012;489(7414):57-74.
25. Schaub MA, Boyle AP, Kundaje A, Batzoglou S, Snyder M. Linking disease associations with regulatory information in the human genome. *Genome Res*. 2012;22(9):1748-1759.
26. Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, et al. Integrative analysis of 111 reference human epigenomes. *Nature*. 2015;518(7539):317-330.
27. Edwards SL, Beesley J, French JD, Dunning AM. Beyond GWASs: illuminating the dark road from association to function. *Am J Hum Genet*. 2013;93(5):779-797.
28. Ivaskevicius V, Jusciute E, Steffens M, et al. gammaAla82Gly represents a common fibrinogen gamma-chain variant in Caucasians. *Blood Coagul Fibrinolysis*. 2005;16(3):205-208.
29. Chaing S, Clarke B, Sridhara S, et al. Severe factor VII deficiency caused by mutations abolishing the cleavage site for activation and altering binding to tissue factor. *Blood*. 1994;83(12):3524-3535.
30. Brautbar A, Pompeii LA, Dehghan A, et al. A genetic risk score based on direct associations with coronary heart disease improves coronary heart disease risk prediction in the Atherosclerosis Risk in Communities (ARIC), but not in the Rotterdam and Framingham Offspring, Studies. *Atherosclerosis*. 2012;223(2):421-426.
31. Hughes MF, Saarela O, Stritzke J, et al. Genetic markers enhance coronary risk prediction in men: the MORGAM prospective cohorts. *PLoS One*. 2012;7(7):e40922.
32. Paynter NP, Chasman DI, Pare G, et al. Association between a literature-based genetic risk score and cardiovascular events in women. *JAMA*. 2010;303(7):631-637.
33. Thanassoulis G, Peloso GM, Pencina MJ, et al. A genetic risk score is associated with incident cardiovascular disease and coronary artery calcium: the Framingham Heart Study. *Circ Cardiovasc Genet*. 2012;5(1):113-121.
34. Ganna A, Magnusson PK, Pedersen NL, et al. Multilocus genetic risk scores for coronary heart disease prediction. *Arterioscler Thromb Vasc Biol*. 2013;33(9):2267-2272.
35. CARDIoGRAMplusC4D Consortium, Deloukas P, Kanoni S, et al. Large-scale association analysis identifies new risk loci for coronary artery disease. *Nat Genet*. 2013;45(1):25-33.
36. Hill G, Connelly J, Hebert R, Lindsay J, Millar W. Neyman's bias re-visited. *J Clin Epidemiol*. 2003;56(4):293-296.
37. Mega JL, Stitziel NO, Smith JG, et al. Genetic risk, coronary heart disease events, and the clinical benefit of statin therapy: an analysis of primary and secondary prevention trials. *Lancet*. 2015;385(9984):2264-2271.
38. Wood AR, Esko T, Yang J, et al. Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat Genet*. 2014;46(11):1173-1186.
39. Huan T, Zhang B, Wang Z, et al. A systems biology framework identifies molecular underpinnings of coronary heart disease. *Arterioscler Thromb Vasc Biol*. 2013;33(6):1427-1434.
40. Rosenberg S, Elashoff MR, Beineke P, et al. Multicenter validation of the diagnostic accuracy of a blood-based gene expression test for assessing obstructive coronary artery disease in nondiabetic patients. *Ann Intern Med*. 2010;153(7):425-434.
41. Sinnaeve PR, Donahue MP, Grass P, et al. Gene expression patterns in peripheral blood correlate with the extent of coronary artery disease. *PLoS One*. 2009;4(9):e7037.

42. Wingrove JA, Daniels SE, Sehnert AJ, et al. Correlation of peripheral-blood gene expression with the extent of coronary artery stenosis. *Circ Cardiovasc Genet*. 2008;1(1):31-38.
43. Smith GD. Mendelian Randomization for Strengthening Causal Inference in Observational Studies: Application to Gene x Environment Interactions. *Perspect Psychol Sci*. 2010;5(5):527-545.
44. Wang Q, Lu Q, Zhao H. A review of study designs and statistical methods for genomic epidemiology studies using next generation sequencing. *Front Genet*. 2015;6:149.
45. Yang J, Bakshi A, Zhu Z, et al. Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat Genet*. 2015.
46. Smith NL, Rice KM, Bovill EG, et al. Genetic variation associated with plasma von Willebrand factor levels and the risk of incident venous thrombosis. *Blood*. 2011;117(22):6007-6011.
47. Sabater-Lleal M, Huang J, Chasman D, et al. Multiethnic meta-analysis of genome-wide association studies in >100 000 subjects identifies 23 fibrinogen-associated Loci but no strong evidence of a causal association between circulating fibrinogen and cardiovascular disease. *Circulation*. 2013;128(12):1310-1324.
48. Germain M, Chasman DI, de Haan H, et al. Meta-analysis of 65,734 individuals identifies TSPAN15 and SLC44A2 as two susceptibility loci for venous thromboembolism. *Am J Hum Genet*. 2015;96(4):532-542.

Chapter 6

Summary & Samenvatting

ENGLISH SUMMARY

Hemostasis, the processes causing bleeding to stop, and thrombosis, the formation of blood clots, are essential processes in the development of coronary heart disease (CHD). Many proteins are involved in hemostasis and thrombosis, and understanding their biology and genetic background could lead to insights relevant to cardiovascular disease. In this thesis we explored five of these proteins, and also studied other genetic influences on atherosclerosis and CHD.

Chapter 2 focuses on genetic association studies of proteins involved in hemostasis: fibrinogen, factor VII, factor VIII, von Willebrand factor (VWF), and ADAMTS13. In traditional genetic association studies, millions of variants are tested for association with a trait of interest. However, only a few hundred thousand variants are directly measured: the remaining variants are estimated, or imputed, using a reference panel that provides information about the correlation structure between the variants. The first widely used reference panel was the HapMap project, which provided information on around 2.5 million genetic variants. Recently, new reference panels such as the 1000 genomes project (1000G) have been released that are expected to improve the imputation process. In **Chapter 2.1** we performed a genome-wide association study, based on 1000G imputation, of circulating fibrinogen concentration in over 120,000 individuals. We identify 18 new loci for fibrinogen, and at the two most strongly associated loci we detected additional low-frequency variants independently associated with fibrinogen.

The use of 1000G imputation as opposed to HapMap imputation was not the only difference between our study and previous studies: our study was also larger. Therefore, to be able to adequately examine the benefit of using 1000G imputation over HapMap imputation, in **Chapter 2.2** we performed a comparison of these two methods in exactly the same individuals, using circulating fibrinogen concentration as a quantitative example trait. We found that all other things remaining the same, using 1000G imputation lead to the discovery of 20% more loci. On the other hand, one locus that was found using HapMap imputation was not found using 1000G imputation.

We then further examined the genetics of fibrinogen, but also factor VII, factor VIII, and VWF, using study designs especially suited for the identification of rare variants. In **Chapter 2.3** we performed an exome-wide study using genotypes obtained from the Ilumina Exome Chip. We identified two low-frequency and ten rare variants associated with fibrinogen, factor VII, factor VIII, and VWF that were independent of known associations. In **Chapter 2.4** we performed a similar study using exome sequencing data. We identified three new rare variants for factor VII and two new rare variants for factor VIII that were not discovered in the exome array study. For

fibrinogen, there was also an overlap between the genome-wide association study and the two exome studies.

In **Chapter 2.5** we combined the genome-wide association study and exome chip approaches to study both common and rare genetic variants associated with ADAMTS13 activity. Using the genome-wide association study approach we identified two variants at the *ADAMTS13* locus and one variant at the *SUPT3H* locus that were independently associated with ADAMTS13 activity. Using the exome chip approach, we identified a further three rare variants that were independently associated with ADAMTS13 activity.

ADAMTS13 has so far primarily been investigated in relation to stroke and CHD. In **Chapter 3** we further characterized ADAMTS13 by examining its association with cardiovascular risk factors. In **Chapter 3.1** we explored the association of ADAMTS13 activity with kidney function decline. We found that VWF-to-ADAMTS13 ratio was related to kidney function decline, an important direct cause of morbidity and mortality, and a strong risk factor for cardiovascular disease. A higher ADAMTS13 activity was protective, as it was associated with a lower decline in kidney function. This finding was consistent with what we know about thrombotic thrombocytopenic purpura, a condition caused by a severe lack of ADAMTS13 that often results in kidney failure.

In **Chapter 3.2** we examined the association of ADAMTS13 activity with incident type 2 diabetes. In contrast to our findings with kidney function decline, we found that high ADAMTS13 activity was associated with a higher risk of incident type 2 diabetes. This association persisted despite adjustment for potential confounders, and for fasting glucose and insulin. High ADAMTS13 activity was also associated with an increased risk of incident prediabetes. Thus, while ADAMTS13 activity may decrease the risk of cardiovascular disease through its antithrombotic effects and its association with chronic kidney disease, it appears to increase the risk of cardiovascular disease through its association with diabetes.

In **Chapter 4** we investigated CHD and the underlying atherosclerosis directly. In **Chapter 4.1** we evaluate the incremental predictive value of genetic risk scores in the risk prediction of incident coronary heart disease. We found that a genetic risk score using 152 genetic variants was not able to meaningfully improve risk prediction of incident CHD. However, when we performed the analysis for prevalent CHD the improvements in prediction were considerably larger. We theorized that this discrepancy may be caused by the use of genetic variants discovered for prevalent rather than incident CHD.

In **Chapter 4.2** we investigated the association of SNPs in the seed sequence of microRNAs with cardiovascular risk factors and disease. The seed sequence consists of 5-6 nucleotides in every microRNA that determine to which target genes the

microRNA can bind. We found that rs2168518, a variant in the seed sequence of miR-4513, was associated with fasting glucose, LDL-cholesterol and total cholesterol, systolic and diastolic blood pressure and the risk of CHD. The direction of the effects was consistent across the different phenotypes, with the mutant allele of rs2168518 leading to an unfavorable cardio-metabolic profile. We experimentally showed that miR-4513 expression is significantly reduced in the presence of the rs2168518 mutant allele, and we highlighted five target genes that may mediate the association between miR-4513 and these cardio-metabolic phenotypes. We validated one of these genes, *GOSR2*, as a target of miR-4513, and demonstrated that the regulation of GOSR2 by miR-4513 varies according to rs2168518.

In a transcriptome-wide association expression levels of genes across the genome are associated with a trait of interest. In **Chapter 4.3** we performed a transcriptome-wide association study of carotid intima media thickness, a measure of atherosclerosis. We identified 3 genes (*TNFAIP3*, *CEBPD*, and *METRNL*) with gene expression levels in blood that were associated with carotid intima media thickness. All of these genes were inversely associated with carotid intima media thickness: high expression levels were associated with less atherosclerosis. *TNFAIP3* and *METRNL* have been described in the literature as anti-inflammatory genes, and *CEBPD* has been described as both pro and anti-inflammatory.

Finally, **Chapter 5** contains an overview of the main findings of this thesis as well as their implications, discusses methodological issues, and explores future directions in molecular epidemiology in general, and in the molecular epidemiology of CHD and hemostasis in particular.

NEDERLANDSE SAMENVATTING

Hemostase, het stoppen van bloeden, en trombose, de formatie van bloedproppen, zijn essentiële processen in de ontwikkeling van hart- en vaat ziekten zoals coronaire hartziekten. De breuk van atherosclerotische plaques leidt immers tot een hartaanval door het uitlokken van bloedstolling: het zijn de bloedproppen in de slagaders van het hart die de bloedtoevoer naar het hart blokkeren. Vele eiwitten spelen een rol in hemostase en trombose. Door de biologie en genetische achtergrond van deze eiwitten beter te begrijpen, kunnen we meer te weten komen over het ontstaan van hart- en vaat ziekten. In deze thesis hebben we vijf hemostase-eiwitten alsook genetische risicofactoren van atherosclerose en coronaire hartziekten bestudeerd.

Hoofdstuk 2 bestaat uit genetische associatie studies van de hemostase-eiwitten fibrinogeen, factor VII, factor VIII, von Willebrand factor (VWF), en ADAMTS13. Genetische associatie studies testen de associatie tussen miljoenen genetische varianten en een fenotype. Echter, enkel een paar honderdduizend van deze varianten zijn direct gemeten: de rest van de varianten wordt geïmputeerd met behulp van een referentie populatie. Op basis van deze referentie populatie kan men de correlatie tussen de genetische varianten schatten. Het HapMap project was de eerste referentie populatie die het mogelijk maakte om de correlatie tussen genetische varianten te schatten en niet-direct gemeten varianten te imputeren. Sinds kort zijn er nieuwe referentie populaties beschikbaar die het imputatieproces naar verwachting verbeteren. Het "1000 genomes project" (1000G) is zo een nieuwe referentie populatie. In **Hoofdstuk 2.1** hebben we een genoomwijde associatiestudie van fibrinogeen uitgevoerd in meer dan 120.000 mensen, gebaseerd op 1000G imputatie. Met gebruik van deze nieuwe referentie populatie vonden we 18 nieuwe genetische loci voor fibrinogeen. Bovendien vonden we dat in de twee sterkste loci voor fibrinogeen meerdere genetische varianten, waaronder zeldzame varianten, onafhankelijk van elkaar geassocieerd waren met fibrinogeen.

Het gebruik van 1000G imputatie was niet het enige verschil tussen onze studie naar genetische factoren voor fibrinogeen levels en voorgaande studies: onze studie was ook groter in vergelijking met eerdere studies. Om het voordeel van het gebruik van 1000G imputatie ten opzichte van HapMap imputatie nader te bekijken, hebben we in **Hoofdstuk 2.2** beide methoden vergeleken in exact dezelfde mensen. We vonden dat 1000G imputatie 20% meer loci identificeert in vergelijking met HapMap imputatie, aannemende dat alle andere factoren hetzelfde blijven. Echter, een locus die we in de HapMap studie vonden, was niet significant geassocieerd in de 1000G geïmputeerde studie.

Vervolgens hebben we de genetica van fibrinogeen, alsook die van factor VII, factor VIII en VWF, bestudeerd met gebruik van een speciaal ontworpen studie

methode voor de identificatie van zeldzame genetische varianten. In **Hoofdstuk 2.3** beschrijven we een exoomwijde studie uitgevoerd met gebruik van de Illumina Exome Chip. We vonden twee varianten met een lage allel frequentie en tien zeldzame varianten die geassocieerd waren met fibrinogeen, factor VII, factor VIII en VWF, onafhankelijk van gekende associaties. In **Hoofdstuk 2.4** hebben we een zelfde soort studie verricht met gebruik van exoom sequencing data. In deze studie vonden we drie nieuwe varianten voor factor VII en twee nieuwe varianten voor factor VIII. Deze varianten waren nieuw ten opzichte van de exome array studie. We vonden een overlap tussen de genoomwijde associatie studie gebaseerd op 1000G imputatie en de twee exoom studies van fibrinogeen.

In **Hoofdstuk 2.5** hebben we de genoomwijde associatie studie methode samen met de exoom chip methode gebruikt om frequente en zeldzame genetische varianten voor ADAMTS13 activiteit te vinden. De genoomwijde associatie studie methode vond twee varianten op de *ADAMTS13* locus en een variant in de *SUPT3H* locus die onafhankelijk van elkaar geassocieerd waren met ADAMTS13 activiteit. De exoom chip methode resulteerde in drie extra varianten voor ADAMTS13 activiteit.

In het verleden hebben onderzoekers vooral de associatie tussen ADAMTS13 en cardiovasculaire ziekte bestudeerd. In **Hoofdstuk 3** hebben we de rol van ADAMTS13 met betrekking tot cardiovasculaire risicofactoren onderzocht. We hebben de associatie tussen ADAMTS13 activiteit en nierfunctie achteruitgang in **Hoofdstuk 3.1** beschreven. We vonden dat de VWF/ADAMTS13 ratio geassocieerd was met nierfunctie achteruitgang, een sterke risicofactor voor cardiovasculair ziekte en een belangrijke directe oorzaak van morbiditeit en mortaliteit. Een hogere ADAMTS13 activiteit was beschermend gezien het geassocieerd was met een tragere nierfunctie achteruitgang. Deze bevinding is in lijn met wat we weten van trombotische trombocytopenische purpura, een ziekte veroorzaakt door een laag ADAMTS13 die zich vaak presenteert met nierfalen.

In **Hoofdstuk 3.2** hebben we de associatie tussen ADAMTS13 activiteit en de incidentie van type 2 diabetes bestudeerd. In tegenstelling tot de bevindingen met nierfunctie achteruitgang vonden we dat hoger ADAMTS13 geassocieerd was met een hoger risico op type 2 diabetes. De associatie veranderde nauwelijks na adjusteren voor mogelijke confounders en vastende glucose en insuline waarden. We concluderen dat ADAMTS13 activiteit het risico op cardiovasculaire ziekte kan verlagen door een mogelijk antitrombotisch en dus protectief effect op nierinsufficientie. Tevens kan ADAMTS13 activiteit het risico op cardiovasculaire ziekte verhogen door de associatie met diabetes.

In **Hoofdstuk 4** hebben we cardiovasculaire ziekte zelf bestudeerd, alsook de onderliggende atherosclerose. In **Hoofdstuk 4.1** hebben we de toegevoegde waarde van een genetische risico score voor het voorspellen van een toekomstig hartinfarct

onderzocht. We vonden geen noemenswaardige verbetering in het voorspellen van een toekomstig hartinfarct met een genetische risico score opgebouwd uit 152 genetische varianten. Desalniettemin, de predictie voor prevalente coronaire hartziekte verbeterde wel substantieel. De discrepantie tussen prevalente en incidente predictie van coronaire hartziekte kan het gevolg zijn van het gebruik van genetische varianten gevonden in studies voor prevalente hartziekte, en niet incidente hartziekte.

In **Hoofdstuk 4.2** hebben we de associatie tussen SNPs in de zogenoemde "seed" sequentie van microRNAs en cardiovasculaire ziekte en zijn risicofactoren onderzocht. Deze "seed" sequentie bestaat uit vijf tot zes nucleotiden en bepaalt aan welke genen het microRNA kan binden. We vonden dat rs2168518, een variant in de sequentie van miR-4513, geassocieerd was met vastende glucose waarden, LDL-cholesterol en totaal cholesterol, alsook systolische en diastolische bloeddruk en het risico op coronaire hartziekte. De richting van het effect was overeenkomstig met de andere fenotypes: het zeldzame allel van rs2168518 was geassocieerd met een slechter cardio-metabool profiel. Middels experimenteel onderzoek toonden we aan dat het zeldzame allel van rs2168518 de expressie van miR-4513 significant verminderde. Daarnaast konden we vijf genen aanwijzen die de associatie tussen miR-4513 en deze cardio-metabole fenotypes zouden kunnen mediëren. We valideerden een van deze genen (*GOSR2*), en we toonden aan dat de regulerende werking van miR-4513 op *GOSR2* varieerde op basis van het genotype van rs2168518.

In transcriptoomwijde associatie studies associeert men expressie levels van alle genen in het genoom met een phenotype. In **Hoofdstuk 4.3** hebben we en transcriptoomwijde associatie studie uitgevoerd op carotis intima media dikte, een maat van atherosclerose. De expressie van drie genen (*TNFAIP3*, *CEBPD* en *METRNL*) was geassocieerd met carotis intima media dikte. Deze drie genen waren alle negatief geassocieerd met carotis intima media dikte: hogere expressie was geassocieerd met minder atherosclerose. *TNAIF3* en *METRNL* zijn beschreven als anti-inflammatoire genen in de literatuur, daar waar *CEBPD* zowel pro- als anti-inflammatoire effecten kan hebben.

Tenslotte bespreken we in **Hoofdstuk 5** de hoofdbevindingen van deze thesis alsook de implicaties en methodologische aspecten. Tevens bespreken we de toekomstige mogelijkheden in de moleculaire epidemiologie, meer specifiek de moleculaire epidemiologie van coronaire hartziekte en hemostase.

Chapter 7

Appendices

ACKNOWLEDGEMENTS

Working on this thesis for the past three years has been an amazing opportunity for me, giving me the chance to have new experiences, meet new people, and visit new places. I am glad to have worked in one of the most collaborative fields in science: genetic epidemiology. Projects usually involve dozens to hundreds of collaborators from almost as many institutions, and communication proceeds by email, teleconference, and the occasional face-to-face meeting. There are therefore many people to thank.

First of all I would like to thank my co-promotor Abbas and my promotor Oscar. I want to thank you both for taking a chance on me. I am very grateful to you for giving me the opportunity to explore genetic epidemiology, and for providing the kind of environment in which I could grow, as well as your trust and support.

These three years have been most affected by my roommates from room 2901: I hope that the 2901 club remains a club with life-long membership and regular meetings. Symen, thank you so much for agreeing to be my paranimph and then going above and beyond your responsibilities. Having you to share this journey with, which we started together in August of 2012, has been a blessing. Dear Sanaz. Thanks to you I did not even notice that genetic epidemiology has quite a steep learning curve: without your guidance and support back then, I am not sure how I would have gotten the ball rolling. It was great to be on the same team as you, even if I occasionally got into trouble for stealing your spot at the lunch table. Layal, the discussions with you have been very interesting and enlightening, instilling in me a dose of healthy skepticism when it comes to global affairs. Please send me your thesis when it is ready: that is, if it is not too heavy for the airplane. Mohsen, you have been a calming presence during these years. It has been a lot of fun learning about your culture, and becoming jealous of your "Mohsen lunches". Jana, it was a pleasure to supervise you during your Master project. I hope that you learnt as much from being supervised by me as I learnt from supervising you.

I would like to thank my old roommates from the Valkenburg library (Anna, Trudy, Lisan, Myrte, Adriana, Cristina, Charlotte, Lisanne, Klodian, Ayesha, and Thirsa), who made a very unpleasant room a very pleasant place. Anna, it has been great to see you find so much happiness with Gerard and Maxi. I would also like to thank the members of the cardiovascular group, and our sister group ErasmusAge. I would especially like to thank Maryam and Maarten, who answered many questions of mine in the early days, as well as Thijs for the pleasant and even fun collaboration. I also thank Ingrid and Mirjam, who have been of great help with all administrative procedures.

An interesting part of these years has been collaborating with other groups of the Erasmus MC. I am very grateful to Carolina, Lennart, and Najaf, who took the time to mentor me as I was starting out. Now that I have more experience I have really come to appreciate the time you took to answer all of my questions. From the department of Hematology I would like to thank Frank, Moniek, Michelle, and Johan. I have really enjoyed collaborating with you: it has gone very smoothly and has resulted in some very nice output. I also want to thank Janine for the interesting discussions and for organizing the MolEpi meetings. From the genetics lab of the department of Internal Medicine I would also like to thank André, Fernando, Jeroen, Marijn, Marjolein and Joyce for their contribution to the project in this thesis and for the good times at the CHARGE meetings. Fernando: thank you for the advice you gave when I had some difficult career choices to make, and for the nice chats about running. From the genetic epidemiology group I would also like to thank Cornelia, Ayse, Maarten, and Sven. Ayse: your guidance on the lipidomics project has been enjoyable, and I have learnt a lot from it. From the Neurology group I would like to thank Arfan: thanks for your input and helpful comments on the ADAMTS13 projects.

Besides collaborations within the Erasmus MC, collaborating with external teams from all over the world has been a true highlight. Most of this has revolved around the CHARGE consortium. Without the infrastructure of the CHARGE consortium much of the research in this thesis would have been inconceivable. There are a few individual people within the CHARGE consortium that I would like to thank in particular: Nick and Chris for their leadership in the Hemostasis working group, David for his role as a mentor in the fibrinogen projects, and Jenny, Nathan, Alanna, and Maria for their essential roles in various chapters of thesis. Alanna: thank you for offering me the opportunity to come and work with you and Eric. I very much look forward to it! Although not featured in this thesis, collaborations within the CARDIoGRAMplusC4D, COMBI-BIO, ENGAGE, EUROSPAN consortia have been very pleasant and educational for me. I would like to thank the COMBI-BIO consortium for giving me the chance to undertake a research visit to Imperial College, where working with Ioanna, Raphaele, and Ali was a great way to experience a new work culture. I also thank Professor Harold Snieder from the University Medical Center Groningen and Professor Eline Slagboom from the Leiden University Medical Center for agreeing to be part of my doctoral committee.

I highly appreciate the effort put in by the investigators and staff of the Rotterdam Study who have put in place an infrastructure that allows PhD students to be very productive. I would especially like to thank Professor Hofman for his role in the creation and management of the Rotterdam Study. I am very grateful for the computer and data management support from Nano and Frank.

Naturally, there are also people in my personal sphere who I would like to take this chance to thank. I would like to acknowledge Ivo for being my paranimph, and Niels for agreeing to help out during the defense. But above all I want to thank you both for being such dependable and awesome friends!

I did not graduate high school with a strong idea of what to do next, and it was my parents who encouraged me to simply study what I found interesting rather than worry too much about career paths. Combined with a bit of serendipity this suggestion has helped me find my way to genetic epidemiology. I am incredible thankful for this, and for your continued support: these years have been immeasurably enriched by your presence. I would also like to thank my grandparents. For the interesting discussions about science, and for being an inspiring example: we will do our best to match your 60+ years of marriage. I am also thankful for the support and encouragement offered by my sister and other family members.

Adjusting to life in Houston without my parents, grandparents, sister and the rest of my family will no doubt be a challenge. Luckily, Lised, I will have you to share that next adventure with, in the same way that you have been there for me throughout these past years. It is hard to overstate how valuable it has been to have you to come home to every night, allowing me to forget about any stress and frustrations I might have had. Your unwavering belief in me manages to rub off on me and give me confidence I wouldn't otherwise have. Of course you have also been on your own journey during this time, and it has been a privilege to watch you adapt to life in the Netherlands.

Finally, I would like to thank the participants of the Rotterdam Study, and all other studies that were used in this thesis. Without your input and dedication it would be impossible for us epidemiologists to accomplish anything at all.

PHD PORTFOLIO

Name of PhD student	Paul Stefan de Vries
Erasmus MC department	Epidemiology
PhD period	August 2012 – January 2016
Promotor	Prof.dr. Oscar H. Franco
Copromotor	Dr. Abbas Dehghan

Training

Courses and workshops	*Year*	*ECTS*
Causal inference, Erasmus MC	2012	0.7
Principles of genetic epidemiology, Erasmus MC	2012	0.7
Genomics in molecular medicine, Erasmus MC	2012	1.4
Master class: advances in genomics research, Erasmus MC	2012	0.4
Genome wide association analysis, Erasmus MC	2012	1.4
Basic course on R, Erasmus MC	2012	1.4
Linux for scientists, Erasmus MC	2012	0.6
SNPs and human diseases, Erasmus MC	2012	1.4
Advances in GWAS, Erasmus MC	2013	1.4
Biomedical English writing, Erasmus MC	2013	1.4
First encounter with NGS data, Erasmus MC	2013	1.4
Metabonomics short course, Imperial College, UK	2013	1.4
Advanced medical writing, Erasmus MC	2014	0.7
Attended conferences		
CHARGE analysis workshop, Boston, MA, USA	2012	0.5
ENGAGE meeting: from genetic discovery to future health, Rotterdam	2012	0.2
CHARGE investigator meeting, Rotterdam	2013	0.5
COMBI-BIO meeting, London, UK	2013	0.5
European Society of Cardiology Congress, Amsterdam	2013	0.5
COMBI-BIO meeting, London, UK	2013	0.5
CHARGE investigator meeting, Los Angeles, CA, USA	2013	1
COMBI-BIO meeting, Rotterdam	2013	0.5
COMBI-BIO meeting, London, UK	2014	0.5
CHARGE investigator meeting, Washington, DC, USA	2015	0.5
Attended seminars		
Seminars of the department of epidemiology	2012-2015	2
2020 meetings	2012-2015	2
Cardiogenetics meetings	2012-2013	1
Nutritional epidemiology meetings (SIGN-E)	2012-2013	0.5

Cardiovascular group meetings	2012-2015	2
MolEpi meetings	2013-2015	1
Genetic epidemiology unit	2012-2013	0.5
Genetics lab	2012- 2013	0.5
Teaching		
Exome chip analysis workshop (organizer, lecturer, and supervisor of practical)	2013	1
Study design (teaching assistant)	2013-2014	1
Methodological topics of study design (teaching assistant)	2013-2014	1
MSc thesis of Jana Nano (supervisor)	2014	2
Other		
Peer review of articles for scientific journals	2014	1
Research visit to Imperial college	2014	3

LIST OF PUBLICATIONS

A meta-analysis of 120,246 individuals identifies 18 new loci for fibrinogen concentration (2015). **de Vries PS**, Chasman DI, Sabater-Lleal M, Chen MH, Huffman JE, Steri M, Tang W, Teumer A, Marioni RE, Grossmann V, Hottenga JJ, Trompet S, Müller-Nurasyid M, Zhao JH, Brody JA, Kleber ME, Guo X, Wang JJ, Auer PL, Attia JR, Yanek LR, Ahluwalia TS, Lahti J, Venturini C, Tanaka T, Bielak LF, Joshi PK, Rocanin-Arjo A, Kolcic I, Navarro P, Rose LM, Oldmeadow C, Riess H, Mazur J, Basu S, Goel A, Yang Q, Ghanbari M, Willemsen G, Rumley A, Fiorillo E, de Craen AJ, Grotevendt A, Scott R, Taylor KD, Delgado GE, Yao J, Kifley A, Kooperberg C, Qayyum R, Lopez LM, Berentzen TL, Räikkönen K, Mangino M, Bandinelli S, Peyser PA, Wild S, Trégouët DA, Wright AF, Marten J, Zemunik T, Morrison AC, Sennblad B, Tofler G, de Maat MP, de Geus EJ, Lowe GD, Zoledziewska M, Sattar N, Binder H, Völker U, Waldenberger M, Khaw KT, McKnight B, Huang J, Jenny NS, Holliday EG, Qi L, McEvoy MG, Becker DM, Starr JM, Sarin AP, Hysi PG, Hernandez DG, Jhun MA, Campbell H, Hamsten A, Rivadeneira F, McArdle WL, Slagboom PE, Zeller T, Koenig W, Psaty BM, Haritunians T, Liu J, Palotie A, Uitterlinden AG, Stott DJ, Hofman A, Franco OH, Polasek O, Rudan I, Morange PE, Wilson JF, Kardia SL, Ferrucci L, Spector TD, Eriksson JG, Hansen T, Deary IJ, Becker LC, Scott RJ, Mitchell P, März W, Wareham NJ, Peters A, Greinacher A, Wild PS, Jukema JW, Boomsma DI, Hayward C, Cucca F, Tracy R, Watkins H, Reiner AP, Folsom AR, Ridker PM, O'Donnell CJ, Smith NL, Strachan DP, Dehghan A. *Human Molecular Genetics*.

A comprehensive 1000 genomes-based GWAS meta-analysis of coronary artery disease (2015). CARDIoGRAMplusC4D Consortium, Nikpay M, Goel A, Won HH, Hall LM, Willenborg C, Kanoni S, Saleheen D, Kyriakou T, Nelson CP, Hopewell JC, Webb TR, Zeng L, Dehghan A, Alver M, Armasu SM, Auro K, Bjonnes A, Chasman DI, Chen S, Ford I, Franceschini N, Gieger C, Grace C, Gustafsson S, Huang J, Hwang SJ, Kim YK, Kleber ME, Lau KW, Lu X, Lu Y, Lyytikäinen LP, Mihailov E, Morrison AC, Pervjakova N, Qu L, Rose LM, Salfati E, Saxena R, Scholz M, Smith AV, Tikkanen E, Uitterlinden A, Yang X, Zhang W, Zhao W, de Andrade M, **de Vries PS**, van Zuydam NR, Anand SS, Bertram L, Beutner F, Dedoussis G, Frossard P, Gauguier D, Goodall AH, Gottesman O, Haber M, Han BG, Huang J, Jalilzadeh S, Kessler T, König IR, Lannfelt L, Lieb W, Lind L, Lindgren CM, Lokki ML, Magnusson PK, Mallick NH, Mehra N, Meitinger T, Memon FU, Morris AP, Nieminen MS, Pedersen NL, Peters A, Rallidis LS, Rasheed A, Samuel M, Shah SH, Sinisalo J, Stirrups KE, Trompet S, Wang L, Zaman KS, Ardissino D, Boerwinkle E, Borecki IB, Bottinger EP, Buring JE, Chambers JC, Collins R, Cupples LA, Danesh J, Demuth I, Elosua R, Epstein SE, Esko T, Feitosa MF, Franco OH, Franzosi MG, Granger CB, Gu D, Gudnason V, Hall AS, Hamsten A, Harris TB, Hazen SL, Hengstenberg C, Hofman A, Ingelsson E, Iribarren C, Jukema JW, Karhunen PJ, Kim BJ, Kooner

JS, Kullo IJ, Lehtimäki T, Loos RJ, Melander O, Metspalu A, März W, Palmer CN, Perola M, Quertermous T, Rader DJ, Ridker PM, Ripatti S, Roberts R, Salomaa V, Sanghera DK, Schwartz SM, Seedorf U, Stewart AF, Stott DJ, Thiery J, Zalloua PA, O'Donnell CJ, Reilly MP, Assimes TL, Thompson JR, Erdmann J, Clarke R, Watkins H, Kathiresan S, McPherson R, Deloukas P, Schunkert H, Samani NJ, Farrall M. *Nature Genetics*.

Rare and low-frequency variants and their association with plasma levels of fibrinogen, FVII, FVIII, and vWF (2015). Huffman JE, **de Vries PS**, Morrison AC, Sabater-Lleal M, Kacprowski T, Auer PL, Brody JA, Chasman DI, Chen MH, Guo X, Lin LA, Marioni RE, Müller-Nurasyid M, Yanek LR, Pankratz N, Grove ML, de Maat MP, Cushman M, Wiggins KL, Qi L, Sennblad B, Harris SE, Polasek O, Riess H, Rivadeneira F, Rose LM, Goel A, Taylor KD, Teumer A, Uitterlinden AG, Vaidya D, Yao J, Tang W, Levy D, Waldenberger M, Becker DM, Folsom AR, Giulianini F, Greinacher A, Hofman A, Huang CC, Kooperberg C, Silveira A, Starr JM, Strauch K, Strawbridge RJ, Wright AF, McKnight B, Franco OH, Zakai N, Mathias RA, Psaty BM, Ridker PM, Tofler GH, Völker U, Watkins H, Fornage M, Hamsten A, Deary IJ, Boerwinkle E, Koenig W, Rotter JI, Hayward C, Dehghan A, Reiner AP, O'Donnell CJ, Smith NL. *Blood*.

Adiposity as a cause of cardiovascular disease: a mendelian randomization study (2015). Hägg S, Fall T, Ploner A, Mägi R, Fischer K, Draisma HH, Kals M, **de Vries PS**, Dehghan A, Willems SM, Sarin AP, Kristiansson K, Nuotio ML, Havulinna AS, de Bruijn RF, Ikram MA, Kuningas M, Stricker BH, Franco OH, Benyamin B, Gieger C, Hall AS, Huikari V, Jula A, Järvelin MR, Kaakinen M, Kaprio J, Kobl M, Mangino M, Nelson CP, Palotie A, Samani NJ, Spector TD, Strachan DP, Tobin MD, Whitfield JB, Uitterlinden AG, Salomaa V, Syvänen AC, Kuulasmaa K, Magnusson PK, Esko T, Hofman A, de Geus EJ, Lind L, Giedraitis V, Perola M, Evans A, Ferrières J, Virtamo J, Kee F, Tregouet DA, Arveiler D, Amouyel P, Gianfagna F, Brambilla P, Ripatti S, van Duijn CM, Metspalu A, Prokopenko I, McCarthy MI, Pedersen NL, Ingelsson E, European Network for Genetic and Genomic Epidemiology Consortium. *International Journal of Epidemiology*.

Genetic variants in the ADAMTS13 and SUPT3H genes are associated with ADAMTS13 activity (2015). **de Vries PS**, Boender J, Sonneveld MA, Rivadeneira F, Ikram MA, Rottensteiner H, Hofman A, Uitterlinden AG, Leebeek FW, Franco OH, Dehghan A, de Maat MP. *Blood*.

Incremental predictive value of 152 single nucleotide polymorphisms in the 10-year risk prediction of incident coronary heart disease: the Rotterdam Study (2015). **de Vries PS**, Kavousi M, Ligthart S, Uitterlinden AG, Hofman A, Franco OH, Dehghan A. *International Journal of Epidemiology*.

Pleiotropy among common genetic loci identified for cardiometabolic disorders and C-reactive protein (2015). Ligthart S, **de Vries PS**, Uitterlinden AG, Hofman A, CHARGE Inflammation working group, Franco OH, Chasman DI, Dehghan A. *PLOS One*.

Age- and sex-specific causal effects of adiposity on cardiovascular risk factors (2015). Fall T, Hägg S, Ploner A, Mägi R, Fischer K, Draisma HH, Sarin AP, Benyamin B, Ladenvall C, Åkerlund M, Kals M, Esko T, Nelson CP, Kaakinen M, Huikari V, Mangino M, Meirhaeghe A, Kristiansson K, Nuotio ML, Kobl M, Grallert H, Dehghan A, Kuningas M, **de Vries PS**, de Bruijn RF, Willems SM, Heikkilä K, Silventoinen K, Pietiläinen KH, Legry V, Giedraitis V, Goumidi L, Syvänen AC, Strauch K, Koenig W, Lichtner P, Herder C, Palotie A, Menni C, Uitterlinden AG, Kuulasmaa K, Havulinna AS, Moreno LA, Gonzalez-Gross M, Evans A, Tregouet DA, Yarnell JW, Virtamo J, Ferrières J, Veronesi G, Perola M, Arveiler D, Brambilla P, Lind L, Kaprio J, Hofman A, Stricker BH, van Duijn CM, Ikram MA, Franco OH, Cottel D, Dallongeville J, Hall AS, Jula A, Tobin MD, Penninx BW, Peters A, Gieger C, Samani NJ, Montgomery GW, Whitfield JB, Martin NG, Groop L, Spector TD, Magnusson PK, Amouyel P, Boomsma DI, Nilsson PM, Järvelin MR, Lyssenko V, Metspalu A, Strachan DP, Salomaa V, Ripatti S, Pedersen NL, Prokopenko I, McCarthy MI, Ingelsson E, ENGAGE Consortium. *Diabetes*.

Association of Rare Loss-Of-Function Alleles in HAL, Serum Histidine Levels and Incident Coronary Heart Disease (2015). Yu B, Li AH, Muzny D, Veeraraghavan N, **de Vries PS**, Bis JC, Musani SK, Alexander D, Morrison AC, Franco OH, Uitterlinden A, Hofman A, Dehghan A, Wilson JG, Psaty BM, Gibbs R, Wei P, Boerwinkle E. *Circulation Cardiovascular Genetics*.

A genetic variant in the seed region of miR-4513 shows pleiotropic effects on lipid and glucose homeostasis, blood pressure, and coronary artery disease (2014). Ghanbari M, **de Vries PS**, de Looper H, Peters MJ, Schurmann C, Yaghootkar H, Dörr M, Frayling TM, Uitterlinden AG, Hofman A, van Meurs JB, Erkeland SJ, Franco OH, Dehghan A. *Human Mutation*.

de Vries PS, Gielen M, Rizopoulos D, Rump P, Godschalk R, Hornstra G, Zeegers MP (2014). Association between polyunsaturated fatty acid concentrations in maternal plasma phospholipids during pregnancy and offspring adiposity at age 7: the MEFAB cohort. *Prostaglandins, Leukotrienes & Essential Fatty Acids*.

ABOUT THE AUTHOR

Paul Stefan de Vries was born in Amsterdam, the Netherlands, on July 15th, 1989. In 2007 he completed his secondary education at the International Secondary School of Eindhoven. He obtained his BSc in Life Sciences at University College Maastricht in 2011, with an emphasis on genetics. During his BSc, his interest moved towards the study of causes of disease in populations, and so he decided to enroll in an MSc in Epidemiology at Maastricht University, which he obtained in 2012. His thesis, supervised by Dr. Marij Gielen, was about the association of polyunsaturated fatty acid concentrations in maternal plasma phospholipids during pregnancy and offspring adiposity at age 7. He was happy to combine aspects of his BSc and MSc in pursuing a PhD with strong elements of both genetics and epidemiology at the Department of Epidemiology of the Erasmus University Medical Center in Rotterdam, the Netherlands. During this time he worked on this thesis within the cardiovascular group under the supervision of Dr. Abbas Dehghan and Professor Oscar H. Franco. He will go on to work as a postdoctoral researcher at the Department of Epidemiology, Human Genetics & Environmental Sciences of the University of Texas Health Science Center at Houston in Houston, Texas, USA, where he will work with Professor Alanna C. Morrison and Professor Eric Boerwinkle on the genetics of cardiovascular disease and its risk factors.