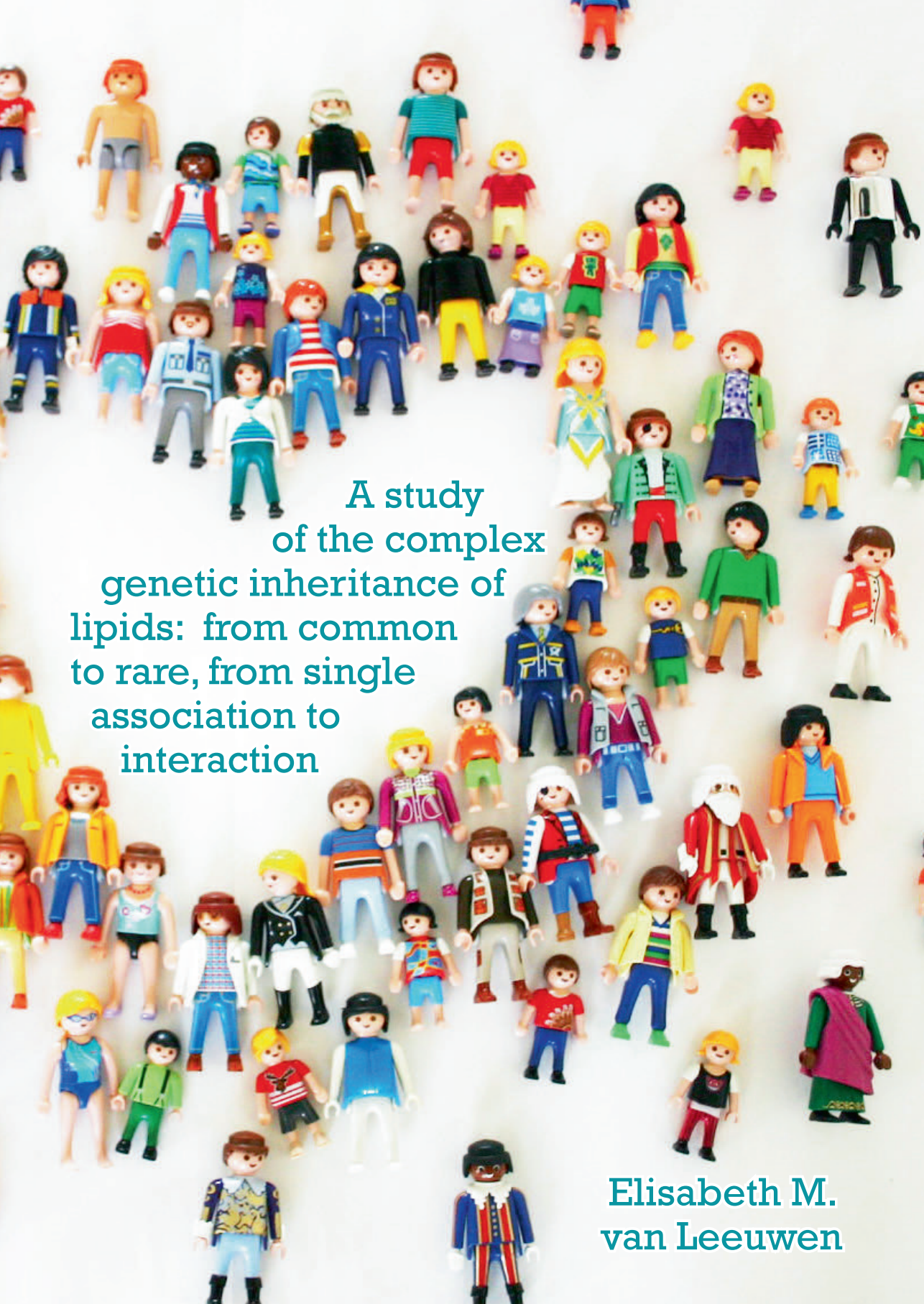


A study of the complex genetic inheritance of lipids: from common to rare, from single association to interaction

Elisabeth M. van Leeuwen

A Study of the Complex Genetic Inheritance of Lipids:

From common to rare, from single association to interaction

Elisabeth M. van Leeuwen

The work presented in this thesis was conducted at the Genetic Epidemiology Unit, Department of Epidemiology, Erasmus Medical Center, Rotterdam, the Netherlands.

The ERF study was supported by the joint grant from the Netherlands Organisation for Scientific Research (NWO, 91203014), the Center for Medical Systems Biology (CMSB), and the Interuniversity Attraction Poles (IUAP) program. The ERF study as a part of EUROSPAN (European Special Populations Research Network) was supported by European Commission FP6 STRP grant number 018947 (LSHG-CT-2006-01947) and also received funding from the European Community's Seventh Framework Programme (FP7/2007-2013)/grant agreement HEALTH-F4-2007-201413 by the European Commission under the program "Quality of Life and Management of the Living Resources" of the Fifth Framework Programme (no. QLG2-CT-2002-01254). High-throughput analysis of the ERF data was supported by a joint grant from Netherlands Organisation for Scientific Research and the Russian Foundation for Basic Research (NWO-RFBR 047.017.043). Exome sequencing in ERF was supported by the ZonMw grant (project 91111025). Exome-chip genotyping was supported by BBMRI-NL.

The generation and management of GWAS genotype data for the Rotterdam Study is supported by the Netherlands Organisation of Scientific Research NOW Investments (nr. 175.010.2005.011, 911-03-012). This study is funded by the Reasearch Institute for Diseases in the Elderly (014-93-015; RIDE2), the Netherlands Genomics Initiative (NGI)/Netherlands Organisation for Scientific Reaserach (NOW) project nr. 050-060-810. The Rotterdam Study is supported by grants from the Netherlands Organisation for Health Research and Development (ZonMw; Priority Medicines Elderly 113102005), by the Earsmus MC and Erasmus University Rotterdam; the Netherlands Organisation for Scientific Research (NOW); the Netherlands Organisation for Health Research and Development (ZonMw); the Reaserach Insitiute for Diseases in the Elderly (RIDE); the Netherlands Genomics Initiative (NGI); the Ministry of Education, Culture and Science; the Ministry of Health Welfare and Sport; the European Commission (DG XII), and the Municipality of Rotterdam.

The publication of this thesis was financially supported by: ABN AMRO, the Erasmus University, Rotterdam and the department of Epidemiology, Erasmus MC, Rotterdam. Financial support by the Dutch Heart Foundation for the publication of this thesis is gratefully acknowledged.

For reasons of consistency within this thesis, some terms and abbreviations have been standardized throughout the text. As a consequence the text may differ in this respect from the articles that have been published.

Cover lay-out: Carin van der Kooi.
Layout and printed by: Gildeprint, Enschede

ISBN: 978-94-6233-249-2

A Study of the Complex Genetic Inheritance of Lipids:

From common to rare, from single association to interaction

Een Studie naar de Complexe Genetisch Natuur van Lipiden:
Van frequent tot zeldzaam, van enkele associatie tot interactie

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 april om 11.30 uur

door

Elisabeth Maria van Leeuwen

geboren te Amsterdam

Erasmus University Rotterdam

PROMOTIECOMMISSIE:

Promotor:	Prof.dr.ir. C.M. van Duijn
Overige leden:	Prof.dr. L.A. Cupples
	Prof.dr. B. Müller-Myhsok
	Prof.dr. E.J.G. Sijbrands

Voor papa en mama

PUBLICATIONS AND MANUSCRIPTS BASED ON THE STUDIES DESCRIBED IN THIS THESIS

1 Elisabeth M. van Leeuwen, Jennifer E. Huffman, Joshua C. Bis, Aaron Isaacs, Monique Mulder, *et al.* **Fine mapping the *CETP* region reveals a common intronic insertion associated to HDL-C.** *NPJ Aging and Mechanisms of Disease (in press).* [Chapter 2.1]

2 Elisabeth M. van Leeuwen, Aniko Sabo, Joshua C. Bis, Jennifer E. Huffman, Ani Manichaikul, *et al.* **Meta-analysis of 49,549 individuals imputed with the 1000 Genomes Project reveals an exonic damaging variant in *ANGPTL4* determining fasting TG levels.** *Journal of Medical Genetics (in press).* [Chapter 2.2]

3 Elisabeth M. van Leeuwen, Alexandros Kanterakis, Patrick Deelen, Mathijs V. Kattenberg, The Genome of the Netherlands Consortium, *et al.* **Population-specific genotype imputations using minimac or IMPUTE2.** *Nature Protocols (Nat Protoc. 2015 Sep;10(9):1285-96).* [Chapter 3.1]

4 Patrick Deelen, Androniki Menelaou, Elisabeth M. van Leeuwen, Alexandros Kanterakis, Freerk van Dijk, *et al.* **Improved imputation quality of low-frequency and rare variants in European samples using the 'Genome of The Netherlands'.** *European Journal of Human Genetics (Eur J Hum Genet, Jun 2014).* [Chapter 3.2]

5 Elisabeth M. van Leeuwen, Lennart C. Karssen, Joris Deelen, Aaron Isaacs, Carolina Medina-Gomez, *et al.* **Genome of the Netherlands population-specific imputations identify an *ABCA6* variant associated with cholesterol levels.** *Nature Communications (Nat Commun, 6:6065, 2015).* [Chapter 3.3]

6 Elisabeth M. van Leeuwen, Françoise A. S. Smouter, Tony Kam-Thong, Nazanin Karbalai, Albert V. Smith, *et al.* **The challenges of genome-wide interaction studies: lessons to learn from the analysis of HDL-C blood levels.** *PLoS One (PLoS One, 9(10):e109290, 2014).* [Chapter 4.1]

7 Elisabeth M. van Leeuwen, Ayşe Demirkan, Najaf Amin, Aaron Isaacs, Jan Bert van Klinken, *et al.* **Identification of rare variants associated with high density lipoprotein cholesterol (HDL-C) by exome sequencing in a family based study.** *Submitted.* [Chapter 4.2]

CONTENT

PART 1
INTRODUCTION AND METHODS

The past century major progress has seen in the understanding of the genetic etiology of lipid metabolism. In this thesis I aim to further dissect the complex genetic nature of circulating lipid levels, in particular four types of lipids: high-density lipoprotein cholesterol (HDL-C), low-density lipoprotein cholesterol (LDL-C), total cholesterol (TC) and triglycerides (TG). Blood concentration of these lipids are highly heritable[1] with genetic heritabilities of 0.485 ± 0.029 for HLD-C, 0.539 ± 0.028 for LDL-C, 0.556 ± 0.028 for TC, and 0.358 ± 0.028 for TG. Circulating lipid levels are determinants of artherosclerosis and cardiovascular disease (CVD)[2,3]. They have been targets for therapeutic intervention of CVD. CVD are the leading cause of morbidity and the number one cause of death worldwide[4]. The goal of this thesis is to identify new variants associated with circulating lipid levels. Ultimately understanding the genetics of lipids may lead to earlier detection, and improved prevention through identification of new targets for therapeutic invention of CVD.

Circulating lipid levels

LDL-C transports cholesterol from the liver to the artery wall. It plays a key role in the pathogenesis of arthesclerosis, and is strongly associated with an increased risk of cardiovascular disease[5,6]. Circulating LDL-C levels are a target for prevention. Statins and other lipid lowering therapy has been successful in reducing the LDL-C levels. LDL-C levels are strongly correlated with total cholesterol levels in humans, which is often used in a clinical setting either by itself or in relation to HDL-C. Although it is preferred to measure the LDL-C levels in the circulation, the LDL-C levels can also be estimated with the Friedewald equation[7]: $LDL\text{-}C = TC - HDL\text{-}C - k \cdot TG$ where k is 0.20 if the quantities are measured in mg/dl and 0.45 if in mmol/l. This method has often been used in epidemiological and genetic studies, though there are limitations to this method, most notably that samples must be obtained after a 12 to 14 hour fasting and that LDL-C cannot be calculated if plasma TG is above 4.52 mmol/L (400 mg/dL).

HDL-C transports various fat molecules including cholesterol out of the artery walls to the liver[8]. In that, its function is the antipode of that of LDL-C and indeed HDL-C has been associated with a decreased risk of CVD. However, HDL-C also is an effective antioxidant and it possesses anti-inflammatory properties. In terms of protecting against the development of cardiovascular disease[9], these antioxidant and anti-inflammatory properties of HDL-C may be as important as its cholesterol efflux function. The key player in the reverse transport of cholesterol from the artery walls to the liver is the protein encoded by the Cholesteryl Ester Transfer Protein (*CETP)* gene[10], as shown by functional analyses in mice[11], hamsters[12] and rabbits[13]. *CETP* has been a target for drug development. *CETP* is one of the genes that has been associated to longevity[14]. Up until now, developments of therapy targeting low HDL-C have failed[15].

Although there is scepticism on the value as HDL-C for prevention of CVD, there are major gaps in our knowledge. Figure 1 gives an overview of the common genetic variants identified to date for HDL-C, LDL-C, TC and TG. Remarkably, *CETP* plays a key role in not only HDL-C, but also LDL-C and TC levels suggesting the gene is a target for various lipids. This is also the case for *TRIB1*, *FADS1-2-3* and *APOA1* (see Figure 1). Genome-wide association studies have also brought to light many new HDL-C genes that are more specific for HDL-C (see Figure 1). A key question to answer is how different genes relate to the various HDL-C particles. There are a large number of sub particles of HDL-C. HDLs are a class of heterogeneous lipoproteins; their heterogeneity is attributable to a different content of apolipoproteins, lipids and enzymes and to the remodelling of HDL-C particles by lipolytic enzymes, lipid transporters and by lipid and apolipoprotein exchange with other circulating lipoproteins and tissues[16,17]. Large HDL-C particles are inversely associated while small HDL-C particles are positively associated with CVD[16,17].

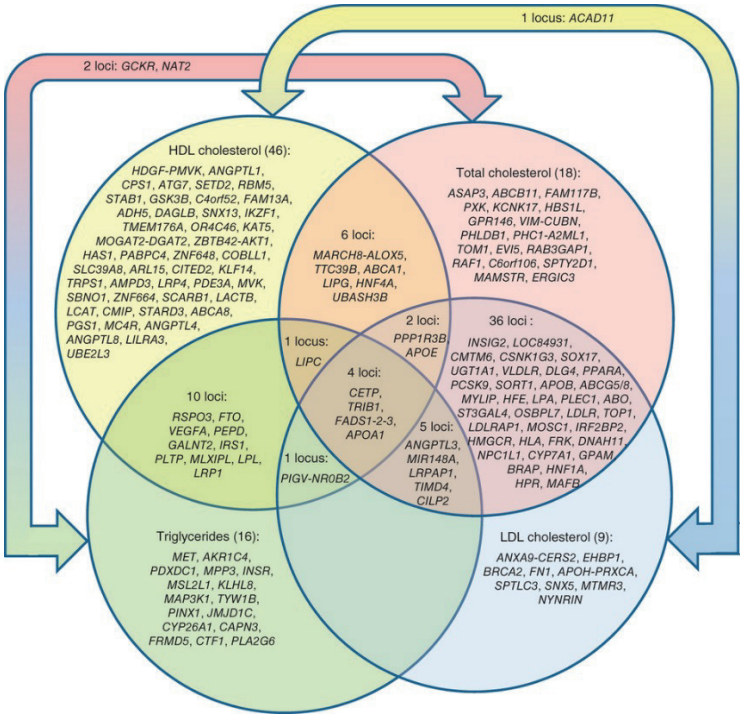


Figure 1: Overlap of loci associated with different lipid traits, as discovered by Teslovich *et al*[32] and GLGC *et al*[33]. The Venn diagram illustrates the number of loci that show association with multiple lipid traits. The number of loci primarily associated with only one trait is listed in parentheses after the trait name, and locus names are listed below. Loci that show association with two or more traits are shown in the appropriate segment. Source: GLGC *et al*[33].

HDL-C levels are strongly inversely related to triglycerides. Triglycerides are esters derived from glycerol and three fatty acids which enables bidirectional transference of adipose fat and circulating glucose from the liver. There are at present 16 genes that are involved in both HDL-C and TG, as expected. This overlap is larger compared to the overlapping genes that are involved in both LDL-C and TC. There has been debated whether lowering TG has not resulted in prevention of CVD[18]. Yet, epidemiological and genetic research found evidence that raised triglycerides are a risk factor for cardiovascular disease also in the general population[19].

The high heritability of circulating lipids[20] has fuelled the great interest in genetic research that has already been successful in the second half of the previous century. These studies have revealed that many lipid related syndromes are caused by a relative rare mutation in a single gene. In these Mendelian forms of dyslipidemia, such as familiar hyperlipidemia, the circulating lipid levels are strongly elevated and carriers have an increased risk of early onset CVD (before the age of 65 years)[21]. This disease is segregating from generation to generation, both as a dominant form with up to 50% of first degree relative affected as well as recessive forms with recurrence risk in siblings of 25%. There are also lipoprotein deficiency disorders in which the circulating lipid levels lead to the pathology, e.g. Tangier disease[22,23]. Tangier disease (TD) is an autosomal recessive disorder of lipid metabolism. It is characterized by absence of plasma HDL-C and deposition of cholesteryl esters in the reticulo-endothelial system with splenomegaly and enlargement of tonsils and lymph nodes. Although low HDL-C is associated with an increased risk for coronary artery disease, this condition is not consistently found in TD pedigrees. Metabolic studies in TD patients have revealed a rapid catabolism of HDL-C and its precursors. The TD locus has been mapped to chromosome 9q31 within the *ABC1* gene[22].

There are many more Mendelian genes involved in lipid related syndromes, of note that these are various genes overlapping with the findings of the GWAS, like *LDL-R* and *APOB*[24,25].

Despite that many rare and common variants found to date, not all heritability is explained yet and thus there are still many variants to be found, probably with even smaller effects and/or smaller frequencies worldwide. There may be various explanations why these have not been found so far. This lack of information is often referred to as the missing heritability[26], or more precisely our missing knowledge of the heritability, which may be attributed to the fact that:

1. Many regions in the genome have been associated with circulating lipid levels, however, the causal variant is still to be identified. Identification of the causal variant may improve the heritability explained. Due to improved technologies, it is now possible to fine-map these regions and locate the causal variant.
2. Mainly common have been studied in the general population and low-frequency variants segregating in families but these have not been studied together with rare variants. However, also rare variants are expected to determine circulating lipid levels in the population.

3. Interactions of genetic variants may explain part of the heritability. This may concern gene environment interactions or gene-gene interactions. Gene environment interactions have been studies as part of the ENGAGE contortium[27]. There has been little work on gene interactions, particular gene-gene interactions. Persistent evidence for interacting loci involved in lipid metabolism comes from experimental animal research in which various loci interact with each other[28].
4. Genetic mechanisms like structural variation, DNA methylation and histone modification are also potential candidates determining circulating lipid levels[29,30].

This thesis focuses on the explanation 1, 2 and 3 and specifically aims to identify variants associated with individual circulating lipid levels in the general population. Common variants typically have small effects compared to the Mendelian variants. Although the variants segregate according to Mendelian principles, there is no typical aggregation of disease in families in contrast to the clinical expression of the Mendelian variants. At present, genome-wide association studies (GWAS) and sequencing of exomes have identified many common variants with small effects on circulating lipid levels[31-33]. I aim to identify common variants integrating the GWAS data with that of large scale sequencing projects such as the 1000 Genomes and genome of the Netherlands (GoNL). This will allow me to finemap regions and search for independent variants explaining the heritability and find new variants. Further I aim to study gene-gene interactions and find rare variants both in families associated with HDL-C.

Genetic epidemiological approaches

In this thesis several genetic epidemiological approaches are used to dissect the complex nature of circulating lipid levels. The genes identified in families have often used a linkage approach which is depicted in figure 2. Linkage occurs in a family when alleles located close together on a chromosome with the disease mutation are inherited together during meiosis. The discovery of the common variants is based on the same principle of inheritance, but used another statistical approach, association. The basic rationale of the methods is that if a variants is causally associated to a disease, the variant is expected to be found more often in cases than controls. However, as genes are segregating from parents to offspring as chromosomes, large pieces of chromosomes may be linked to each other also in the population (linkage disequilibrium) and also nearby non-causal variants are shared. Thus it is not necessary to determine all variants in the genome but for GWAS variants are used to cover the full genome based on linkage disequilibrium[34] (Figure 2). This principle can be applied genome-wide, allowing identification of new variants or regions associated with the trait without any prior hypothesis but also to candidate genes, i.e., genes that based on the protein they encode for as expected to be associated to a trait because the protein has been implicated in the outcome.

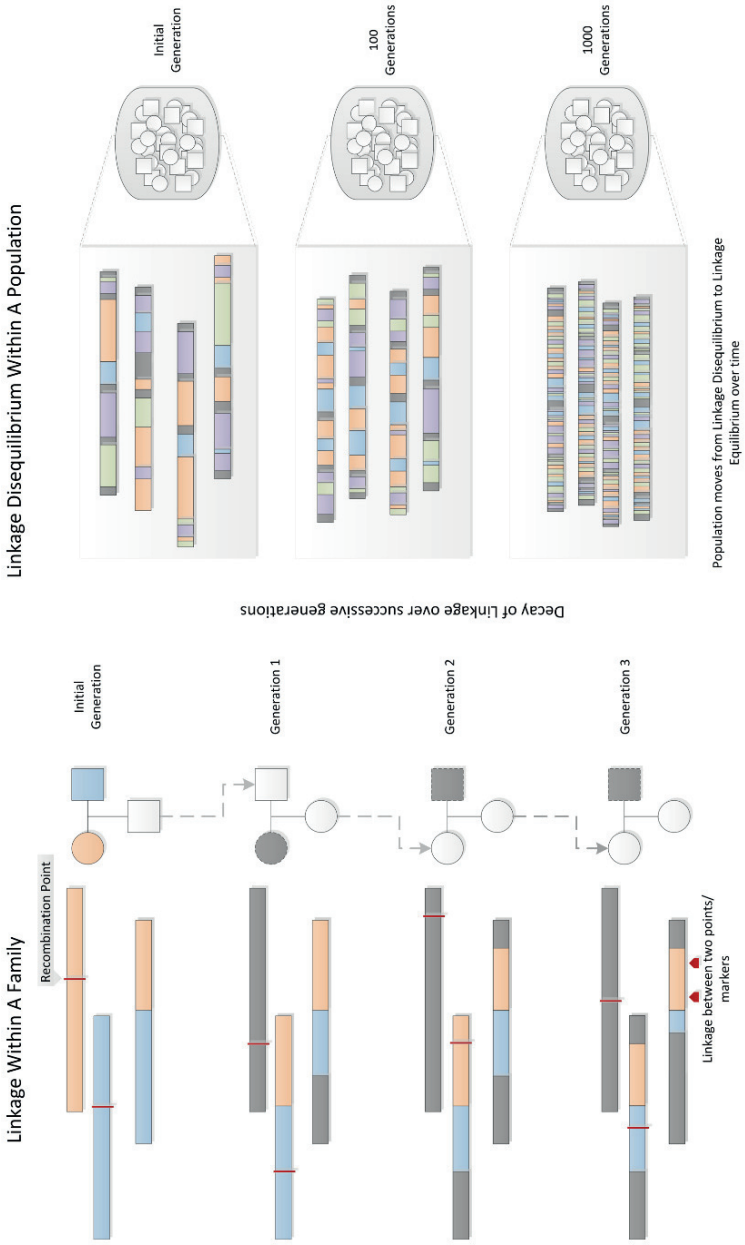


Figure 2: Linkage and Linkage Disequilibrium. Within a family, linkage occurs when two genetic markers (points on a chromosome) remain linked on a chromosome rather than being broken apart by recombination events during meiosis, shown as red lines. In a population, contiguous stretches of founder chromosomes from the initial generation are sequentially reduced in size by recombination events. Over time, a pair of markers or points on a chromosome in the population move from linkage disequilibrium to linkage equilibrium, as recombination events eventually occur between every possible point on the chromosome. Source: Bush and Moore[34].

GWAS have been extensively used to identify variants associated with various traits among which circulating lipid levels[32,33,35-37]. In a GWAS, millions of genetic variants along the genome are tested to be associated with a particular trait[38,39]. However, testing millions of variants, will lead to a large multiple testing correction and therefore the threshold for significance [40] in these studies is $5 \cdot 10^{-8}$. In the early stages of GWAS of circulating lipid levels, only one or a few cohorts were used yielding 6 and 17 variants[35,36]. Subsequently, more to cohorts were meta-analysed due to the imputation strategies and sample sizes increased yielding in total 72, 55, 71 and 42 variants for HDL-C, LDL-C, TC and TG respectively[32,33,41], see Figure 1. Multiple tools have been developed to enable fast and accurate testing of all variants on the genome[37,42,43]. Although GWAS replaced the candidate-gene approach in which a gene is studied in detail as opposed to the full genome, it has produced large number of new candidate regions, where the causal variant remains to be found by sequencing or imputation of clinical studies.

In my thesis I have also used exome-wide association studies (ExWAS). Whereas GWAS makes use of genotype data, mostly imputed to a certain reference panel (see below), ExWAS makes use of exome sequencing data and is thus targeting a subset of the genome, i.e., the region of the genome that encodes for proteins (exome). As the main difference between GWAS and ExWAS is that in an ExWAS only coded variants are tested for association whereas GWAS also test not coding variants, there are no principle differences in the analyses. However, there are major technical differences in how the variants are assesses. In GWAS genotyping arrays are used while exome sequencing uses next generation sequencing. The analysis used for GWAS determine classical genotypes at a locus while they are usually common bi-allelic single nucleotide polymorphism (SNPs). SNPs are DNA sequence variations occurring commonly within a population in which a single nucleotide (A, T, C or G) in the genome differs between individuals. To allow pooling of studies using different arrays, common SNPs are usually imputed. Also rare variants are imputed but these are not always present, depending on the reference panel. The next generation sequencing of exonic variants is a probabilistic approach in which the quality of genotyping depends on the read depth, i.e. how often a variant is seen. Next generation sequencing therefore may results in more reliable results as the variants in the GWAS dataset are only estimates of the variants.

Genome-wide interaction studies (GWIS) are used to discover interactions between genetic variants. GWIS has been hampered by the computation time needed for testing all unique pairs of SNPs on a regular Computer Processing Units (CPUs). In this thesis I make use of the GLIDE software package[44] which makes use of modern Graphics Processing Units (GPUs) to perform linear regression for all pairs of SNPs in a relatively short time period. Just like with GWAS and ExWAS, also in GWIS, the large number of tests that are performed need to be taken into account and therefore the threshold for significance for this GWIS was $1 \cdot 10^{-8}$, which is debatable.

Bioinformatics plays a key role in the current genetic field, where much work is performed by the computer, not only the analysis itself, but also creating the datasets for the analysis. Bioinformatics is needed to create large datasets, harmonizing genetic information over data sets by genetic imputations, develop the tools for the analysis and run the analysis. After the analysis, the results of the genetic epidemiologic research needs to be interpreted. Bioinformatic tools for annotation, for prediction of functional effects, for amino acid substitutions, for pathway analysis can be of help for the interpretation of the results. In my thesis, I have performed various imputations of sequence data using different referent panels, which is discussed below.

Reference panels for imputations

A common approach in GWAS studies to find new variants has been to enlarge the samples by pooling studies. This requires all cohorts to have the same variants in their GWAS. This is not always the case, as different chips of different size and different manufactors are available for genotyping. Using the principles of LD, it has been possible to impute genotypes based on a set of common reference haplotypes. The HapMap reference set[45-47] was the basis of the first genetic imputations. This set has been used in the past intensively to identify variants associated with various phenotypes. The Phase I HapMap version[47] was based on 90 YRI, 90 CEU, 45 CHB and 44 JPT. The CEU and YRI samples were 30 parent-offspring trios. The genotyping goal of the Phase I HapMap Project was to genotype at least one common SNP (MAF≥0.05) every 5 kilobases (kb) across the genome in the 269 samples, resulting in approximately 1.3 million SNPs. In Phase II of the HapMap Project, a further 2.1 million SNPs were successfully genotyped on the same individuals resulting in an SNP density of approximately one per kilobase[45]. The HapMap 3 Project contains 1.6 million common SNPs in 1,184 individuals from 11 global populations (including the 269 individuals from HapMap Phase I and II), and sequenced ten 100-kilobase regions in 692 of these individuals[46]. These populations were included to provide further variation data from each of the three continental regions represented in HapMap Phase I and II, as well as data from some more admixed populations residing in the US. Although HapMap was successful for most European common variants, rare variants (MAF < 0.01) cannot be imputed as too few haplotypes are available in the reference panel.

The 1000 Genomes (1kG) project[48,49] aimed to make a reference panel for rare variants. One of the aims of the pilot phase of the project[48] was to develop and compare different strategies for genome-wide sequencing with high-throughput platforms. Therefore 179 individuals were sequenced low-coverage, 6 individuals in two trios were deep sequenced and 8,140 exons in 697 individuals were sequenced. The phase I of the project contained[49] 1,092 individuals sampled from 14 populations drawn from Europe, East Asia, sub-Saharan Africa and the Americas analysed through a combination of low-coverage (2-6x) whole-genome sequence

data, targeted deep (50-100x) exome sequence data and dense SNP genotype data. The final version of the project (phase 3 version 5) of the 1kG reference set contains a total of about 81.2M polymorphic variants coming from low-coverage whole-genome sequencing of 2,500 unrelated samples from 19 populations.

The main difference between HapMap and 1kG project is that the former has fewer genotypes: as it is based on fewer persons it capture less variation in the population. There also is a difference in quality of haplotypes between HapMap and 1kG. The reliability of the imputations is determinant by: (1) the number of haplotypes in the reference set and (2) the quality of the haplotypes which differs in parent-offspring trios and statistical estimates. A variant may not be represented adequately in the reference data set. This explains why imputations of common variants are more reliable compared to imputations of rare variants. Intially HapMap haplotypes were estimated for different ethnic groups as LD is expected to be different between ethnic groups. The 1kG aims to create maps of genetic variation across multiple populations, but the number of individuals per population is modest in this reference panel. Splitting groups reduces the size of the reference set and as rare variant haplotypes may occur in the different ethnic groups mixing different ethnic populations in one reference set appears to improve imputation quality particular for the rare variants[50-52], even the addition of samples are from ethnic groups are not closely related to the samples in the target set. However, when the percentage of unrelated samples is beyond a certain proportion, the imputation quality does not improve, especially for low-frequency variants[52]. Nowadays, many efforts are ongoing to further increase the sample size of the reference panels. The larger the reference panel, the larger the change it will contain the haplotype of interest and thus the more accurate the imputations will be. As the frequency of rare variants may increase in certain populations due to drift and founder effects[53], the power of searches for rare functional variants may improve by the use of reference sets specific to distinct populations. Such references allow for better quality imputation of rare variants especially those with increased frequency in the population of interest[53-55]. However, to characterize a population, it is crucial to sequence as many individuals as possible to maximize the probability of capturing rare variants[55]. This approach has been applied by the Genome of the Netherlands consortium to develop the GoNL reference panel. For this custom-built reference panel for the Dutch population the whole genome of 250 parent-offspring trios were sequenced at approximately 13x coverage[54,55]. An approximately equal representation from the original 11 Dutch provinces were choosen, and an oversampling from the two major cities, Amsterdam and Rotterdam.

One of the first questions addressed whether there are differences in the genome across the Netherlands. Figure 3[55] shows the results of the Principal Component Analysis (PCA) of all 769 GoNL samples. The three PCs correlated significantly with geographic location and distinguished between: (1) the North and South of the Netherlands; (2) between the East

and West; and (3) between the middle-band of the Netherlands and the rest of the country. The PCs capture the geographical variation in the data very well. Due to the trio design, the phasing quality of the reference panel was better than that of the 1kG Phase 1 panel[56]. The GoNL reference panel was used in this thesis for the imputations of the Dutch biobanks prior to a meta-analysis of circulating lipid levels with the aim to identify low-frequency and rare variants associated with circulating lipid levels.

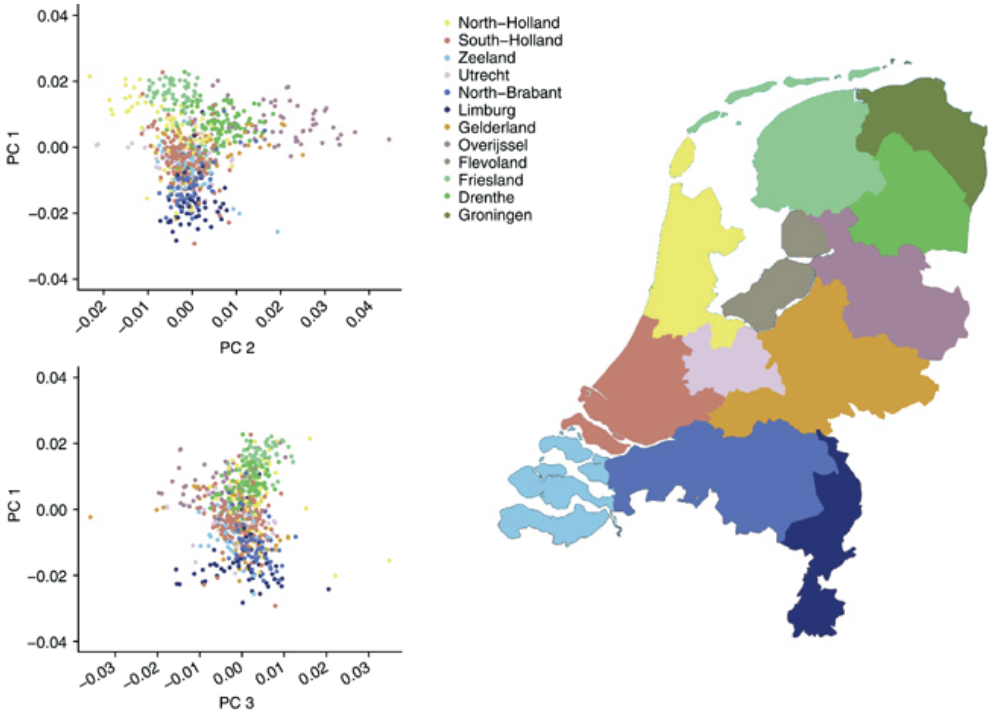


Figure 3: PCA results highlighting differences in genetic make-up across the Netherlands: the plots give PC1 versus PC2, and PC1 versus PC3. Source: Boomsma *et al*[55]

Scope of this thesis

The aim of this thesis is to dissect the complex genetic nature of circulating lipid levels, in particular HDL-C, LDL-C, TC and TG. One of the approaches used in this thesis to identify new variants associated with these circulating lipid levels are meta-analysis of GWAS of multiple cohorts. This requires all cohorts to have the same variants in their GWAS. To this end, genotypes that have not been measured in a given cohort can be imputed based on a set of reference haplotypes. In Chapter 2 the 1000 Genomes reference panel was used for imputations of the cohorts of the CHARGE consortium. Chapter 2.1 describes the fine-mapping of the *CETP* region, a region that has been known to be associated with HDL-C for a long time[32]. Though the causal variant has not been determined so far and by using the

1kG as a reference panel, there will be more power to fine-map the association between *CETP* and HDL-C. Chapter 2.2 focuses on new variants associated with one of the four circulating lipid levels. Chapter 3 focuses on the Genome of the Netherlands reference panel. Chapter 3.1 provides guidelines for performing imputations with this population-specific reference panel and Chapter 3.2 uses this population-specific reference panel there was a significant improvement for rare variants (MAF between 0.05 and 0.5%) compared to the 1000 Genomes, both for Dutch, British and Italian samples. Chapter 3.3 uses GoNL for the identification of novel variants associated with circulating lipid levels after imputations with the GoNL reference panel, followed by a meta-analyses. Meta-analysis of GWAS of multiple cohorts has been applied on various phenotypes before with various reference panels being used for the imputations and although also in this thesis the method showed to identify even more variants associated with circulating lipid levels, I also applied new methods to dissect the complex genetic nature of HDL-C (Chapter 4). In Chapter 4.1 I performed the first GWIS to identify SNPxSNP interactions associated with HDL-C. In Chapter 4.2 I performed an ExWAS to identify rare coding variants associated with HDL-C. Finally, in Chapter 5, I discuss the findings of this thesis, and their implications for future research.

REFERENCES

1. Namboodiri, K. K. *et al.* The Collaborative Lipid Research Clinics Family Study: biological and cultural determinants of familial resemblance for plasma lipids and lipoproteins. *Genet Epidemiol* **2**, 227–254 (1985).
2. Kannel, W. B., Dawber, T. R., Kagan, A., Revotskie, N. & Stokes, J., 3rd. Factors of risk in the development of coronary heart disease–six year follow-up experience. the Framingham study. *Ann Intern Med* **55**, 33–50 (1961).
3. Miller, N. E. & Miller, G. J. Letter: High-density lipoprotein and atherosclerosis. *Lancet* **1**, 1033 (1975).
4. Tóth, P. P., Potter, D. & Ming, E. E. Prevalence of lipid abnormalities in the United States: the National Health and Nutrition Examination Survey 2003-2006. *J Clin Lipidol* **6**, 325–330 (2012).
5. Fraley, A. E. & Tsimikas, S. Clinical applications of circulating oxidized low-density lipoprotein biomarkers in cardiovascular disease. *Curr Opin Lipidol* **17**, 502–509 (2006).
6. Packard, C. J. Small dense low-density lipoprotein and its role as an independent predictor of cardiovascular disease. *Curr Opin Lipidol* **17**, 412–417 (2006).
7. Friedewald, W. T., Levy, R. I. & Fredrickson, D. S. Estimation of the concentration of low-density lipoprotein cholesterol in plasma, without use of the preparative ultracentrifuge. *Clin Chem* **18**, 499–502 (1972).
8. Asztalos, B. F. & HDL Atherosclerosis Treatment Study. High-density lipoprotein metabolism and progression of atherosclerosis: new insights from the HDL Atherosclerosis Treatment Study. *Curr Opin Cardiol* **19**, 385–391 (2004).
9. Barter, P. J. *et al.* Antiinflammatory properties of HDL. *Circ Res* **95**, 764–772 (2004).
10. Zhong, S. *et al.* Increased coronary heart disease in Japanese-American men with mutation in the cholesteryl ester transfer protein gene despite increased HDL levels. *J Clin Invest* **97**, 2917–2923 (1996).
11. Hayek, T. *et al.* Decreased early atherosclerotic lesions in hypertriglyceridemic mice expressing cholesteryl ester transfer protein transgene. *J Clin Invest* **96**, 2071–2074 (1995).
12. Briand, F. *et al.* Anacetrapib and dalcetrapib differentially alters HDL metabolism and macrophage-to-feces reverse cholesterol transport at similar levels of CETP inhibition in hamsters. *Eur J Pharmacol* **740**, 135–143 (2014).
13. Kee, P. *et al.* Effect of inhibiting cholesteryl ester transfer protein on the kinetics of high-density lipoprotein cholesteryl ester transport in plasma: in vivo studies in rabbits. *Arterioscler Thromb Vasc Biol* **26**, 884–890 (2006).
14. Sun, L. *et al.* Gene-gene interaction between CETP and APOE polymorphisms confers higher risk for hypertriglyceridemia in oldest-old Chinese women. *Exp Gerontol* **55**, 129–133 (2014).
15. Santos-Gallego, C. G., Badimon, J. J. & Rosenson, R. S. Beginning to understand high-density lipoproteins. *Endocrinol Metab Clin North Am* **43**, 913–947 (2014).
16. Camont, L., Chapman, M. J. & Kontush, A. Biological activities of hdl subpopulations and their relevance to cardiovascular disease. *Trends Mol Med* **17**, 594–603 (2011).
17. Pirillo, A., Norata, G. D. & Catapano, A. L. High-density lipoprotein subfractions–what the clinicians need to know. *Cardiology* **124**, 116–125 (2013).
18. Miller, M. *et al.* Triglycerides and cardiovascular disease: a scientific statement from the american heart association. *Circulation* **123**, 2292–2333 (2011).
19. Nordestgaard, B. G. & Varbo, A. Triglycerides and cardiovascular disease. *Lancet* **384**, 626–635 (2014).
20. van Dongen, J., Willemsen, G., Chen, W.-M., de Geus, E. J. C. & Boomsma, D. I. Heritability of metabolic syndrome traits in a large population-based sample. *J Lipid Res* **54**, 2914–2923 (2013).

21. Fredrickson, D. S. & Lees, R. S. A system for phenotyping hyperlipoproteinemia. *Circulation* **31**, 321–327 (1965).
22. Bodzioch, M. *et al.* The gene encoding ATP-binding cassette transporter 1 is mutated in Tangier disease. *Nat Genet* **22**, 347–351 (1999).
23. Rust, S. *et al.* Tangier disease is caused by mutations in the gene encoding atp-binding cassette transporter 1. *Nat Genet* **22**, 352–355 (1999).
24. Civeira, F. *et al.* Frequency of low-density lipoprotein receptor gene mutations in patients with a clinical diagnosis of familial combined hyperlipidemia in a clinical setting. *J Am Coll Cardiol* **52**, 1546–1553 (2008).
25. Jarvik, G. P., Brunzell, J. D. & Motulsky, A. G. Frequent detection of familial hypercholesterolemia mutations in familial combined hyperlipidemia. *J Am Coll Cardiol* **52**, 1554–1556 (2008).
26. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).
27. Surakka, I. *et al.* A genome-wide screen for interactions reveals a new locus on 4p15 modifying the effect of waist-to-hip ratio on total cholesterol. *PLoS Genet* **7**, e1002333 (2011).
28. Brockmann, G. A. *et al.* Genetic control of lipids in the mouse cross DU6i x DBA/2. *Mamm Genome* **18**, 757–766 (2007).
29. Guay, S. P. *et al.* DNA methylation variations at CETP and LPL gene promoter loci: new molecular biomarkers associated with blood lipid profile variability. *Atherosclerosis* **228**, 413–420 (2013).
30. Turner, S. D. *et al.* Knowledge-driven multi-locus analysis reveals gene-gene interactions influencing HDL cholesterol level in two independent EMR-linked biobanks. *PLoS One* **6**, e19586 (2011).
31. Lange, L. A. *et al.* Whole-exome sequencing identifies rare and low-frequency coding variants associated with LDL cholesterol. *Am J Hum Genet* **94**, 233–245 (2014).
32. Teslovich, T. M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707–713 (2010).
33. Global Lipids Genetics Consortium *et al.* Discovery and refinement of loci associated with lipid levels. *Nat Genet* **45**, 1274–1283 (2013).
34. Bush, W. S. & Moore, J. H. Chapter 11: Genome-wide association studies. *PLoS Comput Biol* **8**, e1002822 (2012).
35. Kathiresan, S. *et al.* Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. *Nat Genet* **40**, 189–197 (2008).
36. Willer, C. J. *et al.* Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nat Genet* **40**, 161–169 (2008).
37. Aulchenko, Y. S., Ripke, S., Isaacs, A. & van Duijn, C. M. GenABEL: an R library for genome-wide association analysis. *Bioinformatics* **23**, 1294–1296 (2007).
38. Balding, D. J. A tutorial on statistical methods for population association studies. *Nat Rev Genet* **7**, 781–791 (2006).
39. Evans, D. M. & Cardon, L. R. Genome-wide association: a promising start to a long race. *Trends Genet* **22**, 350–354 (2006).
40. Pe'er, I., Yelensky, R., Altshuler, D. & Daly, M. J. Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genet Epidemiol* **32**, 381–385 (2008).
41. Aulchenko, Y. S. *et al.* Loci influencing lipid levels and coronary heart disease risk in 16 European population cohorts. *Nat Genet* **41**, 47–55 (2009).
42. Galesloot, T. E., van Steen, K., Kiemeney, L. A. L. M., Janss, L. L. & Vermeulen, S. H. A comparison of multivariate genome-wide association methods. *PLoS One* **9**, e95923 (2014).

43. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559–575 (2007).
44. Kam-Thong, T. *et al.* GLIDE: GPU-based linear regression for detection of epistasis. *Hum Hered* **73**, 220–236 (2012).
45. International HapMap Consortium *et al.* A second generation human haplotype map of over 3.1 million snps. *Nature* **449**, 851–861 (2007).
46. International HapMap 3 Consortium *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52–58 (2010).
47. International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
48. 1000 Genomes Project Consortium *et al.* A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
49. 1000 Genomes Project Consortium *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
50. Howie, B., Marchini, J. & Stephens, M. Genotype imputation with thousands of genomes. *G3 (Bethesda)* **1**, 457–470 (2011).
51. Hao, K., Chudin, E., McElwee, J. & Schadt, E. E. Accuracy of genome-wide imputation of untyped markers and impacts on statistical power for association studies. *BMC Genet* **10**, 27 (2009).
52. Jostins, L., Morley, K. I. & Barrett, J. C. Imputation of low-frequency variants using the HapMap3 benefits from large, diverse reference sets. *Eur J Hum Genet* **19**, 662–666 (2011).
53. Pardo, L. M., MacKay, I., Oostra, B., van Duijn, C. M. & Aulchenko, Y. S. The effect of genetic drift in a young genetically isolated population. *Ann Hum Genet* **69**, 288–295 (2005).
54. Genome of the Netherlands Consortium. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat Genet* **46**, 818–825 (2014).
55. Boomsma, D. I. *et al.* The Genome of the Netherlands: design, and project goals. *Eur J Hum Genet* **22**, 221–227 (2014).
56. Deelen, P. *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).

PART 2
THE 1000 GENOMES

CHAPTER 2.1

Fine mapping the *CETP* region reveals a common intronic insertion associated to HDL-C

Elisabeth M. van Leeuwen, Jennifer E. Huffman, Joshua C. Bis, Aaron Isaacs, Monique Mulder, Aniko Sabo, Albert V. Smith, Serkalem Demissie, Ani Manichaikul, Jennifer A. Brody, Mary F. Feitosa, Qing Duan, Katharina E. Schraut, Pau Navarro, Jana V. van Vliet-Ostaptchouk, Gu Zhu, Hamdi Mbarek, Stella Trompet, Niek Verweij, Leo-Pekka Lyytikäinen, Joris Deelen, Ilja M. Nolte, Sander W. van der Laan, Gail Davies, Andrea J.M. Vermeij- Verdoold, Andy A.L.J. van Oosterhout, Jeannette M. Vergeer- Drop, Dan E. Arking, Holly Trochet, Generation Scotland, Carolina Medina-Gomez, Fernando Rivadeneira, Andre G. Uitterlinden, Abbas Dehghan, Oscar H. Franco, Eric J. Sijbrands, Albert Hofman, Charles C. White, Josyf C. Mychaleckyj, Gina M. Peloso, Morris A. Swertz, LifeLines Cohort Study, Gonneke Willemsen, Eco J. de Geus, Yuri Milaneschi, Brenda W.J.H. Penninx, Ian Ford, Brendan M. Buckley, Anton J.M. de Craen, John M. Starr, Ian J. Deary, Gerard Pasterkamp, Albertine J. Oldehinkel, Harold Snieder, P. Eline Slagboom, Kjell Nikus, Mika Kähönen, Terho Lehtimäki, Jorma S. Viikari, Olli T. Raitakari, Pim van der Harst, J.Wouter Jukema, Jouke-Jan Hottenga, Dorret I Boomsma, John B. Whitfield, Grant Montgomery, Nicholas G Martin, CHARGE Lipids Working Group, Ozren Polasek, Veronique Vitart, Caroline Hayward, Ivana Kolcic, Alan F Wright, Igor Rudan, Peter K. Joshi, James F. Wilson, Leslie A. Lange, James G. Wilson, Vilmundur Gudnason, Tamar B. Harris, Alanna Morrison, Ingrid B. Borecki, Stephen S. Rich, Sandosh Padmanabhan, Bruce M. Psaty, Jerome I. Rotter, Blair H. Smith, Eric Boerwinkle, L. Adrienne Cupples and Cornelia M. van Duijn.

Accepted for publication in NPJ Aging and Mechanisms of Disease.
The supplemental information for this chapter is available online at:
http://www.nature.com/articles/npjamd201511#s1

ABSTRACT

Background: Individuals with exceptional longevity and their offspring have significantly larger high-density lipoprotein concentrations (HDL-C) particle sizes due to the increased homozygosity for the I405V variant in the cholesteryl ester transfer protein (*CETP)* gene. In this study we investigate the association of *CETP* and HDL-C further to identify novel, independent *CETP* variants associated with HDL-C in humans.

Methods: We performed a meta-analysis of HDL-C within the *CETP* region using 59,432 individuals imputed with 1000 Genomes data. We performed replication in an independent sample of 47,866 individuals and validation was done by Sanger sequencing.

Results: The meta-analysis of HDL-C within the *CETP* region identified five independent variants, including an exonic variant and a common intronic deletion. We replicated these five variants significantly in an independent sample of 47,866 individuals. Sanger sequencing of the deletion within a single family confirmed segregation of this variant.
The strongest reported association between HDL-C and *CETP* variants, was rs3764261; however, after conditioning on the five novel variants we identified, the support for rs3764261 was highly reduced ($\beta_{unadjusted}$=3.179 mg/dL (*p*-value=$5.25 \cdot 10^{-509}$), $\beta_{adjusted}$=0.859 mg/dL (*p*-value=$9.51 \cdot 10^{-25}$)), and this finding suggests that these five novel variants may partly explain the association of *CETP* with HDL-C. Indeed, three of the five novel variants (rs34065661, rs5817082, rs7499892) are independent of rs3764261.

Conclusions: The causal variants in *CETP* that account for the association with HDL-C remain unknown. We used studies imputed to the 1000 Genomes reference panel for fine mapping of the *CETP* region. We identified and validated five variants within this region that may partly account for the association of the known variant (rs3764261) as well as other sources of genetic contribution to HDL-C.

INTRODUCTION

Aging is characterized by a deterioration in the maintenance of homeostatic processes over time, leading to functional decline and increased risk for disease and death[1]. One of the genes linked to healthy aging and longevity is the cholesteryl ester transfer protein (*CETP*) gene[1,2]. Homozygosity in the 405VV variants of *CETP* is associated with lower concentrations of *CETP*, higher concentrations of high-density lipoprotein concentrations (HDL-C) and greater HDL-C particle size, all associated with both protection against cardiovascular disease[3] and exceptional longevity[4].

Functional analyses in mice[5], hamsters[6] and rabbits[7] have revealed that the protein encoded by the *CETP* gene mediates the transfer of cholesteryl esters from HDL-C to other lipoproteins such as atherogenic (V)LDL particle and is a key participant in the reverse transport of cholesterol from the periphery to the liver[8]. Due to the function of *CETP* and the association of the gene with HDL-C in humans[9,10], the *CETP* gene is one of the targets for drug development for dyslipidemia[6,11,12]. *CETP*-inhibition leads to an increase of HDL-C from 30% up to 140% depending on the compound used. The first drug of its class, Torcetrapib was unfortunately associated with an increased mortality and morbidity in patients receiving the CETP-inhibitor in addition to atorvastatin[13,14].

The estimated heritability of HDL-C levels is high in humans: 47-76%[15-23]. Previously published whole-genome sequence data[23] reported that common variants (minor allele frequency (MAF) > 1%) explain up to 61.8% of the variance in HDL-C levels and that rare variants (MAF < 1%) explain an additional 7.8% of the variance. Genome-wide association studies (GWAS) revealed that numerous variants are associated with HDL-C, among which are various common[9,10] and rare[24,25] variants within the *CETP* gene in multiple ancestries[4,8,26-28]. In this paper we investigate the association between *CETP* and HDL-C in humans in further detail to identify variants that are likely to be causal.

To this end, we used a meta-analysis of association studies with imputed genotypes within the *CETP* region. Our study consisted of data from 59,432 samples, of which the genotypes were imputed to the 1000 Genomes project reference panel (version Phase 1 integrated release v3, April 2012, all populations). By using 1000 Genomes imputed data we expected to find more rare or low-frequent variants as well as novel insertions and deletions.

METHODS

Study descriptions

The descriptions of the participating cohorts can be found in the supplemental material. All studies were performed with the approval of the local medical ethics committees, and written informed consent was obtained from all participants.

Study samples and phenotypes

The total number of individuals in the discovery phase was 59,432 and in the replication phase, 47,866. Of the discovery samples, 44,108 individuals (74.21%) were of European ancestry. Of the replication samples, 47,081 individuals (98.36%) were of European ancestry. A summary of the details of both the discovery and replication cohorts participating in this study can be found in Supplemental Table 1.

Genotyping and imputations

All cohorts were genotyped using commercially available Affymetrix or Illumina genotyping arrays, or custom Perlegen arrays. Quality control was performed independently for each study. To facilitate meta-analysis and replication, each discovery and replication cohort performed genotype imputation using IMPUTE2[29] or Minimac[30] with reference to the 1000 Genomes project reference panel (version Phase 1 integrated release v3, April 2012, all populations). The details per cohort can be found in Supplemental Table 2.

Association analysis in discovery cohorts

The lipid measurements were adjusted for sex, age and age^2 in all cohorts and if necessary also for cohort-specific covariates (Supplemental Table 1). Some cohorts included samples using lipid lowering medication; we did not adjust for lipid lowering medication in our analysis because HDL-C levels are only minimally influenced by lipid lowering medication. Each discovery cohort ran association analysis for all variants within the *CETP* region (chromosome 16, 56.99 Mbp – 57.02 Mbp) with HDL-C.

Meta-analysis of discovery cohorts

The association results of all discovery cohorts for all variants within the *CETP* region (chromosome 16, 56.99 Mbp – 57.02 Mbp) were combined using inverse variance weighting as applied by METAL31. This tool also applies genomic control by automatically correcting the test statistics to account for small amounts of population stratification or unaccounted relatedness and the tool also allows for heterogeneity. We used the following filters for the variants: 0.3 < R2 (measurement for the imputation quality) < 1.0 and expected minor allele count (expMAC = 2 · MAF · R2 · sample size) > 10 prior to meta-analysis. After meta-analysis of all available variants, we excluded the variants that were not present in at least 3 cohorts, to prevent false positive findings.

Selection of independent variants

In order to select only variants that were independently associated with HDL-C, we used the Genome-wide Complex Trait Analysis (GCTA) tool, version 1.13[31]. Although this tool currently supports multiple functionalities, we only used the functions for conditional and joint

genome-wide association analysis. This function performs a stepwise selection procedure to select independent SNP associations by a conditional and joint analysis approach. It utilizes summary-level statistics from the meta-analysis and linkage disequilibrium (LD) corrections between SNPs are estimated from the 1000 Genomes (1000G Phase I Integrated Release Version 22 Haplotypes (2010-11 data freeze, 2012-02-14 haplotypes)). GCTA estimates the effective sample size and determines the effect size, the standard error and the p-value from a joint analysis of all the selected SNPs. In this way we select the best associated variants in *CETP*. We subsequently checked whether these variants were in LD within the 1000 Genomes reference panel (1000G Phase I Integrated Release Version 22 Haplotypes (2010-11 data freeze, 2012-02-14 haplotypes)) using PLINK[32] software (Supplemental Table 3).

Replication of independent *CETP* variants

Five variants were selected for replication in a sample of 12 independent cohorts: Athero-Express, CHS, FINCAVAS, LBC1936, Lifelines, LLS, NTR-NESDA, PREVEND, PROSPER, QIMR, TRAILS and YFS. The lipid measurements were adjusted for sex, age and age2 in all cohorts and if necessary also for cohort-specific covariates (Supplemental Table 1b). The details per cohort regarding variant genotyping and imputations can be found in Supplemental Table 2. The association results of all replication cohorts were combined and the standard error based weights were calculated by METAL[33]. Since none of the five variants are in LD (Supplemental Table 3), the Bonferroni-corrected p-value for multiple testing was 0.01.

Test previous published results

The meta-analysis of HDL-C as published by Teslovich *et al.*[9] identified 38 genome-wide significant (p-value < $5 \cdot 10^{-8}$) variants within the *CETP* region (chromosome 16, 56.99 Mbp – 57.02 Mbp). Within all discovery and replication cohorts, we tested these 38 variants, adjusting for the 5 newly identified independent variants to explore whether the new variants explain previously published results. The association results of all cohorts were combined and the standard error based weights were calculated by METAL[33].

We used the genotypes of all 1,092 individuals of the 1000 Genomes project (1000G Phase I Integrated Release Version 22 Haplotypes (2010-11 data freeze, 2012-02-14 haplotypes)) to calculate the correlation between the 38 variants. This correlation matrix was used by matSpDlite[34] which examines the ratio of observed eigenvalue variance to its theoretical maximum to determine the number of independent variables. For these 38 genome-wide significant variants within the *CETP* region, the effective number of independent variables is 18 and therefore the experiment-wide significance threshold required to keep type I error rate at 5% is $2.85 \cdot 10^{-3}$.

Conditional analysis of independent *CETP* variants

The replicated independent variants were selected for conditional analysis in both the discovery and the replication cohorts. In this analysis we adjusted for the lead SNP for this region as reported by Teslovich *et al.*[9] (rs3764261, chromosome 16, position 56,993,324 basepairs). The association results of all discovery and replication cohorts were combined and the standard error based weights were calculated by METAL[33]. The Bonferroni-corrected p-value for multiple testing was 0.01, since none of the five variants is in LD (Supplemental Table 3).

Validation of the new CETP insertion within a family

Within the ERF study, 3,658 individuals have been genotyped on various Illumina and Affymetrix chips, followed by imputations with MaCH (1.0.18c) and Minimac (minimac-beta-2012-03-14) to the 1000 Genomes reference panel (1000G Phase I Integrated Release Version 22 Haplotypes (2010-11 data freeze, 2012-02-14 haplotypes)). Based on the best guess imputed genotypes, we selected one family in which we expected the insertion to segregate.

Validation of the insertion was performed by Sanger sequencing. Genomic DNA was isolated from peripheral blood using standard protocols (salting-out). The intron 2-3 of the *CETP* gene (Supplemental Table 4) was amplified using PCR and the following primer sequences were used to amplify: forward; tgggggactcaggtctctcc; reverse; aaagcacctggcccacaacc; size 409 bp. PCR reactions was performed in 17.5 µl containing 37.5 ng DNA, 10 pmol/ul of each primer, 2.5 mM dNTP's, 10x PCR buffer with Mg+ (Roche) and 5 U/ul FastStart Taq (Roche). Cycle conditions: 7 min at 94°C; 10 cycles of 30 s denaturation at 94°C, 30 s annealing at 70°C to 1°C per cycle and 90s extension at 72°C; followed by 20 cycles of 30 s denaturation at 94°C, 30 s at 60°C and 90 s at 72°C; final extension 10 min at 72°C. Sephadex G50 (Amersham Biosciences) was used to purify the sequenced PCR products. Direct sequencing of both strands was performed using Big Dye Terminator chemistry version 4 (Applied Biosystems). Fragments were loaded on an ABI3100 automated sequencer and analyzed with DNA Sequencing Analysis (version 5.3) and SeqScape (version 2.6) software (Applied Biosystems). All sequence variants are numbered at the nucleotide levels according to the following references: NC_000016.10:g.56963437_56963438insA (NCBI), NM_000078.2:c.233+313_233+314insA, Human Feb. 2009 (GRCh37/hg19) Assembly.

RESULTS

Meta-analysis in all discovery cohorts to select independent variants

The association of all variants within the *CETP* region (chromosome 16, 56.99 Mbp – 57.02 Mbp) to HDL-C was tested in all discovery cohorts. These results were combined using the inverse-variance weights as applied by METAL[33]. After exclusion of the variants that were not present in at least 3 cohorts, 254 variants remained (Figure 1). A conditional and joint analysis of the 254 variants using GCTA identified five independent variants (Figure 2). Three variants were intronic (rs5817082, rs4587963 and rs7499892), one variant was intergenic (rs12920974) and one variant was exonic (rs34065661) (Table 1). Using PLINK software, we calculated the LD between the 5 variants based on the 1000 Genomes reference panel (1000G Phase I Integrated Release Version 22 Haplotypes (2010-11 data freeze, 2012-02-14 haplotypes)), and found that none are in high LD with each other (Supplemental Table 3).

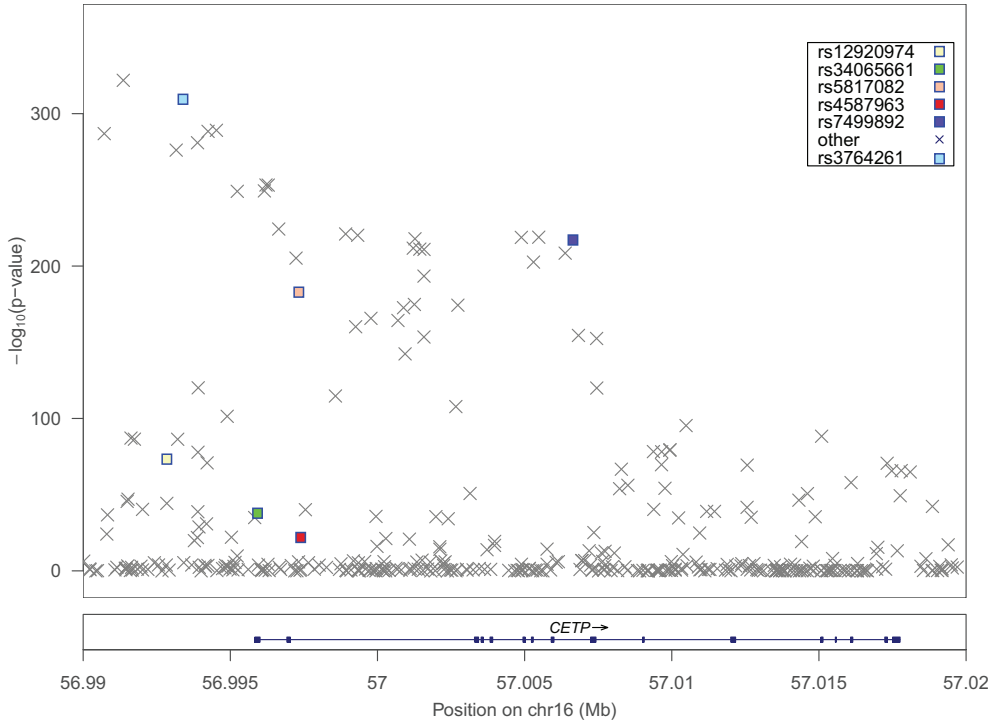
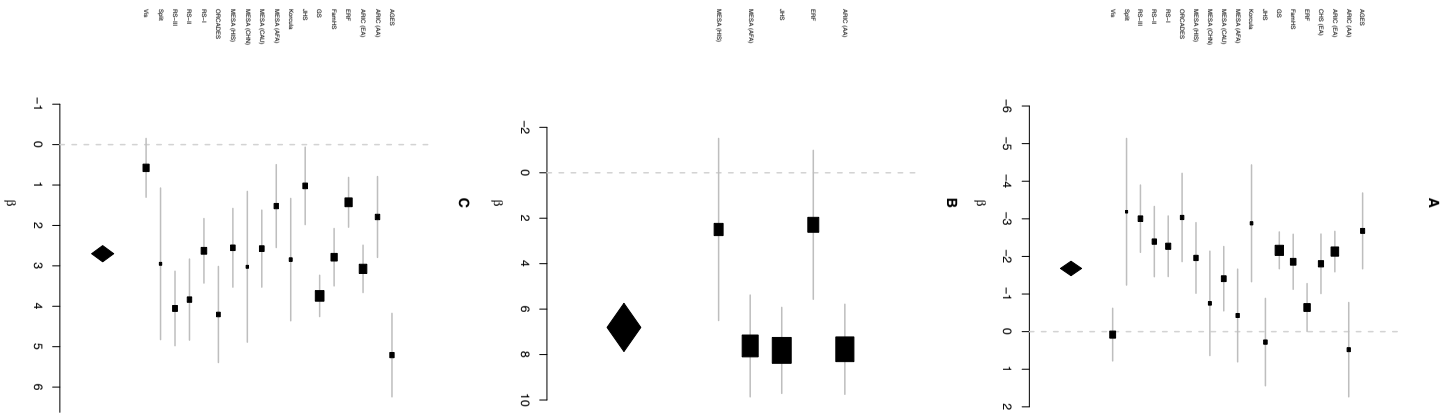


Figure 1. Results of the meta-analysis of all discovery cohorts within the *CETP* region.

Table 1. The five independent variants after meta-analysis in the discovery cohorts.

MarkerName	Chr	Position	EA*	Type	Freq†	After meta-analysis			$Freq_{geno}$ §	After GCTA analysis		
						β‡	SE_{β}	*p*-value		β_j ‡	$SE_{\beta j}$	$p\text{-value}_j$
rs12920974	16	56,993,025	T	SNP	0.271	-1.748	0.096	1.41E-74	0.281	-1.806	0.139	2.40E-38
rs34065661	16	56,995,935	G	SNP	0.058	7.203	0.560	7.04E-38	0.020	6.782	0.582	2.23E-31
rs5817082	16	56,997,349	CA	INDEL	0.285	-2.869	0.098	8.95E-187	0.305	-4.286	0.172	1.55E-137
rs4587963	16	56,997,369	A	SNP	0.240	-0.972	0.101	5.25E-22	0.261	-2.014	0.165	2.11E-34
rs7499892	16	57,006,590	T	SNP	0.209	-3.384	0.107	2.94E-218	0.245	-2.083	0.150	1.31E-43

* EA (effect allele): the allele for which the effect on HDL-C is estimated.
† Freq: the frequency of effect allele in the discovery cohorts.
‡ β: the effect of the effect allele. Bj: the effect of the effect allele after joint analysis of all selected variants by GCTA.
§ $Freq_{geno}$: the frequency of the variant within the reference panel (1000G Phase I Integrated Release Version 22 Haplotypes (2010-11 data freeze, 2012-02-14 haplotypes)).



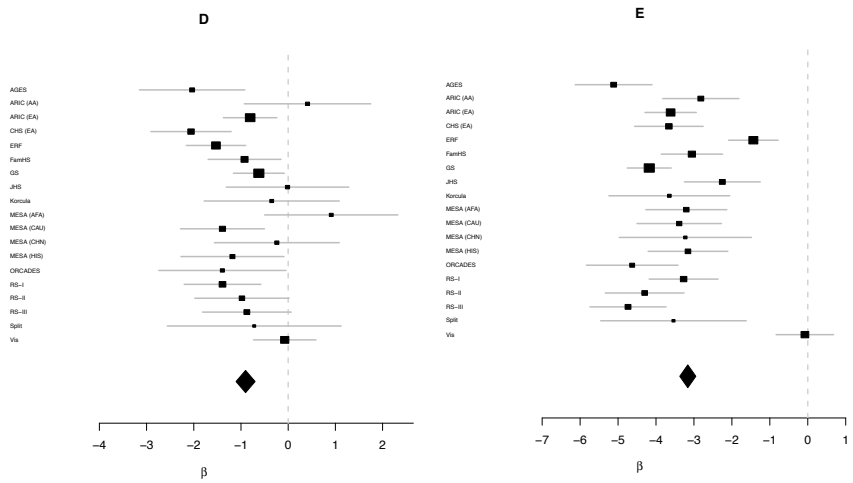


Figure 2. Forest plots from the discovery meta-analysis results for the five independent variants identified within the *CETP* region. Only cohorts in which the variants passed QC are included in the forest plot. A: rs12920974 (chromosome 16, position 56,993,025), B: rs34065661 (chromosome 16, position 56,995,935), C: rs5817082 (chromosome 16, position 56,997,349), D: rs4587963 (chromosome 16, position 56,997,369) and E: rs7499892 (chromosome 16, position 57,006,590).

Replication of the independent *CETP* variants

The five independent variants within the *CETP* region were selected for replication within the following cohorts: Athero-Express, CHS, FINCAVAS, LBC1936, Lifelines, LLS, NTR-NESDA, PREVEND, PROSPER, QIMR, TRAILS and YFS. Five variants were replicated at a *p*-value of $2.99 \cdot 10^{-34}$ (Figure 3 and Table 2).

Test to explain the previously published results

In each discovery and replication cohort we tested if the five independent variants explain the associations within the *CETP* region (chromosome 16, 56.99 Mbp – 57.02 Mbp) as reported in Teslovich *et al.*[9]. We tested a total of 38 genome-wide significant (*p*-value < $5 \cdot 10^{-8}$) SNPs within this region identified by Teslovich *et al.*[9] and conditioned for the five independent variants in all discovery and replication cohorts. All 38 variants were significantly (*p*-value corrected for multiple testing < $2.85 \cdot 10^{-3}$) associated with HDL-C in our joint analyses without adjusting for the five independent variants we identified in this work, and 37 (97.37%) were genome-wide significant (*p*-value < $5 \cdot 10^{-8}$) despite the fact that our sample size is about 65% of the study of Teslovich *et al.*[9] (Table 3). When conditioning on the 5 variants identified in this work, 27 (71.05%) variants remained significant (*p*-value < $2.85 \cdot 10^{-3}$), though the p-values were markedly reduced (Table 3). This finding suggests that the new variants we identified may explain in part the previously reported association. Remarkably, the *p*-value of rs3764261 which was reported as the lead SNP for this *CETP* region by Teslovich *et al.*[9] was highly reduced from $5.25 \cdot 10^{-509}$ to $9.51 \cdot 10^{-25}$ while the β decreased from 3.179 mg/dL to

0.859 mg/dL. This variant is not in LD with any of the 5 new variants. Due to the lack of LD, the standard error of rs3764261 does not change much (SE_{unadj}=0.066, SE_{adj}=0.084), but the effect of rs3764261 does (β_{unadj}=3.179, β_{adj}=0.859) and therefore the chi-square decreases as well, and that results in a higher *p*-value. This indicates that a part of the effect of rs3764261 can be explained by the effect of the 5 new variants.

Conditional analysis of the independent *CETP* variants

Next, we performed conditional analysis of the independent variants in both the discovery and replication cohorts. We conditioned on the lead SNP for the *CETP* region as reported by Teslovich *et al.*[9] (rs3764261, chromosome 16, position 56,993,324 basepairs), see Table 4 and Figure 4. This analysis showed that three out of the five variants (rs34065661, rs5817082, rs7499892) are independent of rs3764261. For all variants the *p*-values and β's decreased, but all *p*-values remained significant. The effect of the single variant rs34065661, of the insertion rs5817082 and of the single variant rs7499892 were reduced by 53.20%, 38.48% and 32.67%, respectively.

Validation of the insertion within a family

We selected based on the best guess imputations of the ERF study, a large family of 30 individuals for Sanger sequencing of rs5817082. Using MERLIN[36] we estimated that the total heritability of HDL-C within this family is 27.47%. DNA was available for 16 individuals. Figure 5 shows the results of the Sanger sequencing for rs5817082 for these 16 individuals within the family. The sequencing of the insertion confirmed the best guess results for ten individuals (62.5%), of which seven were heterozygous for the insertion, one was homozygous for the insertion and two did not carry the insertion. Three individuals that are homozygous for the insertion, were predicted to be heterozygous by the best guess imputations. Three individuals that are heterozygous for the insertion, were not predicted to carry the insertion by the best guess imputations. Furthermore, the Sanger sequencing showed that the insertion segregates with the outcome within this family. The proportion of variance explained by the insertion within this family is 35.50%, while the proportion explained by rs3764261, the lead SNP within the *CETP* region as reported by Teslovich *et al.*[9] is 14.11%.

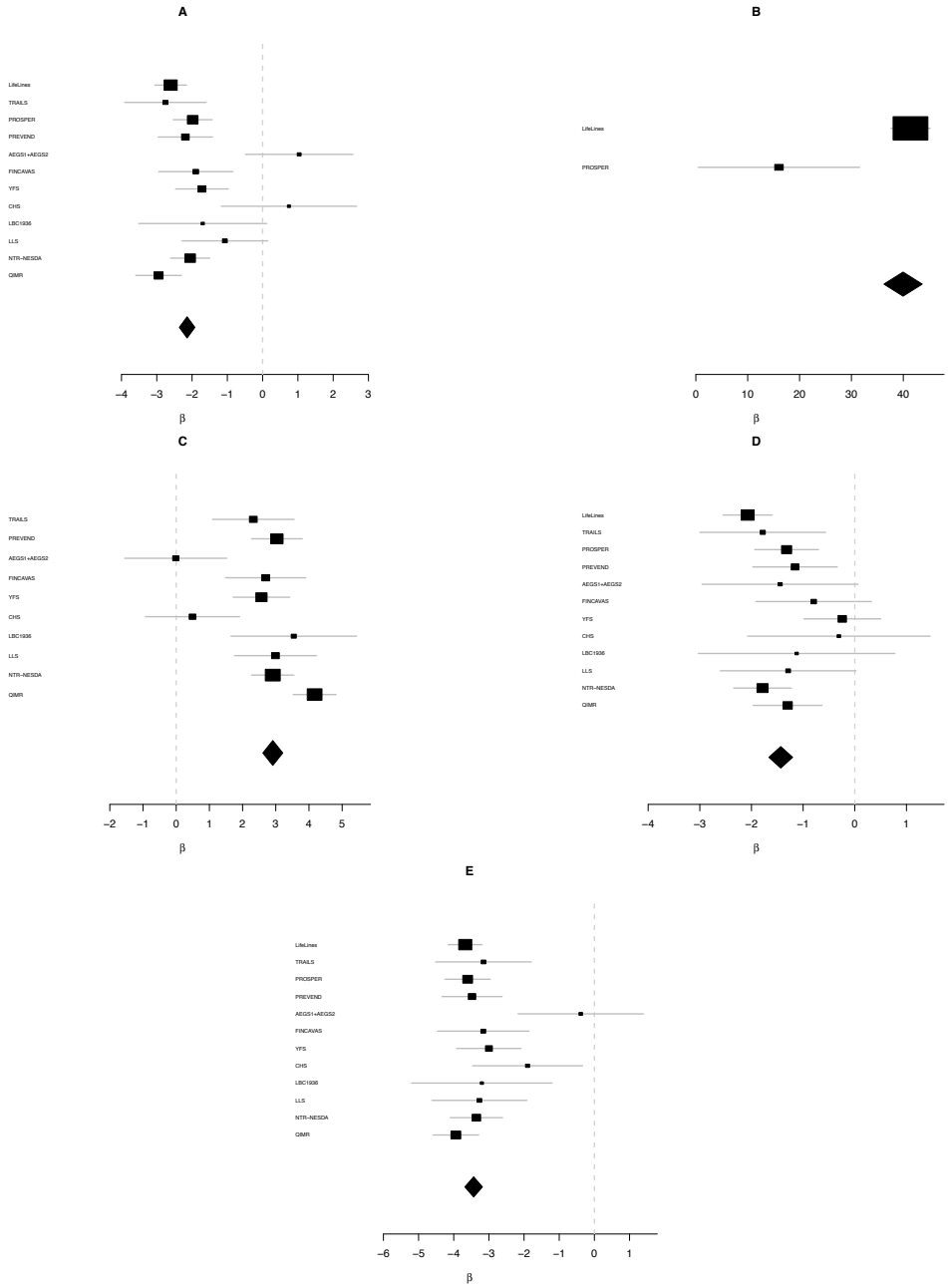


Figure 3. Forest plots of the replication meta-analysis for the five independent variants within the *CETP* region. Only cohorts in which the variants passed QC are included in the forest plot. A: rs12920974 (chromosome 16, position 56,993,025), B: rs34065661 (chromosome 16, position 56,995,935), C: rs5817082 (chromosome 16, position 56,997,349), D: rs4587963 (chromosome 16, position 56,997,369) and E: rs7499892 (chromosome 16, position 57,006,590).

Table 2. Replication of the 5 independent variants within the *CETP* region.

MarkerName	Chr	Position	Effect allele*	Non effect allele	Freq†	β‡	SE_β	*p*-value	Direction§											
rs12920974	16	56,993,025	T	G	0.288	-2.140	0.112	3.36E-81	-	-	-	-	-	-	-	-	-	-	-	-
rs34065661	16	56,995,935	G	C	0.018	39.958	1.884	8.46E-100	?	?	?	?	+	?	?	?	+	?	?	?
rs5817082	16	56,997,349	CA	C	0.229	-2.911	0.153	1.09E-80	+	-	-	-	?	-	-	-	?	-	-	-
rs4587963	16	56,997,369	A	T	0.325	-1.433	0.117	2.99E-34	-	-	-	-	-	-	-	-	-	-	-	-
rs7499892	16	57,006,590	T	C	0.257	-3.434	0.127	5.64E-160	-	-	-	-	-	-	-	-	-	-	-	-

* EA (effect allele): the allele for which the effect on HDL-C is estimated.
† Freq: the frequency of effect allele in the replication cohorts.
‡ β: the effect of the effect allele.
§ Direction of the effect of the effect allele of the following cohorts: AEGS, CHS (AA), FINCAVAS, LBC1936, Lifelines, LLS, NTR-NESDA, PREVEND, PROSPER, QIMR, TRAILS, YFS. The question marks means that the variant was removed prior to meta-analysis due to a low imputation quality and/or expMAC below 10.

Table 3. Unadjusted and conditional analysis of the Teslovich variants on the five independent variants in the combined analysis of all discovery and replication cohorts.

MarkerName	Chr	position	EA*	NEA	Unadjusted analysis				Adjusted analysis			
					Freq†	β‡	SE_β	*p*-value	Freq†	β‡	SE_β	*p*-value
rs6499861	16	56,991,495	C	G	0.758	1.432	0.090	5.63E-57	0.781	1.083	0.106	1.47E-24
rs6499863	16	56,992,017	A	G	0.251	-1.420	0.093	1.02E-52	0.227	-1.162	0.112	2.59E-25
rs12708967	16	56,993,211	T	C	0.726	2.419	0.087	9.61E-170	0.768	-0.363	0.110	9.99E-04
rs3764261	16	56,993,324	A	C	0.409	3.179	0.066	5.25E-509	0.358	0.859	0.084	9.51E-25
rs12447839	16	56,993,935	T	C	0.665	1.215	0.077	1.87E-56	0.738	0.302	0.111	6.35E-03
rs12447924	16	56,994,192	T	C	0.683	1.218	0.077	8.54E-57	0.737	0.321	0.109	3.15E-03
rs4783961	16	56,994,894	A	G	0.496	1.680	0.064	9.60E-152	0.493	0.732	0.073	6.73E-24
rs4783962	16	56,995,038	T	C	0.318	-1.178	0.081	1.51E-48	0.255	-0.288	0.123	1.97E-02
rs1800775	16	56,995,236	A	C	0.471	2.788	0.064	2.12E-416	0.495	0.547	0.088	4.97E-10
rs711752	16	56,996,211	A	G	0.445	2.782	0.064	3.93E-414	0.435	0.396	0.083	1.56E-06
rs1864163	16	56,997,233	A	G	0.311	-2.991	0.076	1.33E-340	0.238	-0.307	0.115	7.75E-03

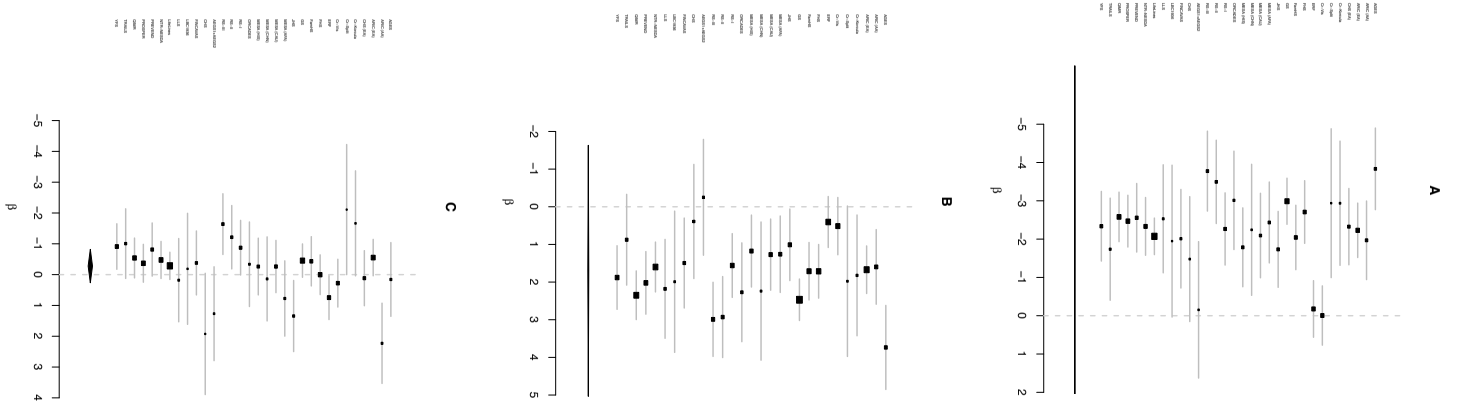
rs9929488	16	56,998,572	C	G	0.338	-2.189	0.075	7.55E-189	0.308	0.125	0.092	1.76E-01
rs7203984	16	56,999,258	A	C	0.693	2.903	0.080	2.44E-287	0.737	0.076	0.112	4.95E-01
rs11508026	16	56,999,328	T	C	0.417	2.703	0.065	1.27E-383	0.407	0.326	0.082	7.60E-05
rs820299	16	57,000,284	A	G	0.578	0.892	0.066	8.60E-42	0.595	0.336	0.084	6.07E-05
rs12597002	16	57,002,404	A	C	0.389	-1.228	0.071	2.02E-66	0.307	-0.481	0.103	3.25E-06
rs9926440	16	57,002,663	C	G	0.371	-2.141	0.072	1.18E-196	0.351	0.131	0.085	1.26E-01
rs9939224	16	57,002,732	T	G	0.288	-2.944	0.080	2.72E-300	0.229	0.051	0.109	6.41E-01
rs11076174	16	57,003,146	T	C	0.797	2.388	0.123	1.70E-83	0.825	0.496	0.133	1.99E-04
rs7205804	16	57,004,889	A	G	0.440	2.644	0.063	1.63E-386	0.422	0.291	0.082	3.51E-04
rs1532624	16	57,005,479	A	C	0.420	2.639	0.063	6.82E-386	0.412	0.291	0.082	3.48E-04
rs11076175	16	57,006,378	A	G	0.740	3.326	0.084	5.05E-342	0.815	-0.031	0.127	8.05E-01
rs7499892	16	57,006,590	T	C	0.323	-3.227	0.084	6.95E-323	0.241	-0.197	0.119	9.74E-02
rs289714	16	57,007,451	A	G	0.669	2.624	0.085	6.46E-208	0.708	0.540	0.101	1.01E-07
rs289715	16	57,008,508	A	T	0.256	2.047	0.106	5.38E-83	0.245	0.420	0.106	7.37E-05
rs289717	16	57,009,388	A	G	0.422	-1.357	0.068	1.39E-89	0.401	-0.353	0.077	4.15E-06
rs289719	16	57,009,941	T	C	0.383	1.701	0.070	2.85E-132	0.374	0.461	0.072	1.32E-10
rs4784744	16	57,011,185	A	G	0.396	-1.319	0.066	1.05E-87	0.386	-0.350	0.074	2.37E-06
rs4784745	16	57,014,875	A	G	0.614	1.327	0.068	5.66E-85	0.626	0.314	0.075	3.21E-05
rs5880	16	57,015,091	C	G	0.135	-4.495	0.175	4.42E-146	0.119	-1.331	0.181	1.92E-13
rs5882	16	57,016,092	A	G	0.613	-1.442	0.067	4.19E-102	0.614	-0.410	0.069	2.39E-09
rs9923854	16	57,017,002	T	G	0.802	-1.391	0.115	1.07E-33	0.805	-0.543	0.117	3.28E-06
rs289741	16	57,017,474	A	G	0.631	-1.547	0.068	3.37E-113	0.633	-0.476	0.070	1.02E-11
rs1801706	16	57,017,662	A	G	0.276	1.040	0.091	1.82E-30	0.270	0.493	0.095	1.92E-07
rs289742	16	57,017,762	C	G	0.295	1.811	0.098	1.21E-76	0.285	0.407	0.098	3.40E-05
rs289744	16	57,018,102	T	G	0.641	-1.544	0.069	4.99E-110	0.643	-0.469	0.071	3.33E-11
rs12720917	16	57,019,392	T	C	0.769	-1.474	0.110	1.15E-40	0.775	-0.377	0.109	5.43E-04
rs289745	16	57,019,532	A	C	0.579	0.276	0.081	6.82E-04	0.581	0.204	0.081	1.12E-02

* EA (effect allele): the allele for which the effect on HDL-C is estimated.
† Freq: the frequency of the effect allele.
‡ β: the effect of effect allele.

Table 4. Analysis of the independent variants within the *CETP* region conditioned on the lead SNP for the *CETP* region as reported by Teslovich *et al.*[9] (rs3764261) in the combined analysis of all discovery and replication cohorts.

MarkerName	Chr	Position	EA*	NEA	Unadjusted analysis				Adjusted analysis			
					Freq†	β‡	SE_β	*p*-value	Freq†	β‡	SE_β	*p*-value
rs12920974	16	56,993,025	T	G	0.344	-1.880	0.074	9.91E-143	0.336	-0.278	0.076	2.82E-04
rs34065661	16	56,995,935	C	G	0.854	-9.333	0.520	6.02E-72	0.838	-4.368	0.550	1.94E-15
rs5817082	16	56,997,349	CA	C	0.360	-2.765	0.085	1.49E-231	0.351	-1.701	0.086	2.16E-86
rs4587963	16	56,997,369	A	T	0.351	-1.133	0.077	1.62E-48	0.339	0.309	0.079	8.81E-05
rs7499892	16	57,006,590	T	C	0.317	-3.275	0.082	2.90E-346	0.304	-2.205	0.083	5.14E-156

* EA (effect allele): the allele for which the effect on HDL-C is estimated.
† Freq: the frequency of the effect allele.
‡ β:the effect of the effect allele.



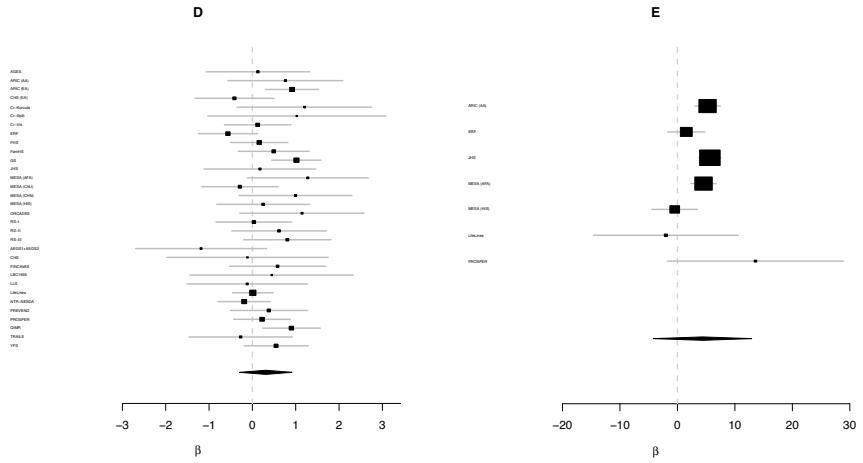


Figure 4. Forest plots of the conditional analysis in the combined discovery and replication cohorts for the five independent variants within the *CETP* region. Only cohorts in which the variants passed QC are included in the forest plot. A: rs12920974 (chromosome 16, position 56,993,025), B: rs34065661 (chromosome 16, position 56,995,935), C: rs5817082 (chromosome 16, position 56,997,349), D: rs4587963 (chromosome 16, position 56,997,369) and E: rs7499892 (chromosome 16, position 57,006,590).

DISCUSSION

We conducted an analysis to fine-map the association between *CETP* genetic variants and HDL-C. To this end, a total of 59,432 samples were imputed to the latest version of the 1000 Genomes (version Phase 1 integrated release v3, April 2012, all populations). We identified and replicated five independent variants within the *CETP* region (chromosome 16, 56.99 Mbp – 57.02 Mbp), of which four are SNPs and one is an insertion. We validated the insertion by Sanger sequencing within a large family, as the largest effect on HDL-C comes from this insertion.

The relationship between the *CETP* gene and HDL-C has been known for a long time[9] and GWAS have revealed many common and rare variants in this region. Although the associated genetic variants are strongly correlated with HDL-C, the causal variants have not been determined. Our study showed that when using the latest 1000 Genomes reference panel, we have more power to fine-map this association. By conditional analysis of the five variants, we were able to reduce the p-values of the genome-wide significant associations published before by Teslovich *et al.*[9]. Furthermore, conditional analysis showed that three out of the five variants are independent of the lead SNP for the *CETP* region as reported by Teslovich *et al.*[9] (rs3764261).

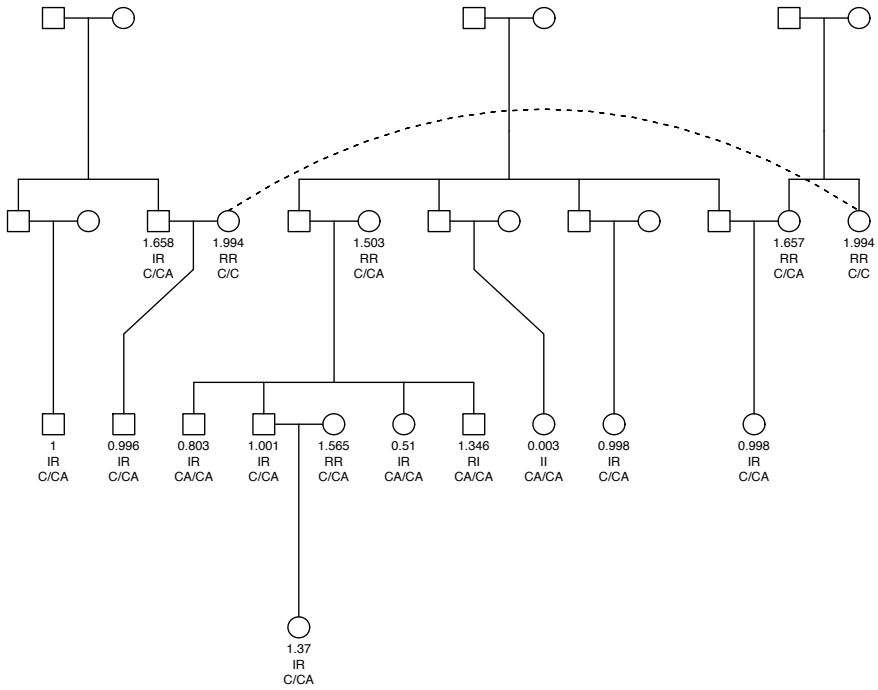


Figure 5. Validation of the deletion (rs5817082) with a large family. The numbers present the dosage for rs5817082 after imputations, second row the best guess result (I is insertion, R is reference) and the third row the genotypes of the deletion from Sanger sequencing.

Several fine-mapping effort have been previously published[37,38] and in all those efforts sequencing was used for the fine-mapping. In our project we did not use sequencing, but imputations using the 1000 Genomes as a reference panel. This method has been widely-used in the past and is much lower in cost. With new reference panels available, we were able to have a revised study of this region. The 1000 Genomes reference panel consists of 30 million variants including a million insertions and deletions. By using this reference panel for imputation, we were able to impute these insertions and deletions in 59,432 samples from various cohorts. This led to the significant association of an insertion within a known region with HDL-C. So far, no association between a structural variation and HDL-C has been found in such a large sample size. Validation of the insertion by Sanger sequencing confirms the correct imputations of this insertion in 62.5% of the individuals, of which seven heterozygous carriers, one homozygous carrier and two did not carry the insertion.

The results of this study showed that by using the 1000 Genomes reference panel, the proportion of the variance explained can be increased and that multiple common variants in the same region may be implicated in a single family of the ERF study. The insertion we identified in this study explains 35.50% of variation in the HDL-C level in a single family of the

ERF study; this is in concordance with the results of the whole-genome sequence data[23]. This is much higher than the proportion of the variance explained (14.11%) in the same family by rs3764261 which was reported before as the lead variant of this region. Fine-mapping of various associations may help us in unravel the genetic background of various phenotypes.

Although rs3764261 was identified by Teslovich *et al.*[9] to be the lead SNP of this region, other variants are used in clinical settings. Three of the classical variants are located in the promoter region of the *CETP* gene: -1337C/T (rs708272 or Taq1B), -971G/A and -629C/A (rs1800775) polymorphisms[39]. Carriers of the B2 allele of the common Taq1B polymorphism exhibit lower plasma CETP levels and higher HDL-C. Furthermore, a recent meta-analysis showed that the B2 allele is associated with a reduced risk for coronary heart disease[40]. One more classical variant is rs5882A (405I/V), which is located outside the promoter region[41]. The -1337C/T and -629C/A are in strong linkage disequilibrium (LD), however, they are in very low LD (r2 of 0.442 for rs708272 and 0.461 for rs1800775) with rs3764261, despite the fact that all three variant are within 3,000 basepairs of each other.

Large HDL-C particle sizes have been associated with exceptional longevity before and with an increased homozygosity for the I405V variant within the *CETP* gene[1-4]. Many of the studies confirm this relationship, however, all are based on genotyping of the I405V variant. Our study however shows that more variants within the *CETP* gene are associated with HDL-C levels in the blood circulation. Therefore we would suggest investigating more variants within the *CETP* gene for its association with longevity and healthy aging.

Some genetic variants identified in our study were published before[42,43], but so far no conditional analyses have been performed with these variants. Our study suggests that various *CETP* variants may be relevant for HDL-levels in the blood circulation and that these may have a substantial role in the heritability of HDL-C in specific families.

REFERENCES

1. Barzilai, N. *et al.* Genetic studies reveal the role of the endocrine and metabolic systems in aging. *J Clin Endocrinol Metab* **95**, 4493–4500 (2010).
2. Barzilai, N., Huffman, D. M., Muzumdar, R. H. & Bartke, A. The critical role of metabolic pathways in aging. *Diabetes* **61**, 1315–1322 (2012).
3. Vergani, C. *et al.* I405v polymorphism of the cholesteryl ester transfer protein (*CETP*) gene in young and very old people. *Arch Gerontol Geriatr* **43**, 213–221 (2006).
4. Barzilai, N. *et al.* Unique lipoprotein phenotype and genotype associated with exceptional longevity. *JAMA* **290**, 2030–2040 (2003).
5. Hayek, T. *et al.* Decreased early atherosclerotic lesions in hypertriglyceridemic mice expressing cholesteryl ester transfer protein transgene. *J Clin Invest* **96**, 2071–2074 (1995).
6. Briand, F. *et al.* Anacetrapib and dalcetrapib differentially alters HDL metabolism and macrophage-to-feces reverse cholesterol transport at similar levels of CETP inhibition in hamsters. *Eur J Pharmacol* **740**, 135–143 (2014).
7. Kee, P. *et al.* Effect of inhibiting cholesteryl ester transfer protein on the kinetics of high-density lipoprotein cholesteryl ester transport in plasma: in vivo studies in rabbits. *Arterioscler Thromb Vasc Biol* **26**, 884–890 (2006).
8. Zhong, S. *et al.* Increased coronary heart disease in Japanese-American men with mutation in the cholesteryl ester transfer protein gene despite increased HDL levels. *J Clin Invest* **97**, 2917–2923 (1996).
9. Teslovich, T. M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707–713 (2010).
10. Global Lipids Genetics Consortium *et al.* Discovery and refinement of loci associated with lipid levels. *Nat Genet* **45**, 1274–1283 (2013).
11. Siebel, A. L. *et al.* Effects of high-density lipoprotein elevation with cholesteryl ester transfer protein inhibition on insulin secretion. *Circ Res* **113**, 167–175 (2013).
12. Remaley, A. T., Norata, G. D. & Catapano, A. L. Novel concepts in HDL pharmacology. *Cardiovasc Res* **103**, 423–428 (2014).
13. Joy, T. R. & Hegele, R. A. The failure of torcetrapib: what have we learned? *Br J Pharmacol* **154**, 1379–1381 (2008).
14. Barter, P. J. *et al.* Effects of torcetrapib in patients at high risk for coronary events. *N Engl J Med* **357**, 2109–2122 (2007).
15. Snieder, H., van Doornen, L. J. & Boomsma, D. I. Dissecting the genetic architecture of lipids, lipoproteins, and apolipoproteins: lessons from twin studies. *Arterioscler Thromb Vasc Biol* **19**, 2826–2834 (1999).
16. Friedlander, Y., Kark, J. D. & Stein, Y. Biological and environmental sources of variation in plasma lipids and lipoproteins: the Jerusalem Lipid Research Clinic. *Hum Hered* **36**, 143–153 (1986).
17. Souren, N. Y. *et al.* Anthropometry, carbohydrate and lipid metabolism in the East Flanders Prospective Twin Survey: heritabilities. *Diabetologia* **50**, 2107–2116 (2007).
18. Sung, J., Lee, K. & Song, Y.-M. Heritabilities of the metabolic syndrome phenotypes and related factors in Korean twins. *J Clin Endocrinol Metab* **94**, 4946–4952 (2009).
19. Almgren, P. *et al.* Heritability and familiality of type 2 diabetes and related quantitative traits in the Botnia Study. *Diabetologia* **54**, 2811–2819 (2011).
20. Vattikuti, S., Guo, J. & Chow, C. C. Heritability and genetic correlations explained by common SNPs for metabolic syndrome traits. *PLoS Genet* **8**, e1002637 (2012).
21. Zhou, X., Carbonetto, P. & Stephens, M. Polygenic modeling with bayesian sparse linear mixed models. *PLoS Genet* **9**, e1003264 (2013).

22. Browning, S. R. & Browning, B. L. Identity-by-descent-based heritability analysis in the Northern Finland Birth Cohort. *Hum Genet* **132**, 129–138 (2013).
23. Morrison, A. C. *et al.* Whole-genome sequence-based analysis of high-density lipoprotein cholesterol. *Nat Genet* **45**, 899–901 (2013).
24. Peloso, G. M. *et al.* Association of low-frequency and rare coding-sequence variants with blood lipids and coronary heart disease in 56,000 whites and blacks. *Am J Hum Genet* **94**, 223–232 (2014).
25. Singaraja, R. R. *et al.* Identification of four novel genes contributing to familial elevated plasma HDL cholesterol in humans. *J Lipid Res* **55**, 1693–1701 (2014).
26. Misra, A. & Shrivastava, U. Obesity and dyslipidemia in South Asians. *Nutrients* **5**, 2708–2733 (2013).
27. Sun, L. *et al.* Gene-gene interaction between CETP and APOE polymorphisms confers higher risk for hypertriglyceridemia in oldest-old Chinese women. *Exp Gerontol* **55**, 129–133 (2014).
28. Walia, G. K. *et al.* Association of common genetic variants with lipid traits in the Indian population. *PLoS One* **9**, e101688 (2014).
29. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* **5**, e1000529 (2009).
30. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & calo R Abecasis, G. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat Genet* **44**, 955–959 (2012).
31. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* **88**, 76–82 (2011).
32. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559–575 (2007).
33. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
34. Li, J. & Ji, L. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity (Edinb)* **95**, 221–227 (2005).
35. Codd, V. *et al.* Identification of seven loci affecting mean telomere length and their association with disease. *Nat Genet* **45**, 422–7, 427e1–2 (2013).
36. Abecasis, G. R., Cherny, S. S., Cookson, W. O. & Cardon, L. R. Merlin–rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet* **30**, 97–101 (2002).
37. Wu, Y. *et al.* Trans-ethnic fine-mapping of lipid loci identifies population-specific signals and allelic heterogeneity that increases the trait variance explained. *PLoS Genet* **9**, e1003379 (2013).
38. Sanna, S. *et al.* Fine mapping of five loci associated with low-density lipoprotein cholesterol detects variants that double the explained heritability. *PLoS Genet* **7**, e1002198 (2011).
39. Le Goff, W. *et al.* A novel cholesteryl ester transfer protein promoter polymorphism (-971g/a) associated with plasma high-density lipoprotein cholesterol levels. interaction with the taqib and -629c/a polymorphisms. *Atherosclerosis* **161**, 269–279 (2002).
40. Boekholdt, S. M. *et al.* Cholesteryl ester transfer protein taqib variant, high-density lipoprotein cholesterol levels, cardiovascular risk, and efficacy of pravastatin treatment: individual patient meta-analysis of 13,677 subjects. *Circulation* **111**, 278–287 (2005).
41. Peloso, G. M. *et al.* Common genetic variation in multiple metabolic pathways influences susceptibility to low HDL-cholesterol and coronary heart disease. *J Lipid Res* **51**, 3524–3532 (2010).
42. Ko, A. *et al.* Amerindian-specific regions under positive selection harbour new lipid variants in Latinos. *Nat Commun* **5**, 3983 (2014).
43. Feitosa, M. F. *et al.* Genetic analysis of long-lived families reveals novel variants influencing high density-lipoprotein cholesterol. *Front Genet* **5**, 159 (2014).

CHAPTER 2.2

Meta-analysis of 49,549 individuals with the 1000 Genomes Project reveals an exonic damaging variant in *ANGPTL4* determining fasting TG levels

Elisabeth M. van Leeuwen, Aniko Sabo, Joshua C. Bis, Jennifer E. Huffman, Ani Manichaikul, Albert V. Smith, Mary F. Feitosa, Serkalem Demissie, Peter K. Joshi, Qing Duan, Jonathan Marten, Jan B. van Klinken, Ida Surakka, Ilja M. Nolte, Weihua Zhang, Hamdi Mbarek, Ruifang Li-Gao, Stella Trompet, Niek Verweij, Evangelos Evangelou, Leo-Pekka Lyytikäinen, Bamidele O. Tayo, Joris Deelen, Peter J. van der Most, Sander W. van der Laan, Dan Arking, Alanna Morrison, Abbas Dehghan, Oscar H. Franco, Albert Hofman, Fernando Rivadeneira, Eric J. Sijbrands, Andre G. Uitterlinden, Josyf C. Mychaleckyj, Archie Campbell, Lynne J. Hocking, Sandosh Padmanabhan, Jennifer A. Brody, Kenneth M. Rice, Charles C. White, Tamara Harris, Aaron Isaacs, Harry Campbell, Leslie A. Lange, Igor Rudan, Ivana Kolcic, Pau Navarro, Tatijana Zemunik, Veikko Salomaa, The LifeLines Cohort Study, Jaspal S. Kooner, Benjamin Lehne, William R. Scott, Sian-Tsung Tan, Eco J. de Geus, Yuri Milaneschi, Brenda W. J. H. Penninx, Gonneke Willemsen, Renée de Mutsert, Ian Ford, Ron T. Gansevoort, Marcelo P. Segura-Lepe, Olli T. Raitakari, Jorma S. Viikari, Kjell Nikus, Terrence Forrester, Colin A. McKenzie, Anton J.M. de Craen, Hester M. de Ruijter, CHARGE Lipids Working Group, Gerard Pasterkamp, Harold Snieder, Albertine J. Oldehinkel, P. Eline Slagboom, Richard S. Cooper, Mika Kähönen, Terho Lehtimäki, Paul Elliott, Pim van der Harst, J. Wouter Jukema, Dennis O. Mook-Kanamori, Dorret I. Boomsma, John C. Chambers, Morris Swertz, Samuli Ripatti, Ko Willems van Dijk, Veronique Vitart, Ozren Polasek, Caroline Hayward, James G. Wilson, James F. Wilson, Vilmundur Gudnason, Stephen S. Rich, Bruce M. Psaty, Ingrid B. Borecki, Eric Boerwinkle, Jerome I. Rotter, L. Adrienne Cupples, Cornelia M. van Duijn.

Accepted for publication in Journal of Medical Genetics.

The first author is willing to distribute the supplemental information for this chapter at your request.

ABSTRACT

So far, more than 170 loci have been associated for circulating lipid levels through genome-wide association studies (GWAS). These associations are largely driven by common loci, their function is often not known, and many are likely to be markers for the causal variants. In order to obtain better estimates for rare functional variants we used the 1000 Genomes Project as a reference panel for the imputations of GWAS data from ~60,000 individuals. Replication in ~90,000 samples resulted in the identification of five new associations with circulating lipid levels at four loci. All four loci are within genes that can be linked biologically to lipid metabolism. One of the variants, rs116843064, is a damaging missense variant within the *ANGPTL4* gene. This study illustrates that GWAS with high-scale imputation may still help us unravel the biological mechanism behind circulating lipid levels.

INTRODUCTION

Genome-wide association studies (GWAS) for circulating lipid levels (high-density lipoprotein cholesterol (HDL-C), low-density lipoprotein cholesterol (LDL-C), total cholesterol (TC) and triglycerides (TG)) have identified over 170 loci[1-3]. These studies have been based on imputations to the HapMap reference panel[2] or primary versions of the 1000 Genomes Project (1kG)[1] or genotyping on the Illumina Exome Chip[3]. None has used imputations with the Phase 1 integrated release v3 of the 1kG which allows the imputation of rare functional variants and structural variations with more precision. Evidence of rare functional variants associated with circulating lipid levels comes from recent studies in which exome sequencing of the *NPC1L1* gene identified rare variants associated with reduced LDL-C levels and reduced risk of coronary heart disease[4]. Moreover, exome sequencing of *LDLR* and *APOA5* identified rare variants associated with an increased LDL-C and increased TG levels[5] and exome sequencing of *APOC3* identified rare variants associated with reduced TG levels and reduced risk of coronary heart disease[6].

Our goal in this study was to identify rare functional variants associated with circulating lipid levels in a larger sample size compared to the exome sequencing of candidate gene approach. To this end, we imputed genotypes for study samples participating in the cohorts of the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) consortium using the Phase 1 integrated release v3 of the 1kG and conducted a meta-analysis of about approximately 60,000 individuals, followed by a replication in an independent set of 90,000 individuals.

METHODS

Please see Supplementary Methods for complete descriptions of the methods. In summary, for the discovery stage of this project, we used the data from 20 cohorts of the CHARGE consortium (Supplemental methods). All cohorts were imputed with reference to the 1kG reference panel (version Phase 1 integrated release v3). The total number of individuals in the discovery stage was 59,409 for HDL-C, 48,780 for LDL-C, 60,024 for TC and 49,549 for TG. Supplemental Table 1 and 2 contain the baseline characteristics per cohort and more details about SNP genotyping and genotype imputations. Within each cohort, each variant was tested for association with each of the lipid traits, assuming an additive genetic model. The association results of all cohorts for all variants were combined using inverse variance weighting. We used the following filters for the variants: $0.3 < R^2$ (measurement for the imputation quality) ≤ 1.0 and expected minor allele count (expMAC=2 · MAF (minor allele frequency) · R^2 · sample size) > 10 prior to meta-analysis. After meta-analysis of all available

variants, we excluded the variants that were not present in at least 4 cohorts, to prevent false positive findings. In order to select only variants that were independently associated with each of the lipid traits, we used the GCTA[7] tool. To identify novel loci we selected from the list of variants identified by GCTA, those variants located more than 0.5Mb away from previously identified loci of the corresponding trait[2,3] and which were significant (p-value < $5 \cdot 10^{-8}$) in the initial discovery stage. To prevent the identification of false positive loci, we added a second replication stage within 23 independent cohorts. The experiment-wide significance threshold required to keep type I error rate within the replication stage at 5% is $2.63 \cdot 10^{-3}$ (Bonferroni correction based on nineteen variants). We also meta-analyzed the individuals of the discovery and replication stage together.

RESULTS

The association of all variants with HDL-C, LDL-C, TC and TG was tested in all discovery cohorts (Supplemental Figure 1 and 2). We significantly replicated 88.1% of the loci described by Teslovich *et al.*[2] despite a sample size of about 80% (Supplemental Figure 5 and Supplemental Table 3). We also significantly replicated 43.4% of the loci described by Global Lipids Genetics Consortium (GLGC)[3] despite a sample size of about 30% (Supplemental Figure 6 and Supplemental Table 4).

A conditional and joint analysis using GCTA identified 186 independent variants for HDL-C, 175 for LDL-C, 215 for TC and 120 for TG. Next, we excluded all variants that were not genome-wide significant (p-value < $5 \cdot 10^{-8}$) in the initial discovery stage as these are probably false positives and we excluded all variants which are within 0.5 Mb of a loci previously published by Teslovich *et al.*[2] or GLGC[3], which resulted in three variants for HDL-C, three for LDL-C, seven for TC and six for TG. These variants are located at seventeen different loci and includes one deletion (Figure 1 and Table 1). These nineteen variants were selected for replication. The total number of individuals in the replication stage was 84,598, 72,486, 83,739 and 73,519 for HDL-C, LDL-C, TC and TG respectively (see Supplemental Table 1 and 2 for baseline characteristics and information about SNP genotyping and imputation details). The sample size in the replication stage was larger than the initial discovery sample for seventeen out of the nineteen variants. The frequencies of the variants were similar between the discovery and replication cohorts. The directions of effect were the same in both the discovery and replication cohorts for sixteen out of the nineteen variants (Supplementary Figure 7). We used a Bonferroni corrected threshold for significance (p-value < $2.63 \cdot 10^{-3}$). Five out of the nineteen variants were significantly replicated (Table 1): rs6457374 (TC), rs186696265 (LDL-C and TC), rs77697917 (HDL-C) and rs116843064 (TG). The frequency of these variants ranging from 0.012 to 0.249 within the discovery sample.

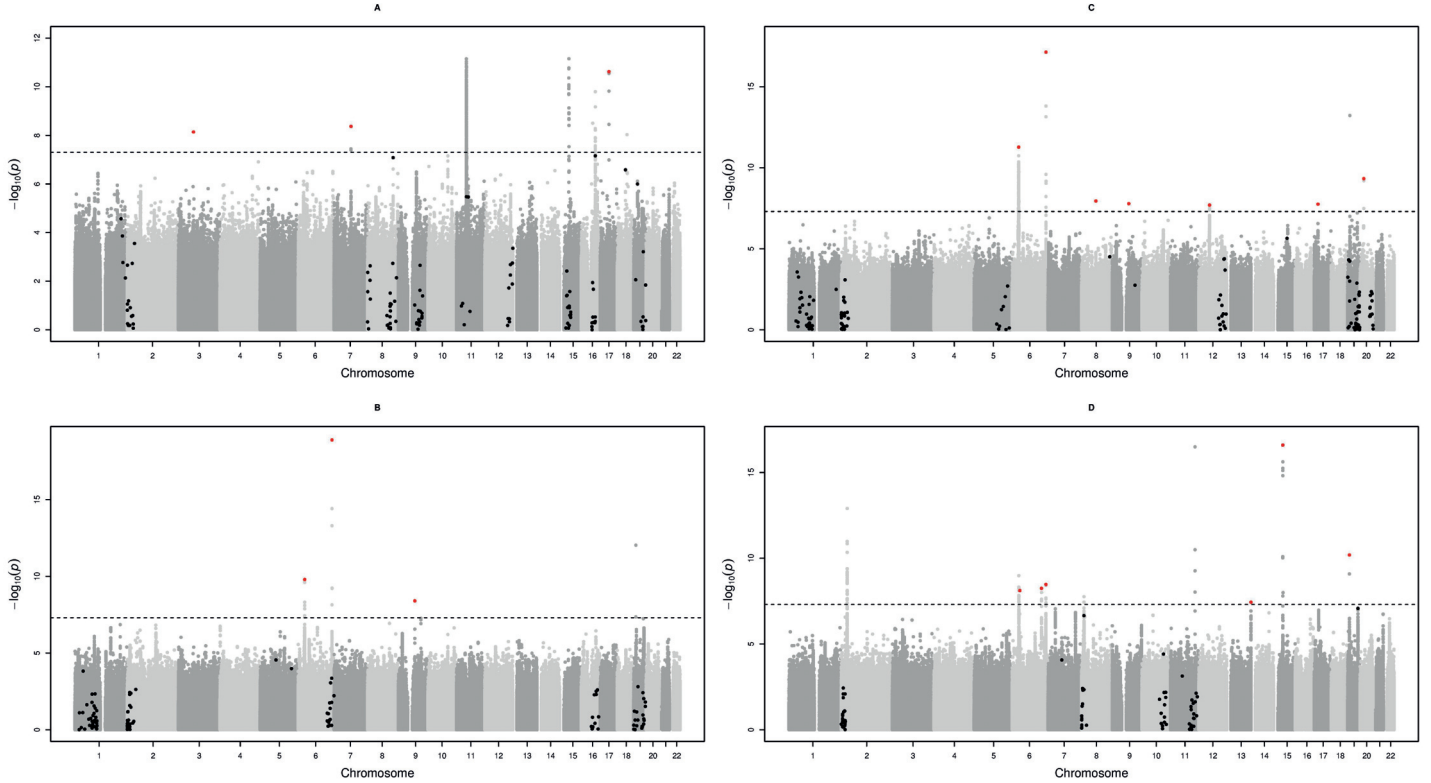


Figure 1: Manhattan plots for HDL-C **(A)**, LDL-C **(B)**, TC **(C)** and TG **(D)** after the meta-analysis of all discovery cohorts. Variants that were present in at least 4 cohorts and that are not within 0.5Mb of a previous published loci[2,3] were included. The black line indicates the genome-wide significant line ($5 \cdot 10^{-8}$), the black and red dots the variants identified by GCTA which are not genome-wide significant and which are genome-wide significant, respectively.

Table 1. The results for the nineteen variants after the meta-analysis of all discovery cohorts, of all replication cohorts and of all cohorts combined. A1 is allele 1 and A2 is allele 2, Freq is the frequency of A1, β is the effect of A1.

Trait	Chr:Position	rs identifier	A1/A2	All cohorts combined			
				Freq	β	SE_β	*p*-value
HDL-C	3:72,067,255	rs75909755	T/C	0.034	0.002	0.031	9.57E-01
TC	6:31,272,261	rs6457374	T/C	0.807	0.062	0.016	1.18E-04
LDL-C	6:31,325,323	rs9266229	C/G	0.411	-0.029	0.014	4.04E-02
TG	6:36,648,275	-	CAG/C	0.464	-0.013	0.003	5.93E-07
TG	6:139,839,498	rs608736	C/G	0.484	-0.013	0.002	9.10E-09
TG	6:160,851,766	rs376563	T/C	0.460	-0.010	0.002	1.36E-05
LDL-C	6:161,111,700	rs186696265	T/C	0.011	0.304	0.076	6.17E-05
TC	6:161,111,700	rs186696265	T/C	0.010	0.278	0.075	1.93E-04
HDL-C	7:80,492,357	rs60839105	T/C	0.070	2.948	0.518	1.25E-08
TC	8:68,351,787	rs151198427	A/G	0.112	4.797	1.035	3.56E-06
LDL-C	9:78,728,065	rs146369471	T/C	0.994	0.110	0.103	2.84E-01
TC	9:78,728,065	rs146369471	T/C	0.994	0.057	0.103	5.79E-01
TC	12:51,207,704	rs829112	A/G	0.732	0.012	0.012	3.18E-01
TG	13:114,544,024	rs7140110	T/C	0.716	-0.015	0.003	5.13E-07
TG	15:43,726,625	rs150844304	A/C	0.961	-0.066	0.008	9.52E-16
TC	17:18,046,290	rs8065026	T/C	0.808	-0.029	0.013	2.66E-02
HDL-C	17:41,840,849	rs77697917	T/C	0.031	-0.241	0.035	1.04E-11
TG	19:8,429,323	rs116843064	A/G	0.031	-0.087	0.012	3.83E-13
TC	20:17,844,684	rs2618566	T/G	0.600	-0.027	0.011	1.38E-02

DISCUSSION

We conducted a GWAS that included GWAS data imputed to the 1kG to identify rare, potentially functional, variants associated with circulating lipid levels. To this end, we imputed genotypes in approximately 60.000 individuals from 20 cohorts in the CHARGE consortium with the 1kG reference panel. The meta-analysis, followed by GCTA analysis revealed nineteen associations with MAF ranging from 0.01 to 0.48. Of the nineteen associations, we were able to replicate five in an independent sample of approximately 90.000 individuals.

One of the five associations we identified is between TG and rs116843064, an exonic variant in the *ANGPTL4* gene on chromosome 19 (Figure 2C). This missense variant changes the amino acid glutamic acid into lysine (Glu40Lys) and is predicted to be damaging for the structure and function of the protein by Polyphen2[8], MutationTaster[9] and LRT[10]. *ANGPTL4* has been associated with HDL-C before using the GWAS approach2 and with TG before using an exome sequencing approach[11] and more recently using the GWAS approach[1]. *ANGPTL4* is significantly associated with the KEGG term fatty acid metabolism, the GO process lipid storage and the GO cellular component lipid particle (*p*-value of $1.10 \cdot 10^{-6}$, $1.31 \cdot 10^{-10}$ and $2.87 \cdot 10^{-18}$, respectively, genenetwork.nl).

Table 1 continued. The results for the nineteen variants after the meta-analysis of all discovery cohorts, of all replication cohorts and of all cohorts combined. A1 is allele 1 and A2 is allele 2, Freq is the frequency of A1, β is the effect of A1.

Trait	rs identifier	A1/A2	Discovery cohorts				
			Freq	*N*	β	SE_β	*p*-value
HDL-C	rs75909755	T/C	0.033	62,607	1.593	0.275	7.27E-09
TC	rs6457374	T/C	0.751	46,839	2.339	0.339	5.32E-12
LDL-C	rs9266229	C/G	0.526	37,981	-2.201	0.344	1.62E-10
TG	(6:36,648,275)	CAG/C	0.451	53,425	-0.019	0.003	7.63E-09
TG	rs608736	C/G	0.481	53,425	-0.019	0.003	5.67E-09
TG	rs376563	T/C	0.459	47,036	-0.020	0.003	3.37E-09
LDL-C	rs186696265	T/C	0.012	49,221	11.247	1.241	1.31E-19
TC	rs186696265	T/C	0.012	59,859	10.004	1.162	7.20E-18
HDL-C	rs60839105	T/C	0.068	7,882	3.355	0.571	4.26E-09
TC	rs151198427	A/G	0.108	17,361	6.552	1.147	1.12E-08
LDL-C	rs146369471	T/C	0.990	43,398	8.529	1.449	3.99E-09
TC	rs146369471	T/C	0.990	53,787	7.978	1.413	1.64E-08
TC	rs829112	A/G	0.681	56,924	1.448	0.258	2.02E-08
TG	rs7140110	T/C	0.713	48,221	-0.021	0.004	3.65E-08
TG	rs150844304	A/C	0.968	52,720	-0.083	0.010	2.52E-17
TC	rs8065026	T/C	0.785	56,924	-1.644	0.292	1.76E-08
HDL-C	rs77697917	T/C	0.023	45,052	-2.717	0.407	2.38E-11
TG	rs116843064	A/G	0.030	35,643	-0.101	0.016	6.46E-11
TC	rs2618566	T/G	0.651	63,300	-1.566	0.251	4.68E-10

The second new findings we identified is the association between TC and rs6457374, an intergenic variant located on chromosome 6 between the genes *HLA-C* and *HLA-B* (Figure 2A). Both genes are associated with the KEGG term ABC transporters (*p*-value of $4.29 \cdot 10^{-5}$ and $3.84 \cdot 10^{-5}$ for *HLA-C* and *HLA-B* respectively, genenetwork.nl) which is in line with among others a previously published association between TC and an exonic variant in the ABCA6 gene which is also an ABC transporter[12]. ABC transporters transport a wide variety of substrates across extra- and intracellular membranes, including lipids[13].

The third finding of this study is the association between HDL-C and rs77697917, an intergenic variant on chromosome 17 between the genes *SOST* and *DUSP3* (Figure 2B). *DUSP3* is associated with the regulation and function of carbohydrate-responsive element-binding protein (ChREBP) in the liver (*p*-value=$3.03 \cdot 10^{-5}$, genenetwork.nl). ChREBP mediates the activation of several regulatory enzymes involved in lipogenesis[14-18]. This variant is in high linkage disequilibrium (D'=0.936) in the 1kG with rs72836561, an exonic variant in the gene *CD300LG* (MAF=0.027, β=-2.437, se_β=0.381, *p*-value=$1.51 \cdot 10^{-10}$ in the discovery stage). This missense variant changes the amino acid arginine into cysteine (Arg82Cys) and is predicted to be damaging for the structure and function of the protein by Polyphen2[8], MutationTaster[9] and LRT[10]. This amino acid polymorphism has been associated with HDL-C in exome-wide association studies[19] and TG in GWAS[1] before.

Table 1 continued. The results for the nineteen variants after the meta-analysis of all discovery cohorts, of all replication cohorts and of all cohorts combined. A1 is allele 1 and A2 is allele 2, Freq is the frequency of A1, β is the effect of A1.

Trait	rs identifier	A1/A2	Replication cohorts				
			Freq	*N*	β	SE_β	*p*-value
HDL-C	rs75909755	T/C	0.034	86,252	-0.019	0.031	5.45E-01
TC	rs6457374	T/C	0.807	74,417	0.057	0.016	4.23E-04
LDL-C	rs9266229	C/G	0.411	61,582	-0.025	0.014	7.37E-02
TG	(6:36,648,275)	CAG/C	0.486	59,018	-0.003	0.004	5.20E-01
TG	rs608736	C/G	0.486	73,512	-0.008	0.003	2.67E-02
TG	rs376563	T/C	0.46	73,512	-0.001	0.003	8.22E-01
LDL-C	rs186696265	T/C	0.011	59,497	0.263	0.076	5.42E-04
TC	rs186696265	T/C	0.01	75,821	0.238	0.075	1.46E-03
HDL-C	rs60839105	T/C	0.078	4,971	1.067	1.228	3.85E-01
TC	rs151198427	A/G	0.128	1,419	-2.858	2.396	2.33E-01
LDL-C	rs146369471	T/C	0.994	51,367	0.068	0.103	5.11E-01
TC	rs146369471	T/C	0.994	70,241	0.015	0.103	8.84E-01
TC	rs829112	A/G	0.732	87,659	0.009	0.012	4.63E-01
TG	rs7140110	T/C	0.721	60,437	-0.006	0.005	2.68E-01
TG	rs150844304	A/C	0.945	63,884	-0.026	0.015	8.85E-02
TC	rs8065026	T/C	0.808	76,913	-0.026	0.013	4.93E-02
HDL-C	rs77697917	T/C	0.031	67,843	-0.222	0.036	4.27E-10
TG	rs116843064	A/G	0.031	44,194	-0.065	0.019	4.53E-04
TC	rs2618566	T/G	0.6	88,946	-0.024	0.011	2.83E-02

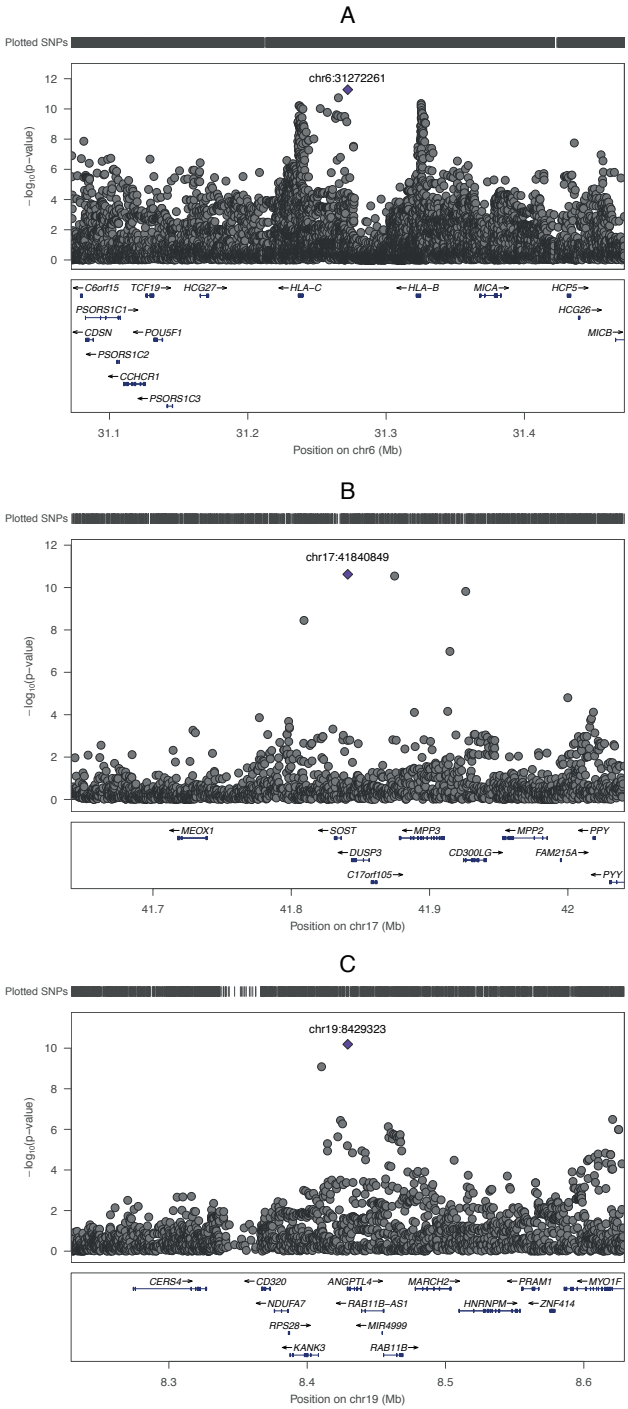
The fourth variant we identified is rs186696265, which is located on chromosome 6 and associated with both LDL-C and TC (Figure 2D and 2E). This intergenic variant is between the *LPA* (Lipoprotein, Lp(A)) gene and the *PLG* (Plasminogen) gene. The *LPA* gene has been associated before with LDL-C and TC before[2]. The reported lead SNP was rs1564348, which is in the newer human genome versions is annotated to the *SLC22A1* (Solute Carrier Family 22 (Organic Cation Transporter), Member 1) gene instead of the *LPA* gene. This explains why we identified again a locus near the *LPA* gene, which has been identified by others as well[1].

Fourteen out of the nineteen variants were not replicated despite similar sample size and similar frequencies within the replication stage as compared to the discovery stage. Of those fourteen variants, eleven exhibited effect sizes in the same direction in both stages. A possible explanation might be that the replication sample size is much larger compared to that of the discovery sample size. Two variants might have lacked significant replication due to small sample size, rs60839105 and rs151198427. Both variants only pass quality control in the cohorts in the discovery stage that contain individuals of African ancestry (Supplementary figure 7). Although there are several cohorts with individuals of African ancestry in the replication stage, both variants did not pass quality control in most cohorts which leads to

the conclusion that these variants might be population-specific. This is also suggested by the 1kG data (Phase 3) as the frequency of the C-allele is 92% in African samples and 100% in the European samples for rs60839105 and the frequency of the G-allele is 86% in the African samples and 100% in the European samples for rs151198427. Imputations of cohorts with individuals of African ancestry with the African Genome Variation Project[20] might confirm the association of rs60839105 with HDL-C and rs151198427 with TC.

To our knowledge, this is the first GWAS of circulating lipid levels using the Phase 1 integrated release v3 of the 1kG, therefore we cannot compare the positive replication rate with other studies. However, we did replicate 88.1% of the findings of Teslovich *et al.*[2] and 43.4% of the findings of GLGC[3] despite our smaller sample. We also tried to replicate findings from exome sequencing of candidate genes. The p.Arg406X mutation in the *NPC1L1* gene (rs145297799), which was reported to be associated with reduced LDL-C levels and reduced risk of coronary heart disease[4], is not available in the 1kG reference panel and, therefore, we were not able to replicate this finding. Do *et al.*[5] described the exome sequencing of the genes *LDLR* and *APOA5* and identified rare variants associated with an increased risk of myocardial infarction, increased LDL-C and TG levels. Of those rare variants, only two in the *LDLR* gene and seven in the APOA5 gene exist in our discovery meta-analysis. Both *LDLR* variants are associated with TG in our discovery meta-analysis (rs34282181, β=-0.093, SE_β=0.023, *p*-value=$4.827 \cdot 10^{-5}$ and rs2075291, β=0.219, SE_β=0.046, *p*-value=$2.092 \cdot 10^{-6}$), but not significantly associated with LDL-C (rs34282181, β=-3.939, SE_β=1.861, *p*-value=0.034 and rs2075291, β=-2.316, SE_β=3.001, *p*-value=0.440). None of the seven *APOA5* variants were significantly associated with TG or LDL-C in our discovery meta-analysis (lowest *p*-value is for LDL-C with rs72658860, β=-18.430, SE_β=7.140, *p*-value=$9.848 \cdot 10^{-3}$). The third published finding we tried to replicate, was the association between *APOC3* and TG levels[6]. Of the seven variants reported, only one existed in our discovery meta-analysis (chromosome 11, position 116,701,354), which is associated with TG (β=-0.343, SE_β=0.113, *p*-value=$2.311 \cdot 10^{-3}$). Those authors also reported an association between an *APOA5* variant (rs3135506) and TG as the most significant finding. This variant was also significantly associated with TG in our discovery meta-analysis (β=0.129, SE_β=0.007, *p*-value=$1.099 \cdot 10^{-87}$). These replication efforts demonstrate that many of the published results of exome sequencing can be replicated through the use of 1kG imputations.

In conclusion, we identified and replicated five variants associated with circulating lipid levels. These variants are in genes that can be linked biologically to lipid metabolism. Although there were a large number of variants that did not replicate at the accepted genome-wide significance threshold, the low-cost, hypothesis-free approach that we applied uncovered five variants. This study, therefore, illustrates that GWAS may still help us unravel the biological mechanisms behind circulating lipid levels.


A
Plotted SNPs
chr6:31272261
−log10(p−value)
C6orf15
TCF19
HCG27
HLA−C
HLA−B
MICA
HCP5
PSORS1C1
HCG26
CDSN
POU5F1
MICB
PSORS1C2
CCHCR1
PSORS1C3
31.1
31.2
31.3
31.4
Position on chr6 (Mb)
B
Plotted SNPs
chr17:41840849
−log10(p−value)
MEOX1
SOST
MPP3
MPP2
PPY
DUSP3
CD300LG
FAM215A
C17orf105
PYY
41.7
41.8
41.9
42
Position on chr17 (Mb)
C
Plotted SNPs
chr19:8429323
−log10(p−value)
CERS4
CD320
ANGPTL4
MARCH2
PRAM1
MYO1F
NDUFA7
RAB11B−AS1
HNRNPM
ZNF414
RPS28
MIR4999
KANK3
RAB11B
8.3
8.4
8.5
8.6
Position on chr19 (Mb)

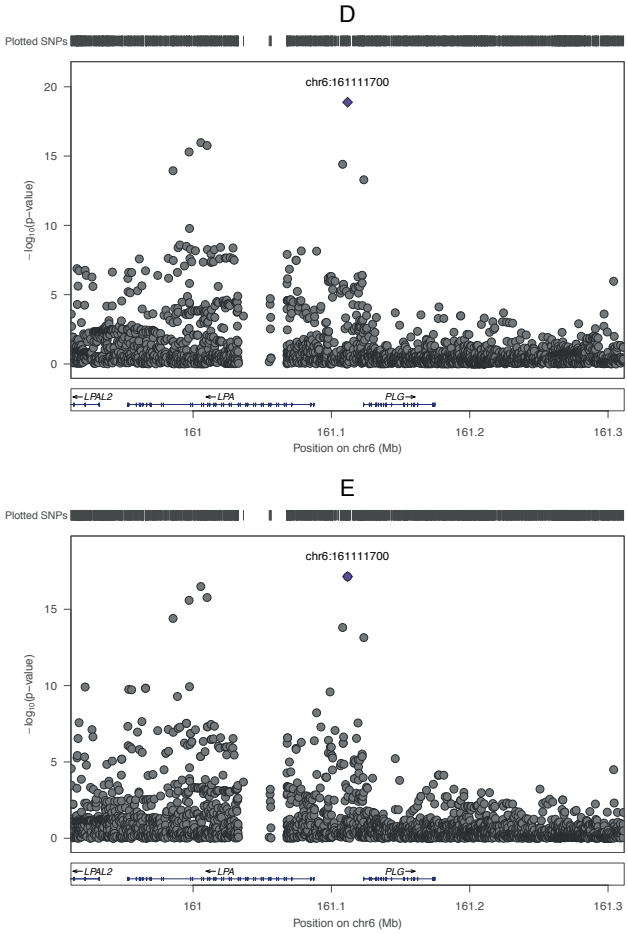


Figure 2: The regional association results of the initial meta-analysis of all discovery cohorts for (A) TC on chromosome 6, (B) HDL-C on chromosome 17, (C) TG on chromosome 19, (D) LDL-C on chromosome 6 and (E) TC on chromosome 6.

REFERENCES

1. Surakka, I. *et al.* The impact of low-frequency and rare variants on lipid levels. *Nat Genet* **47**, 589–597 (2015).
2. Teslovich, T. M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707–713 (2010).
3. Global Lipids Genetics Consortium *et al.* Discovery and refinement of loci associated with lipid levels. *Nat Genet* **45**, 1274–1283 (2013).
4. Myocardial Infarction Genetics Consortium Investigators *et al.* Inactivating mutations in NPC1L1 and protection from coronary heart disease. *N Engl J Med* **371**, 2072–2082 (2014).
5. Do, R. *et al.* Exome sequencing identifies rare LDLR and APOA5 alleles conferring risk for myocardial infarction. *Nature* **518**, 102–106 (2015).
6. The TG and HDL Working Group of the Exome Sequencing Project, National Heart, Lung, and Blood Institute *et al.* Loss-of-function mutations in APOC3, triglycerides, and coronary disease. *N Engl J Med* **371**, 22–31 (2014).
7. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* **88**, 76–82 (2011).
8. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nat Methods* **7**, 248–249 (2010).
9. Schwarz, J. M., Rödelsperger, C., Schuelke, M. & Seelow, D. MutationTaster evaluates disease-causing potential of sequence alterations. *Nat Methods* **7**, 575–576 (2010).
10. Chun, S. & Fay, J. C. Identification of deleterious mutations within three human genomes. *Genome Res* **19**, 1553–1561 (2009).
11. Romeo, S. *et al.* Population-based resequencing of angptl4 uncovers variations that reduce triglycerides and increase hdl. *Nat Genet* **39**, 513–516 (2007).
12. van Leeuwen, E. M. *et al.* Genome of the Netherlands population-specific imputations identify an ABCA6 variant associated with cholesterol levels. *Nat Commun* **6**, 6065 (2015).
13. Jones, P. M. & George, A. M. The ABC transporter structure and mechanism: perspectives on recent research. *Cell Mol Life Sci* **61**, 682–699 (2004).
14. Towle, H. C. Glucose as a regulator of eukaryotic gene transcription. *Trends Endocrinol Metab* **16**, 489–494 (2005).
15. Dentin, R. *et al.* Hepatic glucokinase is required for the synergistic action of chrebp and srebp-1c on glycolytic and lipogenic gene expression. *J Biol Chem* **279**, 20314–20326 (2004).
16. Dentin, R., Girard, J. & Postic, C. Carbohydrate responsive element binding protein (ChREBP) and sterol regulatory element binding protein-1c (SREBP-1c): two key regulators of glucose metabolism and lipid synthesis in liver. *Biochimie* **87**, 81–86 (2005).
17. Ma, L., Robinson, L. N. & Towle, H. C. ChREBP*MLx is the principal mediator of glucose-induced gene expression in the liver. *J Biol Chem* **281**, 28721–28730 (2006).
18. Uyeda, K. & Repa, J. J. Carbohydrate response element binding protein, ChREBP, a transcription factor coupling hepatic glucose utilization and lipid synthesis. *Cell Metab* **4**, 107–110 (2006).
19. Albrechtsen, A. *et al.* Exome sequencing-driven discovery of coding polymorphisms associated with common metabolic phenotypes. *Diabetologia* **56**, 298–310 (2013).
20. Gurdasani, D. *et al.* The african genome variation project shapes medical genetics in africa. *Nature* **517**, 327–332 (2015).

PART 3

GENOME OF THE NETHERLANDS, A POPULATION-SPECIFIC REFERENCE PANEL

CHAPTER 3.1

Population-specific genotype imputations using minimac or IMPUTE2

Elisabeth M. van Leeuwen, Alexandros Kanterakis, Patrick Deelen, Mathijs V. Kattenberg, The Genome of the Netherlands Consortium, P. Eline Slagboom, Paul I.W. de Bakker, Cisca Wijmenga, Morris A. Swertz, Dorret I. Boomsma, Cornelia M. van Duijn, Lennart C. Karssen*, Jouke J. Hottenga*.

* These authors contributed equally to this work.

Published in Nature Protocols (Nat Protoc. 2015 Sep;10(9):1285-96).

The supplemental information for this chapter is available online at:
http://www.nature.com/nprot/journal/v10/n9/extref/nprot.2015.077-S1.pdf

ABSTRACT

In order to meaningfully analyse common and rare genetic variants, results from Genome-Wide Association Studies (GWAS) of multiple cohorts need to be combined in a meta-analysis in order to obtain enough power. This requires all cohorts to have the same Single Nucleotide Polymorphisms (SNPs) in their GWAS. To this end, genotypes that have not been measured in a given cohort can be imputed based on a set of reference haplotypes. This protocol provides guidelines for performing imputations with two widely used tools: minimac and IMPUTE2. These guidelines were developed and used by the Genome of the Netherlands consortium that has created a population-specific reference panel for genetic imputations and used this reference to impute various Dutch biobanks. We also examine several factors that might influence the final imputation quality. This protocol, which has been used by the largest Dutch biobanks should take approximately several days, depending on the sample size of the biobank and the computer resources available.

INTRODUCTION

Data from Genome-Wide Association Studies (GWAS) of different cohorts can be combined into a meta-analysis even when the samples of the cohorts have been typed on different genotyping platforms. By imputing missing genotypes, a homogeneous data set for meta-analysis can be created. Genotype imputation allows estimation of genotypes in a target data set, based on one or more available reference sets of Single Nucleotide Polymorphisms (SNPs) and is based on searching common haplotypes between an individual's genome and a reference panel with a high density of genotyped SNPs, such as those provided by the HapMap[1], 1000 Genomes[2] and the Genome of the Netherlands (GoNL)[3-5] projects. Missing genotypes are then inferred from common haplotypes found in the reference set. Implementation of these methods usually results in estimates of the posterior probability distributions $P_g = (P_{AA}, P_{AB}, P_{BB})$ of the genotypes based on the available data[6].

Weaknesses in both genotype calling and imputation of missing genotypes can lead to biases in GWAS and subsequently in meta-analysis. Therefore, Anderson *et al*.[7] have previously published a protocol dealing with quality control of genotype data, and our work can be seen as an extension of that protocol. A guideline for imputations with the Beagle[8] and IMPUTE2[9] tools, as well as post-imputation quality control has been published by Verma *et al.*[10], and a protocol for doing meta-analysis of GWAS results for large numbers of cohorts is described in Winkler *et al*.[11].

In this protocol, we show how to perform genotype imputations with a population-specific reference panel including how to deal with factors that may adversely affect the imputation result (e.g. how to properly split up large data sets for imputation). This protocol differs to the previous guideline from Verma *et al.*[10], providing instructions for imputations with IMPUTE2[9] and minimac[12]. We describe the different pipelines for imputations using the genome-wide SNP data provided by Anderson *et al*.[7] as a target data set. We will start with the quality control of this target set using the pipeline from Anderson *et al*.[7]. We will show how to lift the target set over to the correct NCBI build and then provide pipelines for imputation using IMPUTE2[9] and minimac[12] (Figure 1). All pipelines are developed for GNU/Linux based computer resources and all commands should be typed at the Bash shell prompt where Bash variables are indicated by ${variablename}. This protocol does not include commands to submit compute intensive tasks to a job scheduling system like OpenPBS (see Section Computer Resources), as different computer clusters may use different scheduling systems.

This protocol has been used to impute the genotypes of individuals of various Dutch biobanks, using the GoNL reference panel. This has resulted in the discovery of five novel associations at four loci for cholesterol levels including a rare missense variant in the *ABCA6* gene which is predicted to be deleterious[13].

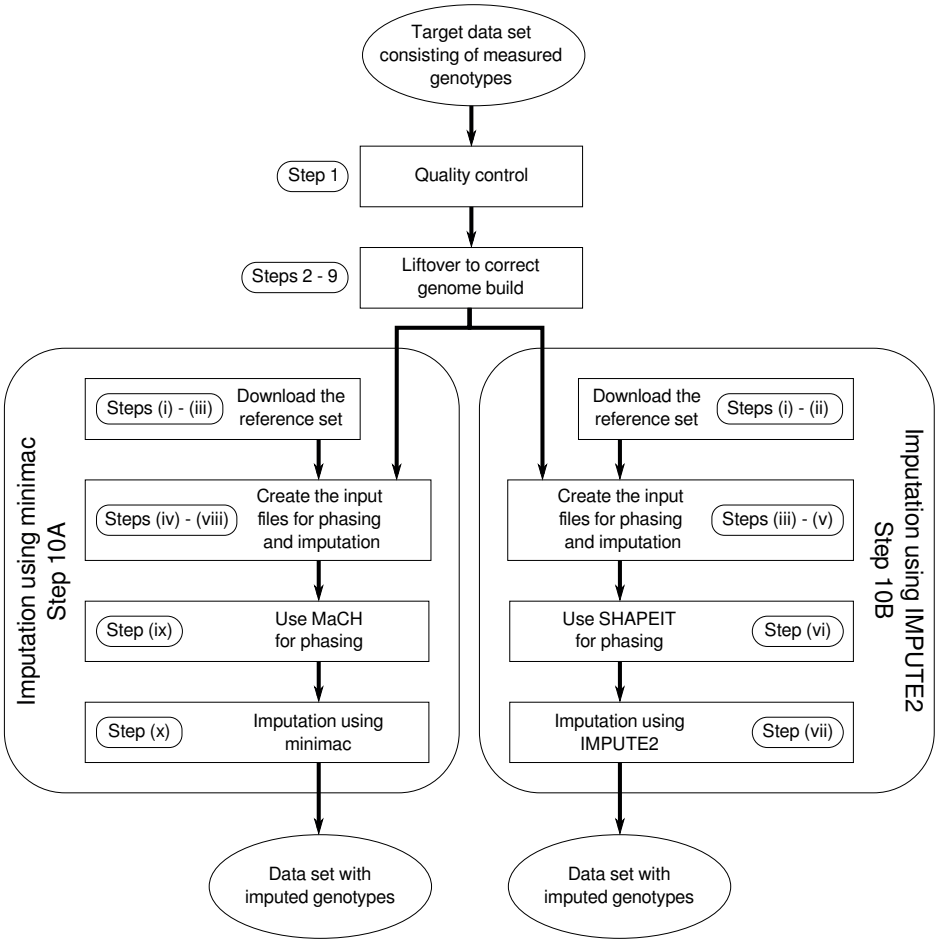


Figure 1: Workflow of the imputation protocol for imputations of unobserved genotypes with the GoNL reference panel. The first stage of the protocol is to perform quality control of the target data set consisting of measured genotypes, followed by liftover to the correct human genome build. The human genome build of the GoNL reference panel is UCSC hg19. These steps are independent of the tools that are used for the actual phasing and imputation. The next step is to download the reference set, which is necessary to create the correct input file for phasing and imputations. The reference set file format are different for each tool. Next, MaCH or SHAPEIT are used for phasing, followed by minimac and IMPUTE2 for the imputations.

Genome of the Netherlands reference set

The construction of a novel imputation reference data set is a complex procedure that requires dense genotyping and accurate estimation of haplotypes from genotype data (known as phasing) of samples from a specific population. The most thoroughly documented and widely available imputation reference sets are coming from the HapMap[1] and 1000 Genomes projects[2]. Both projects contain samples from various populations and consequently a given

genotype of a low-frequency variant may not be represented adequately in the reference data set. Moreover, when the percentage of samples belonging to a different geographical population is beyond a certain proportion, the imputation quality does not improve. Jostins *et al*.[14] found that when imputing samples from the 1,958 British Birth Cohort, the accuracy starts to fall off when the proportion of non-CEU samples exceed 20%, as the effect of increased diversity is outweighed by the effect of mismatching. This relationship is specific to low-frequency variants. Morever, Pistis *et al*.[15] found that the effectiveness of population-specific reference panels can be appreciable for other populations, but will vary depending on the size of the panels and the demographic history of the isolate.

As the interest of the field of genetic epidemiology is shifting towards low-frequency variants, the GoNL consortium created a population-specific reference set for imputation with the goal of identifying associations between various phenotypes and low-frequency genetic variants. To this end, 231 parent-offspring trios and 19 parent-offspring quartets of Dutch descent had their complete genome sequenced with at least 12× coverage[3-5]. The strength of this reference set comes from several factors. First, the trio design which improves the haplotype quality, second, the coverage which is higher than that of the 1000 Genomes Project, and third, the sequencing of samples from a homogeneous population. The quality of the haplotypes boosts imputation accuracy in independent samples, especially for lower frequency alleles[4].

The GoNL reference set is available by applying through http://www.nlgenome.nl/, menu option "Request data", which leads to the application form. After filling in the form, the request will be evaluated by the GoNL steering committee. After positive evaluation, a data access agreement needs to be signed and subsequently, the reference panel can be downloaded in Variant Call Format (VCF). For this protocol the fourth release of the GoNL reference panel was used, which contains 499 individuals of Dutch ancestry and 19,562,004 autosomal SNPs.

Tools for imputation

The three most commonly used tools for genotype imputation are minimac[12], IMPUTE2[9] and Beagle[8]. Multiple aspects of the three tools, e.g. their imputation accuracy, error rates and computational performance have been compared previously[6,10,16,17]. The choice for a given tool depends on the target set that is to be imputed and on the type of computational resources available as discussed in this paper. Within the GoNL[3-5] consortium, only minimac and IMPUTE2 were used for imputations, and therefore Beagle will not be discussed in this manuscript. It is, however, possible to impute samples with the GoNL reference panel using Beagle. Minimac can be downloaded freely from the web, its source code is available under an open source licence. IMPUTE2 is available for download for academic use only, no source code is provided.

IMPUTE2 performs both the phasing and the imputation, whereas minimac only imputes data sets that have been phased by MaCH[18] or SHAPEIT2[19]. However, although IMPUTE2

can perform phasing, its authors recommend to use SHAPEIT2[19] for the phasing followed by using IMPUTE2 for the imputations. Of the three tools, only IMPUTE2 can combine two reference panels. This allows imputation with both the 1000 Genomes reference panel as well as the GoNL reference panel, which has been shown to improve imputation quality[3]. MaCH/minimac make their own recombination map based on input data, IMPUTE2 requires a recombimation map.

The requested file format of the reference set is also different among the tools. Both the GoNL project[3-5], 1000 Genomes project[2] and the HapMap project[1] provide their data in Variant Call Format (VCF)[20]. The VCFtools[20] software package can convert these VCF files into phased haplotypes in IMPUTE2 reference-panel format. The authors of IMPUTE2 also provided a Perl script to perform this conversion. Minimac can handle the original VCF files without conversion.

Both tools produce several output files. The first one is the so-called info file containing the SNP name, the basepair positions, the frequencies of the alleles and the R^2. Here R^2 is the estimated squared correlation (between zero and one) between the allele dosage with highest posterior probability in the genotype probabilities file and the true allele dosage for the marker; larger values of allelic R^2 indicate more accurate genotype imputation. In a second file IMPUTE2 gives the probabilities of the three genotypes *AA*, *AB* and *BB*, whereas minimac gives the probability of a homozygote for allele 1 and the probability of the heterozygote. Only minimac has the option to output best-guess alleles. Dosage files are produced only by minimac, however, it takes only one additional step to convert the genotype probabilities from IMPUTE2 into dosages. If a sample has genotype probabilities (P_{AA}, P_{AB}, P_{BB}) for a marker, then the estimated B-allele dosage (d_B) is $d_B = P_{AB} + 2\,P_{BB}$. All formats can be converted using fcGene[21].

Quality control of the target data set

In order to achieve high-quality imputation standard GWAS quality control filters need to be applied to the target data set and if necessary also to the reference set prior to imputation. The purpose of these filters is to exclude both markers and samples with low-quality data. Anderson *et al*.[7] and Verma *et al.*[10] provide a detailed protocol that deals with both per-maker and per-individual filtering.

Other factors influencing the imputation quality are the type of arrays used for genotyping, strand and build issues. Present day high-density arrays are of high quality, however, the low-density arrays used in the beginning of the GWAS era were less so. It is therefore useful to check the type of array that was used for genotyping of the target set. The genotype calls from the arrays are aligned to a specific strand[22]. In order to obtain high-quality imputations it is important to correct possible strand alignment issues. Although IMPUTE2 and MaCH have options to fix misaligned alleles between study and reference panel by inverting the alleles

when possible, the alignment of the target set should be fixed prior to imputing the target set with for example SHAPEIT2[19]. This only holds for ambiguous strands (AT and TA for example), detecting and correcting the strand of the non-ambiguous SNPs (AT and GC for example) is more of a challenge, Deelen *et al.* have published a method for solving the strand issues of non-ambiguous SNPs[23]. For imputation purposes, the alleles should be aligned to the forward strand, since the imputation tools assume that the target set is on the same strand as the reference panel, which is the forward strand.

It is important for imputation that both the target set and the reference set are on the same NCBI build as SNP names may change or SNPs may be relocated or merged between builds. Release four of the GoNL reference set uses NCBI build 37 (human genome 19, hg19). If the reference and the target set are aligned using a different genome assembly, it is recommend to re-align the target panel to the assembly of the reference rather than the other way around. This is because the phased haplotype structure of the reference panel will be distorted if the position of the markers is altered. Moreover, re-aligning of the target set costs less time compared to re-aligning the reference panel. The liftOver tool from UCSC[24] converts genome positions between different genome builds (see Section *Perform quality control* and `http://genome.sph.umich.edu/wiki/LiftOver`).

A major pitfall of genotype imputation is a difference between groups of individuals which after imputations can be (falsely) associated with a phenotype. Array differences or quality differences (for example call rates) between cases and controls should be avoided. Therefore, the most ideal situation would be to genotype all individuals on the same array. If this is not possible, it is highly advised to apply strict quality control. The type of array is also of influence on the imputations, chunking the observed genotypes of low-density arrays as discussed in Section "Handling large target data sets" may lead to empty chunks. High-density genotype arrays are therefore advised. Other important imputation pitfalls are monomorphic and extremely rare SNPs[25], therefore these should be removed from both the target set and the reference panel.

After performing all quality control steps, the target data set needs to be converted into the correct input format (see *BOX 1*) for the imputation tool of choice.

Quality metrics

The quality of an imputation experiment can be assessed by various metrics[10]. These can be divided into two categories based on whether true genotypes are available or not. The most common imputation metric is the R^2 that represents the correlation between the imputed and the real genotypes.

When the true genotypes are unknown, various statistics can be used to estimate the R^2. Marchini *et al*.[6] present a thorough review of the R^2 metrics used by MaCH, Beagle, SNPTEST and IMPUTE2. Comparison of these measures showed that they are highly correlated.

Another R^2 metric[26] is the ratio of the variance of the imputed allele dosage and the variance of the true allele dosage. Although the variance of the true allele dosage is unknown, it can be estimated as *2p(1-p)* under Hardy-Weinberg equilibrium, where *p* is the estimated allele frequency. To illustrate how well rare and common SNPs were imputed, a plot can be made with the percentage of SNPs at various cut-offs for the R^2 for various minor allele frequency (MAF) bins[8,27].

In case the true genotypes are available, the quality of the imputation can also be evaluated by calculating the false positive and false negative genotypes[4]. False positive genotypes are the ones that have a high imputation R^2, but were in fact imputed incorrectly. False negative genotypes are the ones that have a low $R^{2,}$ but were actually imputed correctly. Another qualitative metric is the concordance between real and imputed genotypes. A graph of the percentage of discordance versus percentage of missing genotypes for various thresholds of the genotype probability can be used to compare different imputation methods[9].

Handling large target data sets

To successfully identify rare variants associated with particular phenotypes large sample sizes are needed. Splitting up the target sets and distributing the computational burden of phasing and imputation over several computers allows imputation of such large sets to finish within a reasonable time frame. Splitting up the target set reduces the time to finish the imputations (see Supplementary Figure S1), however it does require a computer cluster. A target set can be split up in two ways: (1) splitting into subsets of samples and (2) splitting into chunks of chromosomes. The division into groups of samples can be done randomly, although the distribution of cases and controls should be similar in the subgroups. However, since imputations are mostly done once per cohort followed by the subsequent analysis of many phenotypes using the same imputed genotype data, splitting a target set into equal proportions of cases and controls provides a challenge and we therefore do not recommend this. This only holds for the imputations and not for phasing, as the samples do not affect each other in phasing. Splitting up in samples may, however, be helpful to optimize the capacity utilization of a compute cluster.

The second, more useful, strategy of splitting up the target set is to split the chromosomes into chunks of a few Mb. Depending on the imputation tool, the strategy to split up into chunks is different. When using minimac, the ChunkChromosome tool (`http://genome.sph.umich.edu/wiki/ChunkChromosome`) can be used to split each chromosome prior to imputation (see Section *Imputations with MaCH and minimac*). When imputing with IMPUTE2 it is not necessary to first split up the chromosome as one of the command line arguments of IMPUTE2 is the position interval to impute.

To evaluate the quality of the imputations after the chromosome is split into chunks, we imputed chromosome 21 of all 5,974 samples of the Rotterdam Study cohort I with the

European part of the 1000 Genomes reference set (release August 2010) using minimac after phasing with MaCH using two approaches. In both approaches the data set was split up before phasing with MaCH. The first approach was to split the SNPs on chromosome 21 into chunks of 500kb, 1Mb, 2Mb, 3Mb, 4Mb, 5Mb, 7.5Mb and 10Mb, respectively, each with an overlap of 5% on each side of the chunk. The second approach was to split the same chromosome into chunks of 5Mb with an overlap of 2.5% (250kb), 5% (500kb), 7.5% (750kb), 10% (1Mb) and 12.5% (1.25Mb) on each side, respectively. Figures 2a and 2b show that the target set can be split into subsets of at least 5Mb with an overlap of at least 250kb without decreasing imputation quality.

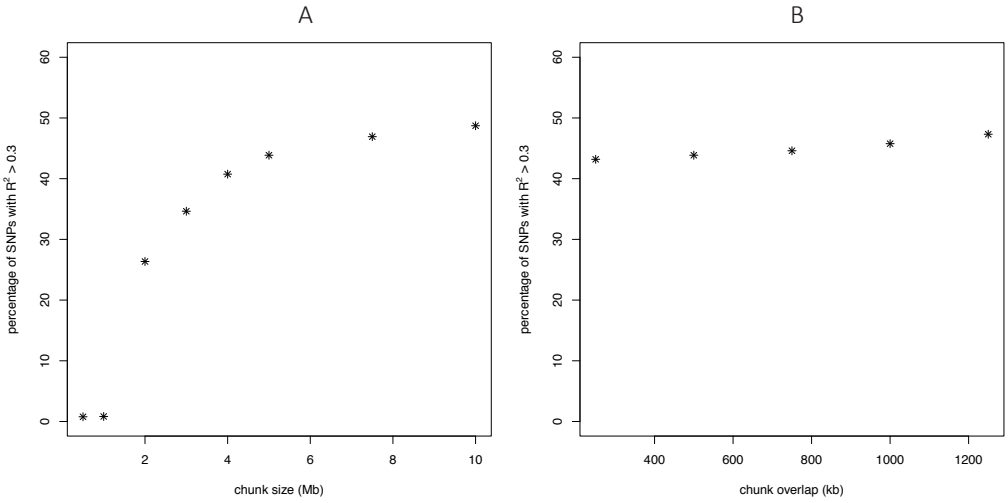


Figure 2: The percentage of SNPs with $R^2 > 0.3$ after imputing chromosome 21 of 5,974 samples of Rotterdam Study cohort I (a) when the target set is split into several chunks of chromosomes and the percentage overlap between chunks is 10% and (b) when the chromosome of the target set is split into 5Mb chunks and the size of the overlap is varied. This figure illustrates that the target set can be split into subsets of at least 5Mb with an overlap of at least 250kb without decreasing imputation quality.

MATERIALS

Equipment

Data

- Genome-wide SNP data (`raw-GWA-data.tgz`). See supplementary data from Anderson *et al.*[7].
- GoNL reference panel for imputations. The reference set is available by applying through `http://www.nlgenome.nl/`.

Software

This protocol assumes that the computer uses GNU/Linux as its operating system (which is the case for most, if not all computer clusters), and the analyst uses Bash as his/her shell (which is the default on most GNU/Linux systems).

- Several tools like gawk, sort, uniq, wget, tar, sed, and head, which are usually installed by default on a GNU/Linux system.
- PLINK v1.07[28]; the binaries compiled for various platforms and installation instructions can be downloaded from http://pngu.mgh.harvard.edu/~purcell/plink/download.shtml.
- liftOver; this tool can be used to lift over from one human genome build to the other and can be downloaded from http://hgdownload.cse.ucsc.edu/admin/exe/linux.x86_64/liftOver
- VCFtools v0.1.12b; this tool can be downloaded from http://sourceforge.net/projects/vcftools/files/latest/download/vcftools_0.1.12b.tar.gz
- ChunkChromosome (release 2014-05-27); this tool can be downloaded from http://www.sph.umich.edu/csg/cfuchsb/generic-ChunkChromosome-2014-05-27.tar.gz
- MaCH (release 1.0); this tool can be downloaded from http://www.sph.umich.edu/csg/abecasis/MaCH/download/mach.1.0.18.Linux.tgz
- Minimac (release 2013.7.17); this tool can be downloaded from http://www.sph.umich.edu/csg/cfuchsb/minimac-beta-2013.7.17.tgz
- SHAPEIT v2.790; this tool can be downloaded from https://mathgen.stats.ox.ac.uk/genetics_software/shapeit/shapeit.v2.r790.RHELS_5.4.static.tar.gz
- IMPUTE2 v2.3.1; this tool can be downloaded from https://mathgen.stats.ox.ac.uk/impute/impute_v2.3.1_x86_64_static.tgz

Computer resources

Imputing SNPs in data sets of several thousands of samples using reference sets consisting of several millions of SNPs (e.g. HapMap[1] up to several tens of millions of SNPs (GoNL project[3-5] or the 1000 Genomes project[2] cannot be done on a commodity desktop computer since that would take months of time and requires more memory (RAM) than is usually available. As discussed earlier, the answer lies in splitting the imputation task into smaller pieces and running these sub-tasks on a computer cluster.

The work described in this paper was done on two such clusters. The Lisa cluster at SARA (`http://www.surfsara.nl/systems/lisa/`) is a heterogeneous cluster consisting of more than 500 machines with a total of more than 6000 cores and 16 to 24 GB of RAM each, running Debian Linux (`http://www.debian.org`). The Millipede cluster at Groningen University is a heterogeneous cluster with 252 nodes with a total of 3216 cores and 24 to 128 GB of RAM each. It runs RedHat Enterprise Linux 5 (`http://www.redhat.com/products/enterprise-linux/`). Both clusters use the OpenPBS (`http://www.mcs.`

anl.gov/research/projects/openpbs/) system to schedule tasks across their nodes. The memory requirements for MaCH are about 100MB and for the minimac protocol 3GB, whereas SHAPEIT requires about 1.5MB and IMPUTE2 about 3GB.

PROCEDURE

Perform quality control (TIMING ~8 hours)

1. The first step is to perform standard quality control on the target set. To do this, complete the protocol for quality control as described by Anderson *et al.*[7]. We assume that the genotypes have been called by a genotyping center and returned in PLINK format named `raw-GWA-data.ped`, `raw-GWA-data.map`. All genotypes are annotated to the forward strand. After performing quality control of this genome-wide SNP data, 1,919 samples and 313,878 markers remain. The resulting files are named `clean-GWA-data.bed`, `clean-GWA-data.bim` and `clean-GWA-data.fam`.

Converting the target set to the correct genome build (TIMING ~20 min)

2. If the target set is on another genome build than the reference set, it is important to lift the target set over to the same build as the reference set. The following protocol shows how to convert the target set from UCSC hg17 (NCBI build 35) to UCSC hg19 (Genome Reference Consortium GRCh37). First download the chain file:
   `wget http://hgdownload.cse.ucsc.edu/goldenPath/hg17/liftOver/hg17ToHg19.over.chain.gz`
   and type `gunzip hg17ToHg19.over.chain.gz` to unzip the chain file.
3. To start the liftover, convert the target set with PLINK to a map and ped file:
   `plink --noweb --bfile clean-GWA-data --recode --out clean-GWA-data`
   This will create the files `clean-GWA-data.map` and `clean-GWA-data.ped`.
4. The next step is to create a BED file based on the map file using the following command:
   `gawk '{print "chr"$1, $4, $4+1, $2}' OFS="\t" clean-GWA-data.map > clean-GWA-data_HG17.BED`
5. Then perform the liftover: `./liftOver -bedPlus=4 clean-GWA-data_HG17.BED hg17ToHg19.over.chain clean-GWA-data.HG19.BED clean-GWA-data_unmapped.txt`
6. Use the resulting file `clean-GWA-data_unmapped.txt` to create a list of unmapped SNPs:
   `gawk '/^[^#]/ {print $4}' clean-GWA-data_unmapped.txt > clean-GWA-data_unmappedSNPs.txt`
7. Create a mapping file using the new BED file:
   `gawk '{print $4, $2}' OFS="\t" clean-GWA-data.HG19.BED > clean-GWA-data.HG19.mapping.txt`

8. Use PLINK to remove the unmapped SNPs from the target data set:

```
plink --noweb --file clean-GWA-data --exclude clean-GWA-data_unmappedSNPs.txt --update-map clean-GWA-data.HG19.mapping.txt --make-bed --out clean-GWA-data.HG19.temp
plink --noweb --bfile clean-GWA-data.HG19.temp --recode --out clean-GWA-data.HG19
```

9. Create a new SNP list for the data set:

```
gawk '{print $2}' clean-GWA-data.HG19.map > clean-GWA-data.HG19.snplist
```

The resulting files produced after quality control and lifting over the data set to the correct build, are named `clean-GWA-data.HG19.map` and `clean-GWA-data.HG19.ped`. In this case the data set was lifted over from build 35 to build 37, however, other liftovers are also possible, the UCSC Genome Browser website provides multiple chain files.

Imputations with minimac or IMPUTE2

10. SNP imputations can be performed using either a combination of MaCH/minimac (Option A) or IMPUTE2 (Option B).

(A) MaCH/minimac (Timing ~60 hours)

(i) Download the reference set for minimac. This pipeline for imputations with MaCH and minimac imputes the target set after quality control and (if necessary) lifted over to the correct build with the GoNL reference panel release 4. First create a new directory for the reference set: `mkdir reference-GoNL-v4`. The zipped VCF files of the GoNL reference panel should be placed in this directory. In this protocol we assume the names of the files are as follows: `gonl.chr{1-22}.release4.gtc.vcf.gz`.

(ii) Use VCFtools to create info files for all chromosomes by running:

```
for chr in {1..22}; do vcftools --gzvcf reference-GoNL-v4/gonl.chr${chr}.release4.gtc.vcf.gz --get-INFO NS --out reference-GoNL-v4/gonl.chr${chr}.release4.gtc; done
```

(iii) Create a file with all the positions that are in the reference set:

```
rm -f snps-reference.txt
for i in reference-GoNL-v4/gonl.chr*.release4.gtc.INFO; do gawk '$1!="CHROM" {print $1"_"$2}' $i >> snps-reference.txt; done
```

(iv) Creating the input files for phasing and imputation. To get a list of positions of SNPs that are in the target set and/or in the reference set:

```
gawk '{print $1"_"$4}' clean-GWA-data.HG19.map > snps-reference-and-rawdata
```

and

```
sort snps-reference.txt | uniq >> snps-reference-and-rawdata
```

To get only those SNPs that are in both the target set and reference set:

```
sort snps-reference-and-rawdata | uniq -d | gawk -F "_" '{$3=$2+1;
print $1, $2, $3, "R"NR}' > snps-reference-and-rawdata-duplicates
```

? TROUBLESHOOTING

(v) The names of the SNPs that are in both the target set and in the reference set need to be extracted from the target set. Using PLINK this can be done as follows:

```
plink --noweb --file clean-GWA-data.HG19 --extract snps-reference-
and-rawdata-duplicates --range --make-bed --out clean-GWA-data.
HG19.for-impute.plink
```

(vi) MaCH and minimac need one file per chromosome. Extract SNPs for each chromosome:

```
for chr in {1..22}; do plink --noweb --bfile clean-GWA-data.HG19.
for-impute.plink --chr ${chr} --recode --out clean-GWA-data.HG19.
for-impute.plink.chr${chr}; done
```

(vii) Convert the resulting PLINK sets into merlin file format since minimac requests this:

```
for chr in {1..22}; do gawk '{$6=0; print $0}' clean-GWA-data.HG19.
for-impute.plink.chr${chr}.ped > clean-GWA-data.HG19.for-impute.
merlin.chr${chr}.ped; echo "T faket1" > clean-GWA-data.HG19.for-
impute.merlin.chr${chr}.dat; gawk '$2="M "$2 {print $2}' clean-
GWA-data.HG19.for-impute.plink.chr${chr}.map >> clean-GWA-data.
HG19.for-impute.merlin.chr${chr}.dat; echo "chromosome markername
position" > clean-GWA-data.HG19.for-impute.merlin.chr${chr}.map;
gawk '{print $1, $2, $4}' clean-GWA-data.HG19.for-impute.plink.
chr${chr}.map >> clean-GWA-data.HG19.for-impute.merlin.chr${chr}.
map; done
```

(viii) Split the merlin files so they contain 2500 markers with a 500 marker overlap using the ChunkChromosome tool:

```
for chr in {1..22}; do ./generic-ChunkChromosome/executables/Chunk
Chromosome -d clean-GWA-data.HG19.for-impute.merlin.chr${chr}.dat
-n 2500 -o 500; done
```

(ix) Using MaCH for phasing. Use MaCH to phase the haplotypes in each chunk:

```
for chunk in chunk*.dat; do machfile="${chunk%.*}"; merlinfile
="${machfile#*-}.ped"; executables/mach1 -d ${chunk} -p ${merlinfile}
--rounds 20 --states 200 --phase --interim 5 --sample 5 --compact
--prefix ${machfile}; done
```

? TROUBLESHOOTING

(x) Imputation with minimac. Execute the following commands to impute all chunks using minimac:

```
for chunk in chunk*.dat; do filename1="${chunk%.*}"; filename2
="${filename1#*-}.ped"; chr=`echo "${filename1##*.}" | sed 's/
```

```
chr//'`; minimac --vcfReference --rs --refHaps reference-GoNL-v4/
gonl.chr${chr}.release4.gtc.vcf.gz --haps ${filename1}.gz --snps
${filename1}.dat.snps --rounds 5 --states 200 --autoClip autoChunk-
clean-GWA-data.HG19.for-impute.merlin.chr${chr}.dat --gzip --phased
--probs --prefix ${filename1}; done
```

? TROUBLESHOOTING

(B) Imputations with IMPUTE2 (Timing ~7 hours)

(i) Download the reference set for IMPUTE2: This pipeline for imputations with IMPUTE2 imputes the target set after quality control and (if necessary) lifted over to the correct build with the GoNL reference panel release 4. First create a new directory for the reference set: `mkdir reference-GoNL-v4`. All files of the GoNL reference panel should be placed in this directory. In this protocol we assume the names of the files are as follows: `gonl.chr{1-22}.release4.gtc.{hap.gz, legend.gz, geneticmap.txt}`.

(ii) Now a file can be created with all the SNP names that are in the reference set:

```
rm –r snps-reference.txt; for chr in {1..22}; do gunzip -c reference-
GoNL-v4/gonl.chr${chr}.release4.gtc.legend.gz | gawk -v chr=${chr}
'$5=="SNP" && $1!="id" {print chr"_"$2}' >> snps-reference.txt;
done
```

(iii) Creating the input files for phasing and imputation. Use the following commands to get a list of positions of SNPs that are in the target set and/or in the reference set:

```
gawk '{print $1"_"$4}' clean-GWA-data.HG19.map > snps-reference-
and-rawdata
```

and

```
sort snps-reference.txt | uniq >> snps-reference-and-rawdata
```

To get only those SNPs that are in both the target set and reference set:

```
sort snps-reference-and-rawdata | uniq -d | gawk -F "_" '{$3=$2+1;
print $1, $2, $3, "R"NR}' > snps-reference-and-rawdata-duplicates
```

? TROUBLESHOOTING

(iv) The names of the SNPs that are in both the target set and in the reference set need to be extracted from the target set. Use PLINK to run:

```
plink --noweb --file clean-GWA-data.HG19 --extract snps-reference-
and-rawdata-duplicates --range --make-bed --out clean-GWA-data.
HG19.for-impute.plink
```

(v) Since we will phase per chromosome, split the PLINK file into 22 files:

```
for chr in {1..22}; do plink --bfile clean-GWA-data.HG19.for-impute.
plink --chr $chr --recode --out clean-GWA-data.HG19.for-impute.
plink.chr${chr}; done
```

This creates the following files per chromosome: `clean-GWA-data.HG19.for-impute.plink.chr${chr}.ped` and `clean-GWA-data.HG19.for-impute.plink.chr${chr}.map`.

? TROUBLESHOOTING

(vi) Using SHAPEIT for phasing. For every chromosome, the haplotypes are phased using SHAPEIT:

```
for chr in {1..22}; do namefile="clean-GWA-data.HG19.for-impute.plink.chr${chr}"; ./shapeit.v2.r790.RHELS_5.4.static --input-ped ${namefile}.ped ${namefile}.map --input-map reference-GoNL-v4/gonl.chr${chr}.release4.gtc.geneticmap.txt --output-max ${namefile}.phased --thread 8 --output-log ${namefile}.phased; done
```

(vii) Imputation with IMPUTE2. For every chromosome, perform imputations in chunks of 5Mb:

```
refdir="reference-GoNL-v4"; for chr in {1..22}; do namefile="clean-GWA-data.HG19.for-impute.plink.chr${chr}.phased"; maxPos=$(gawk '$1!="position" {print $1}' ${refdir}/gonl.chr${chr}.release4.gtc.geneticmap.txt | sort -n | tail -n 1); nrChunk=$(expr ${maxPos} "/" 5000000); nrChunk2=$(expr ${nrChunk} "+" 1); start="0"; for chunk in $(seq 1 $nrChunk2); do endchr=$(expr $start "+" 5000000); startchr=$(expr $start "+" 1); ./impute_v2.3.1_x86_64_static/impute2 -known_haps_g ${namefile}.haps -m ${refdir}/gonl.chr${chr}.release4.gtc.geneticmap.txt -h ${refdir}/gonl.chr${chr}.release4.gtc.hap.gz -l ${refdir}/gonl.chr${chr}.release4.gtc.legend.gz -int ${startchr} ${endchr} -Ne 20000 -o ${namefile}.chunk${chunk}.impute2; start=${endchr}; done done
```

(viii) Convert the files with the probabilities for the three genotypes into dosage files:

```
for chr in {1..22}; do namefile="clean-GWA-data.HG19.for-impute.plink.chr${chr}.phased"; maxPos=$(gawk '$1!="position" {print $1}' ${refdir}/gonl.chr${chr}.release4.gtc.geneticmap.txt | sort -n | tail -n 1); nrChunk=$(expr ${maxPos} "/" 5000000); nrChunk2=$(expr ${nrChunk} "+" 1); for chunk in $(seq 1 $nrChunk2); do gawk '{tp = $1 " " $2 " " $3 " " $4 " " $5; for (i=6; i<=NF; i+=3) tp = tp " " $(i+1) + 2.0*$(i+2); print tp }' ${namefile}.chunk${chunk}.impute2 > ${namefile}.chunk${chunk}.impute2.dosage; done done
```

? TROUBLESHOOTING

Timing
Step 1, Perform quality control: ~8 hours
Step 2 – 9, Converting the target set to the correct build: ~20 min

Imputations with minimac or IMPUTE2:
(A) Minimac: ~ 60 hours
Step i – iii, Download the reference set for minimac: ~15 min
Step iv – viii, Creating the input files for imputation: ~5 min
Step ix, Using MaCH for phasing per chunk: ~15 hours
Step x, Imputation with minimac: ~45 hours
(B) IMPUTE2
Step i – ii, Download the reference set for IMPUTE2: ~10 min
Step iii – v, Creating the input files for imputation: ~10 min
Step vi, Using SHAPEIT for phasing per chromosome: varies per chromosome from 1.5 hours up to 5.5 hours.
Step vii, Imputation with IMPUTE2 per chunk: ~1 hour
Inexperienced analysts will typically require more time. The estimated times and memory requirements are based on the target and reference sets used in this protocol, the estimates may also vary with different cohort designs. Moreover, given the computational nature of this protocol, timing will also heavily depend on the computational resources available to the analyst, and to a lesser extent on the versions of the tools. The phasing and imputation steps are the most time consuming steps.

Troubleshooting
It is likely that many of the tools used in this protocol will be updated as time passes, we therefore recommend checking if there are new versions of the tools each time the protocol is run and what the changes between versions are.
Imputation with MaCH and minimac, step 10A(iv) and imputation with IMPUTE2, step 10B(iii): This step checks the concordance between SNPs within the target set and the reference panel based on position on the chromosome, assuming the SNP names are equal in both. This requires both panels to be aligned to the correct human genome build. Another option is to leave the SNPs which are in the target set and not in the reference panel. In that case, step 10A(iv) and 10A(v) (for MaCH and minimac) or step 10B(iii) and 10B(iv) (for IMPUTE2) can be replaced by `plink --noweb --file clean-GWA-data.HG19 --make-bed --out clean-GWA-data.HG19.for-impute.plink`. It is also important to have both the target set and the reference panel on the same human genome build, as IMPUTE2 links the two panels based on chromosome and position, not on SNP name.

Imputation with MaCH and minimac, step 10A(ix): The command line parameters *--interim 5* (to save intermediate results), *--sample 5* (random (but plausible) sets of haplotypes for each individual should be drawn every 5 iterations) and *--compact* (reduces memory use at the cost of runtime) can be removed from the command line to save time and disk space.
Imputation with MaCH and minimac, step 10A(x): The command line parameter *--rs* allows the use of rs GWAS SNP identifiers in the target set. This command line parameter can be removed if the target set does not include rs identifiers.
Imputation with IMPUTE2, step 10B(v): To increase the speed of the IMPUTE2 protocol, the target set could be reformatted into binary PLINK format (see BOX 1), therefore the `--recode` command should be replaced by `--make-bed`. The follow up steps 10B(vi) and 10B(vii) should be adjusted for binary files in that case.
Imputation with IMPUTE2, step 10B(vii): When the analyst wants to use two phased reference panels, the IMPUTE2 command should be replaced with `./impute_v2.3.1_x86_64_static/impute2 -known_haps_g ${namefile}.haps -m ${refdir}/gonl.chr${chr}.release4.gtc.geneticmap.txt -h ${refdir}/gonl.chr${chr}.release4.gtc.hap.gz ${refdir}/1000g.chr${chr}.release4.gtc.hap.gz -l ${refdir}/gonl.chr${chr}.release4.gtc.legend.gz ${refdir}/1000g.chr${chr}.release4.gtc.legend.gz -int ${startchr} ${endchr} -Ne 20000 -o ${namefile}.chunk${chunk}.impute2;`
When combining several of the commands into Bash shell script files, be sure to add `set -e` and `set -u` as the first two actual commands in the script. This makes sure that the script halts on errors and when undefined variables are being used, respectively. If additional debugging of Bash scripts is required, running a script like this: `bash -x scriptfile.sh` will run the script in debug mode, showing the value of variables, etc. Alternatively, if only a certain part of a Bash script is to be debugged, adding `set -x` before and `set +x` after the problematic part will enable debugging only for that part.

ANTICIPATED RESULTS

Converting the target set to the correct build. The genome-wide SNP data used in this protocol consists of 1,919 samples and 313,878 markers after performing quality control. After lifting this data set over from hg17 to hg19, the data set consists of 1,919 samples and 304,930 markers.

Imputation with MaCH and minimac. Imputation with minimac results in 8 files per chunk. Each files is a compressed (zipped) file. If needed such a file can be decompressed by running `gunzip -c filename.gz > filename`. Given the command for minimac specified earlier, the names of the outputfiles start with `chunk1-clean-GWA-data.HG19.for-impute.merlin.chr1` for chunk 1 of chromosome 1.

- a file with the extension .dose.gz which contains the imputed dosage for each genotype. Each row in the output will include one column per marker.

- a file with the extension .erate.gz which contains the error rate per marker.
- a file with the extension .hapDose.gz which contains the dosage for each haplotype separately.
- a file with the extension .haps.gz which contains the most likely alleles for each haplotype separately.
- a file with the extension .info.draft which contains the reference allele, non reference allele, frequency per marker. It also gives which markers were genotyped.
- a file with the extension .info.gz which contains the information about reference allele, frequencies and quality of imputations per marker. It also lists which markers were genotyped.
- a file with the extension .prob.gz which contains the imputed probabilities for each genotype. Each row in the output will include two columns per marker. The first of these columns denotes the probability of a homozygote for allele 1. The second column denotes the probability of a heterozygote.
- a file with the extension .rec.gz which contains the switch error rate per interval.

Imputation with IMPUTE2. Imputations of IMPUTE2 results in 5 files per chunk. Given the command for IMPUTE2 specified earlier, the names of the outputfiles start with `clean-GWA-data.HG19.for-impute.plink.chr1.phased.chunk1.impute2` for chunk 1 of chromosome 1:

- a file without any extra extension, this file contains the main results of the imputations. The first 5 entries of each line should be the SNP ID, rs ID of the SNP, base-pair position of the SNP, the allele coded *A* and the allele coded *B*. The subsequent columns contain the probabilities for the three genotypes *AA*, *AB* and *BB* for the each individual in the target set. This format allows for genotype uncertainty and therefore the probabilities for a given individual need not sum to 1.
- a file with the extension _info, this file contains the following columns: SNP identifier, rsID, base pair position, expected frequency of allele coded 1, measure of the observed statistical information associated with the allele frequency estimate, average certainty of best-guess genotypes and the internal "type" assigned to SNP.
- A file with the extension _info_by_sample which contains the concordance and the R^2 per sample.
- a file with the extension _summary which contains a summary of the screen output.
- a file with the extension _warnings which contains all warnings generated by IMPUTE2.

Box 1: input files for imputations

The input files for the various imputation tools

For MaCH and minimac, the target set that will be imputed needs to be stored per chromosome in Merlin[29] format. The Merlin pedigree file contains both the relationships, the phenotypes and the genotypes per individual per row. The first columns of the pedigree file contains the family identifier, the individual identifier, the father and mother identifiers, the sex of the individual (with females decoded as 2 and the males decoded as 1). The subsequent columns can encode phenotypes for discrete and quantitative traits followed by the genotypes. The alleles should be coded as 'A', 'C', 'G' or 'T' and missing alleles should be encoded with 'N', 'X' or '0'. Since MaCH and minimac assume samples to be unrelated, both the father and mother identifiers should be zero. The description of the columns is stored in the data file, with one row per column, indicating the data type (encoded as M- marker, A- affection status, T- quantitative trait and C- covariate) and providing a one-word label for each column.

For IMPUTE2 the genotype information should be stored in a one-line-per-SNP format. The first 5 entries of each line should be the SNP ID, rs ID of the SNP, base-pair position of the SNP, the allele coded *A* and the allele coded *B*. The subsequent columns contain the prior probabilities for the three genotypes *AA*, *AB* and *BB* for the each individual in the target set. This format allows for genotype uncertainty and therefore the probabilities for a given individual need not sum to 1. The order of samples in the genotype file should match the order of the samples in the sample file. The sample file has three parts (a) a header line detailing the names of the columns in the file, (b) a line detailing the types of variables stored in each column, and (c) a line for each individual detailing the information for that individual (more details on the IMPUTE2 file formats can be found at `http://www.stats.ox.ac.uk/~marchini/software/gwas/file_format.html`).

PLINK format to store genotyped data

The most commonly used file format for storing genotype data of the samples in the target set is the PLINK format (`http://pngu.mgh.harvard.edu/~purcell/plink/data.shtml#ped`). The pedigree file (extension `.ped`) in PLINK format is a headerless white-space (space or tab) delimited file which contains the pedigree information, the phenotype information and the genotype information for all samples in the data set. Every row corresponds to one individual and contains at least six columns which contain the family identifier, the individual identifier, the paternal and maternal identifier, the sex of the samples (with males encoded as 1 and females encoded as 2) and the phenotype of the sample, just like the Merlin format. Genotypes (column 7 onwards) can be any character (e.g. 1, 2, 3, 4 or A, C, G, T or anything else) except 0 which is, by default, the missing genotype character. All markers should be biallelic. All SNPs (whether haploid or not) must have two alleles specified and either both or neither alleles should be missing. The SNPs are described in the map file (extension `.map`), each line of the this file describes a single marker and must contain exactly 4 columns, the chromosome, the SNP identifier, the genetic distance in Morgans and the base-pair position in bp units. The ped and map file can be converted into a more memory- and time-efficient binary files with the extensions `.bed`, `.bim` and `.fam`.

REFERENCES

1. International HapMap 3 Consortium *et al.* Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52–58 (2010).
2. 1000 Genomes Project Consortium *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
3. Boomsma, D. I. *et al.* The Genome of the Netherlands: design, and project goals. *Eur J Hum Genet* **22**, 221–227 (2014).
4. Deelen, P. *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).
5. Genome of the Netherlands Consortium. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat Genet* **46**, 818–825 (2014).
6. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nat Rev Genet* **11**, 499–511 (2010).
7. Anderson, C. A. *et al.* Data quality control in genetic case-control association studies. *Nat Protoc* **5**, 1564–1573 (2010).
8. Browning, B. L. & Browning, S. R. A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am J Hum Genet* **84**, 210–223 (2009).
9. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* **5**, e1000529 (2009).
10. Verma, S. S. *et al.* Imputation and quality control steps for combining multiple genome-wide datasets. *Front Genet* **5**, 370 (2014).
11. Winkler, T. W. *et al.* Quality control and conduct of genome-wide association meta-analyses. *Nat Protoc* **9**, 1192–1212 (2014).
12. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & calo R Abecasis, G. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat Genet* **44**, 955–959 (2012).
13. van Leeuwen, E. M. *et al.* Genome of the Netherlands population-specific imputations identify an ABCA6 variant associated with cholesterol levels. *Nat Commun* **6**, 6065 (2015).
14. Jostins, L., Morley, K. I. & Barrett, J. C. Imputation of low-frequency variants using the HapMap3 benefits from large, diverse reference sets. *Eur J Hum Genet* **19**, 662–666 (2011).
15. Pistis, G. *et al.* Rare variant genotype imputation with thousands of study-specific whole-genome sequences: implications for cost-effective study designs. *Eur J Hum Genet* (2014).
16. Browning, S. R. & Browning, B. L. Haplotype phasing: existing methods and new developments. *Nat Rev Genet* **12**, 703–714 (2011).
17. Nho, K. *et al.* The effect of reference panels and software tools on genotype imputation. *AMIA Annu Symp Proc* **2011**, 1013–1018 (2011).
18. Li, Y., Willer, C. J., Ding, J., Scheet, P. & calo R Abecasis, G. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet Epidemiol* **34**, 816–834 (2010).
19. Delaneau, O., Zagury, J.-F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat Methods* **10**, 5–6 (2013).
20. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
21. Roshyara, N. R. & Scholz, M. fcgene: a versatile tool for processing and transforming snp datasets. *PLoS One* **9**, e97589 (2014).
22. Nelson, S. C., Doheny, K. F., Laurie, C. C. & Mirel, D. B. Is 'forward' the same as 'plus'?and other adventures in snp allele nomenclature. *Trends Genet* **28**, 361–363 (2012).

23. Deelen, P. *et al.* Genotype harmonizer: automatic strand alignment and format conversion for genotype data integration. *BMC Res Notes* **7**, 901 (2014).
24. Kent, W. J. *et al.* The human genome browser at UCSC. *Genome Res* **12**, 996–1006 (2002).
25. Sulovari, A. & Li, D. Gact: a genome build and allele definition conversion tool for snp imputation and meta-analysis in genetic association studies. *BMC Genomics* **15**, 610 (2014).
26. de Bakker, P. I. W. *et al.* Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Hum Mol Genet* **17**, R122–R128 (2008).
27. Wang, Z. *et al.* Improved imputation of common and uncommon SNPs with a new reference set. *Nat Genet* **44**, 6–7 (2012).
28. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559–575 (2007).
29. Abecasis, G. R., Cherny, S. S., Cookson, W. O. & Cardon, L. R. Merlin–rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet* **30**, 97–101 (2002).

CHAPTER 3.2

Improved imputation quality of low-frequency and rare variants in European samples using the 'Genome of The Netherlands'

Patrick Deelen, Androniki Menelaou, Elisabeth M. van Leeuwen, Alexandros Kanterakis, Freerk van Dijk, Carolina Medina-Gomez, Laurent C. Francioli, Jouke Jan Hottenga, Lennart C. Karssen, Karol Estrada, Eskil Kreiner-Møller, Fernando Rivadeneira, Jessica van Setten, Javier Gutierrez-Achury, Harm-Jan Westra, Lude Franke, David van Enckevort, Martijn Dijkstra, Heorhiy Byelas, Cornelia M. van Duijn, Genome of the Netherlands Consortium, Paul I. W. de Bakker, Cisca Wijmenga, Morris A. Swertz.

Published in European Journal of Human Genetics (Eur J Hum Genet, Jun 2014).
The supplemental information for this chapter is available online at:
http://www.nature.com/ejhg/journal/v22/n11/suppinfo/ejhg201419s1.html?url=/ejhg/journal/v22/n11/full/ejhg201419a.html

ABSTRACT

Although genome-wide association studies (GWAS) have identified many common variants associated with complex traits, low-frequency and rare variants have not been interrogated in a comprehensive manner. Imputation from dense reference panels, such as the 1000 Genomes Project (1000G), enables testing of ungenotyped variants for association. Here we present the results of imputation using a large, new population-specific panel: the Genome of The Netherlands (GoNL). We benchmarked the performance of the 1000G and GoNL reference sets by comparing imputation genotypes with 'true' genotypes typed on ImmunoChip in three European populations (Dutch, British, and Italian). GoNL showed significant improvement in the imputation quality for rare variants (MAF 0.05–0.5%) compared with 1000G. In Dutch samples, the mean observed Pearson correlation, r^2, increased from 0.61 to 0.71. We also saw improved imputation accuracy for other European populations (in the British samples, r^2 improved from 0.58 to 0.65, and in the Italians from 0.43 to 0.47). A combined reference set comprising 1000G and GoNL improved the imputation of rare variants even further. The Italian samples benefitted the most from this combined reference (the mean r^2 increased from 0.47 to 0.50). We conclude that the creation of a large population-specific reference is advantageous for imputing rare variants and that a combined reference panel across multiple populations yields the best imputation results.

INTRODUCTION

Although genome-wide association studies (GWAS) have been very effective in identifying loci associated with diseases or traits[1], it has proved difficult to fine-map the association signals to causal variants[2,3]. To overcome these limitations, there has been increasing interest in the interrogation of less frequent variants, especially given the enrichment of deleterious alleles at low frequencies[4-7]. There are specialized chips that can assess a larger number of rare variants, like the ImmunoChip[8] or Metabochip[9], although they do not provide uniform genome-wide coverage. Hence, most investigators will use statistical imputation from SNP arrays in GWAS using dense reference panels.

Imputation using a densely typed reference set can be performed to infer untyped variants that can be used to improve the power of a GWAS[10], and there are numerous examples in which imputation has effectively enriched the results in GWAS[11,12]. Although most large studies have so far been based on meta-analysis of HapMap-based imputations across cohorts, the primary limitation is that HapMap is essentially restricted to common variation (MAF > 5%). Thanks to the sequencing of larger samples, such as 1000G, more complete reference panels are now being assembled, setting off a new wave of meta-analyses.

The power of detecting an association in a GWAS is determined by its sample size and effective genome-wide coverage of the included variants, among other things[13,14]. The effective coverage depends directly on the number and quality of the imputed genotypes[15]. In turn, the quality of the reference panel will depend largely on the number of samples, the quality of the haplotypes, and the number of variants included[16].

The Genome of The Netherlands (GoNL) has the potential to provide a good imputation reference panel. GoNL is a population based sequencing project, in which 769 Dutch samples were sequenced at, on average, 14x coverage[17]. In particular, the fact that GoNL sequenced trios (231) or quartets (19) has enabled improved haplotype phasing by using one of the children[18]. The GoNL imputation reference set contains 998 unrelated haplotypes. In this paper, we report a quantitative analysis to assess the quality of imputed genotypes from using both GoNL and 1000G in Dutch and other European populations.

We adopted a 'gold standard' approach using samples genotyped on two distinct platforms, HumanHap550 and ImmunoChip. Hap550 is a commonly used genotyping chip designed to tag as many haplotypes as possible using common variants. ImmunoChip, however, is a fine-mapping chip: it contains a large number of low frequency and rare variants for a limited number of loci (primarily selected on the basis of loci identified in immune-related traits). Starting from the Hap550-genotyped SNPs, we were able to impute a large number of variants present on ImmunoChip. We then compared these imputed genotypes with the measured ('gold standard') genotypes on ImmunoChip to quantify the imputation performance. We have such a data set for three European populations: the Dutch, British, and Italians. For each

population we used 745 samples genotyped on both platforms. These three populations allowed us to ascertain population-specific differences in the imputation quality of SNPs.

MATERIAL AND METHODS

Genome of the Netherlands

GoNL is a project in which 769 individuals from different Dutch provinces were sequenced at, on average, 14x coverage[17]. All samples are part of either one of the 231 trios or one of the 19 quartets. The phasing was performed using the trio information[18], and for the quartets one of the children was used to enhance the phasing. Because of sequence failures of two parents, from different trios, these samples were excluded from the imputation reference set. Instead, from these two trios, we used the haplotype of the child that was not present in the other parent. This resulted in an imputation reference set containing 998 unrelated haplotypes. We used GoNL release 4 for all our analyses (see http://www.nlgenome.nl). The current GoNL release 5 also contains over one million indels but did not change the SNPs.

Benchmarking samples

Samples from a celiac disease patient cohort were selected, since they had been genotyped on both the Hap550 and ImmunoChip[19]. The 745 Dutch and the 745 British samples were all cases, while the 745 Italian samples comprised 371 cases and 374 controls. The clustering for the genotype calling of the ImmunoChip data was performed manually in the past, to ensure proper genotyping results.

The Hap550 (516,426 SNPs) data was filtered on MAF > 1% and HWE *p*-value > $1 \cdot 10^{-4}$ for each population separately. The ImmunoChip (113,991 SNPs) data was filtered on MAF > 0.05% and HWE *p*-value of $1 \cdot 10^{-4}$. Both datasets are filtered on variants present in both the 1000G reference set as in the GoNL reference set. After QC the Dutch, British and Italian Hap550 data contain 509,888, 509,984 and 510,225 SNPs. The ImmunoChip data contains in the same order 107,383, 107,212 and 107,611 SNPs.

Combining 1000G and GoNL data

The reference set combining data from 1000G and GoNL was created using the Impute2 option: "--merge_ref_panels". This merged reference set was written to a file and subsequently used for the benchmarking. Since our benchmarking data is filtered for variants present in both reference sets, we did not assess the imputations of variants that are unique to either reference set.

Pre-phasing

The 745 samples for each population were pre-phased using SHAPEIT2[15]. This was done per chromosome using the default settings.

Imputation

The imputations were performed using Impute2 2.3.0[16]. The different populations were imputed separately and in chunks of 5 Mb. For the comparison using an equal number of identical European haplotypes, we performed an imputation using all 379 European 1000G samples and a random selection of 379 GoNL samples. The random selection of GoNL samples was performed stratified on the Dutch provinces. These samples were selected using the Impute2 option: "--exclude_samples_h".

We used MOLGENIS compute to implement the imputation pipeline, run the 8,835 imputation chunks in parallel on a PBS compute cluster, and to keep track of the 15 imputations (five for each population). All pipelines are available as open source via http://www.molgenis.org/wiki/ComputeStart.

Gold standard method

As stated above, we used samples genotyped on two distinct platforms. We imputed the Hap550 genotypes from these samples and compared the imputed genotypes to the SNPs previously only present in the ImmunoChip data. We used the ImmunoChip data as our 'gold standard'. The concordance between imputed genotypes and ImmunoChip genotypes was determined by calculating the Pearson correlation r^2 between the imputed dosage and ImmunoChip observed genotypes. The mean concordances were calculated for three MAF bins: rare (≥ 0.05% and < 0.5%), low-frequency (≥ 0.5% and < 5%) and common (> 5%) SNPs. The MAF used to stratify the SNPs into the bins was calculated separately for each population. The results were plotted using R 2.14.2. The significance of the differences between the reference sets was calculated using the Wilcoxon signed-rank test implementation in R.

Principal component analysis

The principal component analysis (PCA) was performed using the EIGENSOFT 4.2 package[20]. The components were calculated using the European 1000G, GoNL, and the 3 GWAS datasets that we used for benchmarking. Before the components were calculated, all datasets were filtered to only include variants with a MAF > 5%. A joint dataset, featuring variants present in all 5 datasets, was created. This dataset was again filtered for MAF > 5%, the merged data was also filtered on HWE *p*-value > $1 \cdot 10^{-4}$ and a call rate of 95%. This dataset was pruned using PLINK 1.07[21] with the "--indep-pairwise" option, windows: 1000, step: 5, r^2 threshold: 0.2. The first component explained 0.33% of the variation and the second 0.10%. All subsequent components described less than 0.06%.

RESULTS

We stratified our analysis into three groups: common variants (MAF ≥ 5%), low-frequency variants (MAF 0.5%–5%), and rare variants (MAF 0.05%–0.5%). We focused mainly on the rare variants, since these are more difficult to impute and most can be gained in terms of imputation quality when using a better reference set. We observed a large increase in the imputation quality of rare variants when using GoNL as the reference compared to 1000G (Figure 1, Table 1). The mean observed Pearson correlation (r^2) showed a significant increase from 0.61 to 0.71 for Dutch samples (Wilcoxon *p*-value = $7.16 \cdot 10^{-60}$). The British and Italian imputations also showed a significant improvement when imputing rare variants, from 0.58 to 0.65 ($p = 3.70 \cdot 10^{-35}$) and from 0.43 to 0.47 ($p = 2.64 \cdot 10^{-13}$), respectively. GoNL also significantly outperformed the 1000G reference set in the imputation of variants with higher MAFs (Supplementary Figures/Appendices S1, S2, S3).

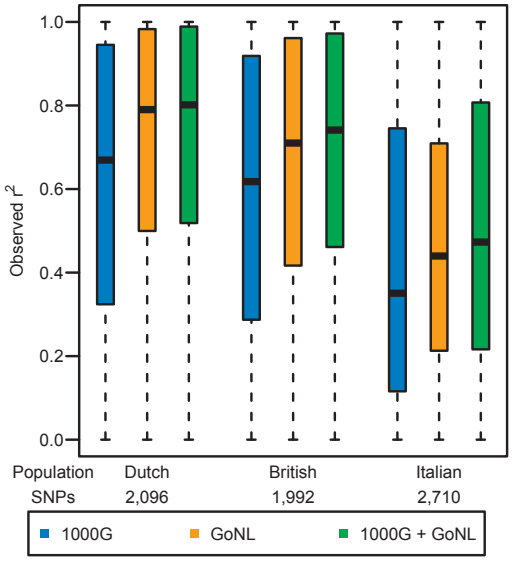


Figure 1. Comparison of imputation quality of rare variants using the 1000G data, GoNL, and the combined reference panel.

Table 1. Mean observed r^2 of rare variants.

Reference set	Dutch	British	Italian
1000G	0.61	0.58	0.43
GoNL	0.71	0.65	0.47
1000G + GoNL	0.72	0.67	0.50

Differences in the mean imputation quality between the reference sets was significant for each population ($p < 0.001$).

Using a combined reference set composed of the 1000G and GoNL samples, we could improve the imputation further. The imputation of rare variants using the combined reference in Dutch and British samples showed a small increase in quality compared to GoNL-only imputation, respectively 0.02 ($p = 1.16 \cdot 10^{-3}$) and 0.02 ($p = 2.70 \cdot 10^{-5}$). The Italians benefitted most from the combined reference with an increase of 0.04 ($p = 3.62 \cdot 10^{-30}$) compared to a GoNL-only reference, resulting in a mean concordance for rare variants of 0.5. The differences in imputation quality when using the combined reference set for more frequent alleles were either very small or not significant (Supplementary Figure S1, Supplementary Tables S2 and S3).

A striking trend in these results is that the imputation quality of rare variants in Italians samples is lower than that in Dutch and British samples. The Dutch and Italian samples were genotyped at the same center and have similar call rates, and there were no indications that the genotyping quality of the Italian samples was lower. However, a principal component analysis (PCA) revealed that the Italian samples were not as well represented by either 1000G or GoNL compared to the Dutch and British GWAS samples used for benchmarking (Figure 2).

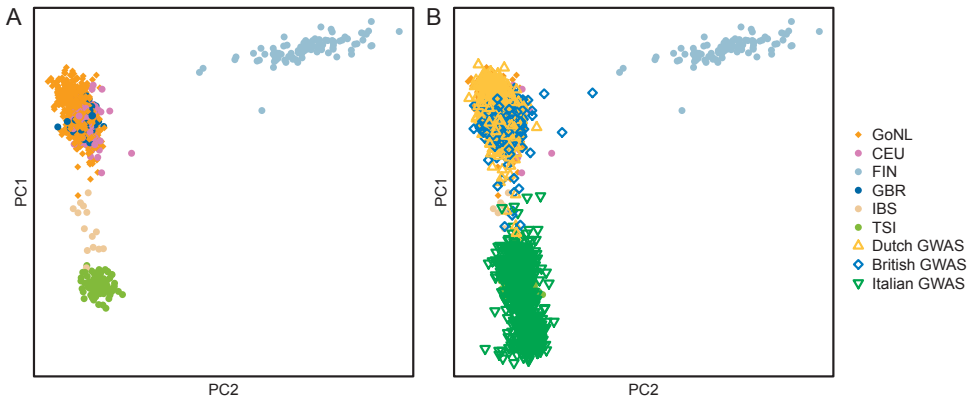


Figure 2. Clustering of reference and study samples. PC1 and PC2 reveal 3 main clusters: Tuscans from Italy (TSI), Finnish (FIN), and a Western European cluster with the CEU (Utah Residents with Northern and Western European ancestry), the GBR (British) and the GoNL samples (Panel A). Panel B shows that most of our GWAS samples clustered in a similar way to the corresponding 1000G/GoNL samples.

We assessed whether the better performance of GoNL compared to 1000G was due to the larger number of European haplotypes in the reference set (998 vs. 578 in 1000G). We did this by performing an imputation using solely the 379 European samples in 1000G and a random subset of 379 GoNL samples. We found that the GoNL subset also significantly outperformed the European 1000G subset (Table 2).

Table 2. Mean observed r^2 of rare variants for reference sets of equal sample size from 1000G and GoNL (all of European descent).

Reference set	Dutch	British	Italian
1000G European	0.59	0.57	0.40
GoNL random subset 379 samples	0.68	0.64	0.45

Differences in the mean imputation quality between the reference sets was significant for each population ($p < 0.001$).

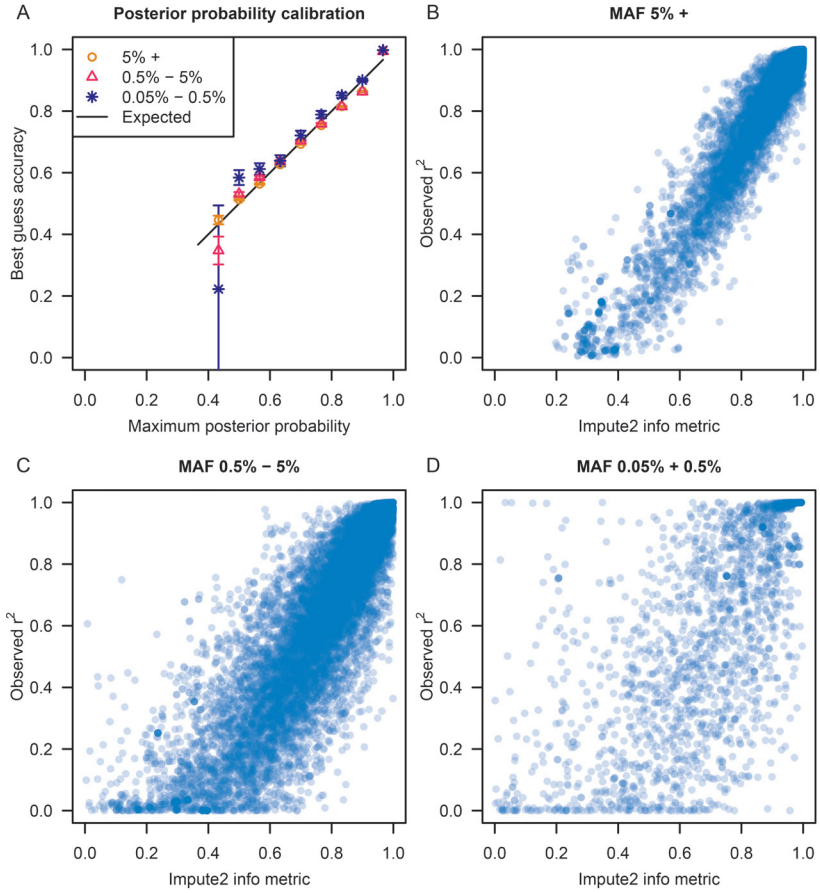


Figure 3. Calibration of posterior probabilities. The posterior probabilities were, in general, well calibrated, although there were a few deviations from the expected accuracy (panel A). For common and low-frequency variants (panels B & C), we observed a strong correlation (r^2 0.97 and 0.91, respectively) between the impute2 info metric and the observed r^2. However, for the rare variants (panel D), the relation between predicted and observed quality was less profound. We also observed a correlation of 0.70 and several large deviations from the diagonal.

Our experimental design also allowed us to assess the calibration of the posterior probabilities of the genotypes as they are output by IMPUTE2. We observed that the posterior probabilities were, in general, well calibrated, although we did observe a few deviations for low-frequency and rare variants (Figure 3A). To ascertain if these deviations in posterior probabilities affect the predicted imputation quality, the IMPUTE2 info metric, we plotted the predicted quality against the observed r^2. This showed a strong correlation between the predicted and observed quality for common variants and low-frequency variants (correlation of 0.97 and 0.91, respectively; Figure 3B & 3C). However, the info metric is not as accurate for rare variants, and the correlation with observed r^2 dropped to 0.70 (Figure 3D). We also observed some discrepancies where a near perfect imputation was predicted while in fact there was poor imputation, and vice versa when assessing rare variants.

DISCUSSION

We have shown that the new GoNL reference set provides higher downstream imputation accuracy than the 1000G reference set, not only for Dutch samples, but also for other European populations studied in this paper. Aside from the increase in imputation quality of rare variants in Dutch samples from 0.61 (1000G) to 0.71 (GoNL), we also observed an increase in imputation quality in British (0.58 to 0.65) and Italian (0.43 to 0.47) samples. We show that GoNL yielded better imputed genotypes for at least these European populations. A combined reference set, of 1000G and GoNL, increased the mean imputation quality of rare variants even further to 0.72, 0.67 and 0.50 for the Dutch, British and Italians, respectively.

By selecting an identical number of European haplotypes from 1000G and from GoNL, we showed a strong added value for GoNL in all the tested populations, confirming that the trio design of GoNL and the resulted accurate haplotypes aid the downstream imputation quality. We also observed a population-specific added value of GoNL when imputing Dutch samples. The added value (i.e. mean increase in imputation quality) was largest when comparing GoNL to 1000G in imputing the Dutch samples. Of course, it was already known that a better matched reference set will result in better imputed genotypes[13], however, the results from this paper were based on low-frequency variants and we show that there is also an inter-European effect of reference sets.

It is important to note that we only assessed variants present on the ImmunoChip. Although these variants were not randomly selected, we have no reason to assume that the imputation quality will be positively biased or that they do not represent low-frequency variants in general. The ImmunoChip was made to fine map loci previously associated to autoimmune diseases using a large number of low-frequency and rare variants.

We were encouraged to observe that the posterior probabilities were, in general, well calibrated with respect to the gold standard genotypes. We observed no adverse effects on the accuracy of the IMPUTE2 info metrics, although for rare variants we did observe a few instances with large deviations between the predicted and observed quality. This is in line with previous observations[22]. This observed inaccuracy also emphasizes the importance of validating associations from imputed genotypes.

It was shown earlier that a larger and more diverse reference set can improve the imputation of low-frequency variants[23]. We observed that a combination of 1000G and GoNL showed limited added value for the imputation of rare variants in the Dutch and British samples. It was, however, interesting to observe that the imputation of the Italian samples was improved more by this combined reference panel, leading us to speculate that populations that are poorly represented in the reference panel benefit more from a large and diverse reference set. Despite the limited added value for the Dutch and British datasets, such a large reference set may still be of interest for consortia aiming to impute cohorts of both European and non-European origin. All these cohorts can be imputed using the same combined reference set and then use IMPUTE2 to automatically select the best matching haplotypes[24]. We should note that we were only able to assess variants present in both reference sets, since there are very few variants on the ImmunoChip that are unique to either GoNL or 1000G. Nonetheless, our results show that population-specific reference sets and cosmopolitan panels, such as 1000G, can augment each other. This even holds true for the imputation of samples with ancestry other than those present in the population-specific reference sets, which provides further motivation for international efforts towards large and integrated reference sets.

REFERENCES

1. Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A* **106**, 9362–9367 (2009).
2. Wellcome Trust Case Control Consortium *et al.* Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nat Genet* **44**, 1294–1301 (2012).
3. Shea, J. *et al.* Comparing strategies to fine-map the association of common snps at chromosome 9p21 with type 2 diabetes and myocardial infarction. *Nat Genet* **43**, 801–805 (2011).
4. Kryukov, G. V., Pennacchio, L. A. & Sunyaev, S. R. Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *Am J Hum Genet* **80**, 727–739 (2007).
5. Cirulli, E. T. & Goldstein, D. B. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat Rev Genet* **11**, 415–425 (2010).
6. Lee, S., Wu, M. C. & Lin, X. Optimal tests for rare variant effects in sequencing association studies. *Biostatistics* **13**, 762–775 (2012).
7. Huyghe, J. R. *et al.* Exome array analysis identifies new loci and low-frequency variants influencing insulin processing and secretion. *Nat Genet* **45**, 197–201 (2013).
8. Cortes, A. & Brown, M. A. Promise and pitfalls of the immunochip. *Arthritis Res Ther* **13**, 101 (2011).
9. Keating, B. J. *et al.* Concept, design and implementation of a cardiovascular gene-centric 50 k snp array for large-scale genomic association studies. *PLoS One* **3**, e3583 (2008).
10. Hao, K., Chudin, E., McElwee, J. & Schadt, E. E. Accuracy of genome-wide imputation of untyped markers and impacts on statistical power for association studies. *BMC Genet* **10**, 27 (2009).
11. Holm, H. *et al.* A rare variant in MYH6 is associated with high risk of sick sinus syndrome. *Nat Genet* **43**, 316–320 (2011).
12. Li, Y., Willer, C., Sanna, S. & calo Abecasis, G. Genotype imputation. *Annu Rev Genomics Hum Genet* **10**, 387–406 (2009).
13. de Bakker, P. I. W. *et al.* Efficiency and power in genetic association studies. *Nat Genet* **37**, 1217–1223 (2005).
14. Flannick, J. *et al.* Efficiency and power as a function of sequence coverage, snp array density, and imputation. *PLoS Comput Biol* **8**, e1002604 (2012).
15. Zheng, J., Li, Y., Abecasis, G. R. & Scheet, P. A comparison of approaches to account for uncertainty in analysis of imputed genotypes. *Genet Epidemiol* **35**, 102–110 (2011).
16. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* **5**, e1000529 (2009).
17. Boomsma, D. I. *et al.* The Genome of the Netherlands: design, and project goals. *Eur J Hum Genet* **22**, 221–227 (2014).
18. Menelaou, A. & Marchini, J. Genotype calling and phasing using next-generation sequencing reads and a haplotype scaffold. *Bioinformatics* **29**, 84–91 (2013).
19. Trynka, G. *et al.* Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. *Nat Genet* **43**, 1193–1201 (2011).
20. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38**, 904–909 (2006).
21. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81**, 559–575 (2007).
22. Li, L. *et al.* Performance of genotype imputation for rare variants identified in exons and flanking regions of genes. *PLoS One* **6**, e24945 (2011).

23. Jostins, L., Morley, K. I. & Barrett, J. C. Imputation of low-frequency variants using the HapMap3 benefits from large, diverse reference sets. *Eur J Hum Genet* **19**, 662–666 (2011).
24. Howie, B., Marchini, J. & Stephens, M. Genotype imputation with thousands of genomes. *G3 (Bethesda)* **1**, 457–470 (2011).

CHAPTER 3.3

Genome of the Netherlands population-specific imputations identify an *ABCA6* variant associated with cholesterol levels

Elisabeth M. van Leeuwen, Lennart C. Karssen, Joris Deelen, Aaron Isaacs, Carolina Medina-Gomez, Hamdi Mbarek, Alexandros Kanterakis, Stella Trompet, Iris Postmus, Niek Verweij, David J. van Enckevort, Jennifer E. Huffman, Charles C. White, Mary F. Feitosa, Traci M. Bartz, Ani Manichaikul, Peter K. Joshi, Gina M. Peloso, Patrick Deelen, Freerk van Dijk, Gonneke Willemsen, Eco J. de Geus, Yuri Milaneschi, Brenda W. J. H. Penninx, Laurent C. Francioli, Androniki Menelaou, Sara L. Pulit, Fernando Rivadeneira, Albert Hofman, Ben A. Oostra, Oscar H. Franco, Irene Mateo Leach, Marian Beekman, Anton J.M. de Craen, Hae-Won Uh, Holly Trochet, Lynne J. Hocking, David J. Porteous, Naveed Sattar, Chris J. Packard, Brendan M. Buckley, Jennifer A. Brody, Joshua C. Bis, Jerome I. Rotter, Josyf C. Mychaleckyj, Harry Campbell, Qing Duan, Leslie A. Lange, James F. Wilson, Caroline Hayward, Ozren Polasek, Veronique Vitart, Igor Rudan, Alan F. Wright, Stephen S. Rich, Bruce M. Psaty, Ingrid B. Borecki, Patricia M. Kearney, David J. Stott, L. Adrienne Cupples, The Genome of the Netherlands Consortium, J. Wouter Jukema, Pim van der Harst, Eric J. Sijbrands, Jouke-Jan Hottenga, Andre G. Uitterlinden, Morris A. Swertz, Gert-Jan B. van Ommen, Paul I.W. de Bakker, P. Eline Slagboom, Dorret I. Boomsma, Cisca Wijmenga and Cornelia M. van Duijn.

Published in Nature Communications (Nat Commun, 6:6065, 2015).

The supplemental information for this chapter is available online at:
http://www.nature.com/ncomms/2015/150120/ncomms7065/abs/ncomms7065.html#supplementary-information

ABSTRACT

Variants associated with blood lipid levels may be population-specific. To identify these low-frequency variants associated with this phenotype, population-specific reference panels may be used. Here we impute nine large Dutch biobanks (~35,000 samples) with the population-specific reference panel created by the Genome of the Netherlands Project and perform association testing with blood lipids levels. We report the discovery of five novel associations at four loci (*p*-value < $6.61 \cdot 10^{-4}$), including a rare missense variant in *ABCA6* (rs77542162, p.Cys1359Arg, frequency 0.034) which is predicted to be deleterious. The frequency of this *ABCA6* variant is 3.65-fold increased in the Dutch and its effect ($\beta_{LDL\text{-}C}$=0.135, β_{TC}=0.140) is estimated to be very similar to those observed for single variants in well-known lipid genes, such as *LDLR*.

INTRODUCTION

Genome-wide association studies (GWAS) have identified a large number of loci associated with blood lipid levels and analysis suggest there are additional susceptibility loci that have not yet been discovered[1-3]. Despite the fact that rare functional variants are known to play a major role in lipid metabolism[1-3], there has been limited success in finding such variants in population-based studies using next generation sequencing. Even if the effect of these variants is expected to be larger than that of common variants, the sample size needed to detect these rare or low frequent variants increases dramatically with variant rarity. As the frequency of rare variants may increase in certain populations due to drift and founder effects[4], the power of searches for rare functional variants may improve by the use of reference sets specific to distinct populations. Such references allow for better quality imputation of rare variants especially those with increased frequency in the population of interest[3,5,6]. Previous studies have successfully detected rare variants by imputation into larger sets of individuals in isolated populations followed by association testing to detect variants associated to the trait of interest[7-9].

Here we describe an imputation-based GWAS for circulating lipid levels using a custom-built reference panel for the Dutch population (Genome of the Netherlands, GoNL, http://www.nlgenome.nl/), in which the whole genome of 250 parent-offspring trios were sequenced at approximately 13x coverage[5,6]. Due to the trio design, the phasing quality of the reference panel was better than that of the 1000 Genomes Phase 1 panel. In this study we show that using this population-specific reference panel we were able to identify five novel associations at four loci.

METHODS

Study descriptions

The descriptions of the including cohorts can be found in the supplementary methods. A written informed consent was obtained from all study participants for all cohorts and local ethical committees at participating institutions approved individual study protocols.

Study samples and phenotypes

A summary of the details of both the discovery and replication cohorts participating in this study can be found in Supplementary Tables 1 and 12.

Only samples of Dutch ancestry were used in the discovery cohorts, the samples in the replication cohorts are from various ancestries, see Supplementary Table 12. In all studies except MESA Whites, all individuals that used lipid lowering medication at the time the

lipid levels were measured, were excluded. In MESA Whites the total cholesterol values for individuals on lipid lowering medication was divided by 0.8. In all studies except for LLS and PREVEND, the subjects were fasting when the lipid levels were measured. In LLS all samples were non-fasted and in PREVEND 2.99% were non-fasted. The LDL-C levels were measured within the ERF, Croatia Korcula, Croatia Split, Croatia Vis, FamHS and Lifelines cohorts, within the other cohorts the Friedewald equation was used to calculate the LDL-C levels[10].
The lipid measurements were adjusted for sex, age and age^2 in all cohorts. Various methods were used to account for family relationships: in ERF grammar-gamma (GenABEL version 1.7.6[11,12], was used, in the Croatia Korcula, Croatia Split, Croatia Vis and Generation Scotland cohorts mmscore (GenABEL[11] was used, and in LLS qt-assoc was used. In CHS the clinic was used as extra covariate, in Lifelines PC1 and PC2, in FamHS the field center, the genotyping array (Illumina 550k, 610k and 1M), PC5 only for TC and PC1 only for LDL, in FHS the cohort (offspring and third generation) and PCs, in MESA Whites 2 PCs and study site, in NTR-NESDA PCs and chip effect, in ORCADES the genotyping array and PC1, PC2 and PC3, in PROSPER-Dutch only PC1 and in both PROSPER-Scottish and PROSPER-Irish PC1-PC4.

Genotyping and imputations

Detailed information about genotyping and imputations per cohort can be found in the supplementary methods. In summary, all cohorts were genotyped using commercially available Affymetrix or Illumina genotyping arrays, or custom Perlegen arrays. Quality control was performed independently for each study. To facilitate meta-analysis, each replication cohort performed genotype imputation using IMPUTE[13], or Minimac[14] with reference to the GoNL project data for the discovery cohorts and with reference to the 1-kG project data for the replication cohorts.

GWAS in all discovery cohorts

All nine discovery cohorts ran separate the genome-wide association study for each of the four traits: HDL-C, LDL-C, TC and TG. Supplementary Table 13 shows the genomic control factor λ per trait per cohort and Supplementary Figs 10-13 show the λ per MAF bin per trait per cohort. We therefore used only the SNPs with a $R^2 > 0.3$, $R^2 < 1.1$ and expected minor allele count (expMAC = $2 \cdot MAF \cdot R^2 \cdot$ sample size) > 10. Most inflation is observed within the ERF study, especially in the lowest-frequency variants, this is probably caused by the family structure in this cohort.

Meta-analysis of discovery cohorts

The association results of all studies were combined and the standard error-based weights were calculated by METAL[15]. This tool also applies genomic control by automatically correcting the test statistics to account for small amounts of population stratification or unaccounted

relatedness. METAL also allows for heterogeneity. We used the following filters: $0.3 < R^2 < 1.1$ and expMAC > 10.

After meta-analyses of all available variants, we excluded the variants that are not present in at least 6 of the 9 cohorts. We also excluded all variants that are labeled as being in the inaccessible genome, since the quality of those SNPs cannot be guaranteed[16]. The remaining variants per trait, see Supplementary Table 14, were used to create Manhattan plots and QQ-plots, see Supplementary Figs 14-15. The meta-analysis resulted in 1,905 SNPs with a *p*-value less than $5 \cdot 10^{-8}$ for HDL-C, 2,626 SNPs for LDL-C, 3,133 SNPs for TC and 1,310 for TG.

Confirmation of known loci

Previously, Teslovich *et al.*[1] and Willer *et al.*[2] identified 157 loci associated with one of more of the lipids. Teslovich *et al.*[1] identified 47, 37, 52 and 32 loci to be associated with HDL-C, LDL-C, TC and TG, respectively. The positions of these loci were reported on human genome build 36, we therefore lifted these positions over to human genome build 37 and checked the association results after the meta-analysis of all discovery cohorts. The effect size of these loci was reported in mg dL^{-1}, whereas in this study we use mmol L^{-1}. We therefore multiplied the effect size for the loci associated with TG with 0.0259 and the other loci with 0.011. Supplementary Fig. 2 and Supplementary Table 6 show the comparison per trait of our meta-analysis of all discovery cohorts with the results of the meta-analysis by Teslovich *et al.*[1]. We did the same for the loci identified by Willer *et al.*[2], see Supplementary Fig. 3 and Supplementary Table 7. The effect size of these loci could not be compared with our results, since trait residuals within each study participating in the meta-analysis of Willer *et al.*[2] were adjusted for sex, age and age^2 and subsequently quantile normalized. Their GWAS was done with the inverse normal transformed trait values.

Selection of independent variants

In order to select only associated variants that were independent of previous findings, we used the GCTA tool[17]. This tool performs a stepwise selection procedure to select multiple associated SNPs by a conditional and joint analysis approach using summary-level statistics from a meta-analysis and linkage disequilibrium (LD) corrections between SNPs estimated from the GoNL reference panel, release 4. This analysis revealed 60 independent variants associated with HDL-C, 142 independent variants associated with LDL-C, 134 independent variants associated with TC and 16 independent variants associated with TG. By using this approach, we were able to identify additional independent variants in known loci. Figure 1 shows that we identified both common and rare variants and more rare variants compared to Teslovich *et al.*[1] and Willer *et al.*[2]. There is overlap between the genome-wide significant SNPs of the different traits, and also between the independent SNPs of the different traits, as shown in Supplementary Fig. 1.

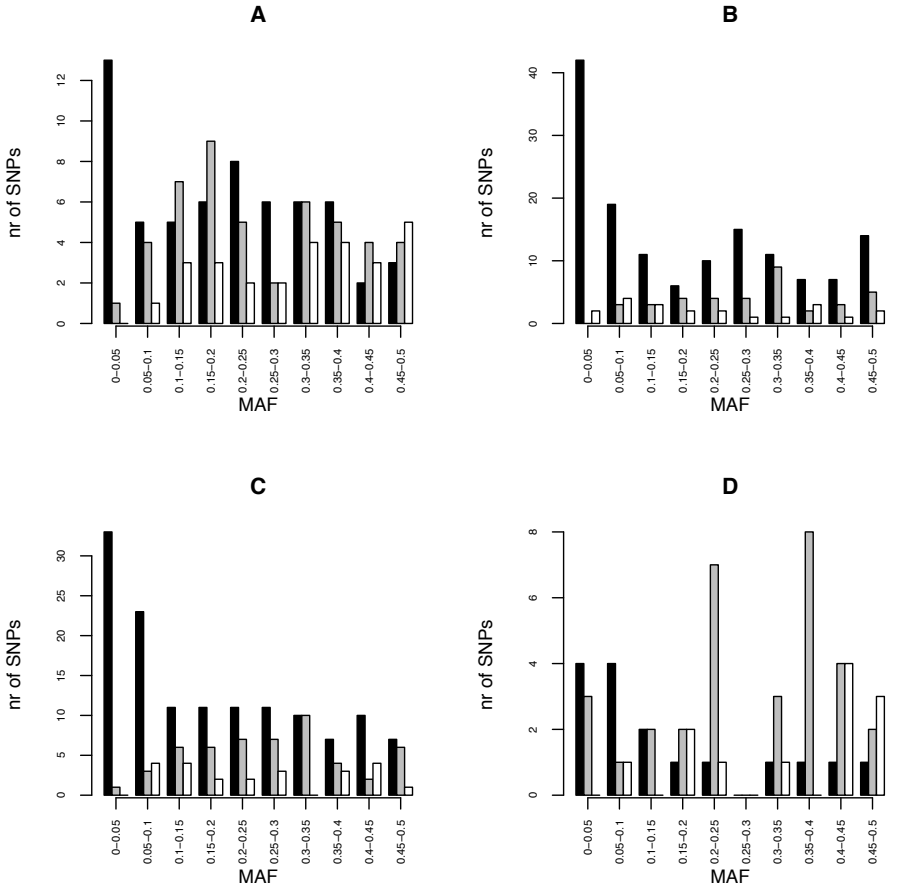


Figure 1. Identified variants for plasma lipid levels. Distribution of the variants identified by conditional analysis implemented by GCTA to be independently associated with the lipid traits (A: HDL-C (60 variants), B: LDL-C (142 variants), C: TC (134 variants) and D: TG (16 variants)) over MAF bins after meta-analysis of discovery cohorts (black). The histograms also includes loci identified by Teslovich *et al.*[1] (gray) and Willer *et al.*[2] (white).

Identification of potential novel variants

To identify potential novel variants, we first excluded all variants within 1 Mb of a known loci from Teslovich *et al.*[1] and from Willer *et al.*[2]. Since the number of loci associated with the four traits differ, we end up with 7,946,245 SNPs for HDL-C, 8,014,693 SNPs for LDL-C, 7,923,530 SNPs for TC and 7,468,790 SNPs for TG. For all traits we do find some genome-wide significant loci, see Supplementary Figs 16 and 17. We used the GCTA tool to select only those variants that are independent associated with the lipid trait. This analysis revealed 2 novel independent variants associated with HDL-C, 1 novel independent variants associated with LDL-C, 2 novel independent variants associated with TC and 1 novel independent variants

associated with TG, see Supplementary Table 8 and Supplementary Fig. 18. We used PLINK to test if these 6 variants are in LD with the knowLn loci from Teslovich *et al.*[1] and from Willer *et al.*[2]. None of the 6 variants are in LD with known loci associated with the same trait on the same chromosome ($R^2 < 0.14$).

Replication of potential novel variants

The 6 potential novel loci were replicated in 11 cohorts: CHS, Croatia-Korcula, Croatia-Split, Croatia-Vis, FamHS, FHS, Generation Scotland, MESA Whites, ORCADES, PROSPER-Scottish and PROSPER-Irish. The association results of all cohorts were combined and the standard error based weights were calculated by METAL[15]. The Bonferroni-correction for multiple testing was $8.33 \cdot 10^{-3}$. This resulted in the significant replication of 5 out of the 6 variants, see Supplementary Fig. 19 and Supplementary Table 11.

Conditional analysis

Within the discovery cohorts we performed a conditional analysis to see if the novel variants are independent of the known loci from Teslovich *et al.*[1] and from Willer *et al.*[2]. Supplementary Table 10 shows the results within these cohorts with and without adjusting for the known loci for the trait in question, if available in the GoNL reference panel. Since the unadjusted and adjusted results are similar, we conclude that the newly identified variant are independent of the known loci.

RESULTS

Nine large Dutch epidemiological cohorts (comprising 36,000 samples in total) were imputed with the GoNL reference panel (~ 19.5 million SNPs) on an identical protocol[6,18]. All cohorts conducted association analysis on the imputed variants assuming an additive genetic effect on high-density lipoprotein cholesterol (HDL-C), low-density lipoprotein cholesterol (LDL-C), total cholesterol (TC) and triglyceride (TG) levels (Methods, Supplementary Methods and Supplementary Table 1) and results were meta-analysed. We used conditional analysis implemented in GCTA[17] to identify variants associated independently with lipid levels.

Both rare (minor allele frequency (MAF) < 0.01), low (0.01 < MAF < 0.05) and common variants (MAF > 0.05) were associated with HDL-C (*N* = 60 variants), LDL-C (*N* = 142 variants), TC (*N* = 134 variants) and TG (*N* = 16 variants) in both known and novel loci (Methods, Supplementary Table 2-5 and Supplementary Fig. 1). In Figure 1 we compare the allele frequencies that reach genome-wide significance in the GCTA analysis (*p*-value < $5 \cdot 10^{-8}$) to those reported by Teslovich *et al.*[1] and Willer *et al.*[2] (Figure 1). The majority of the known HDL-C (31 of 45, 68.9%), LDL-C (24 of 34, 70.6%), TC (33 of 48, 68.6%) and TG (13 of 30, 43.3%) loci described

by Teslovich *et al.*[1] replicated at a *p*-value < $3.18 \cdot 10^{-4}$ (Bonferroni correction based on 157 variants) (Methods, Supplementary Figs 2-3 and Supplementary Tables 6-7). We also confirmed several of the HDL-C (6 of 27, 22.2%), LDL-C (7 of 21, 33.3%), TC (4 of 23, 17.4%) and TG (1 of 12, 8.3%) loci described by Willer *et al.*[2] at a *p*-value < $6.02 \cdot 10^{-4}$ (Bonferroni correction based on 83 variants) despite a sample size of about 20% of the other studies.

To identify novel loci associated with blood lipid levels, we selected from the list of variants identified by GCTA, those variants located more than 1Mb away from previously identified loci. This resulted in six novel associations at five loci (Methods, Table 1 and 2 and Supplementary Table 8). The five loci are not in linkage disequilibrium (LD) with previously described GWAS loci (Methods and Supplementary Table 9). Conditional analysis in the discovery cohorts showed that these new variants were independent from previously identified loci (Supplementary Table 10 and Supplementary Fig. 4). Of the five loci, three (rs149580368, rs77542162 and rs144984216) have an increased frequency in GoNL compared to 1000 Genomes (1-kG, Phase 1 integrated release v3, April 2012, all ancestries; Table 1), suggesting there may have been genetic drift in the Dutch population for these loci[4]. Yet, as each of these loci has a MAF > 0.005, we assumed these alleles also segregate in other populations of European descent[4], such as those of the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) consortium. Therefore, we set out replication in independent samples from the CHARGE cohorts using the 1-kG reference panel (Phase 1 integrated release v3, April 2012, all ancestries). We were able to replicate five out of the six variants using the Bonferroni corrected *p*-value threshold of $8.33 \cdot 10^{-3}$ (Table 2, Methods and Supplementary Table 11).

Of the replicated variants, rs77542162 is the most interesting variant. This missense variant is associated with both LDL-C and TC (Supplementary Figs 5 and 6) and located on chromosome 17 within the *ABCA6* gene (ATP-binding cassette, sub-family A (ABC1), member 6). The frequency of this variant is 1.31-fold higher in the discovery cohorts than in the replication cohorts and even 3.65-fold higher in the GoNL population than in the 1-kG population. This missense variant changes the amino acid cysteine into arginine at position 1359 (Cys1359Arg) and is predicted to be damaging for the structure and function of the protein by Polyphen2[19], MutationTaster[20] and LRT[21]. The effect size of rs77542162 ($\beta_{LDL\text{-}C}$=0.135 and β_{TC}=0.140) is very similar to those observed for other single variants in well-known lipid genes, such as *LDLR* and *CETP* as reported by Teslovich *et al.*[1]. The membrane-associated protein encoded by this gene is a member of the superfamily of ATP-binding cassette (ABC) transporters that transport various molecules across extra- and intracellular membranes. This protein is a member of the ABC1 subfamily, which is the only major ABC subfamily found exclusively in multicellular eukaryotes. *ABCA6* is clustered with four other ABC1 family members on chromosome 17q24 and appears to play a role in macrophage lipid homeostasis.

Table 1: summary descriptions for the variants associated with HDL-C, LDL-C, TC or TG.

SNP	Chr	Position	EA	NEA	Gene	MAF_{GoNL}	$MAF_{1\text{-}kG}$	MAF_{GoNL} / $MAF_{1\text{-}kG}$ (*p*-value for 2 population proportions)
rs4752801	11	47,907,641	G	A	close to the *NUP160*	0.347	0.338	1.027 (0.258)
rs149580368	17	41,874,745	A	C	between *C17orf105* and *MPP3*	0.029	0.015	1.923 (< 0.0001)
rs77542162	17	67,081,278	G	A	*ABCA6*	0.030	0.008	3.647 (< 0.0001)
rs144984216	19	20,479,901	T	C	*ZNF826P*	0.028	0.011	2.555 (< 0.0001)
rs117162033	19	8,627,569	T	C	*MYO1F*	0.007	0.007	0.957 (< 0.0001)

EA: effect allele. NEA: non-effect allele. MAF_{GoNL} and $MAF_{1\text{-}kG}$: the minor allele frequency of the effect allele in the GoNL reference panel and in the 1-kG reference panel (Phase 1 integrated release v3, April 2012, all ancestries), respectively.

Table 2: results for the variants associated with HDL-C, LDL-C, TC or TG.

Trait	SNP	Discovery phase						Replication phase					
		N	MAF	Rsq	β	SE_β	*p*-value	*N*	MAF	Rsq	β	SE_β	*p*-value
HDL-C	rs4752801	33,613	0.355	0.992	-0.023	0.003	1.62E-12	31,422	0.362	0.985	-0.012	0.003	5.63E-05
HDL-C	rs149580368	36,000	0.036	0.674	-0.075	0.010	4.23E-14	21,281	0.023	0.621	-0.079	0.014	5.90E-09
LDL-C	rs77542162	35,624	0.034	0.734	0.135	0.023	6.67E-09	21,969	0.026	0.773	0.125	0.031	4.35E-05
TC	rs77542162	36,109	0.034	0.731	0.140	0.025	1.29E-08	29,196	0.027	0.785	0.095	0.028	6.61E-04
TC	rs144984216	31,622	0.046	0.573	-0.140	0.024	7.88E-09	24,913	0.025	0.632	-0.056	0.036	1.22E-01
TG	rs117162033	26,122	0.016	0.511	-0.143	0.025	8.02E-09	10,296	0.021	0.573	-0.133	0.030	7.98E-06

Trait	SNP	Discovery phase						Combined discovery and replication			
		N	MAF	Rsq	β	SE_β	*p*-value	MAF	β	SE_β	*p*-value
HDL-C	rs4752801	33,613	0.355	0.992	-0.023	0.003	1.62E-12	0.359	-0.017	0.002	8.39E-15
HDL-C	rs149580368	36,000	0.036	0.674	-0.075	0.010	4.23E-14	0.031	-0.077	0.008	1.53E-21
LDL-C	rs77542162	35,624	0.034	0.734	0.135	0.023	6.67E-09	0.031	0.131	0.019	1.33E-12
TC	rs77542162	36,109	0.034	0.731	0.140	0.025	1.29E-08	0.031	0.120	0.019	7.31E-11
TC	rs144984216	31,622	0.046	0.573	-0.140	0.024	7.88E-09	0.039	-0.114	0.020	1.58E-08
TG	rs117162033	26,122	0.016	0.511	-0.143	0.025	8.02E-09	0.018	-0.139	0.019	3.10E-13

MAF: the weighted average of minor allele frequency for the effect allele across all studies in the discovery phase, replication phase or combined, respectively. *N*: sample size after QC. Rsq: mean imputation quality of all cohorts. β is the effect of the effect allele in mmol L^{-1}.

One other replicated variant, rs149580368, is also enriched with a 1.92-fold increase in frequency in the Dutch population compared to the 1-kG population. This intergenic variant (Supplementary Fig. 7) without a significant *cis*-eQTL effect, is located between the protein-coding genes *C17orf105* (chromosome 17 open reading frame 105) and *MPP3* (membrane protein, palmitoylated 3). Two replicated variants have similar frequencies in the GoNL and 1-kG reference sets: rs4752801 (Supplementary Fig. 8), an new intergenic variant with a high frequency (MAF = 0.355) that is located in a region previously identified[1] and rs117162033 (Supplementary Fig. 9), an intronic variant in the myosin F (*MYO1F*) coding gene. *C17orf15, MPP3* and *MYO1F* have no known impact on lipid levels. As the imputation quality of rs117162033 is lower than the other variants, we validated the imputation of this variant using the same approach as published by Scott *et al*[22]. We compared in a random sample of 65 participants of the GoNL reference panel their sequence and best-guess GoNL imputed genotypes and found that the concordance was 100% (all participants were correctly imputed). The association between TG and the intronic variant in the *MYO1F* gene is remarkable because of the low frequency of the variant. This confirms the conclusions as published before about the GoNL reference panel, that the trio-based phasing contributed significantly to the imputation quality of rare variants[5].

In this current study, the GoNL reference panel was used for imputations of the discovery cohorts and the 1-kG reference panel for the imputation of the replication cohorts. Though, it would be interesting to impute with a combined reference panel of both the GoNL data, the 1-kG data and other sequence data, this effort is ongoing.

This study shows that the imputation of a population-specific reference panel into large epidemiological cohorts can reveal both low-frequency and rare variants associated with blood lipid levels using classical association testing approaches. The three variants with increased frequency in the Dutch population as compared to the 1-kG population include a rare, predicted to be deleterious missense variant in *ABCA6,* which has increased frequency 3.65 times larger in the Dutch population. The effect of this variant is comparable to that of variants in the *LDLR* gene, a gene for which several population-based screening programs have been initiated. Our findings suggests that next generation sequencing effort may yield clinically relevant findings. Our paper further shows that next generation sequencing efforts in *specific homogeneous* populations as the Dutch may yield clinically relevant findings *worldwide*.

REFERENCES

1. Teslovich, T. M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707–713 (2010).
2. Global Lipids Genetics Consortium *et al.* Discovery and refinement of loci associated with lipid levels. *Nat Genet* **45**, 1274–1283 (2013).
3. Willer, C. J. & Mohlke, K. L. Finding genes and variants for lipid levels after genome-wide association analysis. *Curr Opin Lipidol* **23**, 98–103 (2012).
4. Pardo, L. M., MacKay, I., Oostra, B., van Duijn, C. M. & Aulchenko, Y. S. The effect of genetic drift in a young genetically isolated population. *Ann Hum Genet* **69**, 288–295 (2005).
5. Genome of the Netherlands Consortium. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat Genet* **46**, 818–825 (2014).
6. Boomsma, D. I. *et al.* The Genome of the Netherlands: design, and project goals. *Eur J Hum Genet* **22**, 221–227 (2014).
7. Jonsson, T. *et al.* A mutation in APP protects against Alzheimer's disease and age-related cognitive decline. *Nature* **488**, 96–99 (2012).
8. Holm, H. *et al.* A rare variant in MYH6 is associated with high risk of sick sinus syndrome. *Nat Genet* **43**, 316–320 (2011).
9. Styrkarsdottir, U. *et al.* Nonsense mutation in the LGR4 gene is associated with several human diseases and other traits. *Nature* **497**, 517–520 (2013).
10. Warnick, G. R., Knopp, R. H., Fitzpatrick, V. & Branson, L. Estimating low-density lipoprotein cholesterol by the Friedewald equation is adequate for classifying patients on the basis of nationally recommended cutpoints. *Clin Chem* **36**, 15–19 (1990).
11. Aulchenko, Y. S., Ripke, S., Isaacs, A. & van Duijn, C. M. GenABEL: an R library for genome-wide association analysis. *Bioinformatics* **23**, 1294–1296 (2007).
12. Svishcheva, G. R., Axenovich, T. I., Belonogova, N. M., van Duijn, C. M. & Aulchenko, Y. S. Rapid variance components-based method for whole-genome association analysis. *Nat Genet* **44**, 1166–1170 (2012).
13. Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* **5**, e1000529 (2009).
14. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & calo R Abecasis, G. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat Genet* **44**, 955–959 (2012).
15. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
16. Browning, S. R. & Browning, B. L. Haplotype phasing: existing methods and new developments. *Nat Rev Genet* **12**, 703–714 (2011).
17. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* **88**, 76–82 (2011).
18. Deelen, P. *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).
19. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nat Methods* **7**, 248–249 (2010).
20. Schwarz, J. M., Rödelsperger, C., Schuelke, M. & Seelow, D. MutationTaster evaluates disease-causing potential of sequence alterations. *Nat Methods* **7**, 575–576 (2010).
21. Chun, S. & Fay, J. C. Identification of deleterious mutations within three human genomes. *Genome Res* **19**, 1553–1561 (2009).
22. Scott, L. J. *et al.* A genome-wide association study of type 2 diabetes in finns detects multiple susceptibility variants. *Science* **316**, 1341–1345 (2007).

PART 4

NEW APPROACHES TO REVEAL VARIANTS ASSOCIATED WITH HDL-C

CHAPTER 4.1

The challenges of genome-wide interaction studies: lessons to learn from the analysis of HDL-C blood levels

Elisabeth M. van Leeuwen, Françoise A. S. Smouter, Tony Kam-Thong, Nazanin Karbalai, Albert V. Smith, Tamara B. Harris, Lenore J. Launer, Colleen M. Sitlani, Guo Li, Jennifer A. Brody, Joshua C. Bis, Charles C. White, Alok Jaiswal, Ben A. Oostra, Albert Hofman, Fernando Rivadeneira, Andre G. Uitterlinden, Eric Boerwinkle, Christie M. Ballantyne, Vilmundur Gudnason, Bruce M. Psaty, L. Adrienne Cupples, Marjo-Riitta Järvelin, Samuli Ripatti, Aaron Isaacs, Bertram Müller-Myhsok, Lennart C. Karssen, Cornelia M. van Duijn.

Published in PLoS One (PLoS One, 9(10):e109290, 2014).
The supplemental information for this chapter is available online at:
http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0109290#s5

ABSTRACT

Genome-wide association studies (GWAS) have revealed 74 single nucleotide polymorphisms (SNPs) associated with high-density lipoprotein cholesterol (HDL-C) blood levels. This study is, to our knowledge, the first genome-wide interaction study (GWIS) to identify SNP×SNP interactions associated with HDL-C levels. We performed a GWIS in the Rotterdam Study (RS) cohort I (RS-I) using the GLIDE tool which leverages the massively parallel computing power of Graphics Processing Units (GPUs) to perform linear regression on all genome-wide pairs of SNPs. By performing a meta-analysis together with Rotterdam Study cohorts II and III (RS-II and RS-III), we were able to filter 181 interaction terms with a *p*-value $< 1 \cdot 10^{-8}$ that replicated in the two independent cohorts. We were not able to replicate any of these interaction term in the AGES, ARIC, CHS, ERF, FHS and NFBC-66 cohorts (N_{total} = 30,011) when adjusting for multiple testing. Our GWIS resulted in the consistent finding of a possible interaction between rs774801 in *ARMC8* and rs12442098 in *SPATA8* being associated with HDL-C levels. However, *p*-values do not reach the preset Bonferroni correction of the *p*-values. Our study suggest that even for highly genetically determined traits such as HDL-C the sample sizes needed to detect SNP×SNP interactions are large and the 2-step filtering approaches do not yield a solution. Here we present our analysis plan and our reservations concerning GWIS.

INTRODUCTION

To date, genome-wide association studies (GWAS) have revealed 95 genetic loci associated with lipid levels in human plasma. Of these, 74 SNPs were associated with high-density lipoprotein cholesterol (HDL-C) levels[1-5]. Together, these 47 SNPs explain approximately 25% of the heritability of HDL-C levels. Although the largest meta-analysis of plasma lipid concentrations[4] to date, already included more than 100,000 individuals of European descent, it is expected that with increasing sample size and larger, better reference panels for imputation, more variants will be found to be associated with HDL-C levels, probably resulting in an increase of the explained heritability. Nevertheless, single SNP effects may not fully explain the heritability of HDL-C levels. Genetic processes like DNA methylation, histone modification and interactions between SNPs are also potential candidates determining HDL-C levels[6-9]. A previous large study did not find evidence of gene-environment interactions influencing HDL-C levels, although this might also play a role with other environmental factors[10]. We defined interactions between SNPs as a departure from a linear statistical model allowing for the additive marginal effects of both SNPs. Persistent evidence for interacting loci involved in lipid metabolism comes from experimental animal research in which various loci interact with each other[11].

Based on the loci for HDL-C levels identified to date, finding evidence for SNP×SNP interactions in humans has proven to be difficult. Ma *et al.*[8] identified a significant association interaction between a locus within the *HMGCR* gene and a locus near the *LIPC* gene in relation to HDL-C cholesterol. Furthermore, Turner *et al.*[9] found 8 SNP×SNP interactions to be associated with HDL-C levels of which the strongest model included an interaction between *LPL* and *ABCA1*. These studies suggest that SNP×SNP interactions can indeed also explain some of the heritability of HDL-C levels in humans. However, only loci were studied that had previously been successfully replicated in GWAS of lipid levels, thus motivating a genome-wide search for interactions associated with HDL-C levels.

Genome-wide searches for associations between phenotypes and SNP×SNP interactions have been hampered by the computation time needed for testing all unique pairs of SNPs, given by $N_{SNPs}(N_{SNPs}-1)/2$, with N_{SNPs} the total number of SNPs. Consequently, the time for testing all interaction terms is proportional to N_{SNPs}^2, translating into months of computation time. Modern Graphics Processing Units (GPUs) are optimised for highly parallel computation tasks and are well-suited to replace regular processors (Central Processing Units or CPUs) for these kind of tasks. The GLIDE software package[12] makes use of GPUs to perform linear regression for all pairs of SNPs. In this study, we aim to identify SNP×SNP interactions for HDL-C levels in the Rotterdam Study cohort I (RS-I) using GLIDE. The most significant interactions terms in RS-I are first filtered by a meta-analysis in cohorts II and III of the Rotterdam Study (RS-II and RS-III, respectively). The resulting interactions were subsequently sent for replication in

the CHARGE cohorts (AGES, ARIC, CHS, ERF, FHS) and the NFBC-66 cohort. We also tested whether the identified interaction terms are associated to dyslipidemia treatment within the cohorts of the Rotterdam Study.

METHODS

Study descriptions

Ethics Statement

The AGES Reykjavik Study Genome Wide Association study was approved by the National Bioethics Committee (00-063) and the Data Protection Authority. The ARIC study was approved by 'The University of Texas Health Science Center at Houston Committee for the Protection of Human Subjects'. The CHS study was approved by the following institutional review boards: Wake Forest University, University of California (Davis), Johns Hopkins University (Bloomberg School of Public Health), University of Pittsburgh, University of Washington, University of Vermont. The ERF study was approved by the Medical Ethics Committee of the Erasmus MC. The committee is constituted according to the WMO (National act medical-scientific research in human beings). The FHS was approved by the Boston University Medical Campus Institutional Review Board. The NFBC66 was approved by the Ethical Committee of the Northern Ostrobothnia Hospital District. 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 study participants for all cohorts.

Discovery cohort

Rotterdam Study cohort I (RS-I). The Rotterdam Study is an ongoing prospective population-based cohort study, focused on chronic disabling conditions of the elderly. The study comprises an outbred ethnically homogenous population of Dutch Caucasian origin. The rationale of the study has been described in detail elsewhere[13]. In summary, 7,983 men and women aged 55 years or older, living in Ommoord, a suburb of Rotterdam, the Netherlands, were invited to participate in the first phase. Fasting blood samples were taken during the participant's third visit to the research center.

Filtering cohorts

Rotterdam Study cohort II (RS-II). The Rotterdam Study cohort II prospective population-based cohort study comprises 3,011 residents aged 55 years and older from the same district of Rotterdam. The rationale and study designs of this cohort is similar to that of the RS-I[13]. The baseline measurements, including the fasting HDL-C measurements, took place during the first visit.

Rotterdam Study cohort III (RS-III). The Rotterdam Study cohort III prospective population-based cohort study comprised 3,932 residents aged 45 years and older from the same district of Rotterdam. The rationale and study designs of this cohort is similar to that of the RS-I[13]. The baseline measurements, including the fasting HDL-C measurements, took place during the first visit.

Replication cohorts

Age, Gene/Environment Susceptibility (AGES Reykjavik) Study. The Age, Gene/Environment Susceptibility (AGES Reykjavik) Study was initiated to examine genetic susceptibility and gene/environment interaction as these contribute to phenotypes common in old age, and represents a continuation of the Reykjavik Study cohort begun in 1967. The study is approved by the Icelandic National Bioethics Committee, (VSN: 00-063) and the Data Protection Authority. The researchers are indebted to the participants for their willingness to participate in the study.

Atherosclerosis Risk in Communities (ARIC) Study. The Atherosclerosis Risk in Communities Study (ARIC), sponsored by the National Heart, Lung, and Blood Institute (NHLBI) is a prospective epidemiologic study conducted in four U.S. communities. ARIC is designed to investigate the causes of atherosclerosis and its clinical outcomes, and variation in cardiovascular risk factors, medical care, and disease by race, gender, location, and date. To date, the ARIC project has published over 800 articles in peer-reviewed journals. ARIC includes two parts: the Cohort Component and Community Surveillance Component.

The ARIC Cohort Component began in 1987, and each ARIC field center randomly selected and recruited a cohort sample of approximately 4,000 individuals aged 45-64 from a defined population in their community, to receive extensive examinations, including medical, social, and demographic data. Follow-up also occurs semi-annually, by telephone, to maintain contact and to assess health status of the cohort.

In the Community Surveillance Component, the four communities are investigated to determine the long term trends in hospitalized myocardial infarction (MI) and coronary heart disease (CHD) deaths in approximately 470,000 men and women aged 35-84 years.

Cardiovascular Health Study (CHS). The CHS[14] is an NHLBI-funded observational study of risk factors for cardiovascular disease in adults 65 years or older. Starting in 1989, and continuing through 1999, participants underwent annual extensive clinical examinations. Measurements included traditional risk factors such as blood pressure and lipids as well as measures of subclinical disease, including echocardiography of the heart, carotid ultrasound, and cranial magnetic-resonance imaging (MRI). At six month intervals between clinic visits, and once clinic visits ended, participants were contacted by phone to ascertain hospitalizations and health status. The main outcomes are coronary heart disease (CHD), angina, heart failure (HF), stroke, transient ischemic attack (TIA), claudication, and mortality. Participants continue

to be followed for these events. CHS participants who were free of cardiovascular disease at the start of the study, and who consented to genetic testing, were included in these analyses.
Erasmus Rucphen Family (ERF) Study. The ERF study has been described in detail previously[15]. A total of approximately 3,000 participants descend from 22 couples who lived in the Rucphen region in The Netherlands in the 19th century. The 2,755 individuals with genotype data and lipid measurements were included in the current analysis.
Framingham Heart Study (FHS). The Framingham Heart Study (FHS), funded by the National Heart Lung and Blood Institute, is an observational population-based cohort study composed of three generations of Framingham (MA) residents predominately of European descent. The Original cohort (*N* = 5,209) was enrolled in 1948. The children and spouses of the Original cohort comprise the Offspring cohort (*N* = 5,124), which was enrolled in 1971-1975[16]. The Third Generation (*N* = 4,095) consists mostly of the children of the Offspring cohort, and was enrolled in 2002 to 2005[17]. All participants were examined every 4-8 years. DNA for surviving participants was collected in the late 1990s and early 2000s (1995-2005). Cholesterol and genetic data from 3,464 Offspring subjects and 3,569 Third Generation subjects contribute to this paper.
Northern Finland Brith Cohort 1966 (NFBC-66). The Northern Finland Birth Cohort 1966 (NFBC-66) study[18] is a longitudinal one-year birth cohort study designed to study the risk factors of perinatal deaths and low birth weight. Mothers living in the two northern-most provinces of Finland were invited to participate if they had expected delivery dates during 1966. Individuals still living in Helsinki area or Northern Finland were asked at age 31 to participate in a detailed examination (*N* = 5,923). Extensive data on intermediate phenotypes related to obesity and behavioral traits have also been collected.

Genotyping and imputation

All cohorts were genotyped using commercially available Affymetrix or Illumina genotyping arrays, or custom Perlegen arrays. Quality control was performed independently for each study. To facilitate meta-analysis, each replication cohort performed genotype imputation using BIMBAM, IMPUTE, or MaCH with reference to HapMap or the 1000 Genomes project data.
The first two cohorts of the Rotterdam Study were genotyped using the Illumina 550K chip, the third cohort was genotyped using the Illumina 610K and 660K chip. The following exclusions were applied to identify a final set of SNPs that was used in this study: MAF < 0.05, SNP callrate < 0.95 and/or HWE *p*-value < $1 \cdot 10^{-7}$. The QC was done per cohort.
In ARIC, genotyping was performed with the Affymetrix 6.0 chip. After genotyping, the following quality control tresholds were applied: (1) comparison of genotype calls to sample replicates, with exclusion of samples with greater than 1% mismatch, (2) exclusion of samples with greater than 5% missing genotypes, (3) exclusion of samples with a mismatch between

reported sex and that determined by genotyping, (4) exclusion of SNPs with greater than 10 % missing genotypes across samples, (5) exclusion of SNPs monomorphic in both races and (6) exclusion of SNPs (MAF > 0.05) with HWE *p*-values of less than $1 \cdot 10^{-6}$. Prior to imputations, principal component analysis was performed to exclude outliers. Imputation to HapMap release 23a was performed using MaCH v.1.0. After imputation SNPs with an imputation quality less than 0.90 were excluded. 26.8% of the SNPs in the replication were genotyped, the rest was imputed.

In AGES only imputed SNPs were used for the replication. The genotypes originated on Illumina Hu370CNV. For imputation, only the SNPs were included which were completed in 97% of individuals and had a MAF above 1%. Imputation was performed by MaCH against HapMap Release 22. Quality of the imputations was evaluated by the MaCH R^2 metric.

In CHS, genotyping was performed at the General Clinical Research Center's Phenotyping/ Genotyping Laboratory at Cedars-Sinai using the Illumina 370CNV BeadChip system. Genotypes were called using the Illumina BeadStudio software. The following exclusions were applied to identify a final set of 306655 autosomal SNPs that were used for imputation: call rate < 97%, HWE $p < 1 \cdot 10^{-5}$, > 2 duplicate errors or Mendelian inconsistencies (for reference CEPH trios), heterozygote frequency = 0 and SNP not found in HapMap. Imputation to HapMap release 22 (build 36) was performed using BimBam v.0.99. Most of the replication SNPs were genotyped (58.4%), the remaining were imputed.

In ERF genotyping was done on various Illumina and Affymetrix chips. QC was done for each chip separately. On average, the following QC criteria were applied: callrate > 0.98, per individual callrate > 0.96, HWE *p*-value > $5 \cdot 10^{-8}$ and MAF > 0.005. IBS checks, sex chromosome checks and ethnicity checks were also performed. The imputation to Hapmap 2 release 22 was performed with MaCH and minimac. All SNPs in the replication were imputed.

In FHS genotyping was done on Affymetrix 250K Nsp and 250K Sty mapping arrays and the Affymetrix 50K supplemental gene-focused array. The following QC criteria were applied before imputations: $p_{HWE} < 1 \cdot 10^{-6}$, callrate > 0.97, mishap test of non-random missingness $p < 1 \cdot 10^{-9}$, < 100 Mendelian errors. The genotyped SNPs were imputed against HapMap (release 22, build 36, CEU population) with MaCH (version 1.0.15). All SNPs in the replication were imputed.

In NFBC-66 genotyping was done on Illumina 370K whole-genome SNP array. The following QC criteria were observed: SNP clustering probability of genotypes > 95%, sample call rate > 95%, SNP call rate > 95%, MAF > 1% and HWE *p*-value > $1 \cdot 10^{-6}$. Heterozygosity, gender check and relatedness checks were performed and any discrepancies were removed. 10 individuals with cryptic relatedness were also excluded from the analysis. To identify a final set of SNPs for imputations, a SNP call rate filter of > 99% was applied to all SNPs with MAF < 5%. The imputation to 1000 Genomes Phase I integrated variant set (Mar 2012) was performed using IMPUTE v2.2.2. After imputation only those variants with info score > 0.9 were analysed. 58.6% of the SNPs in the replication were genotyped, the rest was imputed.

Study samples and phenotypes

A summary of the details of the nine studies participating in this analysis can be found in Table 1. In all studies, the subjects were fasting when the HDL-C levels were measured. The HDL-C measurements were adjusted for sex and age, except for NFBC-66 in which only was adjusted for sex since all individuals are from the same age. In ERF mmscore (GenABEL version 1.7.0[19] was used to account for family relationships. In ARIC, the HDL-C levels were also adjusted for the three ARIC field center with two 0,1 indicator variables. In CHS the HDL-C was adjusted for study clinic site as well and in NFBC-66 HDL-C was also adjusted for 10 PC components. In FHS the HDL-C levels were also adjusted for related individuals with the lmekin function within the coxme package in R (http://cran.r-project.org/web/packages/coxme/) and adjusted for PCs. In the discovery and filtering stage, the HDL-C levels after adjustment for sex and age were normalised around zero as this is a requirement of GLIDE. To compare the β_{int} in the discovery and filtering stage with the β_{int} in the replication stage, we also calculated the β_{int} in the Rotterdam Study cohorts without scaling around zero for the most promising interaction terms.

GWIS with GLIDE in RS-I

To systematically search for the epistatic interactions associated with HDL-C levels in RS-I we used GLIDE[12]. GLIDE makes use of the computational power of consumer-grade graphics cards to detect interactions between SNPs via linear regression. To reduce computation time, we chose to run GLIDE on genotyped SNPs only. In order to run GLIDE, the genotype data of RS-I was stored per chromosome as a text file with one row per SNP and one column per individual. Individuals using lipid-lowering medication were excluded. The file does not contain column headers or row names and the SNPs need to be coded 0 (homozygous for the major allele), 1 (heterozygous) or 2 (homozygous for the minor allele). We only used SNPs with a MAF (Minor Allele Frequency) > 0.05 within the samples of RS-I, RS-II and RS-III which were used in this study, since the sample size is not large enough to investigate low-frequency variants.

The names of the SNPs are stored in a separate one-column text file in the same order as the SNPs in the file with the genotype data. The values of the scaled residuals are stored in a separate text file in the same order as the individuals in the file with the genotype data. GLIDE requires the phenotype to be normalised around zero. GLIDE uses the files with the genotypes and the file with the scaled residuals to perform linear regression for all possible unique SNP×SNP combinations. In order to fit the data into the GPU's memory, GLIDE splits up the genotypes in subsets of SNPs. In this study we chose to split up in subsets of 1000 SNPs. GLIDE outputs a *t*-score for each interaction term and a threshold can be set to only output interactions with a *t*-score above this threshold.

The output of GLIDE does not contain the SNP names, but the number of the chunk and the number of the SNP within a given chunk. With help of the previously created SNP files, we assigned SNP names to the interaction terms output by GLIDE. Since GLIDE handles the data in chunks, interaction terms occur multiple times in the output of GLIDE, consequently, the results had to be filtered on unique interaction terms.

Filtering of interaction terms by meta-analysis of RS-I, RS-II and RS-III.

To reduce the number of false positive interaction terms, we filtered the interaction terms with an absolute value of the *t*-score > 5 (*p*-value < $6.06 \cdot 10^{-7}$) by a meta-analysis of RS-I, RS-II and RS-III. For these interactions, we used linear regression to determine the βs, standard errors and *p*-values in RS-I, RS-II and RS-III. The HDL-C levels after adjustment for sex and age were normalised around zero in all three cohorts. The βs and standard errors of all three cohorts of the Rotterdam Study were subsequently meta-analyzed to filter out only those with a *p*-value less than $1 \cdot 10^{-8}$.

Replication of SNP×SNP interactions

The interaction terms which had a *p*-value less than $1 \cdot 10^{-8}$ after meta-analysis of the three Rotterdam Study cohorts, were replicated in 6 cohorts: AGES, ARIC, CHS, ERF, FHS and NFBC-66. Only individuals that do not use lipid-lowering medication were included, except for AGES. The linear regression model for replication was $HDL_{adj} = \alpha + \beta_1 (SNP1) + \beta_2 (SNP2) + \beta_{int} (SNP1 \times SNP2) + \varepsilon$, where HDL_{adj} are the HDL-C levels adjusted for sex and age. We meta-analysed the β_{int} from all 6 replication cohorts.

To see if the filtered interaction terms effect the probability of using lipid-lowering medication, we performed a case-control study in the three Rotterdam Study cohorts. Those individuals that have HDL-C levels available and use lipid-lowering medication were defined as cases and the individuals in the discovery or filtering stage were defined as controls. The logistic regression model for replication was $Medication_{yes/no} = \alpha + \beta_1 (SNP1) + \beta_2 (SNP2) + \beta_{int} (SNP1 \times SNP2) + \varepsilon$. We performed the analysis in the three cohorts separately, and also in the three cohorts combined, in which we included the cohort number as an additional covariate.

Power calculations

To estimate the effect we could have detected with the current sample size, a certain type I error and various type II errors, we used G*Power, (version 3.1.9.2).

RESULTS

GWIS with GLIDE in RS-I

Figure 1 shows a flow diagram illustrating the analysis plan. A total of 495,508 genotyped SNPs that passed quality control, had a Minor Allele Frequency (MAF) > 0.05 in the sample of 2,996 individuals from RS-I, and were also genotyped in RS-II and RS-III were used to identify SNP×SNP interactions associated with HDL-C using GLIDE. For this analysis the HDL-C levels after adjustment for sex and age were normalized around zero as this is a requirement of GLIDE. This resulted in 84,031 SNP×SNP interactions with an absolute value of the *t*-score > 5 (i.e. $p < 6.06 \cdot 10^{-7}$).

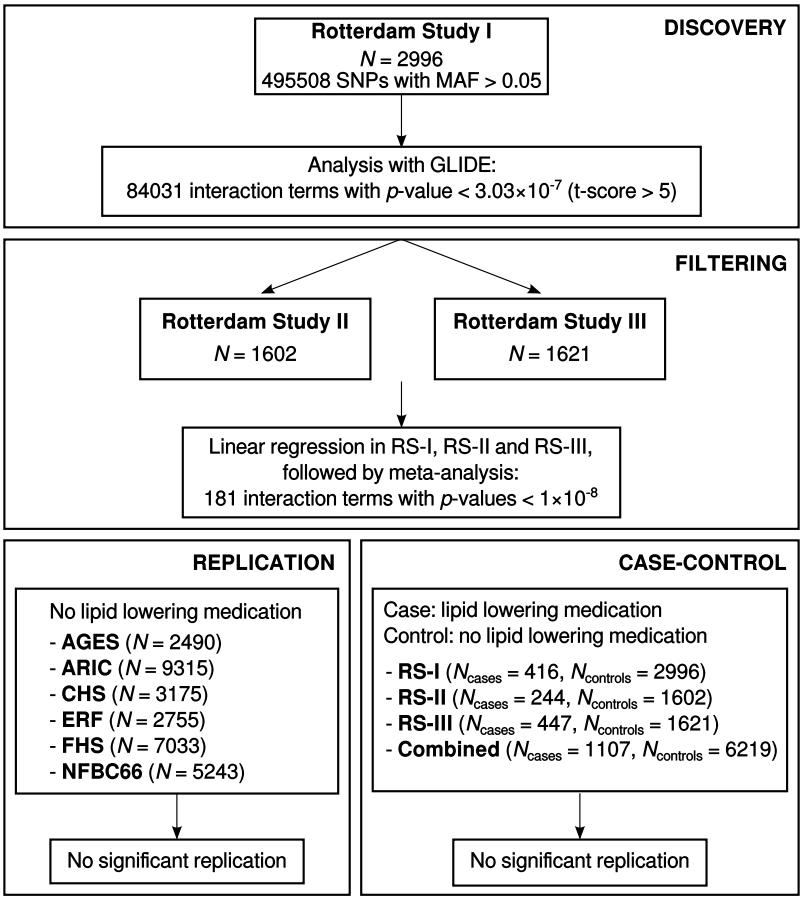


Figure 1: Flow diagram overview of the analysis plan.

Filtering of interaction terms by a meta-analysis of RS-I, RS-II and RS-III

Using linear regression we calculated the regression coefficient β_{int} for the interaction term, the standard errors and the *p*-values for the 84,031 interaction terms in RS-I (*N* = 2,996), RS-II (*N* = 1,602) and in RS-III (*N* = 1,621). For these analyses the HDL-C levels after adjustment for sex and age were normalized around zero since this was done in RS-I in the initial analysis with GLIDE as this is a requirement of GLIDE. The calculated β_{int} and standard errors were used to meta-analyse the association between each of the 84,031 interaction terms and HDL-C levels. After meta-analysis, 181 interaction terms with a *p*-value below $1 \cdot 10^{-8}$ remain, of which 5 interaction terms with a *p*-value less than $1 \cdot 10^{-10}$. The pooled β_{int} for the 84,031 interaction terms range from -0.507 to 0.746. The 181 interaction terms with a *p*-value less than $1 \cdot 10^{-8}$ were taken forward for replication, see Supplementary Table 1. The number of unique interaction terms for replication was reduced to 132 by filtering on linkage disequilibrium (LD) between interaction terms ($R^2 > 0.8$). Consequently, the *p*-value for replication after Bonferroni correction is $3.79 \cdot 10^{-4}$. We also calculated the β_{int} of RS-I, RS-II and RS-III for these 181 interaction terms using linear regression with the unscaled phenotype to compare these with the β_{int} within the replication cohorts.

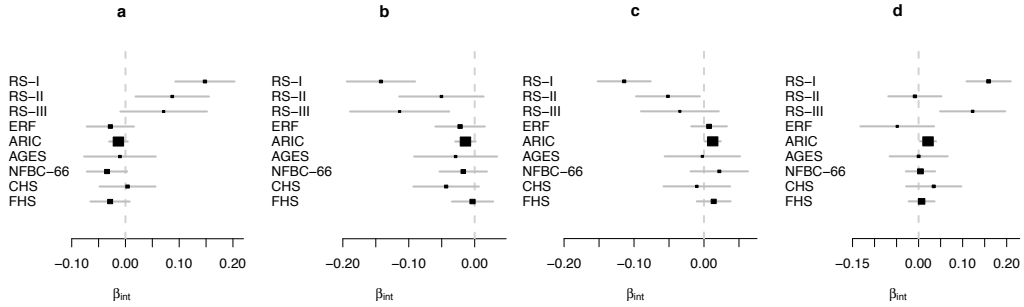


Figure 2: The forest plots for β_{int} of the four most significant interaction terms after meta-analysis of the replication cohorts: rs2315598-rs2853228 (a), rs6848132-rs7863451 (b), rs3756856-rs11758333 (c) and rs4596126-rs11676467 (d). Although the analysis in the discovery and the filtering was done with scaled phenotypes, for these forest plots, the HDL-C levels are not scaled in the Rotterdam Study cohorts.

Replication of SNP×SNP interactions

Replication was conducted in 6 cohorts: AGES, ARIC, CHS, ERF, FHS and NFBC-66. In the replication cohorts only individuals not on lipid-lowering medication were included, with the exception of AGES, see Table 1. In AGES, ARIC, CHS, ERF and FHS, 8, 7, 7, 10 and 7 interaction terms, respectively, could not be tested for replication since one or both of the SNPs in the interaction term had not been genotyped or imputed. In NFBC-66 all interaction terms could be tested for replication. A total of 170 out of the 181 interactions could be tested for replication in all six cohorts. None of the interaction terms reached a significant *p*-value after

Bonferroni correction ($3.79 \cdot 10^{-4}$) in any of the replication cohorts and after meta-analysis of all replication cohorts. Four interaction terms reached nominal significance at $p = 0.05$, see Figure 2. The lowest *p*-value for β_{int} after meta-analysis of all replication cohorts ($N = 30{,}011$) was $7.57 \cdot 10^{-3}$ for the interaction between rs2315598 (chromosome 2, position 132,994,224, gene *GPR39*) and rs2853228 (chromosome 8, position 103,296,258, gene *RRM2B*). The second lowest *p*-value for β_{int} after meta-analysis of all replication cohorts ($N = 30{,}011$) was $8.1 \cdot 10^{-3}$ for the interaction between rs6848132 (chromosome 4, position 93,460,610, gene *GRID2*) and rs7863451 (chromosome 9, position 129,112,065, gene *GARNL3*). The β_{int} is negative in all nine cohorts. Table 2 shows the 20 interaction terms with the lowest *p*-values. Five of these terms are interactions between an intergenic locus at chromosome 6, situated between the *TCP11* and *SCUBE3* genes, and a locus at the same chromosome in the *SOBP* gene which are in LD with each other ($R^2 > 0.872$).

Table 1: Baseline characteristics for discovery and replication cohorts

	Study	Country of origin	N (% male)
RS-I	Rotterdam Study cohort I	Netherlands	2996 (57.7)
RS-II	Rotterdam Study cohort II	Netherlands	1602 (54.9)
RS-III	Rotterdam Study cohort III	Netherlands	1621 (58.3)
AGES	Age, Gene/Environment Susceptibility Study	Iceland	3219 (42.0)
ARIC	Atherosclerosis Risk in Communities Study	United States	9315 (46.9)
CHS	Cardiovascular Health Study	Americans of European descent	3175 (40)
ERF	Erasmus Rucphen Family study	Netherlands	2755 (44.7)
FHS	Framingham Heart Study	Americans of European descent	7033 (46)
NFBC-66	Northern Finland Brith Cohort 1966	Finland	5243 (47.98)

	Mean age (SD), years	HDL-C (SD), mmol/L	lipid lowering medication users
RS-I	66.2 (7.2)	1.39 (0.39)	excluded
RS-II	64.7 (8.1)	1.38 (0.37)	excluded
RS-III	55.6 (5.7)	1.47 (0.44)	excluded
AGES	76.4 (5.5)	1.58 (0.45)	included (22.6%)
ARIC	54.3 (5.7)	1.31 (0.43)	excluded
CHS	72.5 (5.4)	1.43 (0.41)	excluded
ERF	48.9 (14.4)	1.27 (0.36)	excluded
FHS	37.5 (9.6)	1.37 (0.40)	excluded
NFBC-66	31 (0)	1.56 (0.38)	excluded

Table 2: The top 20 interaction terms after replication. The interactions are sorted based on the *p*-value of the interaction term β_{int}. The HDL-C levels are adjusted for sex and age. The phenotypes of the Rotterdam Study are not scaled around zero.

Interaction	Direction of β_{int} **									Meta-analysis of RS-I, RS-II and RS-III			Meta-analysis of replication cohorts		
										β_{int}	SE	*p*-value	β_{int}	SE	*p*-value
rs2315598- rs2853228	+	+	+	-	-	-	-	+	-	0,1123	0,0193	5,91E-9	-0,0178	0,0067	0,0076
rs6848132- rs7863451	-	-	-	-	-	-	-	-	-	-0,1077	0,0181	2,76E-9	-0,0158	0,0060	0,0081
rs3756856- rs11758333*	-	-	-	+	+	-	+	-	+	-0,0769	0,0132	5,33E-9	0,0112	0,0047	0,0160
rs4596126- rs11676467	+	-	+	-	+	+	+	+	+	0,0973	0,0175	2,47E-8	0,0137	0,0068	0,0436
rs871849- rs9672547	-	-	-	-	-	-	-	-	+	-0,0820	0,0141	5,86E-9	-0,0092	0,0048	0,0557
rs5018943- rs7023148	-	-	-	+	-	-	-	-	-	-0,0944	0,0159	2,96E-9	-0,0132	0,0069	0,0558
rs2280518- rs9875407	+	+	+	+	+	+	+	+	+	0,1393	0,0242	9,22E-9	0,0161	0,0085	0,0574
rs754950- rs10926977	+	+	+	-	-	-	+	-	-	0,0876	0,0148	2,92E-9	-0,0103	0,0055	0,0608
rs2301818- rs7537266	-	-	-	-	+	-	+	+	+	-0,0905	0,0156	6,43E-9	0,0098	0,0053	0,0626
rs10816852- rs10911901	+	+	+	+	+	+	+	-	-	0,1491	0,0257	6,45E-9	0,0135	0,0075	0,0734
rs645925- rs12231356	+	+	+	-	-	-	-	+	+	0,1102	0,0184	2,04E-9	-0,0112	0,0063	0,0749
rs820065- rs7762721*	-	-	-	-	-	-	+	-	+	-0,0768	0,0132	5,36E-9	-0,0082	0,0046	0,0782
rs7214582- rs17075071	+	+	+	-	-	+	-	-	-	0,2005	0,0349	9,40E-9	-0,0190	0,0109	0,0817
rs774801- rs12442098	-	-	-	-	-	+	-	+	-	-0,0960	0,0167	8,57E-9	-0,0101	0,0059	0,0866
rs693- rs4677039	+	+	+	+	+	+	+	+	-	0,0570	0,0102	2,07E-8	0,0057	0,0034	0,0928
rs3756856- rs7762721*	-	-	-	-	+	+	-	-	+	-0,0836	0,0131	1,82E-10	0,0078	0,0046	0,0928
rs2242312- rs11190870	+	+	+	+	+	-	+	-	-	0,1238	0,0205	1,65E-9	0,0112	0,0067	0,0951
rs3756856- rs6940398*	-	-	-	-	+	-	-	+	+	-0,0793	0,0131	1,61E-9	0,0078	0,0047	0,0982
rs2316640- rs10973877	+	+	+	+	-	+	-	-	+	0,0729	0,0122	2,08E-9	-0,0064	0,0039	0,0990
rs2234044- rs7762721*	-	-	-	+	+	-	-	-	+	-0,0807	0,0132	1,10E-9	0,0075	0,0046	0,1044

* Interaction terms in LD with each other ($R^2 > 0.872$).
** order of the directions: AGES, ARIC, CHS, ERF, FHS, NFBC-66, RS-I, RS-II, RS-III. The meta-analysis was done with fixed effects.

Individuals with high levels of low-density lipoprotein (LDL) or low levels of HDL-C are treated with lipid-lowering medication. The 181 selected interaction terms were also tested to see whether their presence might explain the use of lipid-lowering medication and therefore the extreme lipid levels. To this end the individuals of the Rotterdam Study in the discovery and filtering stage were used as controls, and the individuals of the Rotterdam Study who use lipid-lowering medication were used as cases. Table 3 shows the 20 interaction terms with the lowest *p*-values for β_{int} after testing in the three cohorts of the Rotterdam Study combined. The interaction between rs6442460 (chromosome 3, position 14,551,071, gene *GRIP2*) and rs10914332 (chromosome 1, position 31,471,589, gene *NKAIN1*) had the lowest *p*-value ($p = 3.98 \cdot 10^{-3}$).

Three interaction terms overlap between the top 20 hits after the replication and the top 20 hits after the case-control test, as shown in Table 4. None of the SNPs of these interaction terms are in high LD with each other ($R^2 > 0.8$). The interaction between rs754950 and rs10926977 has an opposite effect direction after the meta-analysis in the Rotterdam Study cohorts compared to the one after meta-analysis in the replication cohorts and thus will probably be a false-positive finding. The second interaction term (between rs2242312 and rs11190870) had a positive effect on HDL, but increases the risk of lipid lowering medication which is counter-intuitive and consequently this interaction term is likely a false-positive finding as well. The third interaction term, however, between rs774801 (chromosome 3, position 139,413,035, gene *ARMC8*) and rs12442098 (chromosome 15, position 95,385,874, close to gene *SPATA8*) has a negative effect on HDL-C combined with a positive effect on the use of lipid lowering medication. Although this last interaction term is not replicated, the directions of the effects are consistent since this interaction lowers the HDL-C level and increases the chance of using lipid lowering medication.

Table 3: The top 20 interaction terms after case-control studies in RS-I, RS-II and RS-III separate and combined. The interactions are sorted based on the *p*-value of the β_{int} after the case-control study in the combined data set of RS-I, RS-II and RS-III. The HDL-C levels are adjusted for sex and age, the residuals are not scaled around zero. Number of cases: 416 (RS-I), 244 (RS-II) and 447 (RS-III). Number of controls: 2996 (RS-I), 1602 (RS-II) and 1621 (RS-III). *1-*4 mark the interaction terms that are in high LD with each other: $R^2 > 0.913$.

Interaction	Case-control in RS-I			Case-control in RS-II		
	β_{int}	SE	*p*-value	β_{int}	SE	*p*-value
rs6442460- rs10914332	-0,5227	0,286	0,0673	-0,4306	0,396	0,2775
rs2146043- rs11124513	-0,25056	0,172	0,1452	-0,2557	0,249	0,3039
rs774801- rs12442098 *1	0,35367	0,18	0,0497	0,4151	0,245	0,0898
rs3006496- rs17729021	0,25625	0,135	0,0579	0,198	0,186	0,2872
rs754970- rs2306478 *2	-0,17851	0,153	0,2422	-0,17	0,181	0,3469
rs754970- rs3775972 *2	-0,17089	0,152	0,2614	-0,1678	0,18	0,3501
rs5770418- rs17234336	-0,94012	0,422	0,0258	-0,6081	0,53	0,2513
rs7631734- rs12442098 *1	0,31044	0,181	0,0855	0,4119	0,247	0,0959
rs2242312- rs11190870	0,24296	0,226	0,2833	0,4927	0,323	0,1275
rs4861849- rs17123865	-0,04394	0,415	0,9156	0,8862	0,499	0,0757
rs754950- rs4658547 *3	-0,15255	0,162	0,3451	-0,2093	0,232	0,367
rs1203791- rs10496556	-0,29897	0,151	0,0479	-0,0571	0,189	0,7629
rs754950- rs10926977 *3	-0,15562	0,159	0,3275	-0,1923	0,229	0,4002
rs900654- rs10195135 *4	0,07114	0,161	0,6594	0,3194	0,209	0,1267
rs10195135- rs10872670 *4	0,1016	0,16	0,5246	0,2728	0,21	0,1947
rs2919732- rs12759209	-0,18926	0,171	0,2683	-0,3477	0,217	0,1086
rs806454- rs17578868	0,22532	0,17	0,1861	0,3827	0,26	0,1404
rs2013041- rs10511302	0,00345	0,143	0,9808	-0,1725	0,19	0,363
rs10494757- rs11520658	-1,77249	1,013	0,08	0,2238	0,687	0,7447
rs1426588- rs10248926	0,1073	0,141	0,4472	0,0694	0,195	0,7214
rs2013041- rs10511302	0,00345	0,143	0,9808	-0,1725	0,19	0,363
rs10494757- rs11520658	-1,77249	1,013	0,08	0,2238	0,687	0,7447
rs1426588- rs10248926	0,1073	0,141	0,4472	0,0694	0,195	0,7214

Table 3 continued: The top 20 interaction terms after case-control studies in RS-I, RS-II and RS-III separate and combined. The interactions are sorted based on the *p*-value of the β_{int} after the case-control study in the combined data set of RS-I, RS-II and RS-III. The HDL-C levels are adjusted for sex and age, the residuals are not scaled around zero. Number of cases: 416 (RS-I), 244 (RS-II) and 447 (RS-III). Number of controls: 2996 (RS-I), 1602 (RS-II) and 1621 (RS-III). *1-*4 mark the interaction terms that are in high LD with each other: $R^2 > 0.913$.

Interaction	Case-control in RS-III			Case-control RS combined		
	β_{int}	SE	*p*-value	β_{int}	SE	*p*-value
rs6442460- rs10914332	-0,6454	0,305	0,0346	-0,53	0,1841	0,00398
rs2146043- rs11124513	-0,356	0,179	0,0472	-0,302	0,1097	0,00591
rs774801- rs12442098 *1	0,1702	0,21	0,4188	0,293	0,1175	0,01267
rs3006496- rs17729021	0,1344	0,147	0,359	0,208	0,087	0,01694
rs754970- rs2306478 *2	-0,2445	0,137	0,0741	-0,202	0,0871	0,0205
rs754970- rs3775972 *2	-0,2482	0,137	0,0695	-0,199	0,0869	0,02226
rs5770418- rs17234336	-0,2122	0,318	0,5047	-0,489	0,2213	0,02703
rs7631734- rs12442098 *1	0,1156	0,211	0,5845	0,256	0,1179	0,03002
rs2242312- rs11190870	0,2608	0,231	0,2594	0,304	0,1437	0,03462
rs4861849- rs17123865	0,7524	0,338	0,0262	0,446	0,2265	0,04902
rs754950- rs4658547 *3	-0,2082	0,174	0,2314	-0,202	0,1045	0,05368
rs1203791- rs10496556	-0,1164	0,155	0,4514	-0,162	0,0923	0,07865
rs754950- rs10926977 *3	-0,1322	0,17	0,4372	-0,172	0,1026	0,09409
rs900654- rs10195135 *4	0,2456	0,17	0,1477	0,168	0,1006	0,09566
rs10195135- rs10872670 *4	0,2195	0,168	0,192	0,166	0,1003	0,09804
rs2919732- rs12759209	-0,0725	0,164	0,6583	-0,168	0,1022	0,10006
rs806454- rs17578868	0,093	0,177	0,5992	0,18	0,1102	0,10155
rs2013041- rs10511302	-0,305	0,157	0,0518	-0,148	0,0911	0,105
rs10494757- rs11520658	-0,536	0,564	0,3416	-0,608	0,3834	0,11292
rs1426588- rs10248926	0,2176	0,145	0,1332	0,133	0,0889	0,13558
rs2013041- rs10511302	-0,305	0,157	0,0518	-0,148	0,0911	0,105
rs10494757- rs11520658	-0,536	0,564	0,3416	-0,608	0,3834	0,11292
rs1426588- rs10248926	0,2176	0,145	0,1332	0,133	0,0889	0,13558

Table 4: The overlap between the top 20 interaction terms after replication and case-control analysis.

Interaction	Meta-analysis of RS-I, RS-II and RS-III			Meta-analysis of replication cohorts		
	β_{int}	SE	*p*-value	β_{int}	SE	*p*-value
rs754950- rs10926977	0,0876	0,0148	2,92E-009	-0,01028	0,00548	0,06078
rs2242312- rs11190870	0,1238	0,0205	1,65E-009	0,01121	0,00671	0,09511
rs774801- rs12442098	-0,096	0,0167	8,57E-009	-0,01009	0,00588	0,08656

Interaction	Case-control in combined RS		
	β_{int}	SE	*p*-value
rs754950- rs10926977	-0,172	0,1026	0,09409
rs2242312- rs11190870	0,304	0,1437	0,03462
rs774801- rs12442098	0,293	0,1175	0,01267

Power calculations

As none of the findings replicated, we explored the statistical power of our analyses. Figure 3 shows the power calculations using the program G*Power[20,21]. With our current sample size of 2,996 individuals the smallest detectable effect will be 0.11, 0.095 and 0.05 when the type I error is less than $1 \cdot 10^{-7}$ and the type 2 error is 20% (power is 80%), 50% (power is 50%) and 99% (power is 1%), respectively.

DISCUSSION

Here we presented the, to our knowledge, first GWIS of HDL-C levels in blood. Our study shows that in a single population a GWIS results in 84,031 SNP×SNP interactions associated with HDL-C levels (p-value < $6.06 \cdot 10^{-7}$). Our two-step approach to filter these SNP×SNP interactions using two additional cohorts resulted in 181 interactions with a p-value below $1 \cdot 10^{-8}$. Although some reached nominal significance, none of these interactions terms were significantly replicated in a meta-analysis of 30,011 samples when adjusting for multiple testing. We also did not find a significant association between any of the interaction terms and treatment with lipid lowering medication in the cohorts of the Rotterdam Study after adjustment for multiple testing.

To our knowledge, no other GWIS studies with HDL-C exist with which we can compare our results. However, we did try to replicate previously published SNP×SNP interactions. We adjusted for the same covariates as the authors did, except for smoking, which was used as a covariate by Turner *et al.*[9]. Turner *et al.* published an interaction between rs253 and rs2515614 associated with HDL, however, the p-values of β_{int} after testing this interaction term were 0.986, 0.189 and 0.594 in the RS-I, RS-II and RS-III cohorts, respectively. The p-value of β_{int} after meta-analysing this interaction term is 0.614. The interaction term between rs3846662 and rs1532085, as published by Ma *et al.*[8], only replicated in RS-III (p = 0.0214), but not in

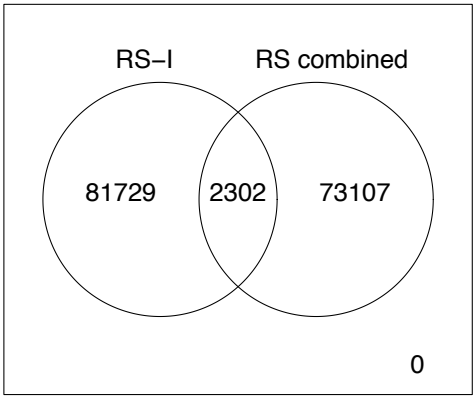


Figure 3: The smallest detectable effect with the current sample size of 2,996 individuals at 80% (a), 50% (b) and 1% (c) power levels.

RS-I (p = 0.212) or RS-II (p = 0.162). The p-value of β_{int} after meta-analysing this interaction term is 0.335.

There can be multiple reasons why we were not able to uncover SNP×SNP interactions using a hypothesis-free approach. First, in this study we selected only common variants (MAF > 0.05) which were genotyped in the Rotterdam Study. We chose these variants to avoid false positive findings in rare variants. Furthermore, the power to detect interaction terms with rare variants is low since our sample size in the two-stage discovery phase was 6,219. A second limitation that we chose to only investigate genotyped SNPs instead of imputed SNPs. Therefore, we may have missed true positive causal SNPs which are not on the genotyping array. However, even with only genotyped SNPs the number of potentially true positive findings is enormous, resulting in 84,031 suggestive hits at $p = 6.06 \cdot 10^{-7}$. This prompted us to use a two-stage discovery phase in which we used the RS-II and RS-III cohorts to filter out the false positives, reducing the number of findings from 84,031 to 181. The total number of individuals in this two-step discovery phase is 6,219. This might be considered low for the identification of SNP×SNP interactions. As a commonly used rule-of-thumb, the sample size within a GWIS should be 3 to 4 times the size of GWAS. As the first GWAS identifying loci associated with HDL-C levels[1] included 2,758 individuals, our study is expected to be underpowered by that rule. To improve power, an alternative approach could have been to combine the three cohorts of the Rotterdam Study into an one-step discovery with GLIDE. This, however, still yielded 75,409 interactions with a p-value below our threshold of $6.06 \cdot 10^{-7}$ as compared to the 84,031 interactions seen in the RS-I only GWIS, see Figure 4. It should be noted that both numbers are well in keeping with expectations.

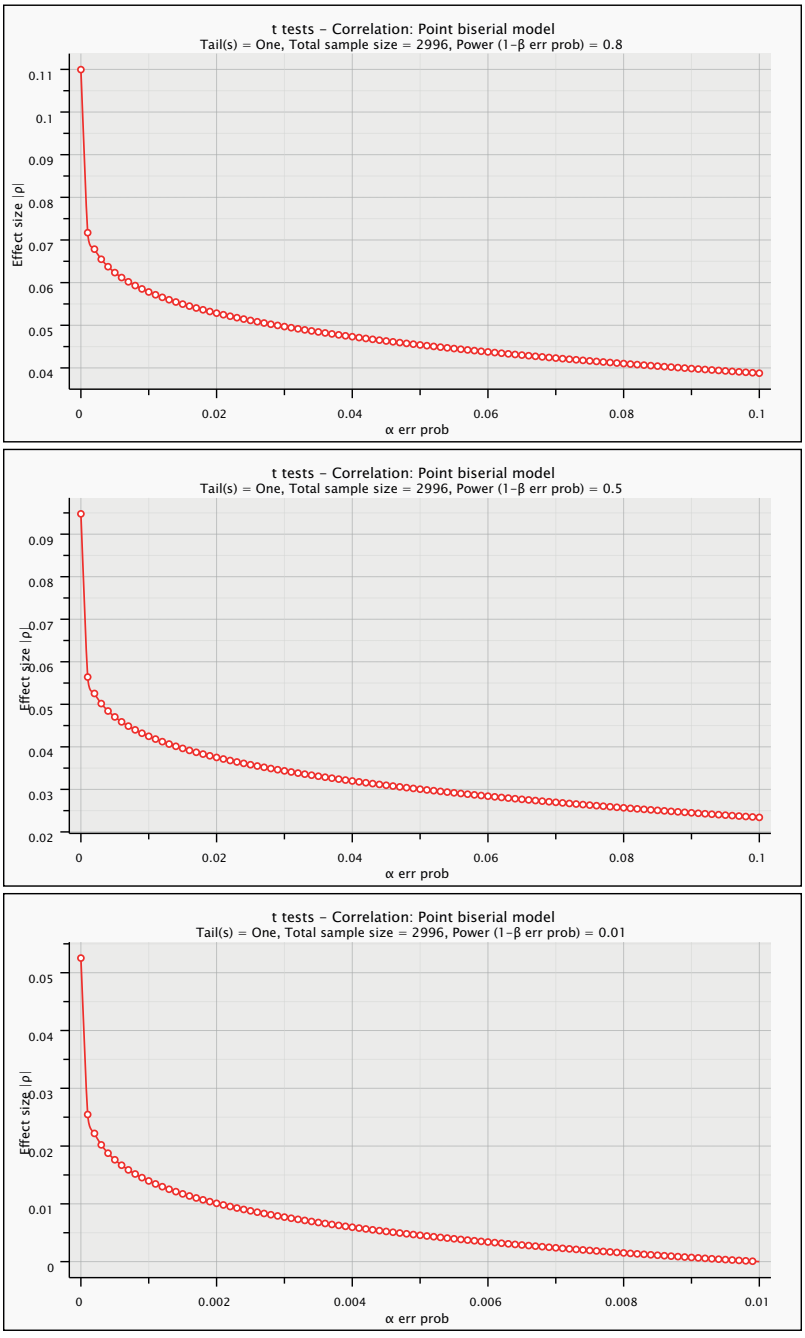


Figure 4: The overlap between the interaction terms with p-value < $3.03 \cdot 10^{-7}$ after a GWIS with GLIDE in RS-I only and after a GWIS with GLIDE in RS-I, RS-II and RS-III combined.

The proposed genome-wide significance level for GWIS is $1 \cdot 10^{-13}$, however, in this study we used all interaction terms with a *p*-value less than $1 \cdot 10^{-8}$ for replication. We chose a much less stringent *p*-value to prevent us from missing true positives due to the relatively small sample size. However, none of the 84,031 interaction terms had a *p*-value below $1 \cdot 10^{-13}$ in the separate Rotterdam Study cohorts and after meta-analysis of the three Rotterdam Study cohorts.

The success of GWAS has been its hypothesis-free approach and this worked well for studying lipids even in studies we consider small by today's standards (1000 – 3000 individuals). A GWIS is now technically feasible but needs larger sample sizes. Our study shows that the number of hits is overwhelming at a *p*-value of $1 \cdot 10^{-8}$. The filtering approach in a similar population did not resolve this problem. Our GWIS resulted in the consistent finding of a possible interaction between rs774801 in *ARMC8* and rs12442098 in *SPATA8* being associated with HDL-C levels, both in the quantitative analysis and the case-control analysis. However, *p*-values do not reach the preset Bonferroni correction of the *p*-values. Other major issues related to the sample size and apparent lack of replication also needs to be overcome.

REFERENCES

1. Kathiresan, S. *et al.* Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. *Nat Genet* **40**, 189–197 (2008).
2. Willer, C. J. *et al.* Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nat Genet* **40**, 161–169 (2008).
3. Aulchenko, Y. S. *et al.* Loci influencing lipid levels and coronary heart disease risk in 16 European population cohorts. *Nat Genet* **41**, 47–55 (2009).
4. Teslovich, T. M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707–713 (2010).
5. Global Lipids Genetics Consortium *et al.* Discovery and refinement of loci associated with lipid levels. *Nat Genet* **45**, 1274–1283 (2013).
6. Guay, S. P. *et al.* DNA methylation variations at CETP and LPL gene promoter loci: new molecular biomarkers associated with blood lipid profile variability. *Atherosclerosis* **228**, 413–420 (2013).
7. Pearce, M. S. *et al.* Global LINE-1 DNA methylation is associated with blood glycaemic and lipid profiles. *Int J Epidemiol* **41**, 210–217 (2012).
8. Ma, L. *et al.* Knowledge-driven analysis identifies a gene-gene interaction affecting high-density lipoprotein cholesterol levels in multi-ethnic populations. *PLoS Genet* **8**, e1002714 (2012).
9. Turner, S. D. *et al.* Knowledge-driven multi-locus analysis reveals gene-gene interactions influencing HDL cholesterol level in two independent EMR-linked biobanks. *PLoS One* **6**, e19586 (2011).
10. Surakka, I. *et al.* A genome-wide screen for interactions reveals a new locus on 4p15 modifying the effect of waist-to-hip ratio on total cholesterol. *PLoS Genet* **7**, e1002333 (2011).
11. Brockmann, G. A. *et al.* Genetic control of lipids in the mouse cross DU6i x DBA/2. *Mamm Genome* **18**, 757–766 (2007).
12. Kam-Thong, T. *et al.* GLIDE: GPU-based linear regression for detection of epistasis. *Hum Hered* **73**, 220–236 (2012).
13. Hofman, A. *et al.* The Rotterdam Study: 2012 objectives and design update. *Eur J Epidemiol* **26**, 657–686 (2011).
14. Fried, L. P. *et al.* The Cardiovascular Health Study: design and rationale. *Ann Epidemiol* **1**, 263–276 (1991).
15. Pardo, L. M., MacKay, I., Oostra, B., van Duijn, C. M. & Aulchenko, Y. S. The effect of genetic drift in a young genetically isolated population. *Ann Hum Genet* **69**, 288–295 (2005).
16. Kannel, W. B., Feinleib, M., McNamara, P. M., Garrison, R. J. & Castelli, W. P. An investigation of coronary heart disease in families. The Framingham offspring study. *Am J Epidemiol* **110**, 281–290 (1979).
17. Splansky, G. L. *et al.* The Third Generation Cohort of the National Heart, Lung, and Blood Institute's Framingham Heart Study: design, recruitment, and initial examination. *Am J Epidemiol* **165**, 1328–1335 (2007).
18. Rantakallio, P. Groups at risk in low birth weight infants and perinatal mortality. *Acta Paediatr Scand* **193**, Suppl 193:1+ (1969).
19. Aulchenko, Y. S., Ripke, S., Isaacs, A. & van Duijn, C. M. GenABEL: an R library for genome-wide association analysis. *Bioinformatics* **23**, 1294–1296 (2007).
20. Faul, F., Erdfelder, E., Lang, A.-G. & Buchner, A. G*power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behav Res Methods* **39**, 175–191 (2007).

21. Faul, F., Erdfelder, E., Buchner, A. & Lang, A.-G. Statistical power analyses using g*power 3.1: tests for correlation and regression analyses. *Behav Res Methods* **41**, 1149–1160 (2009).
22. Becker, T., Herold, C., Meesters, C., Mattheisen, M. & Baur, M. P. Significance levels in genome-wide interaction analysis (GWIA). *Ann Hum Genet* **75**, 29–35 (2011).

CHAPTER 4.2

Identification of rare variants associated with high density lipoprotein cholesterol (HDL-C) by exome sequencing in a family based study

Elisabeth M. van Leeuwen, Ayşe Demirkan, Najaf Amin, Aaron Isaacs, Jan Bert van Klinken, Aswin Verhoeven, Axel Meissner, Rutger W. W. Brouwer, Gina M. Peloso, CHARGE Lipids Working Group, Genome of the Netherlands consortium, Jeroen van Rooij, Morris A. Swertz, Aniko Sabo, Jonathan Marten, Hamdi Mbarek, Serkalem Demissie, Ani Manichaikul, Mary F. Feitosa, Niek Verweij, Jennifer A. Brody, Albert V. Smith, Abbas Dehghan, Fernando Rivadeneira, Marian Beekman, Qing Duan, Peter K. Joshi, Stella Trompet, Igor Rudan, Jennifer Huffman, Veronique Vitart, Ivana Kolcic, Ozren Polasek, Carolina Hayward, J. Wouter Jukema, James F. Wilson, James G. Wilson, P. Eline Slagboom, Oscar H. Franco, Andre G. Uitterlinden, Albert Hofman, Vilmundur Gudnason, Joshua C. Bis, Pim van der Harst, Ingrid B. Borecki, Stephen S. Rich, Charles W. White, Dorret Boomsma, Generation Scotland, Alanna M. Morrison, Lifelines Cohort Study, L. Adrienne Cupples, Ben Oostra, Wilfred F.J. van IJcken, Monique T. Mulder, Eric J. Sijbrands, Thomas Hankemeier, Ko Willems van Dijk and Cornelia M. van Duijn.

The manuscript for this chapter is currently under review.

The first author is willing to distribute the supplemental information for this chapter at your request.

ABSTRACT

Finding rare variants implicated in complex traits has proven to be difficult. Large family-based studies in isolated populations are enriched for rare variants due to founder effects and thus yield increased power for identifying these. We explored the role of rare variants by exome sequencing in determining high-density lipoprotein cholesterol (HDL-C) in the Erasmus Rucphen Family (ERF) study, a family based study. We identified 9 common (MAF > 0.1) and 9 rare variants (MAF < 0.01). The 9 common variants are all located within the *CETP* region, a region which is known to be associated with HDL-C level. We replicated these variants in 85,597 individuals. The 9 rare variants are located within genes not associated with HDL-C before. Carriers of the 9 rare variants have an extremely high HDL-C which is associated to a reduced risk of cardiovascular disease. We validated 7 out of the 9 rare variants by segregation analysis within pedigrees of at least 4 generations. Since HDL-C is a component of the metabolic syndrome, we additionally tested if the variants affecting HDL-C are also associated with several metabolomic compounds. Both rare and common variants were associated with clearly distinct metabolomic compounds in a locus-specific manner, indicating that distinct mechanisms underlie the association of the various loci with HDL-C. The present exome sequencing study shows that power of fine genotyping and phenotyping approaches in family based settings as follow up of genome-wide association studies, provides additional insight in the mechanisms underlying the association between specific loci and HDL-C.

INTRODUCTION

In recent years, various approaches have been successfully applied to unravel the genetic architecture of high density lipoprotein cholesterol (HDL-C) levels in humans. High HDL-C levels are associated with reduced risk of cardiovascular disease[1]. The estimated heritability of HDL-C is high: 47-76%[2-8]. Genome-wide association studies (GWAS) have revealed >70 common variants associated with HDL-C[9-11] while family based linkage studies have identified a large number of rare variants with large effects[12-14]. An extensive effort has been performed to identify variants with an in-between (0.001-0.01) minor allele frequency (MAF) associated with HDL-C. This includes genotyping, exome and whole genome sequencing and imputing the low-frequency and rare coding-sequencing variants[15-18]. However, few of these variants with an in-between frequency have been associated with HDL-C due the fact that these relatively rare single-variants have a modest to small impact on HDL-C and their low frequency requires large sample sizes to obtain sufficient statistical power. The stories of success concern primarily candidate-genes, which are deep sequenced[15-18].

An alternative approach to identify rare variants is to study extended families. Whole exome sequencing in families has been very successful in identifying rare variants with a large effect size[19-22]. However relatively few studies have addressed the contribution of rare variants with a modest effect size to specific traits, in particular circulating blood lipid levels[15,16,18]. In this study, we combined GWAS with whole exome sequencing in a family based population study to identify rare variants with modest effect sizes on HDL-C. To this end, we used the Erasmus Rucphen Family (ERF) study[23], a family-based study which includes a total of approximately 3,000 participants descending from 22 couples who lived in the Rucphen region in the southwest of the Netherlands in the 19th century. Therefore, the participants are not selected for a specific disease, allowing us to study genes that are associated with high and low HDL-C. Family-based studies have the advantage that the frequencies of genomic variants are increased due to founder effects and segregation of these variants with the disease can be studied[23], which increases the power to detect true positive associations. To gain additional insight in the molecular mechanisms underlying the association of specific variants with HDL-C, we determined their association with a variety of metabolomic compounds.

METHODS

Study population

This study as described here was conducted within the ERF study. The ERF study is a family based study that includes inhabitants of a genetically isolated community in the South-West of the Netherlands, studied as part of the Genetic Research in Isolated Population

(GRIP) program[23]. Study population includes approximately 3,000 individuals who are living descendants of 22 couples who had at least six children baptized in the community church. All data were collected between 2002 and 2005. The population shows minimal immigration and high inbreeding, therefore frequency of rare alleles is increased in this population[23]. All participants gave informed consent, and the Medical Ethics Committee of the Erasmus University Medical Centre approved the study.

High density lipoprotein measurements

Fasting blood samples were collected during the participant's visit to the research center. A Synchron LX20 (Beckman Coulter Inc., Fullerton, CA. U.S.A.) spectrophotometric chemical analyzer was utilized for the determination of plasma lipid values, among which HDL-C. Participants were asked to present the medications they used, including lipid-lowering medications. In individuals using statins, to account for the effect of statins on lipids, HDL-C was divided by 1.056. These adjustments are based on the sample-size weighted mean proportional differences in a large prospective meta-analysis including fourteen randomized trials of statins[24].

Exome sequencing

The exomes of 1,336 individuals from the ERF population were sequenced "in-house" at the Center for Biomics of the Department of Cell Biology of the Erasmus MC, The Netherlands. Sequencing was done at a median depth of 57x using the Agilent version V4 capture kit on an Illumina Hiseq2000 sequencer using the TruSeq Version 3 protocol. The sequence reads were aligned to the human genome build 19 (hg19) using BWA and the NARWHAL pipeline[25,26]. Subsequently, the aligned reads were processed further using the IndelRealigner, MarkDuplicates and TableRecalibration tools from the Genome Analysis Toolkit (GATK)[27] and Picard (http://picard.sourceforge.net) to remove systematic biases and to recalibrate the PHRED quality scores in the alignments. After processing, genetic variants were called using the Unified Genotyper tool from the GATK. About 1.4 million Single Nucleotide Variants (SNVs) were called and after removing the low quality variants (QUAL < 150) we retrieved 577,703 SNVs in 1,309 individuals. Further, for comparison and to predict the functionality of the variants, annotations were also performed using the dbNSFP (database of human non-synonymous SNPs and their functional predictions, http://varianttools.sourceforge.net/Annotation/DbNSFP) and Seattle (http://snp.gs.washington.edu/SeattleSeqAnnotation131/) databases. These databases gave functional prediction results from four different programs (polyPhen2, SIFT, MutationTaster and LRT), apart from gene and variant annotations.

Exome-wide association study of exome sequence data

For every SNV we did run a score test for association with HDL-C measures (Figure 1), thereby adjusting for age, age2, sex and family relatedness using mmscore of GenABEL package (version 1.6-7)[28,29]. The Bonferroni corrected significance threshold applied for this step is $2.572 \cdot 10^{-6}$ (0.05/19,438), as Seattle predicted the SNVs to be annotated within 19,438 unique genes. Linkage disequilibrium (LD) between the significant SNV's was estimated using the exome sequence data of the 1,309 individuals of the ERF study to define the independent number of significant SNVs.

Replication of the common SNVs

All SNVs with a MAF above 0.1 are considered common variants. These variants all occur in commonly used reference panels, like the 1000 Genomes reference panel and the Genome of the Netherlands (GoNL)[30]. We therefore replicated the common variants in an independent set of 85,597 individuals (Figure 1). Of these individuals, 33,613 individuals are from Dutch descent and therefore imputed to the GoNL reference panel (Lifelines, LLS, NTR-NESDA, PREVEND, PROSPER, RS-I, RS-II and RS-III). More details can be found in the Supplementary Material of van Leeuwen *et al.*[31]. The remaining 51,984 individuals are not of Dutch descent and therefore imputed to the 1000 Genomes reference panel (AGES, ARIC (African Americans (AA) and European Americans (EA)), CHS (EA), CROATIA KORCULA, CROATIA SPLIT, CROATIA VIS, FHS, FamHS, Generation Scotland, JHS, MESA (AFA, CAU, CHN and HIS) and ORCADES). Cohort descriptions of the individuals imputed to the 1000 Genomes reference panel can be found in Supplementary Methods and Supplementary Table 1 and 2. All studies were performed with the approval of the local medical ethics committees, and written informed consent was obtained from all participants. In most individuals, HDL-C was measured at fasting in subjects. We did not adjust for lipid lowering medication within the replication cohorts.

Replication and validation of rare SNVs

All SNVs with a MAF below 0.1 are considered rare variants. We tried to replicate (Figure 1) the rare SNV findings in the Rotterdam study cohort I (RS-I) exome sequence (*N*=1,387) and GoNL imputations (*N*=2,989). The RS-I is an ongoing prospective population-based cohort study, focused on chronic disabling conditions of the elderly. The study comprises an outbred ethnically homogenous population of Dutch Caucasian origin. The rationale of the study has been described in detail elsewhere[32]. In summary, 7,983 men and women aged 55 years or older, living in Ommoord, a suburb of Rotterdam, the Netherlands, were invited to participate in the first phase. 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

study participants. Fasting blood samples were taken during the participant's third visit to the research center. In the RS-I exomes of 2,628 individuals were sequenced at an average depth of 20x using the Nimblegen SeqCap EZ V2 capture kit on an Illumina Hiseq2000 sequencer and the TrueSeq Version 3 protocol. The sequences reads were aligned to hg19 using BWA[26]. Subsequently, the aligned reads were processed further using Picard, SAMtools[33] and GATK[27]. Genetic variants were called using Unified Genotyper Tool from GATK. Samples with low concordance to genotyping array (< 95%), low transition/transversion ratio (< 2.3), high heterozygote to homozygote ratio (> 2.0) and low call rate (< 90%) were removed from the data. SNVs with a low call rate (< 90%) and out of HWE (p-value < 10^{-8}) were also removed from the data. The final dataset consisted of 600,806 SNVs in 2,356 individuals. File handling and formatting was done using vcftools and PLINK. Annotation of the variants was performed using SeattleSeq annotation 138. The total number of individuals with both fasting HDL-C measurements and exome sequence data available which did not use any lipid lowering medication was 1,387. More details of the imputations with the GoNL reference panel can be found in the supplementary material of van Leeuwen *et al.*[31]. The total number of individuals with both fasting HDL-C measurements and GoNL imputations available which did not use any lipid lowering medication was 2,989.

Next, using the pedigree available for ERF participants, we visualized the pedigrees in which the rare (MAF < 0.1) SNVs are detected (Figure 1). Visual inspection of the pedigrees confirmed if the variants are artifacts or not.

Test for association with metabolomics compounds

For about 1,100 individuals within the ERF study, additional measurements of metabolomics compounds are available on 5 platforms. The first two, include lipid and TG species were quantified either by using liquid chromatography mass spectrometry (LC-MS)[34] or by electrospray ionization tandem mass spectrometry (ESIMS/MS)[35]. In addition to the lipidomics, a third one included aminoacids and acyl-carnitines were analyzed using the AbsoluteIDQTM p150 Kit of Biocrates Life Sciences AG, according to the manufacturer's recommendations and quantified using MetIQ software as integrated a part of the kit[36]. The fourth and fifth include two different extraction windows from nuclear magnetic resonance spectroscopy (NMR); small molecular compounds window as described before[37] and lipoprotein window as extracted using a commercially available algorithm developed by Bruker Corporation, Life Sciences services.

We tested whether the replicated or validated SNVs are also associated with any of the 713 metabolites (Figure 1). The sample sizes varies between the metabolomic compounds, for the analysis of phosphatidylcholine data is available for around 400 individuals, whereas the analysis of HDL-C particles contains about 1,150 samples. We therefore did run a score test for association, thereby adjusting for age[2], sex, lipid lowering medication (binary variable: yes or no) and family relatedness using mmscore of GenABEL package (version 1.6-7)[28,29].

As metabolites are related to each other we used the method of Li and Ji[38] to determine the number of effective number of independent variables. In our study, the number of independent variables was calculated to be 58.0047 and therefore the experiment-wide significance threshold required to keep type I error rate at 5% is $8.84 \cdot 10^{-4}$. Since we test 8 independent variants the final significance threshold for this section is $1.08 \cdot 10^{-5}$.

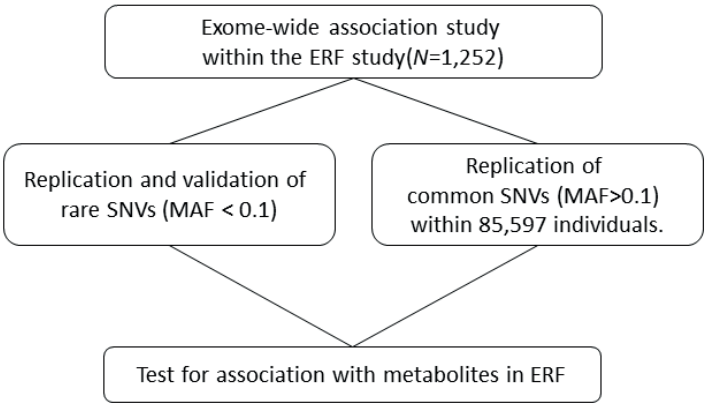


Figure 1. Flow diagram overview of the analysis plan.

Biocrates measurements

Serum samples from 992 individuals were analyzed using the AbsoluteIDQTM p150 Kit of Biocrates Life Sciences AG, according to the manufacturer's recommendations and quantified using MetIQ software as integrated a part of the kit. Liquid handling of serum samples (100 µl) was performed with a Hamilton Star (Hamilton Bonaduz AG) robot. Sample analyses were done on API 4000 Q TRAP LC/MS/MS System (Applied Biosystems) equipped with a Schimadzu Prominence LC20AD pump and a SIL-20AC autosampler. Briefly, the methods include flow injection, ESI-MS/MS detection and extraction. Quantification of the metabolites of the biological sample is achieved by reference to appropriate internal standards. Concentrations of all analyzed metabolites are reported as micromolar concentrations. The kit enables measurement of 14 amino acids, hexose (H1), free carnitine (C0), 40 acylcarnitines (Cx:y), hydroxylacylcarnitines (C(OH)x:y), and dicarboxylacylcarnitines (Cx:y-DC), 15 sphingomyelins (SMx:y) and N-hydroxylacyloylsphingosylphosphocholine (SM (OH)x:y), 77 phosphatidylcholines (PC, aa = diacyl, ae = acyl-alkyl) and 15 lyso-phosphatidylcholines. Lipid side chain composition is abbreviated as Cx:y, where x denotes the number of carbons in the side chain and y the number of double bonds. For example, "PC ae C33:1" denotes an acyl-alkyl phosphatidylcholine with 33 carbons in the two fatty acid side chains and a single double bond in one of them. Five reference samples included in each plate were used calculate the coefficient of variance (CV) and metabolites which have more than 25% of CV were excluded from the analysis. Outlying data points that were 5 standard deviations outside of the mean

were excluded from each variable. For the 27 metabolites which were measured by absolute quantification, 9 (C12, C14, C16, C18, C3, C4, C5, C6 (C4:1-DC) had lower median values than of their experimentally determined lower limit of quantification (LLOQ) and were excluded from the analysis. By definition LLOQ considers the lowest concentration that meets all quality criteria with respect to accuracy and precision according to the FDA guidelines. The precise position of the double bonds and the distribution of the carbon atoms in different fatty acid side chains cannot be determined with this technology.

Bioinformatic analysis

The biological relevance of the findings was validated by bioinformatic analysis with dbSNP, GeneCards and STRING interaction network. Specifically, to facilitate the manual process of assigning genes to a locus, we used an automated workflow developed in-house to generate reports containing the associated protein, enzyme, metabolic reaction, pathway, and disease phenotypes about each gene within a distance of +/- 200 kbp of the locus. In addition, SNVs published in the GWAS catalog[39] and eQTLs from the GTEx-eQTL database (http://www.ncbi.nlm.nih.gov/gtex/GTEX2) were given. In detail, the reports created by our workflow were based on the dbSNP[40], NCBI-Gene (http://www.ncbi.nlm.nih.gov/gene), GTEx-eQTL, GWAS catalog, ConsensusPathDB[41], UniProtKB[42], OMIM[43], Gene Ontology[44], TCDB[45], ExPASy[46] and KEGG database[47]. The databases had been downloaded earlier from the respective ftp servers and have been integrated offline. For the KEGG database the last freely available version was used (30-6-2011).

RESULTS

Exome-wide association study of exome sequence data

Figure 1 shows a flow diagram illustrating the analysis plan. We first conducted an exome-wide association study of HDL-C within 1,252 individuals who had both fasted HDL-C levels and the use of lipid lowering medication available. All individuals are part of the ERF study. Of the 1,252 individuals, 500 are male (39.94%) and 752 are female (60.06%). The mean age of the 1,252 individuals was 47.90 years (standard deviation of 14.19), the mean age of the males was 48.72 years (14.27) and the mean age of the females 47.36 years (14.13). 148 individuals out of the 1,252 individuals indicated using statins and we corrected their HDL-C as described in the methods section. Figure 2 and 3 show the results of the exome-wide association study. Although there is some inflation in the q-q plot (λ=1.05). The inflation is explained by both the common variants (particularly in the cholesteryl ester transfer protein (*CETP*) region) and the rarest variants (MAF < 0.01). There was no evidence for inflation for the low frequency variants (MAF between 0.01 and 0.1). There are 18 SNVs with a *p*-value

below 2.572·10-6, see Table 1. Of note is that the direction of the effect of all 18 SNVs except 2 SNVs in the *CETP* region are positive and thus increase the HDL-C, see Figure 4.

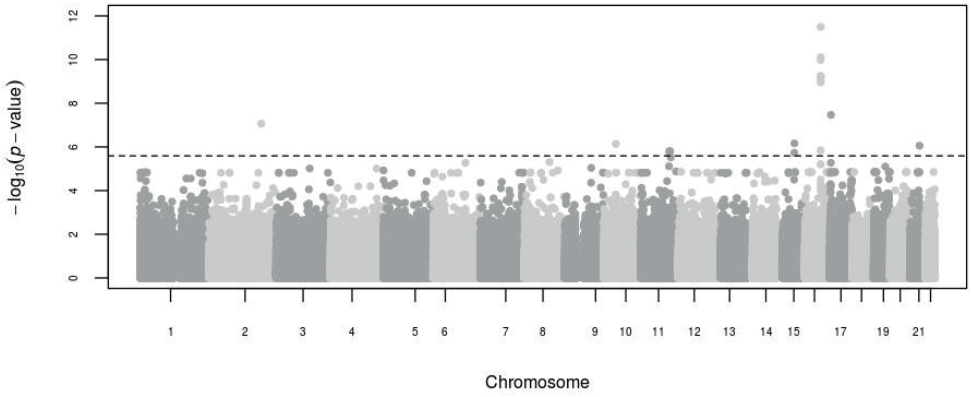


Figure 2. Results of a genome wide association analyses in 1,252 participants of the ERF study. The black line is the exome-wide significance line ($2.572 \cdot 10^{-6}$).

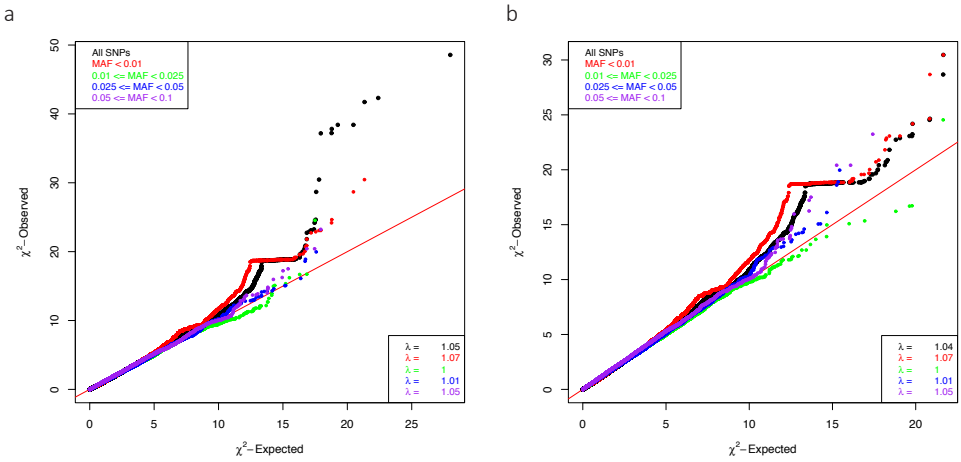


Figure 3. Q-Q plot for the genome wide association analyses in 1,252 participants of the ERF study. Figure a shows the q-q plot including all 563,909 SNVs, figure b shows the q-q plot after the 68 SNV's in the *CETP* region (chromosome 16, 56.99 Mbp – 57.02 Mbp) are removed.

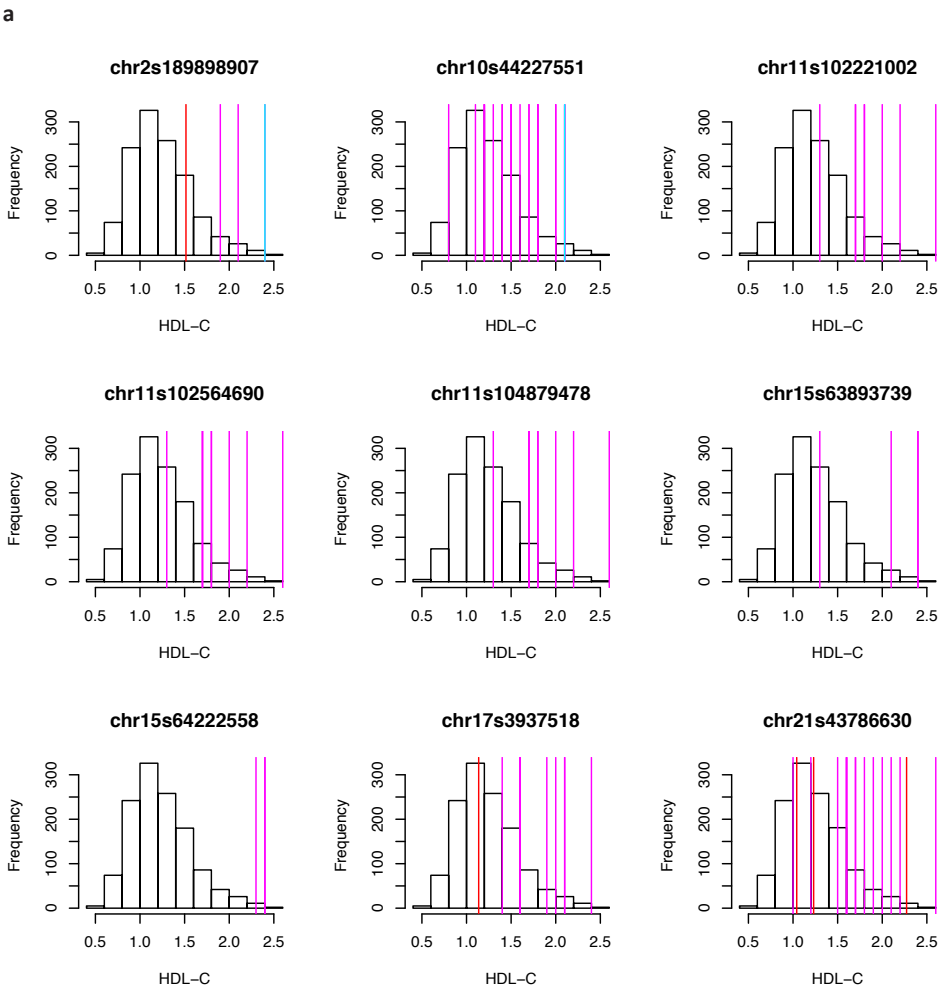


Figure 4. Histogram of (a) the HDL-C and (b) the residuals after adjusting HDL-C for sex, age, age^2 and family relationship in the 1,252 participants of the ERF study marking the carriers of the rare SNVs. The red line indicate the heterozygous carriers that use lipid lowering medication (the HDL-C is corrected for this lipid lowering medication). The magenta lines indicate the heterozygous carriers that do not use lipid lowering medication. The blue line indicate the homozygous carriers, these do not use lipid lowering medication.

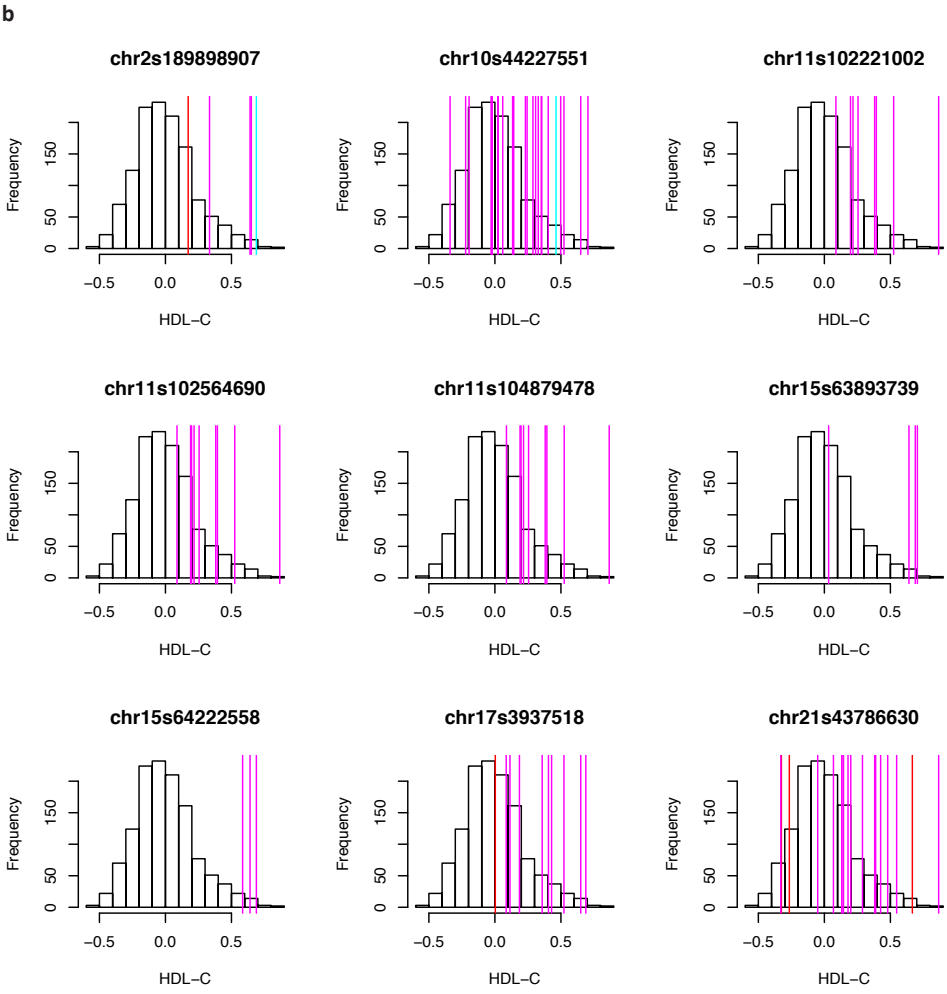


Figure 4. Continued. Histogram of (a) the HDL-C and (b) the residuals after adjusting HDL-C for sex, age, age^2 and family relationship in the 1,252 participants of the ERF study marking the carriers of the rare SNVs. The red line indicate the heterozygous carriers that use lipid lowering medication (the HDL-C is corrected for this lipid lowering medication). The magenta lines indicate the heterozygous carriers that do not use lipid lowering medication. The blue line indicate the homozygous carriers, these do not use lipid lowering medication.

LD analysis between the 18 significant SNV's using PLINK within the 1,309 individuals of the ERF study showed that all SNVs in the CETP region on chromosome 16 are in LD ($R^2 > 0.3$). This is also the case for the SNVs on chromosome 11 and 15 (Supplementary Table 3). Thus, we identified 9 independent loci.

Table 1. The significant SNVs associated to HDL-C after the exome-wide association study in the ERF population. * A1 is allele1, A2 is allele2. †N is the sample size. ‡N11, N12 and N22 are the number of homozygous of A1, the number of heterozygous and the number of homozygous of A2 among the 1,252 samples.

rsID	Chr:position	A1*	A2*	β	se_{β}	*p*-value	$N^{\dagger}$	$N_{11}{}^{\ddagger}$	$N_{12}{}^{\ddagger}$	$N_{22}{}^{\ddagger}$	Function	gene
rs146100075	2:189,898,907	T	C	0.569	0.11	8.56E-08	1251	1245	5	1	coding-synonymous	COL5A2
rs117090827	10:44,227,551	T	C	0.328	0.07	7.26E-07	1218	1194	23	1	intergenic	-
rs150868637	11:102,221,002	A	G	0.560	0.12	1.72E-06	1249	1241	8	0	coding-synonymous	BIRC2
rs141354791	11:102,564,690	G	A	0.530	0.11	1.56E-06	1249	1240	9	0	coding-synonymous	MMP27
11:104879478	11:104,879,478	T	C	0.530	0.11	1.56E-06	1249	1240	9	0	intron	CASP5
rs140242880	15:63,893,739	G	A	0.787	0.17	1.86E-06	1245	1241	4	0	missense	FBXL22
rs143777468	15:64,222,558	C	T	0.922	0.19	6.89E-07	1245	1242	3	0	intron	DAPK2
rs13306677	16:56,926,195	G	A	0.100	0.02	1.43E-06	1242	1019	196	27	intron	SLC12A3
rs17231506	16:56,994,528	C	T	0.096	0.01	3.19E-12	1245	436	602	207	near-gene-5	none
rs1800775	16:56,995,236	A	C	-0.085	0.01	7.77E-10	1245	322	650	273	near-gene-5	none
rs3816117	16:56,996,158	C	T	-0.084	0.01	1.08E-09	1245	326	653	266	intron	CETP
rs711752	16:56,996,211	G	A	0.088	0.01	7.77E-11	1245	331	632	282	intron	CETP
rs708272	16:56,996,288	G	A	0.088	0.01	1.05E-10	1245	331	630	284	intron	CETP
rs7205804	16:57,004,889	G	A	0.084	0.01	5.73E-10	1245	334	627	284	intron	CETP
rs1532625	16:57,005,301	C	T	0.084	0.01	5.73E-10	1245	334	627	284	intron	CETP
rs1532624	16:57,005,479	C	A	0.082	0.01	1.06E-09	1245	334	628	283	intron	CETP
rs35511240	17:3,937,518	G	A	0.532	0.10	3.41E-08	1244	1233	11	0	coding-synonymous	ZZEF1
rs190797467	21:43,786,630	C	A	0.392	0.08	8.75E-07	1242	1224	18	0	utr-5	TFF1

To evaluate whether the findings are influences by the correction of HDL-C in the 148 individuals that used statins, we re-evaluated the 18 variants excluding those treated. There were 4 *CETP* variants that are not significant anymore when we exclude treated individuals (Supplementary table 4). This may in part be explained by the lower statistical power after excluding 148 out of 1,252 individuals. As the effect was in the same direction and very similar as in the initial discovery analyses, we took these variants forward to the replication. All rare variants remained significant when both excluding the individuals using statins.

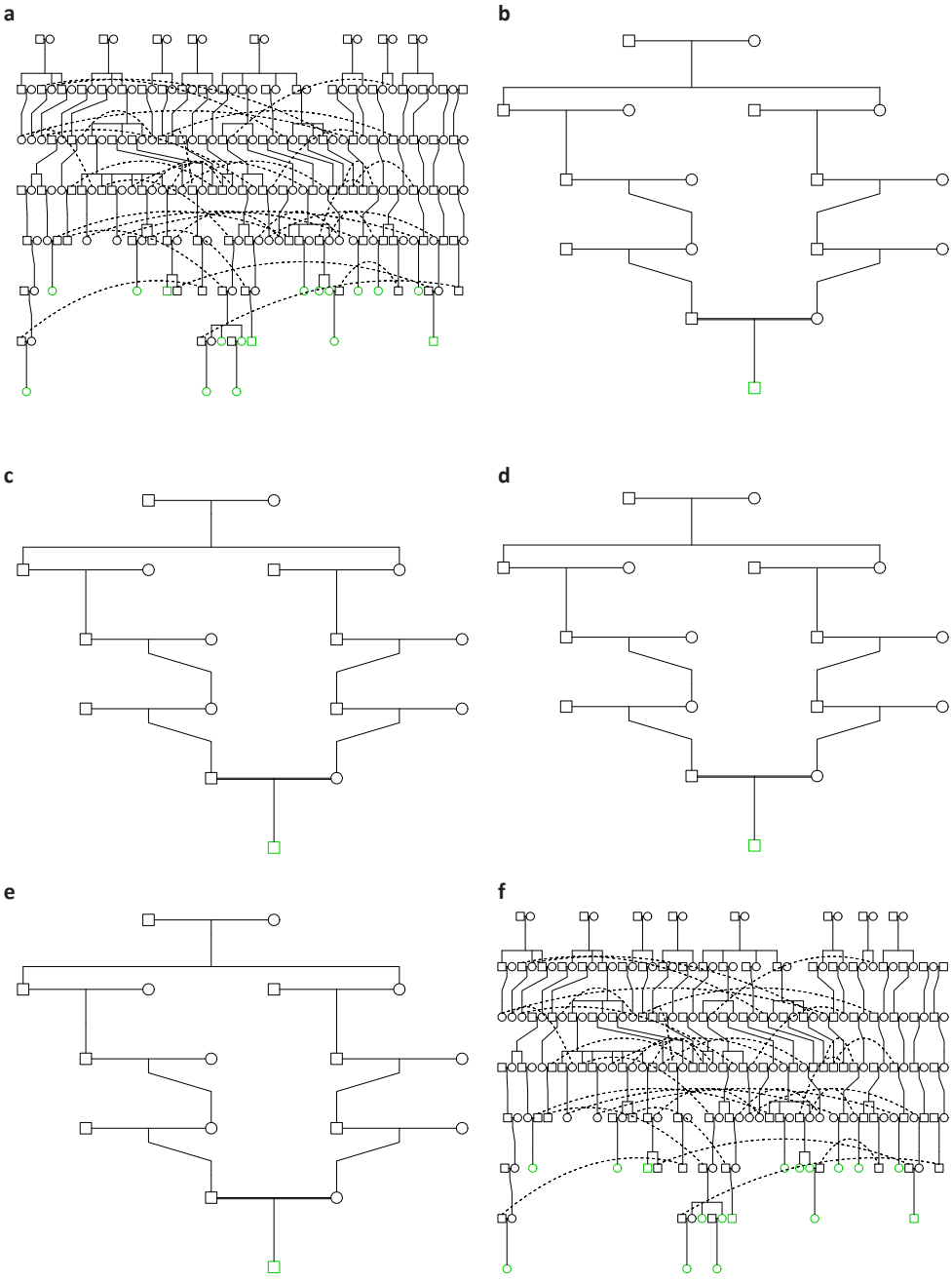
Replication of the common SNVs

Table 2 shows the association results for the common variants after the meta-analysis of 33,613 individuals from Lifelines, LLS, NTR-NESDA, PREVEND, PROSPER, RS-I, RS-II and RS-III imputed to the GoNL reference panel and 51,984 individuals from AGES, ARIC (AA and EA), CHS (EA), CROATIA KORCULA, CROATIA SPLIT, CROATIA VIS, FamHS, FHS, Generation Scotland, JHS, MESA (AFA, CAU, CHN and HIS) and ORCADES imputed to the 1000 Genomes reference panel combined. All common variants are genome-wide significantly replicated including those that lost their significance when excluding those treated.

Table 2. The replication of the significant common SNVs within 85,597 samples.
* A1 is allele1, A2 is allele2. † Freq is the frequency of A1. ‡ β is the effect of A1. § Direction: AGES – ARIC (AA) – ARIC (EA) – CHS (EA) – CROATIA KORCULA – CROATIA SPLIT – CROATIA VIS – FamHS – FHS – Generation Scotland – JHS – Lifelines – LLS – MESA (AFA) – MESA (CAU) – MESA (CHN) – MESA (HIS) – NTR – ORCADES – PRVEND – PROSPER – RS-I – RS-II – RS-III.

rsID	A1*	A2*	Freq†	β‡	se_{β}	*p*-value
rs13306677	A	G	0.075	0.040	0.006	9.03E-12
rs17231506	T	C	0.330	0.095	0.003	6.74E-189
rs1800775	A	C	0.474	0.086	0.003	1.92E-167
rs3816117	C	T	0.477	0.085	0.003	4.30E-166
rs711752	A	G	0.435	0.085	0.003	1.18E-164
rs708272	A	G	0.435	0.085	0.003	1.07E-164
rs7205804	A	G	0.437	0.082	0.003	8.68E-159
rs1532625	T	C	0.438	0.082	0.003	1.92E-158
rs1532624	A	C	0.438	0.083	0.003	7.27E-160

rsID	Direction§																							
rs13306677	+	-	+	+	+	-	-	+	+	+	+	+	+	-	+	-	+	+	+	+	+	+	+	+
rs17231506	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
rs1800775	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
rs3816117	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
rs711752	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
rs708272	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
rs7205804	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+
rs1532625	+	+	+	+	+	+	+	+	+	+	-	+	+	?	?	?	?	+	+	+	+	+	+	+
rs1532624	+	+	+	+	+	+	+	+	+	+	-	+	+	+	+	+	+	+	+	+	+	+	+	+


a
b
c
d
e
f

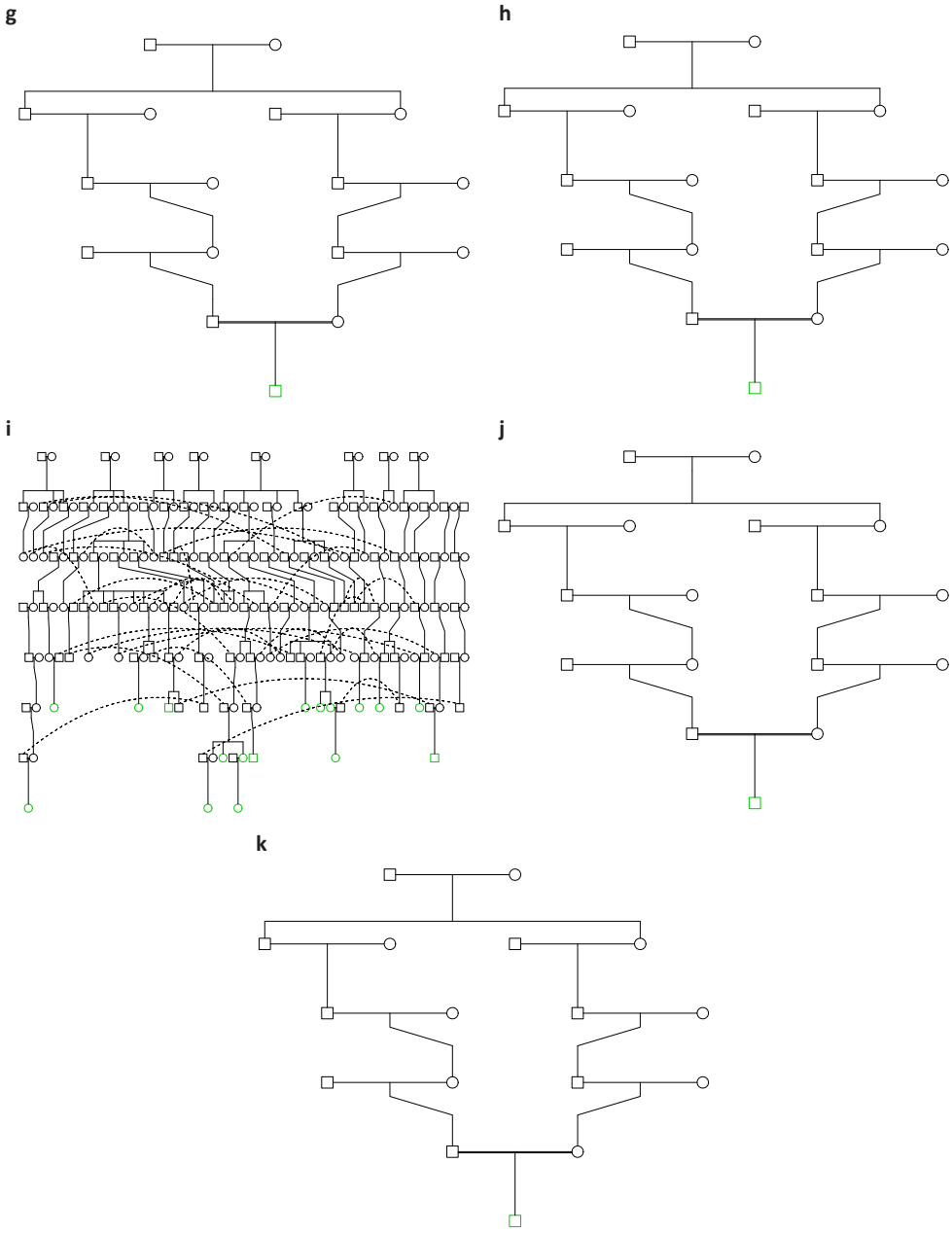


Figure 5. Segregation of the rare variants within families of the ERF studies. (a) chr2s189898907 family 1, (b) chr2s189898907 family 2, (c) chr10s44227551 family 1, (d) chr11s102221002 family 1, (e) chr11s102564690 family 1, chr11s104879478 family 1, (f) chr15s63893739 family 1, (g) chr15s63893739 family 2, (h) chr15s64222558 family 1, (i) chr17s3937518 family 1, (j) chr17s3937518 family 2, (k) chr21s43786630 family 1. No exome data was available for the individuals colored in black. Individuals colored in red do have exome data available, but no variant was detected. Individuals colored green are heterozygous for the variant and individuals colored in blue are homozygous for the variant.

Replication and validation of rare SNVs

Replication efforts within the exome sequencing project of CHARGE failed due to different phenotype definitions and chip differences. We tried to replicate the rare SNV findings in the Rotterdam study cohort I exome sequence (*N*=1,387) and GoNL imputations (*N*=2,989) and found 4 rare SNVs in the exome sequence data (rs146100075, rs150868637, rs141354791 and rs143777468) and 5 rare SNVs in the GoNL imputations (rs146100075, rs117090827, rs140242880, rs143777468 and rs35511240). Supplementary table 5 gives the effect of the variants, none of the rare variants were marginally significant.

As none of the rare variants (MAF < 0.1) could be replicated by imputation or exome sequencing, we studied segregation of these variants within families. SNVs not segregating from one generation to the next generation, might be de novo mutations but more likely are technique errors. Only the two rare loci on chromosome 15 (rs140242880 and rs143777468) did not segregate within pedigrees of multiple generations and may therefore be false positives, see Figure 5f, 5g and 5h. Out of the 6 carriers of rs146100075, 4 could be linked in a single pedigree within 4 generations, the 24 carriers of rs117090827 could be linked within 6 generations, the 8 carriers of rs150868637 could be linked within 5 generations, just like the 9 carriers of rs141354791 and the 9 carriers of the SNV on chromosome 11 without an rs-identifier on position 104,879,478, the 11 carriers of rs35511240 could be linked in 2 families, one including 8 carriers and the other 3 carriers and of the 18 carriers of rs190797467, 17 could be linked within 5 generation, see Figure 5. The 9 carriers of rs141354791 are the same individuals as the 9 carriers of the SNV on chromosome 11 without an rs-identifier on position 104,879,478. Of those 9 carriers, 8 are also carrier of the rs150868637 variant.

Test for association with other phenotypes

The 16 exome wide significant SNVs which were replicated or validated, were tested for association with other related metabolomic compounds. The T-scores (β divided by standard error) and *p*-values of all associations were used to create a heatmap, see Figure 6. In total 47 associations between a SNV and metabolomic compounds were significant after Bonferroni correction (*p*-value < $1.08 \cdot 10^{-4}$), see Table 3. In Figure 6, the column of the dendograms show a clear separation in the common variants on chromosome 16 and the other (rare) variants. This is most likely explained by the smaller effect sizes of the common variants compared to the effect sizes of the rare variants for most metabolomic compounds. As expected, we found association between apolipoprotein A-I (*ApoA1*) and *CETP*[48] and between apolipoprotein A-II (*ApoA2*) and *CETP*[49]. There is a significant cluster of two variants within the *CETP* region (rs3816117 and rs1800775) and M-HDL-ApoA1 (*p*-value of $5.19 \cdot 10^{-5}$ and $5.05 \cdot 10^{-5}$, respectively), L-HDL-ApoA2 (*p*-value of $4.83 \cdot 10^{-5}$ and $3.72 \cdot 10^{-5}$, respectively) and M-HDL-ApoA2 (*p*-value of $6.96 \cdot 10^{-5}$ and $6.89 \cdot 10^{-5}$, respectively). Carriers of the minor allele of the genetic variants on chromosome 16 showed decreased levels of these metabolomics compounds.

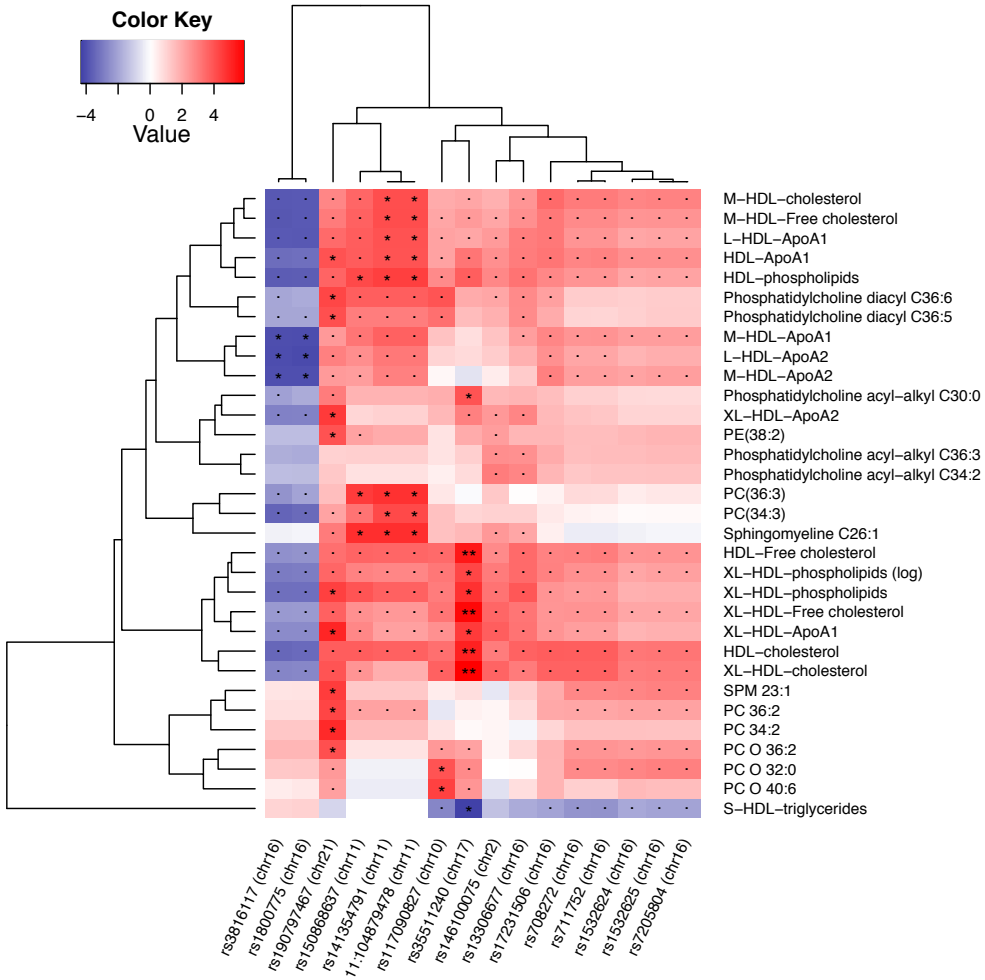


Figure 6. Heatmap based on the T-score of the associations between all replicated or validated SNVs and metabolomic phenotypes. Only the metabolomics phenotypes are shown which are significantly associated with at least one SNV. Associations marked with . have a *p*-value between 0.05 and $1.08 \cdot 10^{-4}$, associations marked with * have a *p*-value between $1.08 \cdot 10^{-4}$ and $5 \cdot 10^{-7}$ and associations marked with ** have a *p*-value between $5 \cdot 10^{-7}$ and $5 \cdot 10^{-9}$.

Table 3. The significant results of the test for association with metabolomic compounds.

rsID	chr	position	metabolite name	β	se_β	*p*-value	*N*
rs35511240	17	3,937,518	XL-HDL-cholesterol	1.661	0.283	5.52E-09	1145
rs35511240	17	3,937,518	XL-HDL-Free cholesterol	1.602	0.282	1.80E-08	1146
rs35511240	17	3,937,518	HDL-Free cholesterol	1.488	0.283	1.75E-07	1151
rs35511240	17	3,937,518	HDL-cholesterol	1.451	0.285	4.27E-07	1151
rs190797467	21	43,786,630	XL-HDL-ApoA1	1.134	0.227	6.67E-07	1146
rs190797467	21	43,786,630	PC 34:2	1.531	0.311	1.29E-06	413
rs35511240	17	3,937,518	XL-HDL-phospholipids	1.373	0.290	2.54E-06	1149
11:104879478	11	104,879,478	PC(36:3)	1.648	0.349	2.56E-06	1074
rs141354791	11	102,564,690	PC(36:3)	1.648	0.349	2.56E-06	1074
11:104879478	11	104,879,478	Sphingomyeline C26:1	4.439	0.933	2.76E-06	382
rs150868637	11	102,221,002	Sphingomyeline C26:1	4.439	0.933	2.76E-06	382
rs141354791	11	102,564,690	Sphingomyeline C26:1	4.439	0.933	2.76E-06	382
rs190797467	21	43,786,630	XL-HDL-ApoA2	1.028	0.227	6.78E-06	1143
rs35511240	17	3,937,518	XL-HDL-ApoA1	1.340	0.298	7.38E-06	1146
rs150868637	11	102,221,002	PC(36:3)	1.673	0.373	7.99E-06	1074
rs190797467	21	43,786,630	SPM 23:1	1.475	0.331	1.09E-05	415
11:104879478	11	104,879,478	HDL-phospholipids	1.399	0.319	1.27E-05	1156
rs141354791	11	102,564,690	HDL-phospholipids	1.399	0.319	1.27E-05	1156
rs190797467	21	43,786,630	XL-HDL-phospholipids	0.969	0.222	1.40E-05	1149
rs35511240	17	3,937,518	S-HDL-triglycerides	-1.216	0.282	1.76E-05	1151
rs117090827	10	44,227,551	PC O 40:6	1.571	0.363	1.90E-05	406
rs190797467	21	43,786,630	PC 36:2	1.272	0.301	2.90E-05	413
rs35511240	17	3,937,518	XL-HDL-phospholipids	1.188	0.283	2.91E-05	1149
rs190797467	21	43,786,630	Phosphatidylcholine diacyl C36:6	1.671	0.395	3.01E-05	378
rs150868637	11	102,221,002	HDL-phospholipids	1.408	0.339	3.49E-05	1156
rs141354791	11	102,564,690	PC(34:3)	1.550	0.374	3.61E-05	1072
11:104879478	11	104,879,478	PC(34:3)	1.550	0.374	3.61E-05	1072
rs1800775	16	56,995,236	L-HDL-ApoA2	-0.278	0.067	3.72E-05	1152
rs3816117	16	56,996,158	L-HDL-ApoA2	-0.273	0.067	4.83E-05	1152
rs141354791	11	102,564,690	M-HDL-Free cholesterol	1.272	0.312	4.92E-05	1157
11:104879478	11	104,879,478	M-HDL-Free cholesterol	1.272	0.312	4.92E-05	1157
rs1800775	16	56,995,236	M-HDL-ApoA1	-0.276	0.068	5.05E-05	1152
rs3816117	16	56,996,158	M-HDL-ApoA1	-0.274	0.067	5.19E-05	1152
rs141354791	11	102,564,690	M-HDL-cholesterol	1.274	0.315	5.55E-05	1156
11:104879478	11	104,879,478	M-HDL-cholesterol	1.274	0.315	5.55E-05	1156
rs1800775	16	56,995,236	M-HDL-ApoA2	-0.269	0.067	6.89E-05	1152
rs3816117	16	56,996,158	M-HDL-ApoA2	-0.267	0.067	6.96E-05	1152
rs190797467	21	43,786,630	Phosphatidylcholine diacyl C36:5	1.541	0.383	6.97E-05	378
rs190797467	21	43,786,630	PC O 36:2	1.377	0.344	7.39E-05	413
rs141354791	11	102,564,690	L-HDL-ApoA1	1.273	0.320	7.56E-05	1156
11:104879478	11	104,879,478	L-HDL-ApoA1	1.273	0.320	7.56E-05	1156
rs190797467	21	43,786,630	PE(38:2)	0.952	0.240	7.60E-05	1068
rs190797467	21	43,786,630	HDL-ApoA1	0.868	0.219	7.82E-05	1150
rs141354791	11	102,564,690	HDL-ApoA1	1.319	0.336	8.99E-05	1155
11:104879478	11	104,879,478	HDL-ApoA1	1.319	0.336	8.99E-05	1155
rs117090827	10	44,227,551	PC O 32:0	1.504	0.384	1.07E-05	406
rs35511240	17	3,937,518	Phosphatidylcholine acyl-alkyl C30:0	2.011	0.514	1.08E-04	380

Another cluster includes the three rare variants located on chromosome 11 (rs150868637, rs141354791 and a SNV without a rs-identifier on position 104,879,478) and the metabolites sphingomyelin C26:1 (p-value of $2.76 \cdot 10^{-6}$ for all three variants), phosphatidylcholine diacyl C34:3 (*p*-value of $1.60 \cdot 10^{-3}$, $3.61 \cdot 10^{-5}$ and $3.61 \cdot 10^{-5}$, respectively) and phosphatidylcholine diacyl C36:6 (*p*-value of $7.99 \cdot 10^{-6}$, $2.56 \cdot 10^{-6}$ and $2.56 \cdot 10^{-6}$, respectively). Both phosphatidylcholine diacyl species can accommodate a linoleic acid (C18:2) moiety. Also rs190797467 on chromosome 21 is significantly associated with multiple metabolites of linoleic acid (C18:2). Carriers of the variants on chromosome 11 and 21 showed increased levels of these metabolomics compounds.

The variant on chromosome 17, rs35511240 clusters strongly with multiple large HDL-C particles including XL-HDL-ApoA1 (*p*-value = $7.38 \cdot 10^{-6}$), XL-HDL-cholesterol (*p*-value = $5.52 \cdot 10^{-9}$), XL-HDL-Free cholesterol (*p*-value = $1.80 \cdot 10^{-8}$), XL-HDL-phospholipids (*p*-value = $2.54 \cdot 10^{-6}$) and XL-HDL-phospholipids (*p*-value = $2.91 \cdot 10^{-5}$). Carriers of this variant showed increased levels of these large HDL-C particles.

DISCUSSION

Combining GWAS with whole exome sequencing and metabolomics in a family-based study, resulted in 18 significant SNVs (*p*-value < $2.572 \cdot 10^{-6}$), among which 9 common variants within the *CETP*-region. These findings provide a bench mark, as this region is known to be associated with HDL-C[11] levels. As expected, the *CETP* clustered with *ApoA1* and *ApoA2* metabolites[48,49] providing a proof-of-principle of the cluster analyses of the new variants with the metabolomics compounds. We found 9 rare variants which were too rare in other populations to replicate. However, with the exception of the two chromosome 15 variants, 7 variants segregated in families. Of interest are the 7 rare variants (MAF < 0.1) associated with HDL-C levels and their association to various metabolomic compounds.

We found two clusters of variants and metabolites The first rare variant cluster involves variants on chromosome 11 and 21 and various metabolomics compounds of linoleic acid (18:2). This finding is in line with earlier publications on the association between linoleic acid and HDL-C[50]. The three variants on chromosome 11 are located in three distinct genes in a 2,6558,476 base pair region: the *BIRC2* (baculoviral IAP repeat containing 2) gene, *MMP27* (matrix metallopeptidase 27) gene and the *CASP5* (caspase 5, apoptosis-related cysteine peptidase) gene. The protein encoded by the *BIRC2* gene is a member of a family of proteins that inhibits apoptosis by binding to tumor necrosis factor receptor-associated factors *TRAF1* and *TRAF2*, probably by interfering with activation of ICE-like proteases. Proteins of the matrix metalloproteinase (*MMP*) family are involved in the breakdown of extracellular matrix in normal physiological processes, such as embryonic development, reproduction, and

tissue remodeling, as well as in disease processes, such as arthritis and metastasis. The *CASP5* gene encodes a member of the cysteine-aspartic acid protease (caspase) family. Sequential activation of caspases plays a central role in the execution-phase of cell apoptosis. There is some evidence that this gene is involved in pantothenate and CoA biosynthesis (*p*-value = $4.34 \cdot 10^{-4}$, genenetwork.nl). The chromosome 21 locus, rs190797467, is located within the *TFF1* (trefoil factor 1) gene. Members of the trefoil family are stable secretory proteins expressed in gastrointestinal mucosa. The function is not defined, but they may protect the mucosa from insults, stabilize the mucus layer, and affect healing of the epithelium. Of note is that the chromosome 21 variant in *TFF1* is 69 kbp downstream of the ATP-binding cassette, subfamily G, member 1 (*ABCG1*) gene. This gene encodes an active lipid transporter and possesses different binding sites for cholesterol[51]. GO annotations related to this gene include phospholipid binding.

The third cluster involves the association of rs35511240 on chromosome 17 with multiple large metabolites like XL-HDL-ApoA1, XL-HDL-cholesterol, XL-HDL-Free cholesterol, XL-HDL-phospholipids and XL-HDL-phospholipids. The chromosome 17 variant (rs35511240) is located within the *ZZEF1* (zinc finger, ZZ-type with EF-hand domain 1) gene. There is some evidence that this locus is involved in phosphatidylinositol signaling system (*p-value* = $1.24 \cdot 10^{-4}$, genenetwork.nl). The chromosome 17 locus is associated with high levels of HDL-C and XL-HDL, which both are associated to a reduced risk of cardiovascular disease[52].

Two rare variants do not cluster with the metabolic products: rs146100075 located on chromosome 2 within the *COL5A2* (collagen, type V, alpha 2) gene and an intergenic variant on chromosome 10 (rs117090827).

A potential limitation of our study is the lack of replication of the rare variants. Replication failed due to the extremely low frequency of these variants and study specific discrepancies in study design and SNV imputation. However, segregation analysis within the ERF study confirmed that all variants, except the two variants on chromosome 15 (rs140242880 and rs143777468), segregate in pedigrees of at least 4 generations. This suggests that these 7 SNVs are not artifacts in the ERF cohort but do not prove a causal association to HDL-C. The strength of our study is the population based design. As opposed to clinical studies, population based studies may yield clues to mechanisms involved in “healthy” individuals.

This study shows that combining GWAS with next-generation sequencing and metabolomics within large family studies can help us unraveling the process from variant into biological processes influencing clinical measurements. By using a family based study instead of a clinical study, this study yielded clues to mechanisms involved in “healthy” individuals. Large population-based samples will be needed to replicate the findings and final replication of our findings await the result of ongoing sequencing efforts. The combination of exome sequencing and metabolomics in the general population allows to identify specific lipid compounds that may be of interest for therapy development.

REFERENCES

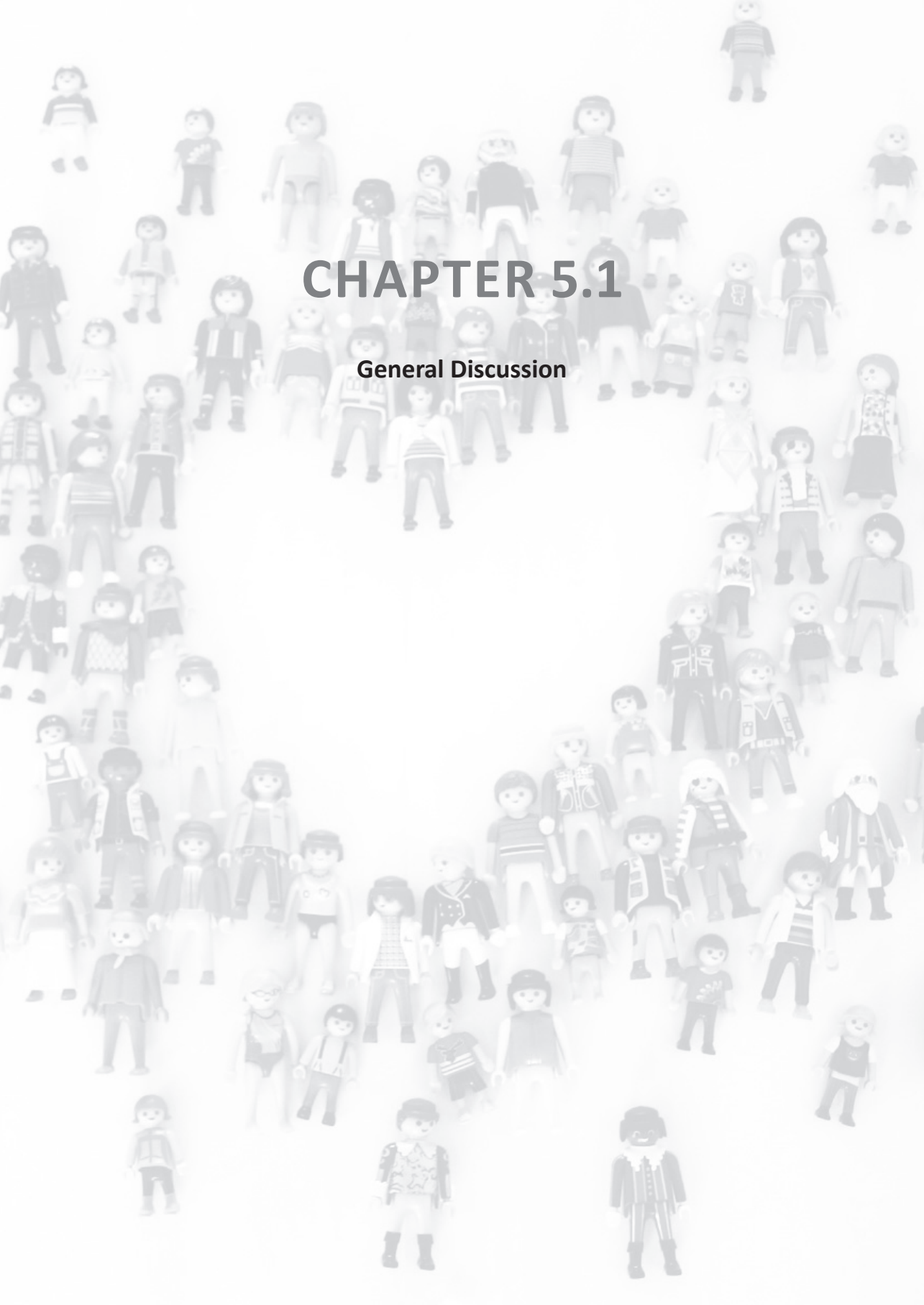
1. Castelli, W. P. *et al.* HDL cholesterol and other lipids in coronary heart disease. the cooperative lipoprotein phenotyping study. *Circulation* **55**, 767–772 (1977).
2. Almgren, P. *et al.* Heritability and familiality of type 2 diabetes and related quantitative traits in the Botnia Study. *Diabetologia* **54**, 2811–2819 (2011).
3. Browning, S. R. & Browning, B. L. Identity-by-descent-based heritability analysis in the Northern Finland Birth Cohort. *Hum Genet* **132**, 129–138 (2013).
4. Friedlander, Y., Kark, J. D. & Stein, Y. Biological and environmental sources of variation in plasma lipids and lipoproteins: the Jerusalem Lipid Research Clinic. *Hum Hered* **36**, 143–153 (1986).
5. Morrison, A. C. *et al.* Whole-genome sequence-based analysis of high-density lipoprotein cholesterol. *Nat Genet* **45**, 899–901 (2013).
6. Sung, J., Lee, K. & Song, Y.-M. Heritabilities of the metabolic syndrome phenotypes and related factors in Korean twins. *J Clin Endocrinol Metab* **94**, 4946–4952 (2009).
7. Vattikuti, S., Guo, J. & Chow, C. C. Heritability and genetic correlations explained by common SNPs for metabolic syndrome traits. *PLoS Genet* **8**, e1002637 (2012).
8. Zhou, X., Carbonetto, P. & Stephens, M. Polygenic modeling with bayesian sparse linear mixed models. *PLoS Genet* **9**, e1003264 (2013).
9. Global Lipids Genetics Consortium *et al.* Discovery and refinement of loci associated with lipid levels. *Nat Genet* **45**, 1274–1283 (2013).
10. Surakka, I. *et al.* The impact of low-frequency and rare variants on lipid levels. *Nat Genet* **47**, 589–597 (2015).
11. Teslovich, T. M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707–713 (2010).
12. Hellwege, J. N. *et al.* Genome-wide family-based linkage analysis of exome chip variants and cardiometabolic risk. *Genet Epidemiol* **38**, 345–352 (2014).
13. Reddy, M. V. P. L. *et al.* Exome sequencing identifies 2 rare variants for low high-density lipoprotein cholesterol in an extended family. *Circ Cardiovasc Genet* **5**, 538–546 (2012).
14. Sanghera, D. K. *et al.* Genome-wide linkage scan to identify loci associated with type 2 diabetes and blood lipid phenotypes in the sikh diabetes study. *PLoS One* **6**, e21188 (2011).
15. Do, R. *et al.* Exome sequencing identifies rare LDLR and APOA5 alleles conferring risk for myocardial infarction. *Nature* **518**, 102–106 (2015).
16. Myocardial Infarction Genetics Consortium Investigators *et al.* Inactivating mutations in NPC1L1 and protection from coronary heart disease. *N Engl J Med* **371**, 2072–2082 (2014).
17. Peloso, G. M. *et al.* Association of low-frequency and rare coding-sequence variants with blood lipids and coronary heart disease in 56,000 whites and blacks. *Am J Hum Genet* **94**, 223–232 (2014).
18. The TG and HDL Working Group of the Exome Sequencing Project, National Heart, Lung, and Blood Institute *et al.* Loss-of-function mutations in APOC3, triglycerides, and coronary disease. *N Engl J Med* **371**, 22–31 (2014).
19. van Bon, B. W. M. *et al.* Cantú syndrome is caused by mutations in ABCC9. *Am J Hum Genet* **90**, 1094–1101 (2012).
20. Boczek, N. J. *et al.* Novel timothy syndrome mutation leading to increase in CACNA1C window current. *Heart Rhythm* **12**, 211–219 (2015).
21. Makrythanasis, P. *et al.* MLL2 mutation detection in 86 patients with Kabuki syndrome: a genotype-phenotype study. *Clin Genet* **84**, 539–545 (2013).

22. Schuurs-Hoeijmakers, J. H. M. *et al.* Identification of pathogenic gene variants in small families with intellectually disabled siblings by exome sequencing. *J Med Genet* **50**, 802–811 (2013).
23. Pardo, L. M., MacKay, I., Oostra, B., van Duijn, C. M. & Aulchenko, Y. S. The effect of genetic drift in a young genetically isolated population. *Ann Hum Genet* **69**, 288–295 (2005).
24. Baigent, C. *et al.* Efficacy and safety of cholesterol-lowering treatment: prospective meta-analysis of data from 90,056 participants in 14 randomised trials of statins. *Lancet* **366**, 1267–1278 (2005).
25. Brouwer, R. W. W., van den Hout, M. C. G. N., Grosveld, F. G. & van Ijcken, W. F. J. NARWHAL, a primary analysis pipeline for NGS data. *Bioinformatics* **28**, 284–285 (2012).
26. Li, Y., Willer, C., Sanna, S. & calo Abecasis, G. Genotype imputation. *Annu Rev Genomics Hum Genet* **10**, 387–406 (2009).
27. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297–1303 (2010).
28. Aulchenko, Y. S., Ripke, S., Isaacs, A. & van Duijn, C. M. GenABEL: an R library for genome-wide association analysis. *Bioinformatics* **23**, 1294–1296 (2007).
29. Chen, W.-M. & Abecasis, G. R. Family-based association tests for genome-wide association scans. *Am J Hum Genet* **81**, 913–926 (2007).
30. Genome of the Netherlands Consortium. Whole-genome sequence variation, population structure and demographic history of the Dutch population. *Nat Genet* **46**, 818–825 (2014).
31. van Leeuwen, E. M. *et al.* Genome of the Netherlands population-specific imputations identify an ABCA6 variant associated with cholesterol levels. *Nat Commun* **6**, 6065 (2015).
32. Hofman, A. *et al.* The Rotterdam Study: 2014 objectives and design update. *Eur J Epidemiol* **28**, 889–926 (2013).
33. Li, H. *et al.* The sequence alignment/map format and samtools. *Bioinformatics* **25**, 2078–2079 (2009).
34. Gonzalez-Covarrubias, V. *et al.* Lipidomics of familial longevity. *Aging Cell* **12**, 426–434 (2013).
35. Demirkan, A. *et al.* Genome-wide association study identifies novel loci associated with circulating phospho- and sphingolipid concentrations. *PLoS Genet* **8**, e1002490 (2012).
36. Demirkan, A. *et al.* Plasma phosphatidylcholine and sphingomyelin concentrations are associated with depression and anxiety symptoms in a Dutch family-based lipidomics study. *J Psychiatr Res* **47**, 357–362 (2013).
37. Demirkan, A. *et al.* Insight in genome-wide association of metabolite quantitative traits by exome sequence analyses. *PLoS Genet* **11**, e1004835 (2015).
38. Li, J. & Ji, L. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity (Edinb)* **95**, 221–227 (2005).
39. Hindorff, L. A. *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A* **106**, 9362–9367 (2009).
40. Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29**, 308–311 (2001).
41. Kamburov, A. *et al.* ConsensusPathDB: toward a more complete picture of cell biology. *Nucleic Acids Res* **39**, D712–D717 (2011).
42. Magrane, M. & Consortium, U. UniProt Knowledgebase: a hub of integrated protein data. *Database (Oxford)* **2011**, bar009 (2011).
43. McKusick, V. A. *Mendelian Inheritance in Man. A Catalog of Human Genes and Genetic Disorders.* (Baltimore, Johns Hopkins University Press., 1998).
44. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet* **25**, 25–29 (2000).

45. Saier, M. H., Jr, Tran, C. V. & Barabote, R. D. TCDB: the Transporter Classification Database for membrane transport protein analyses and information. *Nucleic Acids Res* **34**, D181–D186 (2006).
46. Gasteiger, E. *et al.* ExPASy: The proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res* **31**, 3784–3788 (2003).
47. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28**, 27–30 (2000).
48. Karuna, R. *et al.* Plasma levels of sphingosine-1-phosphate and apolipoprotein m in patients with monogenic disorders of HDL metabolism. *Atherosclerosis* **219**, 855–863 (2011).
49. Maga, S. F., Kalopissis, A.-D. & Chabert, M. Äpolipoprotein A-II is a key regulatory factor of HDL metabolism as appears from studies with transgenic animals and clinical outcomes. *Biochimie* **96**, 56–66 (2014).
50. Rassias, G., Kestin, M. & Nestel, P. J. Linoleic acid lowers LDL cholesterol without a proportionate displacement of saturated fatty acid. *Eur J Clin Nutr* **45**, 315–320 (1991).
51. Schmitz, G., Langmann, T. & Heimerl, S. Role of ABCG1 and other ABCG family members in lipid metabolism. *J Lipid Res* **42**, 1513–1520 (2001).
52. Pascot, A. *et al.* Reduced HDL particle size as an additional feature of the atherogenic dyslipidemia of abdominal obesity. *J Lipid Res* **42**, 2007–2014 (2001).

PART 5

GENERAL DISCUSSION AND SUMMARY



CHAPTER 5.1

General Discussion

Although lifestyle and environmental risk factors such as body weight and nutrition play at key role in circulating lipid regulation, in humans lipid levels are in part determined by genomic variations[1-10], including rare and common coding variants and alternative processes like DNA methylation. This thesis focuses on genetic variations which cause an increased or decreased level of circulating lipid levels in the general population. I investigated the association of common and rare variants, both single associations and interactions of mutations, using various reference databases to impute unmeasured variants. This chapter summarizes the findings of this thesis, addresses methodological issues, links the findings to other related research and discusses the implications of the findings towards the understanding of the genetic background of circulating lipid levels.

The 1000 Genomes reference panel

The two projects described in **Chapter 2** have been conducted in the Cohorts for Heart and Aging Research in Genetic Epidemiology (CHARGE) consortium[11] using the 1000 Genomes (1kG) reference panel[12]. This diverse reference panel is the largest catalogue of human genetic variation available at this moment including 1,092 samples and about 39.7 million bi-allelic polymorphic markers.

In **Chapter 2.1** I used the association analysis of the variants of 59,432 individuals within the *CETP* region to fine-map the association between this gene and high-density lipoprotein cholesterol (HDL-C). The *CETP* gene is a target for drug development for dyslipidemia due to it's association with HDL-C[13-15]. The strongest reported association between HDL-C and *CETP* found by genome-wide association studies (GWAS) was rs3764261[16]. This variant is located 2,8kbp outside the *CETP* gene. The T-allele of this variant is associated with 3.47 mg/dl increase in HDL-C cholesterol. Although rs3764261 was identified by Teslovich *et al.*[16] to be the lead Single Nucleotide Polymorphism (SNP) of this region, other variants are used in clinical settings. Three of the classical variants are located in the promoter region of the *CETP* gene: -1337C/T (rs708272 or Taq1B), -971G/A and -629C/A (rs1800775) polymorphisms[17]. In this project I used the GCTA tool[18] to identify the independent variants associated with HDL-C in the *CETP* region. I discovered and replicated five variants, including an exonic variant and a common intronic deletion in an independent sample of 47,866 individuals. I validated the intronic deletion with Sanger sequencing in a single family from the Erasmus Rucphen Family study (ERF). The association to the variant reported by Teslovich *et al*[16], rs3764261, as measured by the regression coefficient was highly reduced after conditioning on the five novel variants I identified ($\beta_{unadjusted}$ = 3.179 mg/dL (*p*-value = $5.25 \cdot 10^{-509}$), $\beta_{adjusted}$ = 0.859 mg/dL (*p*-value = $9.51 \cdot 10^{-25}$)) but remained highly significant. This finding suggests that these five novel variants may partly explain the association of *CETP* with HDL-C. Moreover, these variants may have an independent effect. The deletion I identified in this study explains 35.50% of variation in the HDL-C level in a single family of the ERF study, which is much higher

than the proportion of the variance explained (14.11%) in the same family by rs3764261. This also suggests that *CETP* may have a major effect on HDL-C in a single family.

Several fine-mapping efforts have been previously published[19,20], including genotyping with the exome chip. Our imputation analysis shows that the 1kG may be an cost effective alternative to finetype regions. Further, using the Phase 1 integrated release v3 of the 1kG I was able to impute successfully a structural variation and associate this variation significantly with HDL-C in a large sample.

In **Chapter 2.2** I performed a meta-analysis of HDL-C, low-density lipoprotein cholesterol (LDL-C), total cholesterol (TC) and triglycerides (TG) genome-wide using the 1kG imputations in approximately 60.000 individuals from the same cohorts of Chapter 2.1. I replicated 88.1% of all loci described by Teslovich *et al*[16] and 43.4% all loci described by Global Lipids Genetics Consortium (GLGC) *et al*[21] despite the reduced sample size. More important, I identified and replicated five new variants: rs6457374 associated with TC, rs186696265 associated with both LDL-C and TC, rs77697917 associated with HDL-C and rs116843064 associated with TG. These variants are all within or nearby genes that can be linked biologically to lipid metabolism (Chapter 2.2).

Of the five variants, rs116843064 is the most interesting finding as it is an exonic missense variant within the *ANGPTL4* that is predicted to be damaging for the structure and the function of the protein by Polyphen2[22], MutationTaster[23] and LRT[24]. *ANGPTL4* has been associated with HDL-C before using the GWAS approach[16] and with TG before using an exome sequencing approach[25] and more recently using the GWAS approach[26]. This missense variant changes the amino acid glutamic acid into lysine at position 40 (Glu40Lys). *ANGPTL4* is associated significantly with the KEGG term fatty acid metabolism, the GO process lipid storage and the GO cellular component lipid particle (*p*-value of $1.10 \cdot 10^{-6}$, $1.31 \cdot 10^{-10}$ and $2.87 \cdot 10^{-18}$, respectively, genenetwork.nl).

rs6457374 is an intergenic variant between the genes *HLA-C* and *HLA-B* which are both associated with the KEGG term ABC transporters (*p*-value of $4.29 \cdot 10^{-5}$ and $3.84 \cdot 10^{-5}$ for *HLA-C* and *HLA-B* respectively, genenetwork.nl). ABC transporters transport a wide variety of substrates across extra- and intracellular membranes, including metabolic products, lipids and drugs. The third finding of this study is the association between HDL-C and rs77697917, an intergenic variant located between the genes *SOST* and *DUSP3*. *DUSP3* is associated with regulation and function of ChREBP in the liver (*p*-value of $3.03 \cdot 10^{-5}$, genenetwork.nl). ChREBP mediates activation of several regulatory enzymes of lipogenesis. This variant is in high linkage disequilibrium (D'=0.936) in the 1kG reference panel with rs72836561, an exonic variant within the gene *CD300LG* which is predicted to be damaging for the structure and function of the protein and has been associated with HDL-C in exome-wide association studies[27] and TG in GWAS[26] before. The fourth finding of this study was the association between rs186696265 and both LDL-C and TC. This intergenic loci is between the *LPA* (Lipoprotein, Lp(A)) gene and

the *PLG* (Plasminogen) gene. The *LPA* gene has been associated before with LDL-C and TC before[16,21].

Remarkable in Chapter 2.2 is the high number of variants that were not significantly replicated despite the similar sample size and frequencies and direction of effect within the replication phase as compared to the discovery phase. Although not extremely low, the frequency of the variants that were not replicated varied between 0.01 and 0.48. One explanation may be that the more rare variants are spurious findings but it also includes a few common variants, rs9266229, rs608736 and rs376563 with a frequency above 45%. Non-replication occurred despite a high imputation quality. Only for two variants (rs60839105 and rs151198427), the sample sizes in both the discovery and the replication phase were much lower as compared to the other variants. An explanation for the smaller sample size might be the lack of African populations in the discovery. As these variants are specific for the African population as suggested by the 1kG data (Phase 3) in which the frequency of the C-allele is 92% in African samples and 100% in the European samples for rs60839105 and the frequency of the G-allele is 86% in the African samples and 100% in the European samples for rs151198427, many studies were not informative. Imputations of cohorts with individuals of African ancestry with the African Genome Variation Project[28] might confirm the association of rs60839105 with HDL-C and rs151198427 with TC.

Both projects described in **Chapter 2** show that GWAS based on the 1kG reference panel are crucial in finding new loci and fine-map known loci for circulating lipid levels and thus help us unraveling the biological mechanism behind circulating lipid levels.

The Genome of the Netherlands reference panel

The two projects described in **Chapter 2** made use of the data of the 1kG project. Before finalizing the 1kG, there has been growing awareness that many more rare variants are population specific. The 1kG contains human variants from various populations, however, the sample size per population in this reference panel is limited. The expected power to find for rare variants which are specific for a population, is therefore low when using the 1kG reference panel. In **Chapter 3** I made use of a reference panel for the Dutch population, the Genome of the Netherlands (GoNL) reference panel with the goal to identify rare variants associated with circulating lipid levels.

The GoNL consortium enabled many researchers of the Netherlands to collaborate. A population-specific reference set for imputation was created by the consortium with the goal of identifying associations between various phenotypes and low-frequency genetic variants. To this end, 231 parent-offspring trios and 19 parent-offspring quartets of Dutch descent had their complete genome sequenced with at least 12× coverage. The strength of this reference set comes from several factors. First, the trio design which improves the haplotype quality, second, the coverage which is higher than that of the 1kG Project, and third, the sequencing

of samples from a homogeneous population. The quality of the haplotypes boosts imputation accuracy in independent samples, especially for lower frequency alleles.

The collaboration resulted in a pipeline for imputations with the GoNL reference panel (**Chapter 3.1**) which is now used by all main Dutch biobanks for imputations. **Chapter 3.2** shows that using the population-specific reference panel there was a significant improvement for rare variants (Minor Allele Frequency (MAF) between 0.05 and 0.5) compared to the 1kG. Of note is that the improved imputation accuracy is also seen for British and Italian samples. A combined reference set comprising both the 1kG and the GoNL improves the imputation of rare variants even further in both Dutch, British and Italian samples. This raises the question to what extent the reference panel for imputations should be enlarged to impute even the rarest variants with high quality.

To illustrate the advantage of a population-specific reference panel for the identification of variants associated with a particular phenotype, I HDL-C, LDL-C, TC and TG. The meta-analysis of all four traits confirmed the previously reported associations[16,21] and revealed five new associations at four loci (**Chapter 3.3**). Among the five loci is a missense variant (rs77542162) which is associated with both LDL-C and TC. This exonic variant changes within the *ABCA6* gene the amino acid cysteine into arginine and is predicted to be damaging for the structure and function of the protein. Of the five loci, three have an increased frequency in the GoNL compared with 1kG, suggesting genetic drift in the Dutch population and confirming the benefit of a population-specific reference panel. Replication in European samples from the CHARGE cohorts resulted in Bonferroni-corrected significant *p*-values, but four associations were not genome-wide significant replicated, which again confirms that these loci would not have been found by using the 1kG.

Of the five loci identified for circulating lipid levels using the GoNL reference panel, three rare variants (rs149580368, rs77542162 and rs144984216) are heavily enriched in the Dutch population. Again these variants are relatively rare (MAF between 0.02 and 0.03). The enrichment of rare variants may be due to founder effects and drift in the Netherlands. Such effects are seen for instance for rare variants in *LDL-R* and *APOB*, which are known to be highly population specific not only in the Netherlands but also elsewhere. Even in a small country as the Netherland, there are geographical differences in frequencies of rare variants. The enrichment of specific variants is highly relevant for discovery of rare variants. It has proven to be difficult to identify population-specific variant(s) associated with circulating lipid levels, because: (1) the sample sizes in a single population are usually not large enough to significantly associate rare variants with circulating lipid levels as the number of carriers are very low. The imputation of the variant in the large Dutch population cohorts boosted our power. (2) It is difficult to replicate population-specific variants. The variants identified were relatively rare but are also found outside the Netherland and studying an universal outcome as lipid levels made it possible to replicate the findings rapidly. (3) Last but not least, one

might argue that the better imputations are partly explained by the improved haplotyping in trios. In conclusion, chapter 3.3 shows that high quality population-specific reference panels are valuable to identify rare variants associated with circulating lipid levels.

New approaches to reveal variants associated with HDL-C

Imputations with reference panels, followed by a GWAS and finally a meta-analysis have been successful approaches to identify associations between traits and single variants for many traits, among which also circulating lipid levels[16,21,29-31]. Also in this thesis, this approach has revealed many new loci associated with the HDL-C, LDL-C, TC and TG. In **Chapter 4**, less commonly used genetic approaches are used; in **Chapter 4.1** I conducted a genome-wide interaction study and in **Chapter 4.2** I conducted an exome-wide association study. As is the case for GWAS, these approaches are hypothesis free, which means that we search for new variants or interactions associated with HDL-C. The statistical approach used is association, though in Chapter 4.2 this approach is combined with segregation analysis in families.

Chapter 4.1 describes the first genome-wide interaction study. Persistent evidence for interacting loci involved in lipid metabolism comes from experimental animal research in which various loci interact with each other[32]. Finding evidence for SNPxSNP interaction in humans has proven to be difficult as this has so far only been based on the common variants known to be associated with circulating lipid levels[33,34]. This motivated a hypothesis free genome-wide search for SNPxSNP interactions. However, these searches were hampered by computational time needed for testing all unique pairs of SNPs. In this thesis, I therefore used the GLIDE software package[35], which makes uses of Graphic Processing Units (GPUs) to perform linear regression for all pairs of SNPs. Although the computational issues are now solved, I was not able to significantly replicate any SNPxSNP interaction that were genome-wide significant in the discovery in the Rotterdam Study. This might be because I only included genotyped, common variants in this project and thus limited ourself to 495,508 genotyped variants. Also the sample size in our project might have been too low as I used only 2,996 individuals who did not received lipid lowering medication. The question is not only why I did not identify and replicate SNPxSNP interactions, but also how to improve the approach. First, the sample size should be enlarged, as I stated in Chapter 4.1, as a rule-of-thumb, the sample size within a genome-wide interaction studies (GWIS) should be 3 to 4 times the size of a GWAS. However, with sequence data become available, also rare variants, both intergenic, intronic and exonic, should be included. This willrequire more computational power, but will also further increase the sample size needed. Although the computational problem may be solved by improving computer resources which enable fast and parallel computing, the increase in sample size may be the major limiting factor in classical GWIS. This raises the question whether alternative approaches such as Random Forrest and machine learning will be more powerful. Up until now, also these have failed to yield replicatable findings.

The approach of an exome-wide association study (ExWAS) as described in **Chapter 4.2** is related to the GWAS approach. In the ExWAS, all exonic variants are tested for association with circulating lipid levels. The exonic variants are not imputed as is done in a classical GWAS, but sequenced. I conducted this study in the ERF. Participants (*N*=1,252) were exome sequences. I identified 18 variants, nine common (MAF ≥ 0.1) and nine rare variants (MAF < 0.1). The common variants are all located with the *CETP* region. The association between HDL-C and this gene has been extensively studied in Chapter 2.1 and this finding therefore provide a bench mark of the approach[16]. All common variants were replicated in an independent set of 85,597 individuals. I further studied the relation of the newly identified variants with other metabolites in the circulation. As expected, the *CETP* variants clustered with *ApoA1* and *ApoA2* metabolites[36,37] providing again a proof-of-principle of the cluster analyses of the new variants with the metabolomics compounds. The nine rare variants are located on multiple chromosomes within genes that have not been associated with HDL-C before. I was not able to replicate these variants in an independent sample due to the extremely low frequency of these variants and study specific discrepancies in study design and imputations. However, segregation analysis within the ERF study validated that 7 out of the 9 variants segregate in pedigrees of at least 4 generations. This suggests that these 7 Single Nucleotide Variants (SNVs) are not artifacts in the ERF cohort but do not prove a causal association to HDL-C. Of interest are the 7 rare variants (MAF < 0.1) associated with HDL-C levels and their association to various metabolomic compounds. I found two clusters of variants and metabolites. The first rare variant cluster involves variants on chromosome 11 and 21 and various metabolomics compounds of linoleic acid (18:2). This finding is in line with earlier publications on the association between linoleic acid and HDL-C[38]. The second cluster involves the association of rs35511240 on chromosome 17 with multiple large metabolites like XL-HDL-ApoA1, XL-HDL-cholesterol, XL-HDL-Free cholesterol, XL-HDL-phospholipids and XL-HDL-phospholipids. High levels of HDL-C and XL-HDL are both associated to a reduced risk of cardiovascular disease[39]. Although the findings on the rare variants are of interest, the findings await replication. This chapter highlight the problem of replicating rare variants. The lack of replication of rare variants is a major problem when more (extremely) rare variants are identified in specific families. To validate the findings in the general population asks for extremely large replication studies in which the variant is either imputed with high precision (not often the case for extremely rare variants) or assessed by direct genotyping (which is rather costly).

The end of the GWAS era?

Recent exome sequence analysis have revealed variants in *NPC1L1*[40], *LDLR*[41], *APOA5*[41] and *APOC3*[42] using a classical association approach. Of interest is the fact that these were all candidate genes that were known to be associated to lipid metabolism before, suggesting that the candidate gene study may make a comeback in the era of whole genome sequencing.

Using a hypothesis-free approach like GWAS approach, it is possible to discover new loci associated to lipid metabolism. It has been long speculated that GWAS has reached its limit in identifying variants with large effects. The efforts described in Chapter 2.1, 2.2 and 3.3 show that the era of GWAS is not over and that this method can still help us in unraveling the genetic background of circulating lipid levels, both for fine-mapping known regions and for the discovery of new loci. Enlarging sample size has resulted in new findings. The first results of GWAS of a few cohorts with circulating lipid levels were published in 2008[29-31] identifying a few common loci associated with HDL-C, LDL-C and TC. Later on in 2010, Teslovich *et al.*[16] published a genome-wide meta-analyses using more than 100,000 individuals of European ancestry which has resulted in 95 common loci for HLD-C, LDL-C, TC and TG, of which 59 show genome-wide significant associations for the first time. A follow-up meta-analysis of circulating lipid levels by the GLGC contained 188,577 individuals of various ancestry and revealed an additional 62 loci[21]. The latest published meta-analysis of circulating lipid levels was published by ENGAGE, it contained 62,166 samples of European ancestry and identified 10 new loci associated with circulating lipid levels[26]. All these projects have revealed mostly common loci associated with HLD-C, LDL-C, TC and TG and underscore that the statistical power of GWAS has not been optimal and thus many of relatively common variants (0.05 < MAF < 0.20) have not surfaced yet in the GWAS conducted to date. The reference panel used for these first lipid GWAS, the HapMap reference panel, is most likely the reason why only common variants are identified as this reference panel mainly contains predominantly common variants. Reference panels that were larger in numbers of variants as they were based on larger population, revealed new variants but are also crucial to fine-map a region identified earlier. As more and more populations are sequenced, reference panels will improve and new imputations into studies with GWAS are likely to lead to the discovery of more rare variants. How to proceed: do we still need new GWAS to be conducted in the age of next generation sequencing? Without a doubt, GWAS is still cheaper than next generation sequencing and therefore the most cost efficient way to increase the sample size. There still is an urgency to study population that are not of European ancestry, which have been overrepresented so far in GWAS. Using more samples of multiple ancestries may increase the power of findings an association of a variants with an increases frequency in a particular ancestry. GWAS is expected to find new loci by the use of improved reference panels for imputations. Imputing improved reference panels into current GWAS may be sufficient to identify new relatively rare variants. In this thesis I did not use the HapMap reference panel like the previously published meta-analysis. I used in Chapter 2.1 and 2.2 the 1kG reference panel and in Chapter 3.3 the GoNL reference panel. The improved reference panels contain much more rare variants then the HapMap reference panel. Is one reference panel better than the other? I showed that using the GoNL reference panel significantly improved the imputations of the rare variants compared to the 1kG, but both successfully mediated the identification of rare variants

associated with circulating lipid levels. However, if population specific reference panels are not available, one may also argue that including as many reference panels as possible may be the most powerful approach to impute rare variants. This approach is currently followed for imputations of the 1kG mixing samples of European, Asian and African descent. Chapter 3.2 shows that a combined reference set improves the imputation of rare variants further. Another question to be answered is what is the most powerful approach to find new rare variants in the general population: sequencing the general population, not selected for any phenotype or sequencing those with dyslipidemia, i.e., those within the extremes in the lipid distribution and imputing their variants into the large population based studies. The latter approach is likely to be most powerful in that a smaller number of persons will have to undergow sequencing and the probability of finding a predicted damaging mutation[22-24] is higher.

Future research

In 2008, Maher commented on that genetic components of common traits and diseases were not found, although that was expected once the human genome was unraveled[43]. Up to now, for HDL-C, LDL-C, TC and TG indeed, about ~25-30% of the genetic components have been unraveled. In this thesis I aimed to identify new variants and fine-map known loci. In this way, I did identify new rare, low-frequency and common variants, in new and known loci combining the GWAS, the ExWAS and imputation approaches. However, also this thesis does not resolve the case of the missing heritability of circulating lipid levels and many new loci remain to be identified. Although this thesis shows that GWAS has not reached its limits, there are also other genetic mechanisms that contribute to the total heritability. For example, structural variants, DNA alterations and gene-gene and gene-environment interactions. Investigation these might also be very helpful in unraveling the biological mechanism behind circulating lipid levels.

There is increasing interest in DNA methylation in lipid research. From an epidemiological perspective the methylation of the DNA is of interest, particularly in relation to environmental metal exposure related to lipid levels. For example, TC and HDL-C levels in very young children are associated with epigenetic metabolic programming, which may affect their vulnerability for developing cardiovascular disease (CVD) in later life[44]. Tissue-specific methylation patterns of the *APOA1/C3/A4/A5* cluster on chromosome 11q23-24 regulate liver-specific expression of the genes which are associated with blood lipid levels[45]. Within this thesis, I found several associations between circulating lipid levels and ABC transporters. DNA methylation studies indeed confirmed that DNA methylation changes at the *ABCA1* gene locus is one of the molecular mechanisms involved in HDL-C interindividual variability[46].

Besides DNA alterations, also gene-gene interactions are yet to be discovered to be associated with circulating lipid levels. In this thesis I present the first genome-wide interaction study.

Although, no gene-gene interactions were identified, I expect large meta-analysis of genome-wide interaction studies may identify gene-gene interaction associated with circulating lipid levels in humans as interacting loci have been seen in experimental animal research[32].

Although many of the genetic components of HDL-C, LDL-C, TC and TG are not found, the question has already been raised, how to translate the genetic components into a pharmaceutical solution for CVD[47]. The benchmark of GWAS is the proprotein Convertase Subtilisin/Kexin Type 9 (*PCSK9*) gene. This gene plays a crucial role in the regulation of plasma cholesterol homeostatis and fatty acid metabolism[48-51]. Development of drugs targeting *PCSK9* has resulted in a drug for patients with high cholesterol at risk of CVD[52]. This drug inhibits *PCSK9* and thereby preventing the binding of *PCSK9* to an *LDLR* which will therefore be able to remove LDL-C from the blood. This drug lowers LDL-C. There is also interest in drugs which increase HDL-C. So far, development of medication that raises HDL-C has failed. One way to prevent the costly failure of medication is to use Mendelian Randomization in the setting of therapy development. Mendelian Randomization is used to estimate the causal effect. Although Mendelian Randomizaton has its limitation, particular if pleiotropy occurs, it has yielded interesting findings. For instance GWAS challenged the view that TG are not important for CVD by showing variants in TG levels that are also relevant for CVD[53]. It has been speculated that, besides LDL-C lowering and HLD-C raising medication, the pharmaceutical companies may also try to develop TG lowering medication in the future.

As CVD is still the leading cause of mobility and the number one cause of death worldwide[54], future research may also focus on new targets for therapeutic intervention of CVD. Therapeutic intervention is most straight forward for variants with large effect. An important problem to solve in GWAS is that the functionally relevant variant has often not been discovered. This limits the use of GWAS as a way to discover new drug targets. The first requirement for the development of a new medicine, is to prioritize the findings of GWAS in terms of the likelihood of a functional effect. Within this thesis, three genes have been described in more detail: *CETP, ABCA6* and *ANGPTL4*. There are several functional variants within these genes. One of the methods to prioritize these variants, is to look at the C score. The C score is a single measure resulting from the Combined Annotation-Dependent Depletion (CADD) method[55] which objectively integrates many diverse annotations. The higher the C score of a particular variant, the more predicted to be deleterious. A C-score of greater of equal 10 indicates that the variant is predicted to be the 10% most deleterious substitutions that can occur within the human genome, a score of greater or equal 20 indicates the 1% most deleterious and so on. Table 1 shows the C scores for the variants within the *CETP, ABCA6* and *ANGPTL4* genes within the 1kG data. Although there are also variants within these genes which are not predicted to be deleterious, there do exists some variants within these gene that are predicted to be the 1% most deleterious substitutions that you can do to the human genome. Table 2 shows the top five highest C scores identified within this thesis. More investigation in these variants may results in new targets for pharmaceutical developments.

Table 1: the C scores for the variants within the *CETP, ABCA6* and *ANGPTL4* genes within the 1kG data.

Gene	Position	Variants within the 1kG	Range C score
CETP	16:56,961,850-56,983,845	312	0.001-34.000
ABCA6	17:69,078,691-69,143,262	987	0.001-22.100
ANGPTL4	19:8,363,289-8,374,375	161	0.001-26.200

Table 2: the top five highest C scores for the variants identified within this thesis within the 1kG data.

rs identifier	Gene	Position	C score
rs116843064	*ANGPTL4*	19:8,429,323	35
rs77542162	*ABCA6*	17:67,081,278	29.1
rs35511240	*ZZEF1*	17:3,937,518	10.26
rs34065661	*CETP*	16:56,995,935	9.047
rs711752	*CETP*	16:56,996,211	8.457

Conclusion

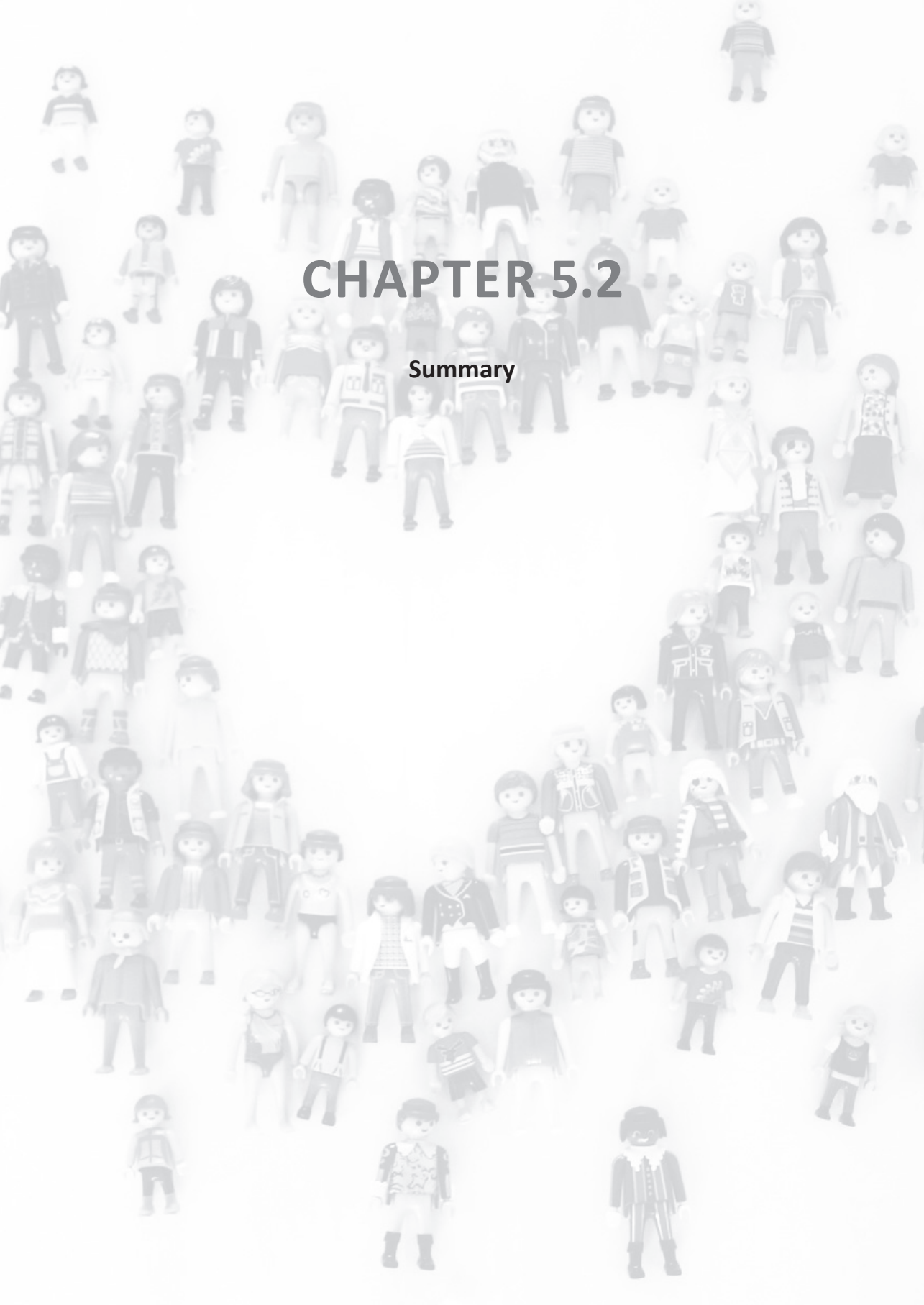
In conclusion, this thesis describes the search for differences in the human genome that cause a change in the level of circulating lipid levels. I used therefore the GWAS, GWIS and ExWAS approach. Although the GWIS did not reveal new significant SNPxSNP interactions associated with HDL-C, the project gave us some lessons for follow-up GWIS projects for the future. For the GWAS I used both the 1kG reference panel and the GoNL reference panel for imputations to improve the power of our studies which were relatively small samples for GWAS. The 1kG has not been used before for GWAS of circulating lipid levels, just like the population-specific reference panel. The fact that I found new variants in new and known regions, suggest that the era of GWAS is far from over. As the sample sizes of the projects described in this thesis are relatively small, it might be expected that there is still a lot of missing heritability to be found using GWAS. A question to be answered is whether the genes fall into novel pathways or fall into the same ones. The next frontier will be whole genome sequencing. Up until now, whole genome sequencing idenfied rare variants within candidate geness, but time will tell whether also hypothesis free approaches will work for these rare variants.

REFERENCES

1. Snieder, H., van Doornen, L. J. & Boomsma, D. I. Dissecting the genetic architecture of lipids, lipoproteins, and apolipoproteins: lessons from twin studies. *Arterioscler Thromb Vasc Biol* **19**, 2826–2834 (1999).
2. Friedlander, Y., Kark, J. D. & Stein, Y. Biological and environmental sources of variation in plasma lipids and lipoproteins: the Jerusalem Lipid Research Clinic. *Hum Hered* **36**, 143–153 (1986).
3. Souren, N. Y. *et al.* Anthropometry, carbohydrate and lipid metabolism in the East Flanders Prospective Twin Survey: heritabilities. *Diabetologia* **50**, 2107–2116 (2007).
4. Sung, J., Lee, K. & Song, Y.-M. Heritabilities of the metabolic syndrome phenotypes and related factors in Korean twins. *J Clin Endocrinol Metab* **94**, 4946–4952 (2009).
5. Almgren, P. *et al.* Heritability and familiality of type 2 diabetes and related quantitative traits in the Botnia Study. *Diabetologia* **54**, 2811–2819 (2011).
6. Vattikuti, S., Guo, J. & Chow, C. C. Heritability and genetic correlations explained by common SNPs for metabolic syndrome traits. *PLoS Genet* **8**, e1002637 (2012).
7. Zhou, X., Carbonetto, P. & Stephens, M. Polygenic modeling with bayesian sparse linear mixed models. *PLoS Genet* **9**, e1003264 (2013).
8. Browning, S. R. & Browning, B. L. Identity-by-descent-based heritability analysis in the Northern Finland Birth Cohort. *Hum Genet* **132**, 129–138 (2013).
9. Morrison, A. C. *et al.* Whole-genome sequence-based analysis of high-density lipoprotein cholesterol. *Nat Genet* **45**, 899–901 (2013).
10. Namboodiri, K. K. *et al.* The Collaborative Lipid Research Clinics Family Study: biological and cultural determinants of familial resemblance for plasma lipids and lipoproteins. *Genet Epidemiol* **2**, 227–254 (1985).
11. Psaty, B. M. *et al.* Cohorts for heart and aging research in genomic epidemiology (CHARGE) consortium: Design of prospective meta-analyses of genome-wide association studies from 5 cohorts. *Circ Cardiovasc Genet* **2**, 73–80 (2009).
12. 1000 Genomes Project Consortium *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
13. Briand, F. *et al.* Anacetrapib and dalcetrapib differentially alters HDL metabolism and macrophage-to-feces reverse cholesterol transport at similar levels of CETP inhibition in hamsters. *Eur J Pharmacol* **740**, 135–143 (2014).
14. Siebel, A. L. *et al.* Effects of high-density lipoprotein elevation with cholesteryl ester transfer protein inhibition on insulin secretion. *Circ Res* **113**, 167–175 (2013).
15. Remaley, A. T., Norata, G. D. & Catapano, A. L. Novel concepts in HDL pharmacology. *Cardiovasc Res* **103**, 423–428 (2014).
16. Teslovich, T. M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707–713 (2010).
17. Le Goff, W. *et al.* A novel cholesteryl ester transfer protein promoter polymorphism (-971g/a) associated with plasma high-density lipoprotein cholesterol levels. interaction with the taqib and -629c/a polymorphisms. *Atherosclerosis* **161**, 269–279 (2002).
18. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* **88**, 76–82 (2011).
19. Wu, Y. *et al.* Trans-ethnic fine-mapping of lipid loci identifies population-specific signals and allelic heterogeneity that increases the trait variance explained. *PLoS Genet* **9**, e1003379 (2013).
20. Sanna, S. *et al.* Fine mapping of five loci associated with low-density lipoprotein cholesterol detects variants that double the explained heritability. *PLoS Genet* **7**, e1002198 (2011).

21. Global Lipids Genetics Consortium *et al.* Discovery and refinement of loci associated with lipid levels. *Nat Genet* **45**, 1274–1283 (2013).
22. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nat Methods* **7**, 248–249 (2010).
23. Schwarz, J. M., Rödelsperger, C., Schuelke, M. & Seelow, D. MutationTaster evaluates disease-causing potential of sequence alterations. *Nat Methods* **7**, 575–576 (2010).
24. Chun, S. & Fay, J. C. Identification of deleterious mutations within three human genomes. *Genome Res* **19**, 1553–1561 (2009).
25. Romeo, S. *et al.* Population-based resequencing of angptl4 uncovers variations that reduce triglycerides and increase hdl. *Nat Genet* **39**, 513–516 (2007).
26. Surakka, I. *et al.* The impact of low-frequency and rare variants on lipid levels. *Nat Genet* **47**, 589–597 (2015).
27. Albrechtsen, A. *et al.* Exome sequencing-driven discovery of coding polymorphisms associated with common metabolic phenotypes. *Diabetologia* **56**, 298–310 (2013).
28. Gurdasani, D. *et al.* The african genome variation project shapes medical genetics in africa. *Nature* **517**, 327–332 (2015).
29. Kathiresan, S. *et al.* Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. *Nat Genet* **40**, 189–197 (2008).
30. Aulchenko, Y. S. *et al.* Loci influencing lipid levels and coronary heart disease risk in 16 European population cohorts. *Nat Genet* **41**, 47–55 (2009).
31. Willer, C. J. *et al.* Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nat Genet* **40**, 161–169 (2008).
32. Brockmann, G. A. *et al.* Genetic control of lipids in the mouse cross DU6i x DBA/2. *Mamm Genome* **18**, 757–766 (2007).
33. Ma, L. *et al.* Knowledge-driven analysis identifies a gene-gene interaction affecting high-density lipoprotein cholesterol levels in multi-ethnic populations. *PLoS Genet* **8**, e1002714 (2012).
34. Turner, S. D. *et al.* Knowledge-driven multi-locus analysis reveals gene-gene interactions influencing HDL cholesterol level in two independent EMR-linked biobanks. *PLoS One* **6**, e19586 (2011).
35. Kam-Thong, T. *et al.* GLIDE: GPU-based linear regression for detection of epistasis. *Hum Hered* **73**, 220–236 (2012).
36. Maga, S. F., Kalopissis, A.-D. & Chabert, M. Äpolipoprotein a-ii is a key regulatory factor of hdl metabolism as appears from studies with transgenic animals and clinical outcomes. *Biochimie* **96**, 56–66 (2014).
37. Karuna, R. *et al.* Plasma levels of sphingosine-1-phosphate and apolipoprotein m in patients with monogenic disorders of HDL metabolism. *Atherosclerosis* **219**, 855–863 (2011).
38. Rassias, G., Kestin, M. & Nestel, P. J. Linoleic acid lowers LDL cholesterol without a proportionate displacement of saturated fatty acid. *Eur J Clin Nutr* **45**, 315–320 (1991).
39. Pascot, A. *et al.* Reduced HDL particle size as an additional feature of the atherogenic dyslipidemia of abdominal obesity. *J Lipid Res* **42**, 2007–2014 (2001).
40. Myocardial Infarction Genetics Consortium Investigators *et al.* Inactivating mutations in NPC1L1 and protection from coronary heart disease. *N Engl J Med* **371**, 2072–2082 (2014).
41. Do, R. *et al.* Exome sequencing identifies rare LDLR and APOA5 alleles conferring risk for myocardial infarction. *Nature* **518**, 102–106 (2015).
42. The TG and HDL Working Group of the Exome Sequencing Project, National Heart, Lung, and Blood Institute *et al.* Loss-of-function mutations in APOC3, triglycerides, and coronary disease. *N Engl J Med* **371**, 22–31 (2014).

43. Maher, B. Personal genomes: The case of the missing heritability. *Nature* **456**, 18–21 (2008).
44. Wijnands, K. P. J., Obermann-Borst, S. A. & Steegers-Theunissen, R. P. M. Early life lipid profile and metabolic programming in very young children. *Nutr Metab Cardiovasc Dis* (2015).
45. Guardiola, M. *et al.* Tissue-specific DNA methylation profiles regulate liver-specific expression of the APOA1/C3/A4/A5 cluster and can be manipulated with demethylating agents on intestinal cells. *Atherosclerosis* **237**, 528–535 (2014).
46. Guay, S. P. *et al.* DNA methylation variations at CETP and LPL gene promoter loci: new molecular biomarkers associated with blood lipid profile variability. *Atherosclerosis* **228**, 413–420 (2013).
47. Christoffersen, M. & Tybjærg-Hansen, A. Novel genes in LDL metabolism□a comprehensive overview. *Curr Opin Lipidol* **26**, 179–187 (2015).
48. Huang, C.-C. *et al.* Longitudinal association of PCSK9 sequence variations with low-density lipoprotein cholesterol levels: the Coronary Artery Risk Development in Young Adults Study. *Circ Cardiovasc Genet* **2**, 354–361 (2009).
49. Folsom, A. R., Peacock, J. M., Boerwinkle, E. & Atherosclerosis Risk in Communities (A.R.I.C.) Study Investigators. Variation in PCSK9, low LDL cholesterol, and risk of peripheral arterial disease. *Atherosclerosis* **202**, 211–215 (2009).
50. Hallman, D. M., Srinivasan, S. R., Chen, W., Boerwinkle, E. & Berenson, G. S. Relation of PCSK9 mutations to serum low-density lipoprotein cholesterol in childhood and adulthood (from The Bogalusa Heart Study). *Am J Cardiol* **100**, 69–72 (2007).
51. Cohen, J. C., Boerwinkle, E., Mosley, T. H., Jr & Hobbs, H. H. Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. *N Engl J Med* **354**, 1264–1272 (2006).
52. Marian, A. J. Clinical significance of single nucleotide polymorphisms in PCSK9. *Curr Atheroscler Rep* **9**, 175–176 (2007).
53. Nordestgaard, B. G. & Varbo, A. Triglycerides and cardiovascular disease. *Lancet* **384**, 626–635 (2014).
54. Tóth, P. P., Potter, D. & Ming, E. E. Prevalence of lipid abnormalities in the United States: the National Health and Nutrition Examination Survey 2003-2006. *J Clin Lipidol* **6**, 325–330 (2012).
55. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* **46**, 310–315 (2014).



CHAPTER 5.2

Summary

Cardiovascular disease (CVD) are the leading cause of morbidity and the number one cause of death worldwide. Risk factors for CVD are four types of circulating lipid levels: high-density lipoprotein cholesterol (HDL-C), low-density lipoprotein cholesterol (LDL-C), total cholesterol (TC) and triglycerides (TG). These four types of circulating lipid levels are highly heritable. Despite the large number of research that has been performed about circulating lipid levels, the genetic variations driving this heritability are still largely unknown. Most genetic variations discovered today come from genome-wide association studies (GWAS), however, these variations are mostly common and the effect of these variations on circulating lipid levels are small.

In this thesis I aimed to identify new variants associated with circulating lipid levels. Therefore I used several genetic epidemiological approaches to dissect the complex nature of circulating lipid levels: GWAS, genome-wide interaction studies (GWIS) and exome-wide association studies (EWAS). The GWAS was applied on genomic data of individuals imputed to both the Genome of the Netherlands reference panel and the 1000 Genomes reference panel. These individuals are part of several population-based and family-based cohorts. The hypothesis free GWIS was applied on the Rotterdam Study, an ongoing prospective population-based cohort study and the EWAS was applied on the Erasmus Rucphen Family study, an isolated family-based population.

In Chapter 2 individuals of several cohorts of the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) consortium imputed with the 1000 Genomes reference panel were used. I first run a GWAS for HDL-C in the *CETP*-region to identify the causal variant for the association between the *CETP* gene and HLD-C in Chapter 2.1. I identified and replicated five variants, including an exonic variant and a common intronic deletion. These variant explain most of the effect of a previously reported variant within this region. In Chapter 2.2 the same individuals were used, but run a GWAS for HDL-C, LDL-C, TC and TG genome-wide. This resulted in the discovery and replication of new variants; rs6457374 for TC, rs77697917 for HDL-C, rs116843064 for TG and rs186696265 for both LDL-C and TC.

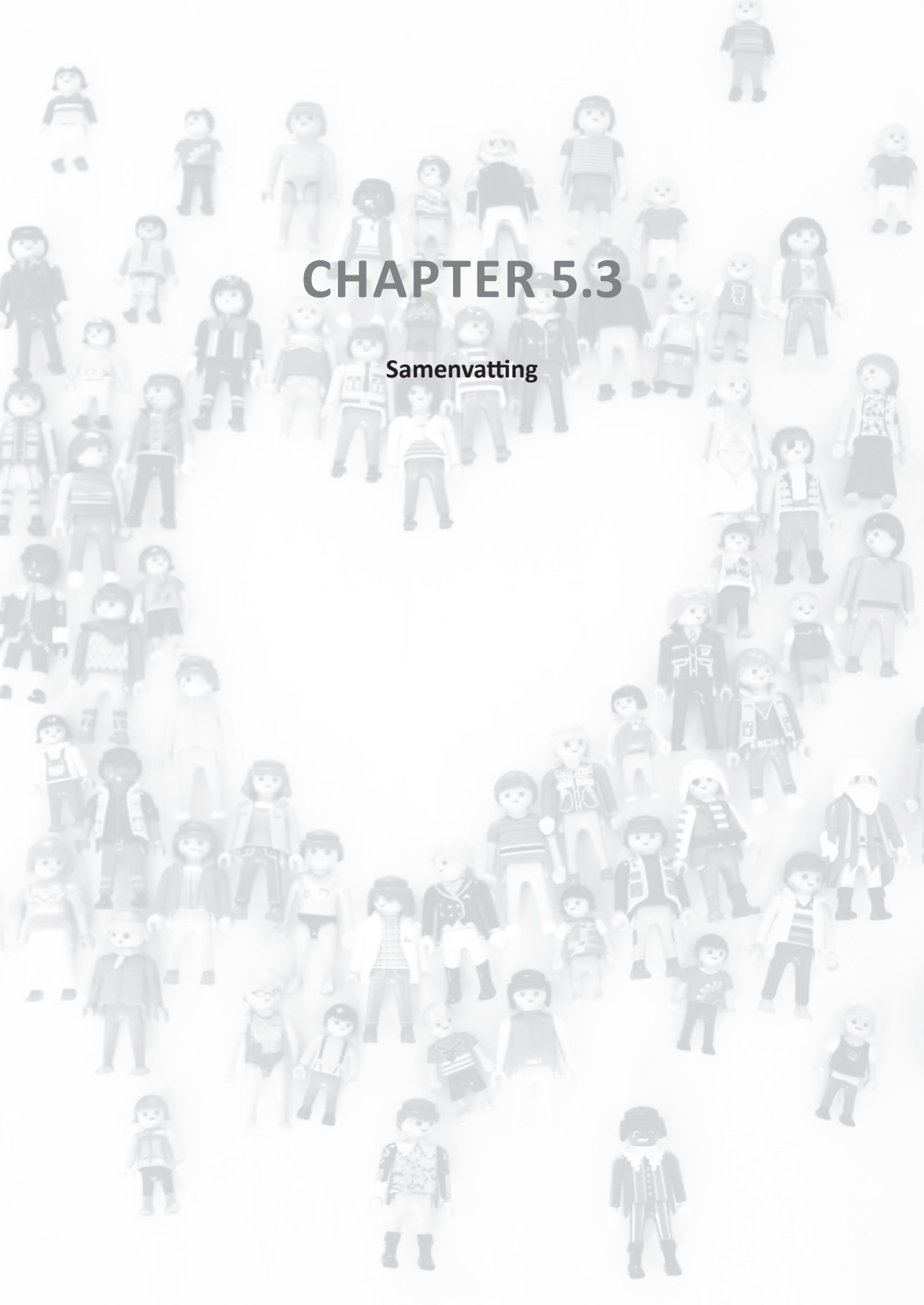
In Chapter 3 the Genome of the Netherlands reference panel was introduced. I first present a protocol for imputations with this population-specific reference panel in Chapter 3.1 and compared this reference panel with the 1000 Genomes reference panel in Chapter 3.2. This comparison shows that using the GoNL reference panel there was a significant improvement in imputation accuracy for rare variants, not only for Dutch samples, but also for British and Italian samples. The GoNL reference panel in Chapter 3.3 was used to impute the nine largest Dutch biobanks followed by a GWAS per cohort with HDL-C, LDL-C, TC and TG. The meta-analysis of all four traits revealed five new associations at four loci among which a missense

variant (rs77542162) within the *ABCA6* gene which is predicted to be damaging for the structure and function of the protein.

In Chapter 4 a less commonly used genetic approaches was applied to search hypothesis-free for new variants or interactions between variants associated with HDL-C. To this end, I performed the, to my knowledge, first GWIS (Chapter 4.1). As these were hampered by computational time need for testing all unique pairs of SNPs, I used the GLIDE software package which makes uses of Graphic Processing Units (GPUs) to perform linear regression for all pairs of SNP. Although the computational issues are now solved, I was not able to identify and significantly replicate any SNPxSNP interaction. The lack of replication might be because of the sample size and because I only included genotyped, common variants in this project.

The second less commonly used genetic approach I applied was an exome-wide analysis study in an isolated population, the ERF study. Testing all exonic SNPs for an association with HDL-C resulted in identification of 18 variants, nine common variants within the *CETP* region and nine rare variants located on multiple chromosomes within genes that have not been associated with HDL-C before. I replicated the common variants in an independent sample of 85,597 individuals and confirmed seven of the rare variants by segregation analysis within pedigrees of the ERF study of at least 4 generations.

To summarize, in this thesis I first applied the commonly used genetic approach, GWAS with two new reference panels: the 1000 Genomes reference panel and the Genome of the Netherlands reference panel. This enabled me to fine-map a known region and to discover new regions associated with circulating lipid levels. This showed me that the era of GWAS is not over as using a larger reference panel provided more information about the genetic background of circulating lipid levels. Secondly, I applied two less commonly used genetic approaches; GWIS and EWAS. Although the first was not successful in identification of common SNPxSNP interactions, the second resulted in nine new variants associated with HDL-C. The conclusion of the work as described in this thesis is that I was able to reveal some of the genetic variations driving the heritability of circulating lipid levels, but there is still much more work to do.



CHAPTER 5.3

Samenvatting

Hart- en vaatziekten vormen wereldwijd de belangrijkste oorzaak van morbiditeit en mortaliteit. Risicofactoren voor hart- en vaatziekten zijn vier typen circulerende lipide levels: high-density lipoproteïne cholesterol (HLD-C), low-density lipoproteïne cholesterol (LDL-C), totaal cholesterol (TC) en triglyceriden (TG). Deze vier typen circulerende lipide levels zijn in hoge mate erfelijk. Ondanks het vele onderzoek dat is gedaan naar circulerende lipide levels, is de meeste genetische variatie die de hoge erfelijkheid bepaald, nog steeds grotendeels onbekend. De meeste genetische variatie die tot op heden bekend is, is ontdekt in genoomwijde associatiestudies (GWAS), echter, deze variaties zijn voornamelijk veel voorkomende variaties en de effecten van deze variaties op circulerende lipide levels zijn erg klein.

In dit proefschrift streef ik ernaar nieuwe variaties te ontdekken die geassocieerd kunnen worden met circulerende lipide levels. Daarvoor gebruikte ik meerdere genetische epidemiologische benaderingen: GWAS, genoomwijde interactie studies (GWIS) en exoomwijde associatie studies (ExWAS). GWAS heb ik toegepast op genoom-data van individuals die geimputeerd zijn met ofwel het Genoom van Nederland referentie panel ofwel het 1000 Genomes referentie panel. Deze individuen zijn deels afkomstig uit populatie gebaseerde cohorten en deels afkomstig uit familie gebaseerde cohorten. De hypothese vrije GWIS benadering was enkel toegepast op de Rotterdam Study, een lopende studie en de ExWAS benadering werd toegepast op een geisoleerde familie gebaseerde populatie.

In hoofdstuk 2 heb ik meerdere individuen gebruikt uit verschillende cohorts die behoorde tot the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) consortium. Deze individuen zijn allen geimputeerd met het 1000 Genomes referentie panel. Allereest heb ik in hoofdstuk 2.1 een GWAS uitgevoerd voor HDL-C in de *CETP*-regio om de variant te identificeren die de associatie tussen het *CETP* gen en HDL-C veroorzaakt. Hierbij zijn vijf varianten gevonden waaronder een exonische variant en een veelvoorkomende intronische deletie. Deze vijf varianten verklaren grotendeels het effect van een variant binnen deze regio die voorheen is gepubliceerd. In hoofdstuk 2.2 heb ik dezelfde individuen gebruikt maar dit keer een GWAS uitgevoerd voor HDL-C, LDL-C, TC en TG genoomwijd. Dit resulteerde in de ontdekking en replicatie van drie nieuwe varianten: rs6457374 voor TC, rs77697917 voor HDL-C, rs116843064 voor TG en rs186696265 voor zowel LDL-C als TC.

In hoofdstuk 3 besprak ik het Genoom van Nederland referentie panel. Allereerst beschreef hoofdstuk 3.1 een protocol om de genotype data van individuen te imputeren met deze populatie-specifieke referentie panel. Ook vergeleken we dit referentie panel met het 1000 Genomes referentie panel in hoofdstuk 3.2. Deze vergelijking liet zien dat wanneer de GoNL referentie panel gebruikt word, er een significante verbetering is in de imputatie

kwaliteit voor zeldzame varianten in zowel Nederlandse als Italiaanse en Britse individuen. In hoofdstuk 3.3 laat ik een GWAS zien voor HDL-C, LDL-C, TC en TG waarbij de individuen van de negen grootste Nederlandse biobanken geimputeerd zijn met het Genoom van Nederland referentie panel. De meta-analyse van alle vier de phenotypes resulteerde in vijf significante associaties op vier posities waaronder een missense variant (rs77542162) binnen het *ABCA6* gen van welke voorspelt is dat deze schadelijk is voor de structuur en de functie van het eiwit dat door het gen geproduceerd wordt.

In hoofdstuk 4 hebben we minder voorkomende genetische benaderingen toegepast om hypothese-vrij te zoeken naar nieuwe variaties of interaction die met HDL-C geassocieerd kunnen worden. Ik voerde de eerste GWIS voor HDL-C uit in hoofdstuk 4.1. Deze methode werd tot dus ver tegengehouden door de computationele tijd die ervoor nodig is om alle unieke SNP paren te testen. Daarom gebruikte ik de GLIDE software die gebruik maakt van Graphic Processing Units (GPUs) om lineare regressies voor alle SNP paren te testen. Ondanks dat de computationele problemen nu konden worden opgelost, werden de bevindingen niet significant gerepliceerd. De oorzaak hiervan is mogelijk het aantal individuen of de veelvoorkomende varianten die gebruikt zijn in de studie.

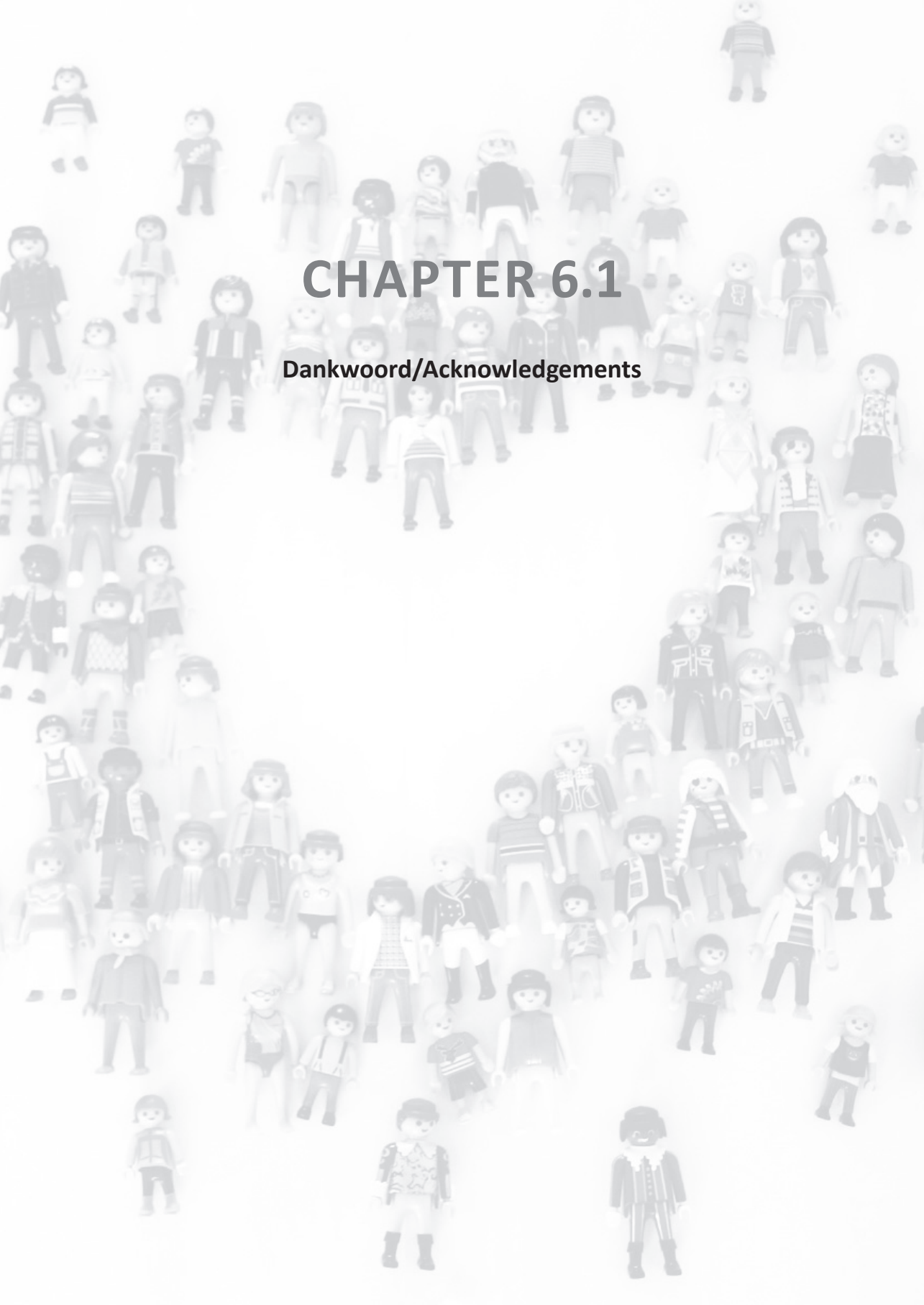
The tweede minder voorkomende genetische benadering die ik heb toegepast was een exoomwijde assocatiestudie in een geisoleerde populatie, de ERF studie. Door alle exonische SNPs te testen voor een associatie met HDL-C, identificeerde ik 19 varianten waaronder negen veelvoorkomende varianten binnen het *CETP* gen en negen zeldzame varianten op verschillende chromosomen binnen genen die nog niet met HDL-C geassocieerd zijn. We repliceerde de veelvoorkomende varianten in een onafhankelijke set van 85,597 individuen en bevestigden zeven van de negen zeldzame varianten aan de hand van seggregatie analyse binnen stambomen van de ERF studie van tenminste vier generaties.

Samenvattend, in dit proefschrift heb ik een veelvoorkomende genetisch epidemiologische aanpak toegepast, namelijk een GWAS met twee verschillende referentie panels: het 1000 Genomes referentie panel en het Genoom van Nederland referentie panel. Hierdoor hebben we bekende regio's nader kunnen besturen en nieuwe regio's kunnen identificeren die geassocieerd zijn met circulerende lipide levels. Dit laat zien dat het tijdperk van GWAS nog niet over is aangezien grotere referentie panels ons meer informatie heeft opgeleverd over de genetische achtergrond van circulerende lipide levels. Daarnaast hebben we twee minder veelvoorkomende genetische benaderingen toegepast; GWIS en ExWAS. Ook al was de eerste niet succesvol in het identificeren van veelvoorkomende SNPxSNP interacties, de tweede heeft wel zeven nieuwe variaten geassocieerd met HDL-C opgeleverd. De conclusie van het werk beschreven in dit proefschrift is dan ook dat ik een deel van de genetische

variatie die leidt tot de hoge erfelijkheid van circulerende lipide levels heb kunnen opsporen maar dat er nog een hoop werk te verzetten is.

PART 6

EPILOGUE



CHAPTER 6.1

Dankwoord/Acknowledgements

Zo, na 5 jaar dan eindelijk toe aan het belangrijkste van alles: het dankwoord. Want ook al staat mijn naam voorop dit boek, een proefschrift schrijf je niet alleen en ook kun je niet alleen wetenschap uitvoeren. Daarom zijn er een heleboel mensen zowel in Nederland als erbuiten, zowel op de werkplek als erbuiten die ik moet bedanken voor hun bijdrage. En mocht ik iemand vergeten: sorry, sorry en nog eens sorry!

Allereerst mijn promotor, professor Cornelia van Duijn. Beste Cock, samen hebben we gestreden, tegen elkaar en met elkaar, met dit boekwerk als resultaat. Je hebt altijd een doel voor ogen, voor de afdeling en voor al je werknemers. En ook ik heb daarvan mogen profiteren. Je hebt mij kansen gegeven om op nationaal en internationaal niveau samen te werken, om congressen bij te wonen en om te mogen proeven van de wetenschappelijke wereld. Daar vroeg je wel wat voor terug, je zei onlangs dat je zelf veeleisend bent en dat klopt! Ondanks dat was je er ook in de minder vrolijke tijden, je hield rekening met de soms erg moeilijke omstandigheden thuis en op je eigen wijze vroeg je me hoe het met me ging en dat heb ik erg enorm gewaardeerd. Dank je wel!

Daarnaast wil ik ook graag alle leden van de kleine commissie bedanken: professor Betram Müller-Myhsok, professor Adrienne Cupples en professor Eric Sijbrands. Dear Bertram, it was a pleasure working with you on the SNPxSNP interactions, unfortunately did not resulted in any significant replicated findings, but it did give me the oppurtunity to work together with you and to learn a lot from you. Dear Adrienne, thank you for leading the Lipids Working Group of the CHARGE consortium, thank you for sharing so much knowledge about lipids and last but not least, thank you for giving me an amazing experience in your lab. Beste Eric, ook u wil ik graag bedanken voor het lezen van mijn proefschrift en uw bijdrage aan de meeste van mijn artikelen. Uw kennis van lipiden was zeer waardevol!
Ook alle leden van de grote commissie wil ik graag hartelijk bedanken: professor Eline Slagboom, professor Ko Willems- van Dijk, professor Oscar Franco, professort Robert Hofstra en professor Andre Uitterlinden. Beste Eline, hartelijk dank voor de fijne samenwerking de afgelopen jaren binnen het Genoom van Nederland consortium. Beste Ko, uw suggesties voor hoofdstuk 4.2 zijn zeer leerzaam geweest, dank u wel dat u de tijd nam mijn manuscript te lezen en mijn proefschrift te beoordelen. Dear Oscar, thank you for being part of the committee. Beste Robert, ook u wil ik graag hartelijk bedanken voor het lezen van mijn proefschrift. Beste Andre, dank u wel voor alle samenwerking van de afgelopen jaren, ik hoop van harte dat voortaan iedereen het asfalt optimaler zal benutten zodat u zich niet meer hoeft te storen aan ongebruikte asfalt in Nederland ;).

Ook professor Ben Oostra wil ik bedanken. Beste Ben, toen ik mijn PhD baan niet meer zag zitten en het liefste wilde stoppen, liet jij me zien dat het eigenlijk allemaal zo slecht niet was.

Je sprak me weer moed in en gaf me tips om door te gaan. Die zijn me altijd bij gebleven en daar ben ik je dan ook erg dankbaar voor. Ik zie een grote glimlach als u praat over uw “nieuwe” leven, genieten met een grote G van reizen en de kleinkinderen. Geniet ervan, het is u gegund!

Beste Lennart, je hebt mij de eerste jaren van mijn PhD begeleidt. Je leerde me op alle (kleine) details te letten, in papers, in presentaties en tijdens het programmeren. Samen zijn we op meerdere plekken geweest voor congressen en meeting: Nürnberg, Lunteren, München, Boston en New York. Met een grote glimlach denk ik terug aan het terrasje in New York waar we iets te veel cocktails dronken om de zon achter de skyline van New York te zien zakken! Helaas kon je me de laatste jaren niet meer begeleiden maar ben blij dat we toch in contact bleven en ik altijd op je kon blijven rekenen. Bedankt voor je begeleiding, je geduld en de gezelligheid! Dear Aaron, I also have to thank you for serving as a second supervisor. Unfortunately you have only been my supervisor for a few months, but I would like to let you know that your knowledge about lipids and experience within this field, have really moved me forward. I appreciate our conversations, also the ones about science ;). Your enthusiasm when you start talking about Noah, is priceless!

Wat ben ik een bofkont om drie paranimfen te hebben, om dus drie meiden te mogen kennen die er altijd voor me zijn geweest de laatste jaren! Lieve Sara, dank je wel voor zoveel, het was leeg in de trein toe je naar Cambridge verhuisde, niemand om mandarijntjes mee te eten, niemand om kritisch maar rechtvaardig de dag mee door te nemen. Je huilde mee toen ik slecht nieuws over mijn moeder kreeg en juichte van harte mee toen ik zwanger van Lotte bleek te zijn. Binnenkort in Cambridge maar eens de nabeschouwing van al die jaren genetische epidemiologie. Lieve Lieke, mijn langste en kleinste vriendinnetje: na jaren van alleen maar lol en feestjes in Leiden, werd t voor ons “serieus business”, allebei een PhD, Amsterdam en Rotterdam. Jullie huis, mijn mama, Lotte, wat er ook allemaal voor zaken waren, wij bleven geregeld samen eten (en drinken), gewoon even ontspannen en alle frustraties eruit gooien! Nu we beide onze PhD afgerond hebben, hoop ik dat er weer nieuwe frustraties komen, zodat we nog vaak samen een excuus hebben om uit eten te gaan. Lieve Nikkie, alhoewel we nu niet meer wekelijks elkaar spreken, weet ik dat jij er altijd voor me bent, altijd voor mij klaar staat en dat is echt heel fijn! Geen feestje is een success zonder jouw hulp en hopelijk als alle feestjes straks gevierd zijn, is er vast weer meer tijd voor winkelen, theetjes drinken en uit eten gaan. Lieve Nikkie, Sara en Lieke, drie vriendinnen die ook een PhD doen/gedaan hebben, die de tegenslagen kennen, die de frustraties kennen en met je meeleven in goede en slechte tijden zowel op t werk als thuis, is enorm waardevol!

Stress has nothing to do with how many hours you work, and everything to do with how you feel during those hours. Dear colleagues from genetic epidemiology, dear Adriana, Andrea, Andy, Annelies, Ashley, Ayse, Bernadette, Carla, Claudia, Constanza, Dina, Dream, Elena, Elena, Eline, Elza, Fizzah, Ivana, Jeannette, Linda, Maaike, Maarten, Maksim, Najaf, Natalia, Petra, Revanius, Robert, Sara, Shazad, Sofia, Sven and Yurii: thank you for being my colleagues, thank you for the nice moments. Sushi on a boat, not in a boat ;). Dear Adriana, no matter how busy you are, you always have time for anyone else. Thank you for your interest in my work and life. I really hope that the fishes will have the phenotype you want them to have and that your thesis will be “een eitje” and that “je niet uit het raam hoeft te springen”! Dear Dina, thank you for the nice conversations, you are a hard working person and I would like to give you once again the advice: take a break some now and then ;). Beste Sven, je positiviteit was vanaf het begin opvallend, dat je die maar mag behouden en iets van dat chaotische mag verliezen. Succes met de laatste loodjes richting thesis. Beste Carla en Linda, al enige tijd weg van de afdeling, toch heb ik een hoop van jullie geleerd, mijn dank daarvoor. Succes op jullie nieuwe plekken. Beste Ashley, wat fijn dat jij tegenover me kwam te zitten, een heel fijn lief kletsmaatje! Ik weet zeker dat jij er ook wel komt!

Daarnaast wil ik ook de afdeling Epidemiologie van het Erasmus MC bedanken. Beste professor Albert Hofman, de gestructureerde opzet van de Rotterdam Studie heeft ook mij de mogelijkheid geboden om deel te nemen in consortia en om onderzoek te doen naar lipiden is zoveel individuen. Drie dames van de Epidemiologie verdienen ook een plekje in het dankwoord: Gabriëlle, Virginie en Henriëtte. Dank jullie wel voor de gezelligheid, tussen het werk door, tijdens congressen en tijdens cursussen. Ik mis de snoeppot! Also some individuals of the Internal Medicine of the Erasmus MC should be thanked: Carolina, Karol and Fernando.

Mijn onderzoek zou niet hebben plaatsgevonden als zoveel mensen vrijwillig hadden bijgedragen aan de Rotterdam Studie en de Erasmus Rucphen Familie Studie, maar ook aan alle andere studies waar ik mee heb samengewerkt. Ook al weet ik niet wie jullie zijn, jullie deelname is van enorme waarde!

Alle deelnemers van het GoNL consortium wil ik hierbij ook bedanken. Beste Cisca, Dorret, Gert-Jan, Eline, Morris en Paul: dank jullie wel dat jullie dit consortium hebben opgezet en geleidt hebben naar verschillende publicaties. Daarnaast gaat mijn dank ook uit naar de andere leden van het consortium waaronder Androniki, Freerk, Jeroen, Jessica, Jouke-Jan, Kai, Laurent, Martijn, Mathijs, Patrick, Pieter en Sara. Dank jullie wel voor de samenwerking, voor de input in mijn werk, dank jullie wel voor de kritische vragen. De samenkomsten in Utrecht waren altijd weer bijzonder en zorgden voor een hoop nieuwe wetenschappelijke vraagstukken. Menig één van jullie ben ik tegen gekomen bij conferenties en dat was altijd

weer gezellig. Een speciaal word van dank gaat uit naar Cisca en Paul. Beste Cisca, graag wil ik je bedanken voor je steun voor mijn werk binnen het consortium, je nam altijd de tijd om mijn werk te beoordelen en hierdoor heb ik me altijd enorm gesteund gevoeld in mijn werk voor dit consortium. Beste Paul, ik wil je graag bedanken voor al je kritische vragen die je mij de afgelopen jaren over mijn werk binnen het Genoom van Nederland stelde. Je liet me hierdoor nog meer nadenken over mijn eigen projecten en dat heeft toch mooie manuscripten opgeleverd.

It has also been a privilege to work within the CHARGE consortium, in particular the Lipids Working Group. Dear Prof. Bruce Psaty, your comments and suggestions for my analysis plan, papers and grant proposal have been very helpful and I learned very much from this, thank you. Dear CHARGE RSC, thank you for giving me the opportunity to go to Boston for two month, a fantastic experience. Dear Gina, thank you for organizing the Lipids Working Group and your input in my projects. Dear Jennifer, thank you for always replying so fast on emails, thank you for the input in my projects and thanks for the nice collaboration during the CHARGE commons project. Thank you for showing me around in Boston and Framingham, you have been a great hostess!

My papers would not have existed without the help of so many co-authors. I could fill a whole chapter with the names of my co-authors: thank you all for the collaboration, the suggestions for my papers and for keeping fingers crossed after submission.

Collega's van de afdeling oog-epi, na een half jaar thuis met de kinderen, is het heerlijk om weer uitgedaagd te worden, om weer aan het werk te zijn. Dank jullie wel voor jullie warme welkom! Op naar de toekomst met mooie projecten!

Maar zonder vrienden buiten het werk zou het leven ook maar saai zijn! Lieve Jan, je opgewektheid is zo inspirerend, gewoon doen waar je zin in hebt, dat zou ik ook meer moeten doen! Lieve Lisa, je bent een lieverd, altijd een luisterend oor. Jij weet als geen ander wie ik op dit moment heel erg mis, hoe het voelt en dat ik dat met je kan delen, is heel waardevol. Dank je wel dat we altijd welkom zijn in Almere/Arnhem, dikke kus, ook voor (al) je mannen. Lieve Bjel, al je apjes met zoveel uitroeptekens maakt me altijd glimlachen, zoveel interesse als je altijd toont is echt heel fijn. Ook al is Maastricht niet om de hoek, je lijkt altijd dichtbij! Lieve mama's uit Gouda: niet altijd kan wetenschap de vragen over Lotte en Gijs beantwoorden, jullie wel! Samen zwemmen, naar de kinderboerderij, naar de speeltuin, of een kopje thee en dan even al die verhalen over Lotte en Gijs kunnen vertellen, alle vragen kunnen stellen, zo fijn! Niettemin wil ik natuurlijk ook Pim's mannen bedanken, lieve Hein, Johan, Martijn, Rob, Robert, dank je wel dat jullie er altijd zijn voor Pim, hem de nodige afleiding van thuis

geven. Dan neem ik wel op de koop toe hoe hij vaak weer afgeleverd word ;). Maar ook Myrthe, Rens, Remco, Loes, Arno, Silvie: even geen boekje schrijven, gewoon een drankje, een kletspraatje, even ontspanning.

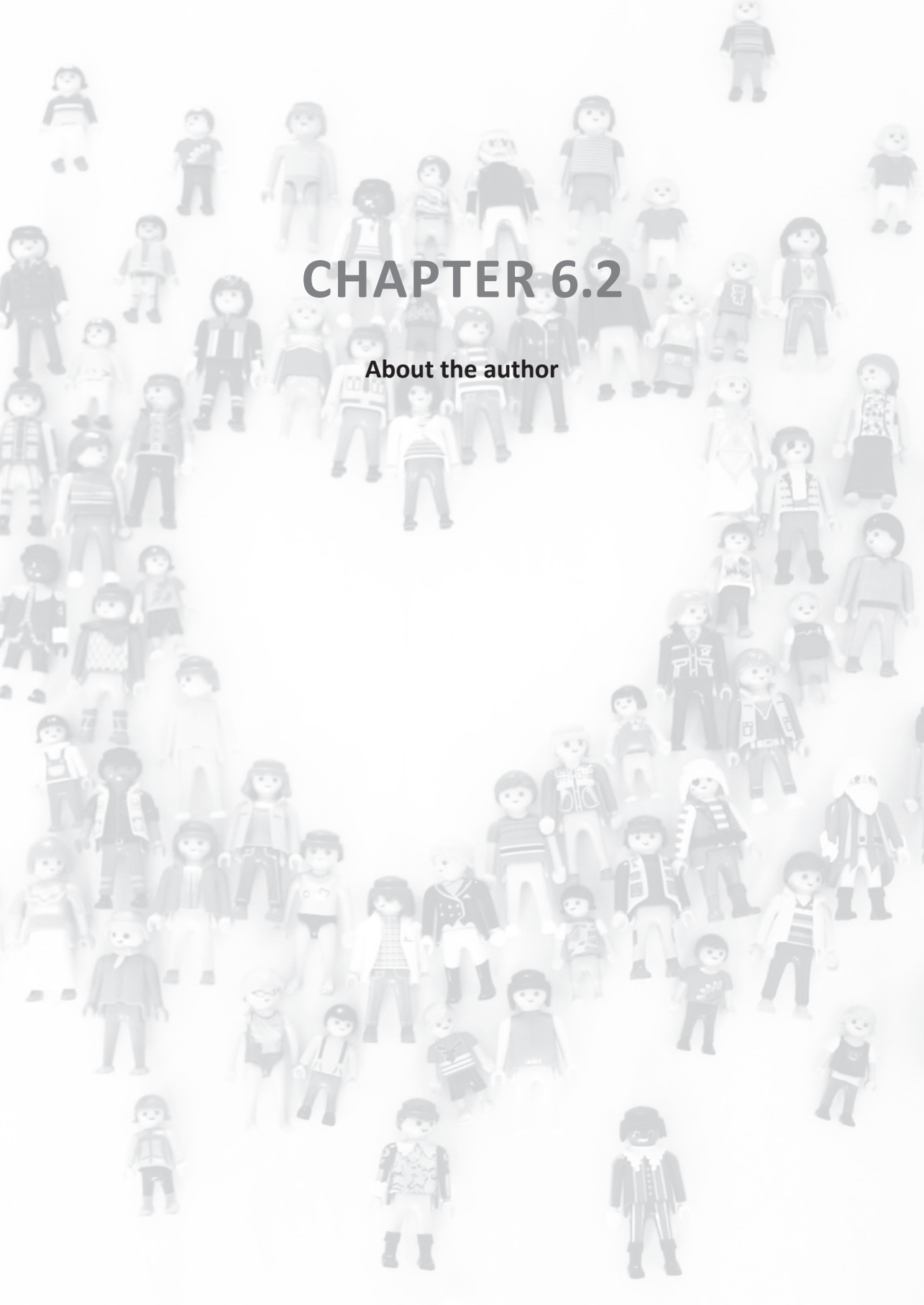
Hierbij wil ik ook alle ooms, tantes, neven en nichten van de familie van Leeuwen, de familie Jonker, de familie Tempelaars en de familie van der Velden bedanken. Familie: je kan ze niet uitzoeken dus bof ik zeker met zoveel lieve familieleden! Weliswaar zie ik niet iedereen even veel maar als we elkaar zien, is het altijd gezellig. Lieve oma, wat een eer dat u hierbij kunt zijn. Lieve Gertie, Carlo, Coen, Marleen, Rob en Maaike: nou, dat genen tellen zit erop ;). Dank jullie wel dat jullie er altijd gewoon zijn en ons steunen in onze plannen.

Lieve broertjes, voor jullie ook een belangrijk plekje in mijn dankwoord. Ik ben apetrots op twee van die grote broers, broers die altijd vragen hoe het gaat, die er altijd zijn op de belangrijke momenten, die zelf hun eigen dromen achternagaan en die geweldige ooms zijn voor kleine Lotte en Gijs. Het was zeker niet altijd even makkelijk de afgelopen jaren voor ons alledrie maar door het verdriet met jullie te kunnen delen, was het verdriet een stuk beter te dragen. Lieve Michiel, het gaat goed met je, als gepromoveerde in Denemarken en dat heb jij allemaal zelf bereikt door hard te werken en je eigen doelen na te schreven. Je mag trots op jezelf zijn daarvoor en anders ben ik het in ieder geval! Lieve Alexander, je bent en blijft mijn kleine broertje maar ik heb enorm veel respect voor hoe je je zaakjes altijd op orde hebt, alles probeert te regelen. Jij gaat vast een hele mooie carriere tegemoet. Lieve Steffie, dank je wel dat je zo'n leuke lieve meid bent, altijd opgewekt en enthousiast!

Lieve papa en mama: voor jullie is dit boek. Om jullie te bedanken voor alle onvoorwaardelijke steun die ik al mijn hele leven van jullie krijg, of ik nu voor t eerst moet gaan lopen, fietsen, boeken moet lezen, proefwerken maken of artikelen schrijven, jullie waren er altijd naast de zijlijn. Jullie hebben me geleerd dat door hard werken je heel veel kan bereiken maar vooral ook veel moet genieten, dank jullie wel daarvoor, wat een belangrijke en waardevolle lessen! Lieve papa, je schreef je proefschrift op je papadagen en nu deed ik dat op m'n mamadagen. Je verteld trots dat je door "al die artikelen" van je dochter, je eigen artikelen niet meer kan vinden op pubmed! Ik ben ook trots op jouw, het waren geen makkelijke jaren maar je doet het geweldig! Lieve mama, het mocht niet zo zijn dat je trots dit boekje aan de wereld kunt showen, dat je op de eerste rij zit bij de verdediging. Dat we samen kleding kopen voor tijdens de verdediging. Dat je even tegen me zegt dat het vast wel gaat lukken. Ik mis je enorm veel, elke dag, zoveel momenten die ik met je zou willen delen, even met je zou willen bespreken, maar toch ben je erbij, in mijn gedachte!

Lieve Lotte en Gijs, dank je wel dat jullie bestaan, dat jullie er zijn, dat jullie altijd zo vrolijk en lief zijn en dat jullie me laten zien dat het leven zo simpel, zo leuk kan zijn. Vanavond even geen werk, geen onderzoek maar gewoon een boekje lezen voor het slapengaan!

Lieve Pim, we hebben zoveel meegemaakt de afgelopen jaren, geweldige dagen en vreselijk verdrietige dagen maar altijd samen. Je bent er altijd voor me geweest, stond altijd klaar, ik mocht bij je uithuilen en samen vierden we de leuke dingen. Zoals op onze trouwkaart: samen met jouw aan mijn zijde, kon en kan ik het allemaal aan. Sorry dat ik soms geen aandacht voor je had, geen tijd voor je had, je bent altijd in mijn hoofd, en komt altijd op de eerste plaats. Jij bent mijn thuis, waar ik elke dag het liefste weer naar terugkeer. Dus lieve Pimmie, dank je wel dat je er bent, ik hou van je.



CHAPTER 6.2

About the author

Elisabeth Maria van Leeuwen was born in Amsterdam, the Netherlands, on July 1st, 1984. In 2003 she finished her atheneum (pre-university) education at the Kennemer Lyceum, Overveen, the Netherlands. After one year of Business Administration at the Erasmus University, Rotterdam, the Netherlands, she switched in 2004 to Biopharmaceutical Science at the University of Leiden, Leiden, the Netherlands. She finished her bachelor degree in 2009, however, in the meanwhile, she also completed the entrance program at the Radboud University, Nijmegen, the Netherlands. This allowed her to enter the master Bioinformatics Track of Molecular Life Science at the Radboud University, Nijmegen, the Netherlands which she finished in 2010. As part of the master, she performed two internships: a six-month internship at the Computation Drug Discovery group of the Molecular Design and Informatics Department of Schering-Plough, Oss, the Netherlands and a six-months internship at the Genomic Disorders Group of the Department of Human Genetics, Nijmegen Centre of Molecular Life Sciences, Nijmegen, The Netherlands and the Department of Physiology, Anatomy and Genetics of the MRC Functional Genomics Unit, Oxford, United Kingdom. After finalization of the master's degree, Elisa worked as a research assistant at the Department of Pharmacokinetics of the University of Groningen, Groningen, the Netherlands. In January 2011, Elisa started the work presented in this PhD thesis under supervision of Prof. Cornelia van Duijn at the Genetic Epidemiology Unit, Department of Epidemiology, Erasmus Medical Center Rotterdam, the Netherlands. Elisa completed during her PhD in 2013 her second Master of Science degree in Health Sciences at the Netherlands Institute of Health Sciences (NIHES), Rotterdam, the Netherlands with specification in Genetic Epidemiology. In 2014 she received a grant from the Cohorts for Heart and Aging Research in Genomic Epidemiology (CHARGE) Consortium to visit Boston for two months under supervision of Prof. L. Adrienne Cupples at the Department of Biostatistics, Boston University School of Public Health, Boston, United States of America. Elisa will to continue her genetic research for better understanding of complex human genetic traits at the Department of Epidemiology, Erasmus Medical Center Rotterdam, the Netherlands. Besides scientist, Elisa is also the wife of Pim and together they take care of their beautiful children, Lotte and Gijs.

CHAPTER 6.3

List of publications

1 M. Horikoshi, H. Yaghootkar, D.O. Mook-Kanamori, U. Sovio, H.R. Taal, B.J. Hennig, J.P. Bradfield, B. St Pourcain, D.M. Evans, P. Charoen, M. Kaakinen, D.L. Cousminer, T. Lehtimäki, E. Kreiner-Møller, N.M. Warrington, M. Bustamante, B. Feenstra, D.J. Berry, E. Thiering, T. Pfab, S.J. Barton, B.M. Shields, M. Kerkhof, E.M. van Leeuwen, A.J. Fulford, Z. Kutalik, J. Hua Zhao, M. den Hoed, A. Mahajan, V. Lindi, L.-K. Goh, J.-J. Hottenga, Y. Wu, O.T. Raitakari, M.N. Harder, A. Meirhaeghe, I. Ntalla, R.M. Salem, K.A. Jameson, K. Zhou, D.M. Monies, V. Lagou, M. Kirin, J. Heikkinen, L.S. Adair, F.S. Alkuraya, A. Al-Odaib, P. Amouyel, E.A. Andersson, A.J. Bennett, A.I.F. Blakemore, J.L. Buxton, J. Dallongeville, S. Das, E.J.C. de Geus, X. Estivill, C. Flexeder, P. Froguel, F. Geller, K.M. Godfrey, F. Gottrand, C.J. Groves, T. Hansen, J.N. Hirschhorn, A. Hofman, M.V. Hollegaard, D.M. Hougaard, E. Hyppönen, H.M. Inskip, A. Isaacs, T. Jørgensen, C. Kanaka-Gantenbein, J.P. Kemp, W. Kiess, T.O. Kilpeläinen, N. Klopp, B.A. Knight, C.W. Kuzawa, G. McMahon, J.P. Newnham, H. Niinikoski, B.A. Oostra, L. Pedersen, D.S. Postma, S.M. Ring, F. Rivadeneira, N.R. Robertson, S. Sebert, O. Simell, T. Slowinski, C. Tiesler, A. Tönjes, A. Vaag, J.S. Viikari, J.M. Vink, N.H. Vissing, N.J. Wareham, G. Willemsen, D.R. Witte, H. Zhang, J. Zhao, MAGIC, J.F. Wilson, M. Stumvoll, A.M. Prentice, B.F. Meyer, E.R. Pearson, C. Boreham, C. Cooper, M.W. Gillman, G.V. Dedoussis, L.A. Moreno, O. Pedersen, M. Saarinen, K.L. Mohlke, D.I. Boomsma, S.-M. Saw, T.A. Lakka, A. Körner, R. Loos, K.K. Ong, P. Vollenweider, C.M. van Duijn, G.H. Koppelman, A.T. Hattersley, J.W. Holloway, B. Hocher, J. Heinrich, C. Power, M. Melbye, M. Guxens, C.E. Pennell, K. Bønnelykke, H. Bisgaard, J.G. Eriksson, E. Widén, H. Hakonarson, A.G. Uitterlinden, A. Pouta, D.A. Lawlor, G.D. Smith, T.M. Frayling, M.I. McCarthy, S. Grant, V. Jaddoe, M.-R. Jarvelin, N.J. Timpson, I. Prokopenko, R.M. Freathy, and EGG Consortium. **New loci associated with birth weight identify genetic links between intrauterine growth and adult height and metabolism.** *Nat Genet, 45(1):76–82, Jan 2013*.

2 V. Codd, C.P. Nelson, E. Albrecht, M. Mangino, J. Deelen, J.L. Buxton, J. J. Hottenga, K. Fischer, T. Esko, I. Surakka, L. Broer, D.R. Nyholt, I. Mateo Leach, P. Salo, S. Hägg, M.K. Matthews, J. Palmen, G.D. Norata, P.F. O'Reilly, D. Saleheen, N. Amin, A.J. Balmforth, M. Beekman, R.A. de Boer, S. Böhringer, P.S. Braund, P.R. Burton, A.J.M. de Craen, M. Denniff, Y. Dong, K. Douroudis, E. Dubinina, J.G. Eriksson, K. Garlaschelli, D. Guo, A.-L. Hartikainen, A.K. Henders, J.J. Houwing-Duistermaat, L. Kananen, L.C. Karssen, J. Kettunen, N. Klopp, V. Lagou, E.M. van Leeuwen, P.A. Madden, R. Mägi, P.K.E. Magnusson, S. Männistö, M.I. McCarthy, S.E. Medland, E. Mihailov, G.W. Montgomery, B.A. Oostra, A. Palotie, A. Peters, H. Pollard, A. Pouta, I. Prokopenko, Samuli Ripatti, V. Salomaa, H.E.D. Suchiman, A.M. Valdes, N. Verweij, A. Viñuela, X. Wang, H.-E. Wichmann, E. Widen, G. Willemsen, M.J. Wright, K. Xia, X. Xiao, D.J. van Veldhuisen, A.L. Catapano, M.D. Tobin, A.S. Hall, A.I.F. Blakemore, W.H. van Gilst, H. Zhu, Cardiogram Consortium, J. Erdmann, M.P. Reilly, S. Kathiresan, H. Schunkert, P.J. Talmud, N.L. Pedersen, M. Perola, W. Ouwehand, J. Kaprio, N.G. Martin, C.M. van Duijn, I. Hovatta, C. Gieger, A. Metspalu, D.I. Boomsma, M.-R.

Jarvelin, P.E. Slagboom, J.R. Thompson, T.D. Spector, P. van der Harst, and N.J. Samani. **Identification of seven loci affecting mean telomere length and their association with disease.** *Nat Genet, 45(4):422–7, 427e1–2, Apr 2013*.

3 C.-Y. Cheng, M. Schache, M. K. Ikram, T.L. Young, J.A. Guggenheim, V. Vitart, S. MacGregor, V.J.M. Verhoeven, V.A. Barathi, J. Liao, P.G. Hysi, J.E. Bailey-Wilson, B. St Pourcain, J.P. Kemp, G. McMahon, N.J. Timpson, D.M. Evans, G.W. Montgomery, A. Mishra, Y.X. Wang, J.J. Wang, E. Rochtchina, O. Polasek, A.F. Wright, N. Amin, E.M. van Leeuwen, J.F. Wilson, C.E. Pennell, C.M. van Duijn, P. de Jong, J.R. Vingerling, X. Zhou, P. Chen, R. Li, W.-T. Tay, Y. Zheng, M. Chew, Consortium for Refractive Error and Myopia, K.P. Burdon, J.E. Craig, S.K. Iyengar, R.P. Igo Jr, J.H. Lass Jr, Fuchs' Genetics Multi-Center Study Group, E.Y. Chew, T. Haller, E. Mihailov, A. Metspalu, J. Wedenoja, C.L. Simpson, R. Wojciechowski, R. Höhn, A. Mirshahi, T. Zeller, N. Pfeiffer, K.J. Lackner, Wellcome Trust Case Control Consortium 2, T. Bettecken, T. Meitinger, K. Oexle, M. Pirastu, L. Portas, A. Nag, K.M. Williams, E. Yonova-Doing, R. Klein, B.E. Klein, S.M. Hosseini, A.D. Paterson, Diabetes Control and Complications Trial/Epidemiology of Diabetes Interventions, Complications Research Group, K.-M. Makela, T. Lehtimaki, M. Kahonen, O. Raitakari, N. Yoshimura, F. Matsuda, L.J. Chen, C.P. Pang, S.P. Yip, M.K.H. Yap, A. Meguro, N. Mizuki, H. Inoko, P.J. Foster, J.H. Zhao, E. Vithana, E.-S. Tai, Q. Fan, L. Xu, H. Campbell, B. Fleck, I. Rudan, T. Aung, A. Hofman, A.G. Uitterlinden, G. Bencic, C.-C. Khor, H. Forward, O. Pärssinen, P. Mitchell, F. Rivadeneira, A.W. Hewitt, C. Williams, B.A. Oostra, Y.-Y. Teo, C.J. Hammond, D. Stambolian, D.A. Mackey, C.C.W. Klaver, T.-Y. Wong, S.-M. Saw, and P.N. Baird. **Nine loci for ocular axial length identified through genome-wide association studies, including shared loci with refractive error.** *Am J Hum Genet, 93(2):264–277, Aug 2013*.

4 E.M. van Leeuwen, F. Smouter, T. Kam-Thong, N. Karbalai, A.V. Smith, T.B. Harris, L.J. Launer, C.M. Sitlani, G. Li, J.A. Brody, J.C. Bis, C.C. White, A. Jaiswal, B.A. Oostra, A. Hofman, F. Rivadeneira, A.G. Uitterlinden, E. Boerwinkle, C.M. Ballantyne, V. Gudnason, B.M. Psaty, L.A. Cupples, M.-R. Järvelin, S. Ripatti, A. Isaacs, B. Müller-Myhsok, L.C. Karssen, and C.M. van Duijn. **The challenges of genome-wide interaction studies: lessons to learn from the analysis of HDL blood levels.** *PLoS One, 9(10):e109290, 2014*.

5 A. Kiezun, S.L. Pulit, L.C. Francioli, F. van Dijk, M. Swertz, D.I. Boomsma, C.M. van Duijn, P.E. Slagboom, G.J.B. van Ommen, C. Wijmenga, Genome of the Netherlands Consortium, P.I.W. de Bakker, and S.R. Sunyaev. **Deleterious alleles in the human genome are on average younger than neutral alleles of the same frequency.** *PLoS Genet, 9(2):e1003301, 2013*.

6 D.I. Boomsma, C. Wijmenga, P.E. Slagboom, M.A. Swertz, L.C. Karssen, A. Abdellaoui, K. Ye, V. Guryev, M. Vermaat, F. van Dijk, L.C. Francioli, J.J. Hottenga, J.F.J. Laros, Q. Li, Y. Li, H. Cao, R. Chen, Y. Du, N. Li, S. Cao, J. van Setten, A. Menelaou, S.L. Pulit, J.Y. Hehir-Kwa, M. Beekman, C.C. Elbers, H. Byelas, A.J.M. de Craen, P. Deelen, M. Dijkstra, J.T. den Dunnen,

P. de Knijff, J. Houwing-Duistermaat, V. Koval, K. Estrada, A. Hofman, A. Kanterakis, D. van Enckevort, H. Mai, M. Kattenberg, E.M. van Leeuwen, P.B.T. Neerincx, B. Oostra, F. Rivadeneira, E.H..D. Suchiman, A.G. Uitterlinden, G. Willemsen, B.H. Wolffenbuttel, J. Wang, P.I.W. de Bakker, G.-J. van Ommen, and C.M. van Duijn. **The Genome of the Netherlands: design, and project goals.** *Eur J Hum Genet, 22(2):221–227, Feb 2014*.

7 K.M. Walsh, V. Codd, I.V. Smirnov, T. Rice, P.A. Decker, H.M. Hansen, T. Kollmeyer, M.L. Kosel, A.M. Molinaro, L.S. McCoy, P.M. Bracci, B.S. Cabriga, M. Pekmezci, S. Zheng, J.L. Wiemels, A.R. Pico, T. Tihan, M.S. Berger, S.M. Chang, M.D. Prados, D.H. Lachance, B.P. O'Neill, H. Sicotte, J.E. Eckel-Passow, ENGAGE Consortium Telomere Group, P. van der Harst, J.K. Wiencke, N.J. Samani, R.B. Jenkins, and M.R. Wrensch. **Variants near TERT and TERC influencing telomere length are associated with high-grade glioma risk.** *Nat Genet, 46(7):731–735, Jul 2014*.

8 Genome of the Netherlands Consortium. **Whole-genome sequence variation, population structure and demographic history of the Dutch population.** *Nat Genet, 46(8):818–825, Aug 2014*.

9 H. Springelkamp, R. Höhn, A. Mishra, P.G. Hysi, C.-C. Khor, S.J. Loomis, J.N Cooke Bailey, J. Gibson, G. Thorleifsson, S.F. Janssen, X. Luo, W.D. Ramdas, E. Vithana, M.E. Nongpiur, G.W. Montgomery, L. Xu, J.E. Mountain, P. Gharahkhani, Y. Lu, N. Amin, L.C. Karssen, K.-S. Sim, E.M. van Leeuwen, A.I. Iglesias, V.J.M. Verhoeven, M.A. Hauser, S.-C. Loon, D.D.G. Despriet, A. Nag, C. Venturini, P.G. Sanfilippo, A. Schillert, J.H. Kang, J. Landers, F. Jonasson, A.J. Cree, L.M.E. van Koolwijk, F. Rivadeneira, E. Souzeau, V. Jonsson, G. Menon, R.N. Weinreb, P. de Jong, B.A. Oostra, A.G. Uitterlinden, A. Hofman, S. Ennis, U. Thorsteinsdottir, K.P. Burdon, T.D. Spector, A. Mirshahi, S.-M. Saw, J.R. Vingerling, Y.-Y. Teo, J.L. Haines, R.C.W. Wolfs, H.G. Lemij, E.-S. Tai, N.M. Jansonius, J.B. Jonas, C.-Y. Cheng, T.Aung, A.C. Viswanathan, C.C.W. Klaver, J.E. Craig, S. Macgregor, D.A. Mackey, A.J. Lotery, K. Stefansson, A.A.B. Bergen, T.L. Young, J.L. Wiggs, N. Pfeiffer, T.-Y. Wong, L.R. Pasquale, A.W. Hewitt, C.M. van Duijn, C.J. Hammond, Blue Mountains Eye Study-GWAS group, NEIGHBORHOOD Consortium, and Wellcome Trust Case Control Consortium 2 (WTCCC2). **Meta-analysis of genome-wide association studies identifies novel loci that influence cupping and the glaucomatous process.** *Nat Commun, 5:4883, 2014*.

10 P.G. Hysi, C.-Y. Cheng, H. Springelkamp, S. Macgregor, J.N. Cooke Bailey, R. Wojciechowski, V. Vitart, A. Nag, A.W. Hewitt, R. Höhn, C. Venturini, A. Mirshahi, W.D. Ramdas, G. Thorleifsson, E. Vithana, C.-C. Khor, A.B. Stefansson, J. Liao, J.L. Haines, N. Amin, Y. Xing Wang, P.S. Wild, A.B. Ozel, J.Z. Li, B.W. Fleck, T. Zeller, S.E. Staffieri, Y.-Y. Teo, G. Cuellar-Partida, X. Luo, R. Rand Allingham, J.E. Richards, A. Senft, L.C. Karssen, Y. Zheng, C. Bellenguez, L. Xu, A.I. Iglesias, J.F. Wilson, J.H. Kang, E.M. van Leeuwen, V. Jonsson, U. Thorsteinsdottir, D.D.G. Despriet, S. Ennis, S.E. Moroi, N.G. Martin, N.M. Jansonius, S. Yazar, E.-S. Tai, P. Amouyel, J. Kirwan, L.M.E. van Koolwijk, M.A. Hauser, F. Jonasson, P.

Leo, S.J. Loomis, R. Fogarty, F. Rivadeneira, L. Kearns, K.J. Lackner, P.T.V.M. de Jong, C.L. Simpson, C.E. Pennell, B.A. Oostra, A.G. Uitterlinden, S.-M. Saw, A.J. Lotery, J.E. Bailey-Wilson, A. Hofman, J.R. Vingerling, C. Maubaret, N. Pfeiffer, R.C.W. Wolfs, H.G. Lemij, T.L. Young, L.R. Pasquale, C. Delcourt, T.D. Spector, C.C.W. Klaver, K.S. Small, K.P. Burdon, K. Stefansson, T.-Y. Wong, BMES GWAS Group, NEIGHBORHOOD Consortium, Wellcome Trust Case Control Consortium 2, A. Viswanathan, D.A. Mackey, J.E. Craig, J.L. Wiggs, C.M. van Duijn, C.J. Hammond, and T. Aung. **Genome-wide analysis of multi-ancestry cohorts identifies new loci influencing intraocular pressure and susceptibility to glaucoma.** *Nat Genet, 46(10):1126–1130, Oct 2014*.

11 P. Deelen, A. Menelaou, E.M. van Leeuwen, A. Kanterakis, F. van Dijk, C. Medina-Gomez, L.C. Francioli, J.J. Hottenga, L.C. Karssen, K. Estrada, E. Kreiner-Møller, F. Rivadeneira, J. van Setten, J. Gutierrez-Achury, H.-J. Westra, L. Franke, D. van Enckevort, M. Dijkstra, H. Byelas, C.M. van Duijn, Genome of the Netherlands Consortium, P.I.W. de Bakker, C. Wijmenga, and M.A. Swertz. **Improved imputation quality of low-frequency and rare variants in European samples using the 'Genome of The Netherlands'.** *Eur J Hum Genet, Jun 2014*.

12 H. Springelkamp, A. Mishra, P.G. Hysi, P. Gharahkhani, R. Höhn, C.-C. Khor, J.N. Cooke Bailey, X. Luo, W.D. Ramdas, E. Vithana, V. Koh, S. Yazar, L. Xu, H. Forward, L.S. Kearns, N. Amin, A.I. Iglesias, K.-S. Sim, E.M. van Leeuwen, A. Demirkan, S. van der Lee, S.-C. Loon, F. Rivadeneira, A. Nag, P.G. Sanfilippo, A. Schillert, P.T.V.M. de Jong, B.A. Oostra, A.G. Uitterlinden, A. Hofman, NEIGHBORHOOD Consortium, T. Zhou, K.P. Burdon, T.D. Spector, K.J. Lackner, S.-M. Saw, J.R. Vingerling, Y.-Y. Teo, L.R. Pasquale, R.C.W. Wolfs, H.G. Lemij, E.-S. Tai, J.B. Jonas, C.-Y. Cheng, T. Aung, N.M. Jansonius, C.C.W. Klaver, J.E. Craig, T.L. Young, J.L. Haines, S. MacGregor, D.A. Mackey, N. Pfeiffer, T.-Y. Wong, J.L. Wiggs, A.W. Hewitt, C.M. van Duijn, and C.J. Hammond. **Meta-analysis of genome-wide association studies identifies novel loci associated with optic disc morphology.** *Genet Epidemiol, 39(3):207–216, Mar 2015*.

13 E.M. van Leeuwen, L.C. Karssen, J. Deelen, A. Isaacs, C. Medina-Gomez, H. Mbarek, A. Kanterakis, S. Trompet, I. Postmus, N. Verweij, D.J. van Enckevort, J.E. Huffman, C.C. White, M.F. Feitosa, T.M. Bartz, A. Manichaikul, P.K. Joshi, G.M. Peloso, P. Deelen, F. van Dijk, G. Willemsen, E.J. de Geus, Y. Milaneschi, B.W.J.H. Penninx, L.C. Francioli, A. Menelaou, S.L. Pulit, F. Rivadeneira, A. Hofman, B.A. Oostra, O.H. Franco, I. Mateo Leach, M. Beekman, A.J.M. de Craen, H.-W. Uh, H. Trochet, L.J. Hocking, D.J. Porteous, N. Sattar, C.J. Packard, B.M. Buckley, J.A. Brody, J.C. Bis, J.I. Rotter, J.C. Mychaleckyj, H. Campbell, Q. Duan, L.A. Lange, J.F. Wilson, C. Hayward, O. Polasek, V. Vitart, I. Rudan, A.F. Wright, S.S. Rich, B.M. Psaty, I.B. Borecki, P.M. Kearney, D.J. Stott, L.A. Cupples, Genome of the Netherlands Consortium, J.W. Jukema, P. van der Harst, E.J. Sijbrands, J.-J. Hottenga, A.G. Uitterlinden, M.A. Swertz, G.-J.B. van Ommen, P.I.W. de Bakker, P.E. Slagboom, D.I.

Boomsma, C. Wijmenga, and C.M. van Duijn. **Genome of the Netherlands population-specific imputations identify an ABCA6 variant associated with cholesterol levels.** *Nat Commun, 6:6065, 2015*.

14 H. Springelkamp, A.I. Iglesias, G. Cuellar-Partida, N. Amin, K.P. Burdon, E.M. van Leeuwen, P. Gharahkhani, A. Mishra, S.J. van der Lee, A.W. Hewitt, F. Rivadeneira, A.C. Viswanathan, R.C.W. Wolfs, N.G. Martin, W.D. Ramdas, L.M. van Koolwijk, C.E. Pennell, J.R. Vingerling, J.E. Mountain, A.G. Uitterlinden, A. Hofman, P. Mitchell, H.G. Lemij, J. Jin Wang, C.C.W. Klaver, D.A. Mackey, J.E. Craig, C.M. van Duijn, and S. MacGregor. **ARHGEF12 influences the risk of glaucoma by increasing intraocular pressure.** *Hum Mol Genet, Jan 2015*.

15 I. Surakka, M. Horikoshi, R. Mägi, A.-P. Sarin, A. Mahajan, V. Lagou, L. Marullo, T. Ferreira, B. Miraglio, S. Timonen, J. Kettunen, M. Pirinen, J. Karjalainen, G. Thorleifsson, S. Hägg, J.-J. Hottenga, A. Isaacs, C. Ladenvall, M. Beekman, T. Esko, J.S. Ried, C.P. Nelson, C. Willenborg, S. Gustafsson, H.-J. Westra, M. Blades, A.J.M. de Craen, E.J. de Geus, J. Deelen, H. Grallert, A. Hamsten, A.S. Havulinna, C. Hengstenberg, J.J. Houwing-Duistermaat, E. Hyppönen, L.C. Karssen, T. Lehtimäki, V. Lyssenko, P.K.E. Magnusson, E. Mihailov, M. Müller-Nurasyid, J.-P. Mpindi, N.L. Pedersen, B.W.J.H. Penninx, M. Perola, T.H. Pers, A. Peters, J. Rung, J.H. Smit, V. Steinthorsdottir, M.D. Tobin, N. Tsernikova, E.M. van Leeuwen, J.S. Viikari, S.M. Willems, G. Willemsen, H. Schunkert, J. Erdmann, N.J. Samani, J. Kaprio, L. Lind, C. Gieger, A. Metspalu, P.E. Slagboom, L. Groop, C.M. van Duijn, J.G. Eriksson, A. Jula, V. Salomaa, D.I. Boomsma, C. Power, O.T. Raitakari, E. Ingelsson, M.-R. Järvelin, U. Thorsteinsdottir, L. Franke, E. Ikonen, O. Kallioniemi, V. Pietiäinen, C.M. Lindgren, K. Stefansson, A. Palotie, M.I. McCarthy, A.P. Morris, I. Prokopenko, S. Ripatti, and ENGAGE Consortium. **The impact of low-frequency and rare variants on lipid levels.** *Nat Genet, 47(6):589–597, Jun 2015*.

16 W.P. Kloosterman, L.C. Francioli, F. Hormozdiari, T. Marschall, J.Y. Hehir-Kwa, A. Abdellaoui, E.-W. Lameijer, M.H. Moed, V. Koval, I. Renkens, M.J. van Roosmalen, P. Arp, L.C. Karssen, B.P. Coe, R.E. Handsaker, E.D. Suchiman, E. Cuppen, D.T. Thung, M. McVey, M.C. Wendl, Genome of Netherlands Consortium, A. Uitterlinden, C.M. van Duijn, M.A. Swertz, C. Wijmenga, G.B. van Ommen, P.E. Slagboom, D.I. Boomsma, A. Schönhuth, E.E. Eichler, P.I.W. de Bakker, K. Ye, and V. Guryev. **Characteristics of de novo structural changes in the human genome.** *Genome Res, 25(6):792–801, Jun 2015*.

17 H.H.M. Draisma, R. Pool, M. Kobl, R. Jansen, A.-K. Petersen, A.A.M. Vaarhorst, I. Yet, T. Haller, A. Demirkan, T. Esko, G. Zhu, S. Böhringer, M. Beekman, J.B. van Klinken, W. Römisch-Margl, C. Prehn, J. Adamski, A.J.M. de Craen, E.M. van Leeuwen, N. Amin, H. Dharuri, H.-J. Westra, L. Franke, E.J.C. de Geus, J.J. Hottenga, G. Willemsen, A.K. Henders, G.W. Montgomery, D.R. Nyholt, J.B. Whitfield, B.W. Penninx, T.D. Spector, A. Metspalu, P.E. Slagboom, K. Willems van Dijk, P.A.C. 't Hoen, K. Strauch, N.G. Martin, G.-J.B. van Ommen, T. Illig, J.T. Bell, M. Mangino, K. Suhre, M.I. McCarthy, C. Gieger, A. Isaacs, C.M.

van Duijn, and D.I. Boomsma. **Genome-wide association study identifies novel genetic variants contributing to variation in blood metabolite levels.** *Nat Commun, 6:7208, 2015*.

18 M. Horikoshi, R. Magi, M. van de Bunt, I. Surakka, A.-P. Sarin, A. Mahajan, L. Marullo, G. Thorleifsson, S. Hagg, J.-J. Hottenga, C. Ladenvall, J.S. Ried, T.W. Winkler, S.M. Willems, N. Pervjakova, T. Esko, M. Beekman, C.P. Nelson, C. Willenborg, S. Wiltshire, T. Ferreira, J. Fernandez, K.J. Gaulton, V. Steinthorsdottir, A. Hamsten, P.K.E. Magnusson, G. Willemsen, Y. Milaneschi, N.R. Robertson, C.J. Groves, A.J. Bennett, T. Lehtimäki, J.S. Viikari, J. Rung, V. Lyssenko, M. Perola, I.M. Heid, C. Herder, H. Grallert, M. Müller-Nurasyid, M. Roden, E. Hypponen, A. Isaacs, E.M. van Leeuwen, L.C. Karssen, E. Mihailov, J.J. Houwing-Duistermaat, A.J.M. de Craen, J. Deelen, A.S. Havulinna, M. Blades, C. Hengstenberg, J. Erdmann, H. Schunkert, J. Kaprio, M.D. Tobin, N.J. Samani, L. Lind, V. Salomaa, C.M. Lindgren, P.E. Slagboom, A. Metspalu, C.M. van Duijn, J.G. Eriksson, A. Peters, C. Gieger, A. Jula, L. Groop, O.T. Raitakari, C. Power, B.W.J.H. Penninx, E. de Geus, J.H. Smit, D.I. Boomsma, N.L. Pedersen, E. Ingelsson, U. Thorsteinsdottir, K. Stefansson, S. Ripatti, I. Prokopenko, M.I. McCarthy, A.P. Morris, and ENGAGE Consortium. **Discovery and fine-mapping of glycaemic and obesity-related trait loci using high-density imputation.** *PLoS Genet, 11(7):e1005230, Jul 2015*.

19 G. Cuellar-Partida, H. Springelkamp, S.E.M. Lucas, S. Yazar, A.W. Hewitt, A.I. Iglesias, G.W. Montgomery, N.G. Martin, C.E. Pennell, E.M. van Leeuwen, V.J.M. Verhoeven, A. Hofman, A.G. Uitterlinden, W.D. Ramdas, R.C.W. Wolfs, J.R. Vingerling, M.A. Brown, R.A. Mills, J.E. Craig, C.C.W. Klaver, C.M. van Duijn, K.P. Burdon, S. MacGregor, and D.A. Mackey. **WNT10A exonic variant increases the risk of keratoconus by decreasing corneal thickness.** *Hum Mol Genet, 24(17):5060–5068, Sep 2015*.

20 E.M. van Leeuwen, A. Kanterakis, P. Deelen, M. van Kattenberg, The Genome of the Netherlands Consortium, P.E. Slagboom, P.I.W. de Bakker, C. Wijmenga, M.A. Swertz, D.I. Boomsma, C.M. van Duijn, L.C. Karssen, J.J. Hottenga. **Population-specific genotype imputations using minimac or IMPUTE2**. *Nat Protocols (Nat Protoc. 2015 Sep;10(9):1285-96)*.

21 J. Huang, B. Howie, S. McCarthy, Y. Memari, K. Walter, J.L. Min, P. Danecek, G. Malerba, E. Trabetti, H.-F. Zheng, UK10K Consortium, G. Gambaro, J.B. Richards, R. Durbin, N.J. Timpson, J. Marchini, and N. Soranzo. **Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel.** *Nat Commun, 6:8111, 2015*.

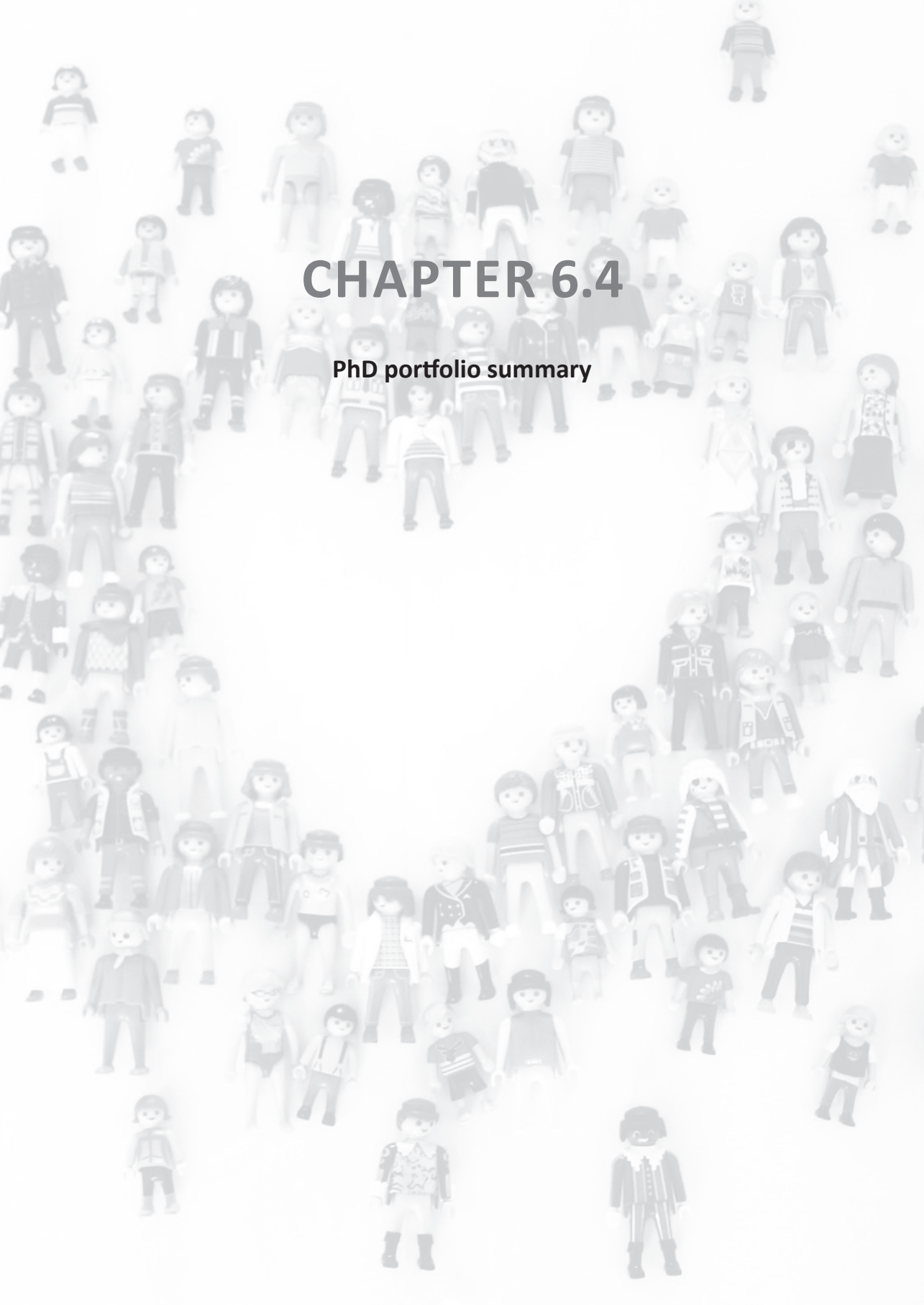
22 UK10K Consortium, K. Walter, J.L. Min, J. Huang, L. Crooks, Y. Memari, S. McCarthy, J.R.B. Perry, C.J. Xu, M. Futema, D. Lawson, V. Iotchkova, S. Schiffels, A.E. Hendricks, P. Danecek, R. Li, J. Floyd, L.V. Wain, I. Barroso, S.E. Humphries, M.E. Hurles, E. Zeggini, J.C. Barrett, V. Plagnol, J.B. Richards, C.M.T. Greenwood, N.J. Timpson, R. Durbin, and N. Soranzo. **The UK10K project identifies rare variants in health and disease.** *Nature, 526(7571):82–90, Oct 2015*.

23 K.J. Gaulton, T. Ferreira, Y. Lee, A. Raimondo, R. Mägi, M.E. Reschen, A. Mahajan, A. Locke, N. William Rayner, N. Robertson, R.A. Scott, I. Prokopenko, L.J. Scott, T. Green, T. Sparso, D. Thuillier, L. Yengo, H. Grallert, S. Wahl, M. Frånberg, R.J. Strawbridge, H. Kestler, H. Chheda, L. Eisele, S. Gustafsson, V. Steinthorsdottir, G. Thorleifsson, L. Qi, L.C. Karssen, E.M. van Leeuwen, S.M. Willems, M. Li, H. Chen, C. Fuchsberger, P. Kwan, C. Ma, M. Linderman, Y. Lu, S.K. Thomsen, J.K. Rundle, N.L. Beer, M. van de Bunt, A. Chalisey, H. Min Kang, B.F. Voight, G.R. Abecasis, P. Almgren, D. Baldassarre, B. Balkau, R. Benediktsson, M. Blüher, H. Boeing, L.L. Bonnycastle, E.P. Bottinger, N.P. Burtt, J. Carey, G. Charpentier, P.S. Chines, M.C. Cornelis, D.J. Couper, A.T. Crenshaw, R.M. van Dam, A.S.F. Doney, M. Dorkhan, S. Edkins, J.G. Eriksson, T. Esko, E. Eury, J. Fadista, J. Flannick, P. Fontanillas, C. Fox, P.W. Franks, K. Gertow, C. Gieger, B. Gigante, O. Gottesman, G.B. Grant, N. Grarup, C.J. Groves, M. Hassinen, C.T. Have, C. Herder, O.L. Holmen, A.B. Hreidarsson, S.E. Humphries, D.J. Hunter, A.U. Jackson, A. Jonsson, M.E. Jørgensen, T. Jørgensen, W.-H.L. Kao, N.D. Kerrison, L. Kinnunen, N. Klopp, A. Kong, P. Kovacs, P. Kraft, J. Kravic, C. Langford, K. Leander, L. Liang, P. Lichtner, C.M. Lindgren, E. Lindholm, A. Linneberg, C.-T. Liu, S. Lobbens, J. Luan, V. Lyssenko, S. Männistö, O. McLeod, J. Meyer, E. Mihailov, G. Mirza, T.W. Mühleisen, M. Müller-Nurasyid, C. Navarro, M.M. Nöthen, N.N. Oskolkov, K.R. Owen, D. Palli, S. Pechlivanis, L. Peltonen, J.R.B. Perry, C.G.P. Platou, M. Roden, D. Ruderfer, D. Rybin, Y.T. van der Schouw, B. Sennblad, G. Sigursson, A. Stancáková, G. Steinbach, P. Storm, K. Strauch, H.M. Stringham, Q. Sun, B. Thorand, E. Tikkanen, A. Tonjes, J. Trakalo, E. Tremoli, T. Tuomi, R. Wennauer, S. Wiltshire, A.R. Wood, E. Zeggini, I. Dunham, E. Birney, L. Pasquali, J. Ferrer, R.J.F. Loos, J. Dupuis, J.C. Florez, E. Boerwinkle, J.S. Pankow, C. van Duijn, E. Sijbrands, J.B. Meigs, F.B. Hu, U. Thorsteinsdottir, K. Stefansson, T.A. Lakka, R. Rauramaa, M. Stumvoll, N.L. Pedersen, L. Lind, S.M. Keinanen-Kiukaanniemi, E. Korpi-Hyövälti, T.E. Saaristo, J. Saltevo, J. Kuusisto, M. Laakso, A. Metspalu, R. Erbel, K.-H. Jöcke, S. Moebus, S. Ripatti, V. Salomaa, E. Ingelsson, B.O. Boehm, R.N. Bergman, F.S. Collins, K.L. Mohlke, H. Koistinen, J. Tuomilehto, K. Hveem, I. Njølstad, P. Deloukas, P.J. Donnelly, T.M. Frayling, A.T. Hattersley, U. de Faire, A. Hamsten, T. Illig, A. Peters, S. Cauchi, R. Sladek, P. Froguel, T. Hansen, O. Pedersen, A.D. Morris, C.N.A. Palmer, S. Kathiresan, O. Melander, P.M. Nilsson, L.C. Groop, I. Barroso, C. Langenberg, N.J. Wareham, C.A. O'Callaghan, A.L. Gloyn, D. Altshuler, M. Boehnke, T.M. Teslovich, M.I. McCarthy, A.P. Morris, and DIAGRAM Consortium. **Genetic fine mapping and genomic annotation defines causal mechanisms at type 2 diabetes susceptibility loci.** *Nat Genet, 47(12):1415–1425, Dec 2015*.

24 E.M. van Leeuwen, J.E. Huffman, J.C. Bis, A. Isaacs, M. Mulder, A. Sabo, A.V. Smith, S. Demissie, A. Manichaikul, J.A. Brody, M.F. Feitosa, Q. Duan, K.E. Schraut, P. Navarro, J.V. van Vliet-Ostaptchouk, G. Zhu, H. Mbarek, S. Trompet, N. Verweij, L.-P. Lyytikäinen, J. Deelen, I.M. Nolte, S.W. van der Laan, G. Davies, A.J.M. Vermeij- Verdoold, A.A.L.J. van

Oosterhout, J.M. Vergeer- Drop, D.E. Arking, H. Trochet, Generation Scotland, C. Medina-Gomez, F. Rivadeneira, A.G. Uitterlinden, A. Dehghan, O.H. Franco, E.J. Sijbrands, A. Hofman, C.C. White, J.C. Mychaleckyj, G.M. Peloso, M.A. Swertz, LifeLines Cohort Study, G. Willemsen, E.J. de Geus, Y. Milaneschi, B.W.J.H. Penninx, I. Ford, B.M. Buckley, A.J.M. de Craen, J.M. Starr, I.J. Deary, G. Pasterkamp, A.J. Oldehinkel, H. Snieder, P.E. Slagboom, K. Nikus, M. Kähönen, T. Lehtimäki, J.S. Viikari, O.T. Raitakari, P. van der Harst, J.E. Jukema, J.-J. Hottenga, D.I. Boomsma, J.B. Whitfield, G. Montgomery, N.G. Martin, CHARGE Lipids Working Group, O. Polasek, V. Vitart, C. Hayward, I. Kolcic, A.F. Wright, I. Rudan, P.K. Joshi, J.F. Wilson, L.A. Lange, J.G. Wilson, V. Gudnason, T.B. Harris, A. Morrison, I.B. Borecki, S.S. Rich, S. Padmanabhan, B.M. Psaty, J.I. Rotter, B.H. Smith, E. Boerwinkle, L.A. Cupples and C.M. van Duijn. **Fine mapping the CETP region reveals a common intronic insertion associated to HDL-C.** *NPJ Aging and Mechanisms of Disease, in press*.

25 E.M. van Leeuwen, A. Sabo, J.C. Bis, J.E. Huffman, A. Manichaikul, A.V. Smith, M.F. Feitosa, S. Demissie, P.K. Joshi, Q. Duan, J. Marten, J.B. van Klinken, I. Surakka, I.M. Nolte, W. Zhang, H. Mbarek, R. Li-Gao, S. Trompet, N. Verweij, E. Evangelou, L.-P. Lyytikäinen, B.O. Tayo, J. Deelen, P.J. van der Most, S.W. van der Laan, D. Arking, A. Morrison, A. Dehghan, O.H. Franco, A. Hofman, F. Rivadeneira, E.J. Sijbrands, A.G. Uitterlinden, J.C. Mychaleckyj, A. Campbell, L.J. Hocking, S. Padmanabhan, J.A. Brody, K.M. Rice, C.C. White, T. Harris, A. Isaacs, H. Campbell, L.A. Lange, I. Rudan, I. Kolcic, P. Navarro, T. Zemunik, V. Salomaa, The LifeLines Cohort Study, J.S. Kooner, B. Lehne, W.R. Scott, S.-T. Tan, E.J. de Geus, Y. Milaneschi, B.W.J.H. Penninx, G. Willemsen, R. de Mutsert, I. Ford, R.T. Gansevoort, M.P. Segura-Lepe, O.T. Raitakari, J.S. Viikari, K. Nikus, T. Forrester, C.A. McKenzie, A.J.M. de Craen, H.M. de Ruijter, CHARGE Lipids Working Group, G. Pasterkamp, H. Snieder, A.J. Oldehinkel, P.E. Slagboom, R.S. Cooper, M. Kähönen, T. Lehtimäki, P. Elliott, P. van der Harst, J.W. Jukema, D.O. Mook-Kanamori, D.I. Boomsma, J.C. Chambers, M. Swertz, S. Ripatti, K. Willems van Dijk, V. Vitart, O. Polasek, C. Hayward, J.G. Wilson, J.F. Wilson, V. Gudnason, S.S. Rich, B.M. Psaty, I.B. Borecki, E. Boerwinkle, J.I. Rotter, L.A. Cupples, C.M. van Duijn. **Meta-analysis of 49,549 individuals imputed with the 1000 Genomes Project reveals an exonic damaging variant in ANGPTL4 determining fasting TG levels.** *Journal of Medical Genetics, in press*.



CHAPTER 6.4

PhD portfolio summary

Name PhD student: Elisabeth van Leeuwen	PhD period: 2011-2015
Erasmus MC Department: Epidemiology	Promotor: C.M. van Duijn
Research School: NIHES	

1. PhD training

	Year	Workload (Hours/ECTS)
In-depth courses		
NIHES Master of Science in Genetic Epidemiology	2011-2013	
- Principles of research in medicine		0.7
- Genome-wide association analysis		1.4
- Principles of genetic epidemiology		0.7
- Genomics in molecular medicine		1.4
- Advances in genomics research		0.4
- Study design		4.3
- Biostatistics for clinicians		0.7
- Regression analysis for clinicians		1.4
- Survival analysis for clinicians		1.9
- Biostatistical methods II: classical regression models		4.3
- Genetic-Epidemiologic research methods		5.7
- SNP's and human diseases		1.4
- Repeated measurements in clinical studies		1.4
- Missing values in clinical research		0.7
- Introduction to clinical and public health genomics		1.4
- Advances in genome-wide association analysis		1.4
- Family-based genetic analysis		1.4
- Mendelian randomization		0.9
- A first encounter with next-generation sequencing data		1.4
Course Basic association, next-generation sequencing and linkage at the Rutgers University, Unites States of America	2011	1.4
Presentations		
International conferences and meetings:		
- Netherlands Bioinformatics Centre Conference, Lunteren, the Netherlands: "Detecting Epistatis Using GPUs" (poster)	2011	1
- Netherlands Bioinformatics Centre Conference, Lunteren, the Netherlands: "Detecting SNP interactions associated with HDL using GPUs" (poster)	2012	1
- European Mathematical Genetics Meeting, Göttingen, Germany: "Detecting SNP interactions associated with HDL using GPUs" (oral)	2012	1
- European Human Genetics Conference, Nürnberg, Germany: "Detecting SNP interactions associated with HDL using GPUs" (poster)	2012	1
- ENGAGE investigator meeting, Rotterdam, the Netherlands: "A genome-wide analysis of SNP-SNP interactions associated with HDL" (poster)	2012	1
- BBMRI-NL annual conference: "Imputation of genomic data: best practices and pipelines" (poster)	2012	1

- European Mathematical Genetics Meeting, Göttingen, Germany: "A genome-wide analysis of SNP-SNP interactions associated with HDL" (oral)	2013	1
- Netherlands Bioinformatics Centre Conference, Lunteren, the Netherlands: "A meta-analysis for blood lipids after imputing with the GoNL reference panel" (oral)	2013	1
- American Society of Human Genetics Conference, San Diego, Unites States of America: "Population-specific imputations identify a variant in the *ABCA6* gene associated with cholesterol levels" (oral) and "Exome sequencing in an isolated population reveals multiple rare variants affecting both high-density lipoprotein cholesterol and the levels of certain blood metabolites" (poster)	2014	1
- CHARGE investigators meeting, Washington, Unites States of America: "Revisiting the association between HLD-C and *CETP* reveals the presence of a common intronic deletion" (poster)	2014	1
- Dutch Techcentre for Life Sciences kickoff event, Amersfoort, the Netherlands: "Exome sequencing in an isolated population reveals multiple rare variants affecting both high-density lipoprotein cholesterol and the levels of certain blood metabolites" (poster)	2014	1
Oral presentations at lab meetings:		
- Semi-annual presentations at the Genetic Epidemiology Unit, Department of Epidemiology, Erasmus MC, Rotterdam, the Netherlands.	2011-2015	1
- 2020 meeting of the Department of Epidemiology, Erasmus MC, Rotterdam, the Netherlands.	2013	1
Seminars, symposia and workshops		
- Weekly seminars and 2020 meetings at the Department of Epidemiology, Erasmus MC, Rotterdam, the Netherlands.	2011-2015	2
- Bridge meetings, Erasmus MC, Rotterdam, the Netherlands.	2011-2015	1
- Annual Centre for Medical Systems Biology Symposium	2011-2013	3
- European Human Genetics Conference, Amsterdam, the Netherlands.	2011	1
- CHARGE investigators meeting, Rotterdam, the Netherlands	2013	1
- Genome of the Netherlands consortium meetings, Utrecht, the Netherlands	2011-2012	1
Other		
- Research fellow at the Department of Biostatistics, Boston University School of Public Health, Boston, MA, USA. Supervised by Prof. Adrienne Cupples and supported by the CHARGE consortium	2015	2 months

2. Teaching		
	Year	**Workload (Hours/ECTS)**
Lecturing		
- Advances in Genome-Wide Association Studies	2015	1
Supervising practicals and excursions		
- Assisting in the “Principles of Genetic Epidemiology” course	2012	1
Supervising Master’s theses		
- Supervision Annelies Smouter (Bachelor student Bioinformatics): “The search for high-density lipoprotein cholesterol correlated SNP-SNP interactions in the Rotterdam Study cohort I”.	2011-2012	3