Genetic risk factors of cardiovascular disease

van Iperen, E.P.A.

Publication date
2018
Document Version
Final published version
License
Other

Genetic Risk factors of Cardiovascular disease

Genetic Risk factors of Cardiovascular disease

Erik P.A. van Iperen

Erik van Iperen

GENETIC RISK FACTORS OF CARDIOVASCULAR DISEASE

Erik Pieter Adriaan van Iperen

Genetic Risk Factors of Cardiovascular Disease
PhD thesis, University of Amsterdam, The Netherlands
ISBN: 978-94-6375-090-5

Cover design: Erik van Iperen
Lay-out: Erik van Iperen
Printing: Ridderprint BV

The author of this thesis can be contacted by email: erik.vaniperen@gmail.com

GENETIC RISK FACTORS OF CARDIOVASCULAR DISEASE

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. K. I. J. Maex
ten overstaan van een door het College voor Promoties ingestelde commissie,
in het openbaar te verdedigen in de Agnietenkapel
op woensdag 10 oktober 2018, te 14:00 uur

door **Erik Pieter Adriaan van Iperen**

geboren te Schiedam

PROMOTIECOMMISSIE:

PROMOTORES:	prof. dr. A. H. Zwinderman	AMC-UvA
	prof. dr. F. W. Asselbergs	Universiteit Utrecht
CO-PROMOTOR:	prof. dr. G. K. Hovingh	AMC-UvA
OVERIGE LEDEN:		
	prof. dr. E. J. Meijers-Heijboer	AMC-UvA
	prof. dr. J. W. Jukema	Universiteit Leiden
	prof. dr. M. A. Swertz	Rijksuniversiteit Groningen
	prof. dr. O. H. Franco Durán	Erasmus Universiteit Rotterdam
	dr. J. Hamann	AMC-UvA
	dr. P. Henneman	AMC-UvA

Faculteit der Geneeskunde

Financial support by the Dutch Heart Foundation for the publication of this thesis is gratefully acknowledged

CONTENTS

III Imputation

IV (Genetic) Risk Factors

1 INTRODUCTION

CARDIOVASCULAR DISEASE

Cardiovascular disease (CVD) is the leading cause of mortality in the Western Societies[1]. CVD encompasses a number of different disease entities, many of which are related to atherosclerosis. Atherosclerosis is the process of development of plaque formation in the subendothelium of arterial walls. These plaques narrow the arterial lumen, thereby disabling blood flow. Upon rupture of the fibrous cap, a blood clot forms, which results in total obstruction of blood flow, which, depending of the site, may lead to myocardial infarction (MI), stroke or peripheral artery disease (PAD). Atherosclerosis has a multifactorial origin involving abnormalities[2] in lipid metabolism, hypertension, obesity, diabetes mellitus, smoking, inflammation, coagulation, and fibrinolysis, amongst others. At present, we lack a complete understanding of the relevance of these individual risk factors and their interplay in the disease process[3]. It has been suggested that genetic factors contribute to the risk of CVD[4].
One of the genetic epidemiology tools is the concept of heritability. Heritability is a parameter that can help understand the genetic architecture of complex traits within the population. Heritability is usually defined as the proportion of total phenotypic variation that is due to additive genetic factors[5]. Estimates for the heritability of CVD vary between 40 and 60% [6]. These different estimates of CVD heritability stem from studies using genetic variants that are ”common” (occurring quite often in the population) or using rare genetic variants[7]. Genetic variation that occurs in more than 1% of the population are called Single Nucleotide Polymorphisms (SNPs). They are considered common enough to be called ”normal variation” in the DNA. SNPs are responsible for many of the differences between individuals such as eye color, hair color, and blood type. SNPs that occur in less than 1% of the population are called rare variants. Although many SNPs have no effects on a persons health, some of these variations may influence the risk of CVD.

FAMILY STUDIES:

In family studies multiple generations of individuals with CVD and controls without CVD are investigated, usually by means of so-called linkage analyses. The identified private and rare mutations often have a large effect on the disease risk. Examples of rare diseases that have an impact on CVD risk are Familial hypercholesterolemia (FH), hypertrophic cardiomyopathy (HCM)[8] and Long QT syndrome[9]. FH is studied in one of the chapters in this thesis (**chapter 4**). FH is an autosomal dominant disease characterized by increased plasma levels of Low density lipoprotein cholesterol (LDL-C), FH is caused by mutations in the LDLR, APOB or PCSK9 gene. The high cholesterol levels lead to development of arterial plaques, which will ultimately lead to CVD events. The onset and severity of CVD varies considerably between FH patients, even among individuals who share an identical gene defect[10].

GENOME WIDE ASSOCIATION STUDIES:

Genome Wide Association Studies (GWAS) are an alternative to family studies. The rationale underlying the concept of GWAS is the ”common disease, common variant” hypothesis in which a limited number of common genetic variants with a high frequency (typically above 5%) in the general population contribute to the susceptibility for disease[7]. GWAS were made possible by means of major technological advances in genetics and molecular biology. In the HapMap project[11] over 10 million common SNPs with a minor allele frequency (MAF) greater than 5% were identified.
It was found that a large number of these 10 million SNPs were in linkage disequilibrium (LD). If two genes are in LD, it means that alleles of both genes are inherited together more often than would be expected by chance. A tagging SNP is a SNP that represents a group of SNPs in high LD. The principle of tagging SNPs made it possible to develop cheap DNA microarrays with a limited but still large number of SNPs to investigate large cohorts of patients with disease and controls without disease. By using tagging SNPs not all SNPs have to be on the DNA microarrays to cover the genome of a person. In the last decade due to technological improvements the number of SNPs covered on a GWAS array increased from ten thousand to more than one million SNPs.

In the past ten years many initiatives arose to combine different GWAS in meta- analyses. By virtue of the large power that was achieved by combining data, new SNPs were found to be associated with CVD. The Global Lipids Genetics Consortium (GLGC)[12], Coronary ARtery DIsease Genome wide Replication and Meta-analysis Consortium (CAR-

DIoGRAM)[13], CARDIoGRAMplusC4D[14] and the IBC CardioChip consortium are 4 examples of such consortia. The aim of the GLGC was to study the genetic determinants of LDL, High-Density Lipoprotein (HDL), Total cholesterol (TC) and triglycerides (TG). A total of 95 common variants were found to be associated with variation in blood lipid levels in the first meta-analysis[15]. In a second analysis an additional 62 common variants were shown to be associated with blood lipid levels[12].
The role of rare variants remained unknown, therefore gene centric genotyping platforms were developed such as the Metabochip[16], IBC CardioChip (Illumina, San Diego CA)[17] and the ExomeChip[18].

The MetaboChip[16] is a custom Illumina iSelect genotyping array designed to test, in a cost-effective manner, 200,000 SNPs of interest for metabolic and atherosclerotic / cardiovascular disease traits. The SNPs on the chip were selected on the basis of large scale meta-analyses (including up to 100,000 individuals), HapMap[11] and the 1000 Genomes Project SNP[19].

The MetaboChip was designed by representatives of different GWAS meta-analysis consortia: CARDIoGRAM (coronary artery disease)[13], DIAGRAM (type 2 diabetes)[20], GIANT (Genetic Investigation of ANthropometric Traits)[21], MAGIC (glycemic traits)[22], Lipids (lipids), ICBP-GWAS (blood pressure)[23], and QT-IGC (QT interval).

A total of 217,695 SNPs were selected for the MetaboChip and 20,970 of these SNPs (9.6%) failed during the assay manufacturing process, resulting in 196,725 SNPs available for genotyping.

The 50K gene-centric Human CVD BeadChip contains approximately 50,000 SNPs in about 2000 genes in relevant loci across a range of cardiovascular, metabolic and inflammatory syndromes[17].

The HumanExome BeadChip contains about 250,000 variants based on the data of 12,000 sequenced genomes and exomes. Each variant on the chip has been identified at least >3 times in at least 2 different data sets[18].

The Coronary ARtery DIsease Genome-wide Replication And Meta-analysis (CARDIoGRAM) consortium was initiated to maximize the chance of finding novel susceptibility loci for CVD. CARDIoGRAM combined data from all published and several unpublished GWAS in individuals with European ancestry; CARDIOGRAM included $>$22,000 cases with Coronary Artery Disease (CAD), Myocardial Infarction (MI), or both and $>$60,000 controls without CAD and MI. In addition 15,420 CHD cases and 15,062 controls from the C4D GWAS meta-analysis were added to the CARDIOGRAM GWAS resulting in the CARDIoGRAMplusC4D dataset, a two stage meta-analysis was performed

within the CARDIoGRAMplusC4D consortium involving 63,746 cases and 130,681 controls. The meta-analyses of these consortia have led to the identification of a total of 46 new loci associated with CVD[14]. Amongst which were loci in or close to the well-known 9p21 locus, the APOA5-APOA1, APOE-APOC1 and LPL genes.

The meta-analysis of these gene-centric platforms in big consortia have resulted in the identification of common variants for either blood lipids or CVD. During my thesis I contributed to these meta-analyses. In **chapter 2** of this thesis I performed a meta-analysis with data from the IBC cardiochip consortium on 4 different lipid traits (LDL-C, HDL-C, TG, TC) in 66,240 individuals from 32 different studies and found 21 new common variants associated with one or more lipid traits: PPARG, GP1HBP1, DGAT2, HCAR2, FTO, VLDLR, SPTY2D1, BRCA2, SOCS3, APOH, C4B, LPAL2, GCK, GATA4, SERPINF2, INSR, FCGR2A, INSIG2, UGT1A1, CHUK, UBE3B[24]. In **chapter 3** we performed a meta-analysis in three different ethnic groups. This trans-ethnic meta-analysis was performed to identify candidate genes with an effect on lipid levels in admixed populations[25]. We found and confirmed two novel signals, ICAM1 and CD36 for LDL-C and HDL-C, and replicated these findings in a cohort of 7,000 African Americans.

GWAS CATALOG

Results from published GWAS studies are published in the GWAS-catalog[26]. The GWAS catalog is a quality controlled, manually curated, literature-derived collection of all published genome-wide association studies assaying at least 100,000 SNPs and all SNP-trait associations with p-values $< 1.0x10^{-5}$. The GWAS catalog contains as of 10-04-2018, 3349 studies and 59,967 unique SNP-trait associations. The GWAS catalog is mostly used as a lookup source.

IMPUTATION

Imputation methods are widely used in GWAS as they facilitate association studies with variants that are not directly genotyped. Using imputation methods, we can extend the number of SNPs from 1 million to about 30 million, using the 1000 genomes reference set or sets from other population sequencing studies. Recently a new version of the 1000 Genomes Project was published[19], using the whole genome data of 2504 individuals while the initial publication comprised 1000 genomes from 26 different populations. More than 88 million variants were identified. Since the start of the HapMap and 1000 Genomes projects many other local initiatives like The Genome of the Netherlands

project (GoNL) project[27] and the UK10K project[28] were started, the aim is to characterize DNA sequence variation in the Dutch and the UK population. The GoNL project is a whole-genome-sequencing project in a representative sample consisting of 250 trio-families from all provinces in the Netherlands, and aims to characterize DNA sequence variation in the Dutch population[27]. A total of 19.5 million novel sequence variants were found. The UK10K project performed whole-genome-sequencing or whole-exome-sequencing of nearly 10,000 individuals from population-based and disease collections. A total of 24 million novel sequence variants were identified[28]. Another application of imputation methods is to combine and analyze data genotyped on different genotyping arrays. In **chapter 6**, I investigated whether the number of high quality SNPs vary while using/applying two different imputation methods. The number of high quality SNPs available were analyzed while first combining different data sets and subsequently perform imputation on the combined dataset. In our second approach we first performed imputation on the individual datasets before combining them. Our results suggest that first performing imputation on the individual datasets and then combining them result in more SNPs of good quality compared to the method where we combine the different datasets before imputation.

RISK FACTORS:

The major cardiovascular risk factors: male sex, hypertension, increased cholesterol, smoking, and diabetes mellitus, have been acknowledged for 50 years[29]. Based on these risk factors, a number of risk prediction scores have been developed, including the Framingham risk score (FRS). The FRS provides an estimate of the 30-year risk of CVD[30, 31, 32, 33]. These risk scores have been validated in many populations. However, the risk (re)classification and clinical utility of many of these scores have been less well studied and further research is needed to investigate the clinical utility[4]. The current risk scores explain a modest proportion of CVD incidence in the general population, only atmost 50% of the incidence of CVD is explained by the traditional risk factors[4]. None of the traditional risk factors is present in 15% to 20% of the patients who suffer from an Acute Myocardial Infarction (AMI). Based on the current risk predictions scores, these patients would have been considered as ”low risk”[34]. Therefore, novel risk factors of preclinical disease are urgently needed to refine the current risk prediction algorithms. Current research has a major focus on emerging genetic risk factors. Many SNPs associated with common risk factors and potential new risk factors have been detected by GWAS. Adding these genetic information to existing risk scores did not much improve the CVD risk prediction until now[4]. A relatively new tool to im-

prove CVD risk prediction is the development of Genetic Risk Scores (GRS). Genetic risk scores summarize risk-associated variation across the genome by counting the number of disease-associated alleles. Because GRS uses information from multiple SNPs, each individual SNP is less important to the summary measurement and the "signal" from the GRS as a whole is more robust. A GRS is therefore an efficient and effective way of constructing genome-wide risk measurements from GWAS findings[35].

In **chapter 4** we evaluated if the 46 known CAD SNPs for the general population also play a role in the event free survival of patients with FH. We constructed a GRS consisting of these 46 SNPs. Our GRS was however not associated with a higher risk of developing a CVD event[36].

In **chapter 5** we performed the same analysis in a cohort of patients with established CVD treated with atorvastatin and we evaluated whether our GRS of 46 known CVD loci hold predictive value for a secondary CVD event. We again did not find an association between our GRS and the risk of a secondary CVD event [ARTICLE NOT PUBLISHED].

In **chapter 7** we set out to determine the independent effect of risk factors for CVD, by testing the association of SNPs in these RF with CAD. We tested known RF for CAD and new potential RFs using genetic risk scores based on published data in the GWAS Catalog[26]and the public available summary level data of the CARDIoGRAM consortium. We confirmed the association of a large number of known RF for CVD. In particular, we identified Coronary Artery Calcification (CAC), Lipoprotein a (LP(a)), Height and Plaque as RFs and of being potential treatment targets and thereby testing their causal effect on CVD[37].

Part I

Lipids

2 LARGE-SCALE GENE-CENTRIC META-ANALYSIS ACROSS 32 STUDIES IDENTIFIES MULTIPLE LIPID LOCI

Genome wide association studies (GWAS) have identified many single-nucleotide polymorphisms (SNPs) underlying variations in plasma lipid levels. We explore whether additional loci associated with plasma lipid phenotypes, such as high-density lipoprotein cholesterol (HDL-C), low-density lipoprotein cholesterol (LDL-C), total cholesterol (TC) and triglycerides (TG) can be identified by a dense gene-centric approach.

Our meta-analysis of 32 studies in 66,240 individuals of European ancestry was based on the custom ~50,000 SNP genotyping array covering ~2,000 candidate genes (the ITMAT-Broad-CARe (IBC) array). SNP-lipid associations were replicated in a cohort comprising an additional 24,736 samples or within the Global Lipid Genetic Consortium.

We identified 4, 6, 10 and 4 unreported SNPs in established lipid genes for HDL-C, LDL-C, TC and TG respectively. We also identified several lipid-related SNPs in previously unreported genes: DGAT2, HCAR2, GPIHBP1, PPARG, and FTO for HDL-C; SOCS3, APOH, SPTY2D1, BRCA2 and VLDLR for LDL-C; SOCS3, UGT1A1, BRCA2, UBE3B, FCGR2A, CHUK, and INSIG2 for TC; and SERPINF2, C4B, GCK, GATA4, INSR and LPAL2 for TG. The proportion of phenotypic variance explained in the subset of studies providing individual-level data was 9.9% for HDL-C, 9.5% for LDL-C, 10.3% for TC and 8.0% for TG.

This large meta-analysis of lipid phenotypes using a dense gene-centric approach identified multiple SNPs not previously described in established lipid genes and several previously unknown loci. The explained phenotypic variance using this approach was comparable to meta-analysis of GWAS data suggesting that a focused genotyping approach can further increase the understanding of heritability of plasma lipids.

INTRODUCTION

Cardiovascular disease (CVD) is one of the leading causes of disability and death worldwide[38]. Atherosclerosis is the major underlying pathological process of CVD and it is highly prevalent in western societies. Atherogenesis has numerous genetic and environmental risk factors[39] with abnormalities of plasma lipids and lipoproteins accounting for ~50% of the population attributable risk of developing CVD[40, 41]. Plasma lipid and lipoprotein levels are themselves highly heritable, with estimates ranging from 40-60% for total cholesterol (TC), low-density lipoprotein cholesterol (LDL-C), high-density lipoprotein cholesterol (HDL-C) and triglycerides (TG)[42].

In a large-scale meta-analysis of genome wide association studies (GWAS) it was shown that common genetic variants in 95 loci affect plasma lipid levels, of which 59 were previously unreported[15]. Taken together, variation at these loci explains 10-12% of the total variance and 25-30% of the genetic variability in plasma lipid phenotypes[15]. This means that while a portion of the genetic contribution to variation in plasma lipids and lipoproteins has been characterized, there is still variance that remains unattributed[7].

To further identify genetic associations underlying variation in plasma lipid phenotypes, we performed a large meta-analysis of 32 studies comprising 66,240 individuals of European ancestry using the candidate gene ITMAT-Broad-CARe (IBC) array (Illumina), also known as the CardioChip or the Human CVD BeadArray. The IBC array was designed to capture genetic diversity using ~50,000 SNPs across ~2,000 candidate gene regions primarily related to cardiovascular, inflammatory and metabolic phenotypes[17]. Prior reports using this array have confirmed previously established associations and identified unreported associations of SNPs with several phenotypes and disease outcomes, including coronary artery disease[43, 44], plasma lipids[45, 46], blood pressure[47, 48], cardiomyopathy[49], type 2 diabetes (T2D)[50, 51] and height[52]. The majority of loci on the IBC array are captured with a marker density equal to or greater than that seen on genome-wide arrays. Compared to the agnostic design of GWAS arrays, gene-centric genotyping with this array may permit a better identification of multiple functional polymorphisms, or their proxies, at each locus. Indeed, this approach has the potential to capture a more detailed genetic architecture in selected high priority regions and increase the total variance explained.

We sought to contribute to the current literature by using a dense gene-centric approach with the IBC array to identify novel loci associated with lipid traits that have not been

discovered using more conventional approaches. A flow diagram of the performed analyses is illustrated in Figure 2.1.

MATERIALS AND METHODS

Participating studies

We analyzed individual-level phenotype and genotype data from 22,471 individuals of European descent in seven cohorts and an additional 25 cohorts contributed summary-level results for 43,769 individuals, yielding a total sample size of 66,240 (Supplementary Table 1a). Five additional cohorts containing data from a total of 25,282 individuals were used for replication (Supplementary Table 1b). Further replication was sought through the GWAS meta-analysis described by the GLGC[15]. In addition to the genotype data, we obtained data on body-mass index (BMI), age, gender, type 2 diabetes (T2D) status, smoking history and, where available, on any treatment for dyslipidemia. Informed consent for DNA analysis was received from each respective local institutional/national ethical review board.

Lipid phenotype definitions and correction for lipid-lowering drug use

Lipid measurements from blood samples collected at baseline or first measurement of each study were used for analysis. We restricted the analyses to those older than 21 years, as lipid levels are unstable prior to this age[53]. Lipid samples were categorized as "known fasting", "non-fasting" or "undefined". Concentrations were converted from mg/dL to mmol/L by dividing by 38.67 for TC, LDL-C and HDL-C measurements and by dividing by 88.57 for TG measurements. With the exception of the PROCARDIS study, where direct LDL-C assay was used (CHOD/PAP assay in an Olympus AU5430[54]), LDL-C concentration was calculated according to Friedewald's formula (LDL-C = TC -HDL-C -kTG) where k is 0.45 for mmol/L (or 0.20 if TG were measured in mg/dl). LDL-C was treated as a missing value if TG values were $>$ 4.51 mmol/L ($>$400 mg/dL)[55]. Prior to analysis, TG levels were transformed using the natural logarithm (ln) to normalize its distribution. For individuals receiving lipid-lowering therapy, we multiplied recorded lipid values by a constant: TC was multiplied by 1.271; LDL-C by 1.352; HDL-C by 0.949, and TG by 1.210, prior to transformation. The multiplicative correction

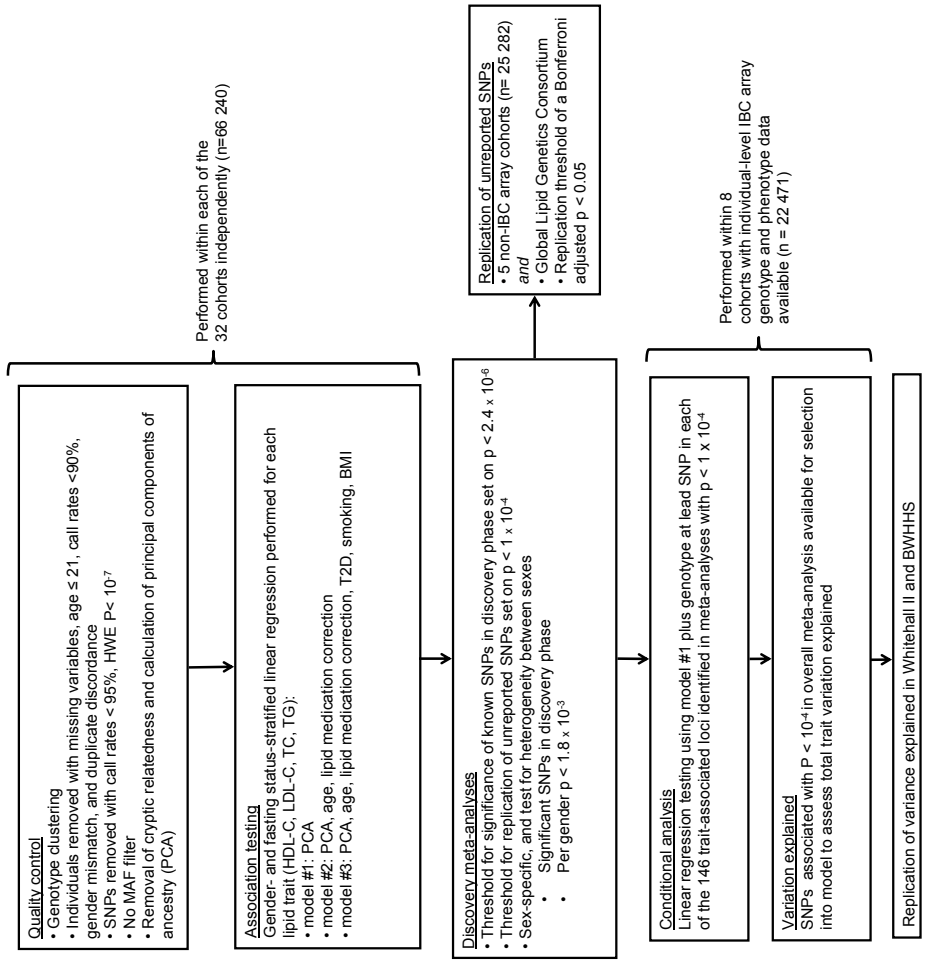

Figure 2.1: Summary of the Design Used and the Number of Individuals Involved and p Value Thresholds Used in Each Step

factors were based on analysis of repeatedly measured lipid levels, including levels measured before and after lipid-lowering treatment, in the Whitehall II (WHII) study[56] as follows. The expected difference between two data collection time points 5 years apart, for each lipid phenotype was estimated among participants of the WHII study who were not on lipid lowering therapy. The mean difference between the expected and observed values for those receiving medication at the latter but not former collection phase was calculated and used as the respective correction factor. The correction factors used here are comparable to published estimates for the effects of statins on lipid values from treatment trials[57].

Genotyping and quality control

Genotyping was performed using the gene-centric IBC array (Illumina HumanCVD)[17]. We used genotyping data from the first two versions of the IBC array. Version 1 of the array captures 45,238 SNPs while version 2 contains an additional 3,989 SNPs comprising a total of 49,227 SNPs. These were clustered into genotypes using the Illumina BeadStudio software. Quality control filters were applied within each cohort at the sample and SNP levels. The filter requirements for the meta-analysis, as sent to each study, required the removal of individuals with call rate $<90\%$, gender mismatch or duplicate discordance. SNPs with call rates $<95\%$ or Hardy-Weinberg equilibrium (HWE) deviation at chi-squared $p < 10^{-7}$ were also removed. No filtering was performed on low minor allele frequency (MAF) variants at this stage to take advantage of the rare variants captured on the array and the large number of samples available.

Evaluation of cryptic relatedness

Only founders within cohorts with recorded family structure were included in the analysis, with the exception of the GRAPHIC, HAPI and PROCARDIS studies in which family structure was maintained but adjusted for in the association analysis. To ensure removal of cryptic relatedness and duplicate samples, pi-hat, a measure of identity-by-descent (IBD), was estimated from the pairwise identity-by-state (IBS) using PLINK[58]. Plink is a computationally efficient open-source analysis toolset for genetic data, able to perform a series of basic, large-scale analysis. For each set of duplicates or monozygotic twins, and for those with a pairwise pi-hat > 0.3, the sample with the highest genotyping call rate was retained for analysis.

Evaluation of population stratification

For the primary analysis, only individuals of European ancestry were included. Self-reported ethnicity was verified by multidimensional scaling (MDS) analysis of IBS distances as implemented in PLINK, using HapMap panels as reference standards. After removing SNPs in pairwise linkage disequilibrium ($r^2 > 0.3$), EIGENSTRAT software was used to compute principal components on the subset of non-excluded individuals for use as covariates in the regression analyses. EIGENSTRAT, part of the EIGENSOFT package, through the use of principal components, corrects for variation in the frequency across ancestral populations, minimizing potential false positives signals due to population structure while increasing the power to detect true associations[59, 60]. Analysis in non-European ancestry participants is reported in an accompanying paper[61].

Thresholds for declaration of statistical significance

Taking linkage disequilibrium (LD) into account, it has been previously calculated that genotyping with the IBC array generates ~20,500 independent tests for individuals of European descent[62]. To maintain the conventional 5% false positive rate, the appropriate multiple testing corrected threshold for statistical significance was set at p = $1.23x10^{-6}$ for the primary analysis[50, 63]. When the individual level data were used in the analysis, as in the conditional analysis and variable selection, or replication was available, we used a more permissive p-value threshold of p $< 1.0x10^{-4}$. To maintain our statistical power of ~80% for a SNP with an effect size r^2 of 0.05% during the gender specific analysis, we used a gender specific threshold of p $< 1.8x10^{-3}$. Since the SNPs included in the gender specific analysis were previously considered significant in the primary main effect analysis, this choice has little effect on false positives. The GWAS threshold of p $< 5x10^{-8}$ was referenced as a comparison to common GWAS practice.

Genomic control estimates reflected by lambda (λ), a method to quantify and adjust population stratification from population-based samples[64], were derived for each study before the meta-analysis. To avoid the problem of λ estimates inflation due to the high proportion of positive variants, based on the selection criteria of the included SNPs and loci, we excluded the upper 10% of the most statistically significant signals during the estimation of λ[65]. METAL used the option to adjust each study with its corresponding λ before the meta-analysis.

Association testing

Association analysis was performed using an additive genetic model with one degree of freedom for all cohorts. We performed gender stratified analysis within each study for the following three models: Model 1 corrected only for population stratification to filter out any artificial association related to population differences; Model 2 corrected for population stratification, age and lipid-lowering medication, using the correction factors described above because those two extra variables are believed to affect the relationship between the traits and the genotypes tested; and Model 3 corrected for population stratification, age, T2D, smoking, BMI, and lipid-lowering medication as described earlier to further control for additional variables able to influence the observed associations. The main results and conditional analysis were reported based on model 1. Variable selection used signals from all 3 models and these were maintained for the variance explained section. All three models were also considered in our scan for previously unreported signals and the replication of previously published associations. The three models were also used as means to understand the associations observed when additional factors were controlled. Meta-analysis was performed with METAL[66] and the results were verified using MANTEL[67] and the Metafor package in R[68]. METAL was run with the options to use the p-values for the meta-analysis taking sample size and direction of effect into account, while MANTEL used the classical approach of meta-analysis with a fixed-effects model[67] and Metafor used a random effects scheme with the Hunter-Schmidt estimator[69]. Reported p-values are based on METAL unless otherwise stated. The use of the probability combination option in METAL does not include the meta-analysis of beta coefficients, although it is able to overcome the problems of differences in phenotype distribution and gender between the studies combined[66]. Metafor used a random effects model that considered differences between studies as part of the heterogeneity adjusted in the model[70], hence given the number of available studies and the difference between them, the beta coefficients from Metafor were considered as the most accurate estimations of the underlying "true" effects of the SNPs, and are presented throughout.
Following the main analysis, we tested for gender specific signals of associations performing the meta-analysis separately for males and females and combining their results. Only SNPs deemed statistically significant in the overall analysis were compared between genders for evidence of heterogeneity of effect. Heterogeneity of the meta-analysis was assessed using the I^2 statistic, which describes the percentage of total variation in the study estimates that is due to the differences between studies. The statistical significance of the heterogeneity was tested by the chi-squared heterogeneity statistic[71]. The criteria for selection of SNPs were: heterogeneity p-value < 0.05 between males and females

and gender-specific p-value $< 1.8x10^{-3}$. When SNPs were in LD, $r^2 > 0.3$, only the strongest associated SNP is presented.

Conditional analysis

Loci harbouring significant evidence for association with p $< 10^{-4}$ in Model 1 were examined for additional signals using conditional analyses in PLINK[58] in data from seven cohorts of European ancestry in which individual-level genotype data were available. A term was added to the regression model including the lead SNP as a covariate, and SNPs within the same genomic region (within 1Mb of the lead SNP) were evaluated for significance. A locus-specific Bonferroni correction, based on the number of tests performed, was then applied to determine the significance of independent signals[50]. For loci harbouring more than one independent signal, we continued the process until no unreported SNP associations were found.

Variable selection and variance explained

Variable selection was used to identify the most informative SNPs to estimate the total phenotypic variance in the lipid phenotype after age and gender adjustment. To avoid removing individuals with missing data from the analysis variable selection was performed in the individual-level data after imputation of any missing genotypes using fastPHASE, a package for haplotypic reconstruction and estimation of missing genotypes[72]. All SNPs with lipids associations at p $< 1.0x10^{-4}$ for any of the meta-analysis algorithms were included in the selection procedure. The previously reported GWAS SNPs for each lipid phenotype were obtained from both the NHGRI Catalogue of Published GWAS[73] and the Global Lipid Genetics Consortium (GLGC) publication[15]. All HumanCVD Beadarray SNPs within 500kb of the reported SNPs were identified using the SNAP[48] web tool and SNPs with the highest LD for each single reported polymorphism were forced into the model. The stepwise selection scheme with the Akaike's Information Criterion (AIC)[74] was implemented in R, separately for each chromosome.

Given that the SNP selection was performed in the available individual level data, including information on previously reported polymorphisms, an estimate of association in the same sample may lead to overestimation of the true effect. Therefore, unbiased estimates of the true variance explained were obtained in the Whitehall II study (WHII) and British Women Heart Health Study (BWHHS) which did not contribute individual level data.

The ratio of phenotypic variance explained by our results, taking into account the number of SNPs used and the number of observations, was further compared to that estimated using only the previously reported SNPs. For comparison with previous GWAS, we also estimated the variance explained using the top SNP at each locus plus the independent SNPs in the region as identified through conditional analysis.

Replication of non-previously described signals/Signals not previously reported

We report two categories of associations: firstly SNPs at established loci that have not been reported previously and secondly non-previously described loci, using the less stringent statistical threshold of p $< 1.0x10^{-4}$. Loci were designated novel if they had not been reported in the NHGRI GWAS database or in GLGC [15], and novel SNPs were those that were not reported in GLGC[15] and with $r^2 \leq 0.3$ to any of the GLGC lead SNPs. Loci within 500kb of reported signals were not considered novel. To attain the final list of novel SNPs, we checked for LD between SNPs within the list itself. In groups of SNPs in LD $r^2 \leq 0.3$, the SNP with the lowest p-value was reported.

Replication

Independent replication was then sought for all associations not previously reported. Look-ups were performed in five additional cohorts containing data from a total of 25,282 individuals. Characteristics and methodological details for cohorts (referred to as '25K cohort') are listed in Supplementary Table 1b. Additional replication was sought through the GLGC GWAS meta-analysis[15]. A signal was considered successfully replicated when its Bonferroni adjusted p-value in the replication sample was lower than 0.05, and its estimate directionally consistent with the discovery meta-analysis. Four of the studies used for this meta-analysis (KORA F3, ARIC, PennCATH, CHS and BRIGHT) had previously contributed data to GLGC. These studies were thus removed from the meta-analysis of the discovery with both replication studies.

RESULTS

Characteristics of study samples

A total of 49,227 SNPs were tested in a meta-analysis of 32 cohorts of 66,240 individuals of European ancestry (Supplementary Table 1a). The ratio of the observed to the null median test statistic, λ, was ≤ 1.1 for all studies, except for GRAPHIC, HAPI and PROCARDIS, where related individuals were included. GRAPHIC had a λ of 1.2 for all phenotypes considered, which decreased to ~1.06 when rare variants (MAF<0.1%) were excluded from the data. Both HAPI and PROCARDIS had λ values of ~1.10 and 1.12 respectively for LDL-C and TC, but these again decreased to ≤ 1.1 when rare variants (MAF<0.1%) were excluded.

Meta-analysis

We observed 598 statistically significant associations with HDL-C, 491 with LDL-C, 575 for total cholesterol and 609 for TG, at p $< 2.4x10^{-6}$ using Model 1 in METAL (Supplementary Table 2). After excluding SNPs present in less than 80% of the studies and filtering associations with a meta-analysis I^2 value for heterogeneity $> 35\%$, the number of statistically significant SNPs was reduced to 276 for HDL-C, 158 for LDL-C, 269 for TC and 242 for TG (Supplementary Table 3).

Of the 2273 statistically significant associations before filtering, 1094 were with SNPs of MAF $< 1\%$ of which 1088 had an I^2 value of $> 35\%$ (Supplementary Table 2). In total, given that several SNPs were associated with more than one phenotype and SNPs clustered tightly at certain loci, we identified 549 study-wide significant SNPs in and around 114 different genes. Supplementary figure 1 shows the overlap of the identified signals between traits; Manhattan plots for each phenotype are shown in Figure 2.2.

We also analyzed the data using MANTEL and the Metafor package in R. Although each algorithm used a slightly different method for the meta-analysis, of the 100 top signals for all phenotypes 98% were also significant in Metafor and 95% significant in MANTEL. Of all the 945 filtered significant associations observed in METAL 78% were also significant in Metafor and 80% in MANTEL (Supplementary Table 3). The differences in the results between the three packages were mainly observed either in associations with SNPs of low frequency, high heterogeneity, or minor differences in statistical significance which were close to our cut-off thresholds.

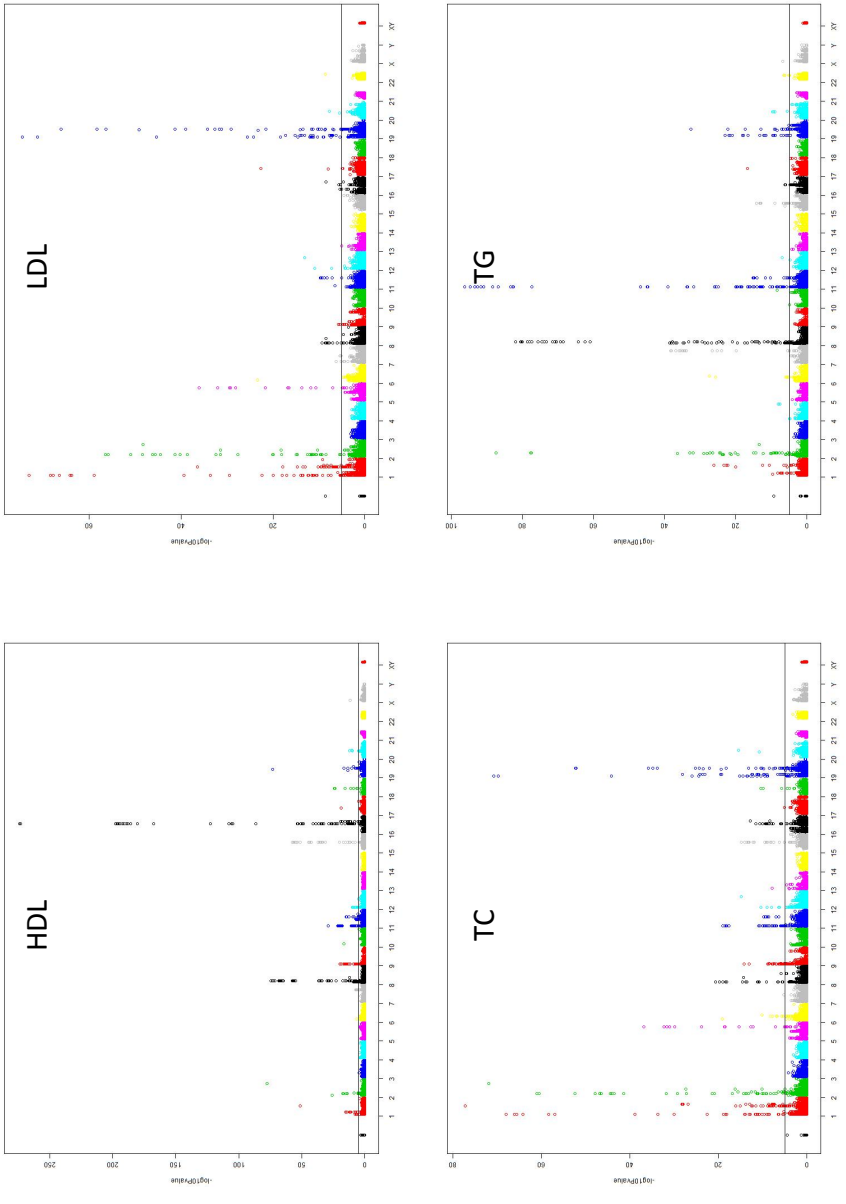

Figure 2.2: Manhattan plots for HDL-C, LDL-C, TC and TG data from the International IBC lipid meta-analysis based on estimates obtained using a p-value based meta-analysis in METAL.

The meta-analysis of plasma lipid levels, corrected for lipid-lowering medication, and adjusted for age (Model 2) or BMI, T2D and current smoking status (Model 3) gave similar results. A summary of the differences between the three models in terms of gene loci identified is shown in Figure 2.3. The Pearson correlation coefficients of the test statistics between Model 1 and Model 2 varied between 0.86 and 0.92 and between 0.85 and 0.89 for Models 1 and 3 for the four phenotypes. The correlation between Models 2 and 3 was 0.94-0.97.

Conditional analysis

Because of the dense gene-centric nature of the IBC array, SNPs showing association unsurprisingly formed tight clusters. We examined 39 loci for HDL-C, 34 loci for LDL-C, 41 loci for TC and 32 for TG with $p < 10^{-4}$. Conditioning on the SNP with the strongest p-value, we identified independent signals for four traits (Supplementary Table 4). Although four independent signals were observed in the LPL (MIM 609708) locus and three signals in the BUD13-APOA5 (MIM 606368) cluster for HDL and TG, only two SNPs in the LPL locus, rs268 and rs3289, were overlapping between the phenotypes.

Gender specific analysis

Of the 66,240 individuals included in the meta-analysis, 31,513 were males and 34,727 females. The data were analysed with stratification by gender in each cohort and the first stage meta-analysis included both genders. We also tested each gender separately and the results were compared for concordance between genders. All of the SNPs showing heterogeneity of effect between males and females had the same direction of effect with the overall analysis, but with one gender showing a significantly weaker association compared to the other. Gender-specific differences were found for the individual lipid traits (HDL-C; 14 SNPs, LDL-C; 9 SNPs, TG; 14 SNPs and TC; 9 SNPs, Supplementary Table 5).

Variance explained

Variable selection used all unfiltered signals, including 1,156 SNPs for HDL-C, 1,063 for LDL-C, 1,173 for TC, and 1,139 for TG, identified as significant at the $p < 10^{-4}$ threshold from any of the three meta-analysis algorithms. Additionally, previously reported

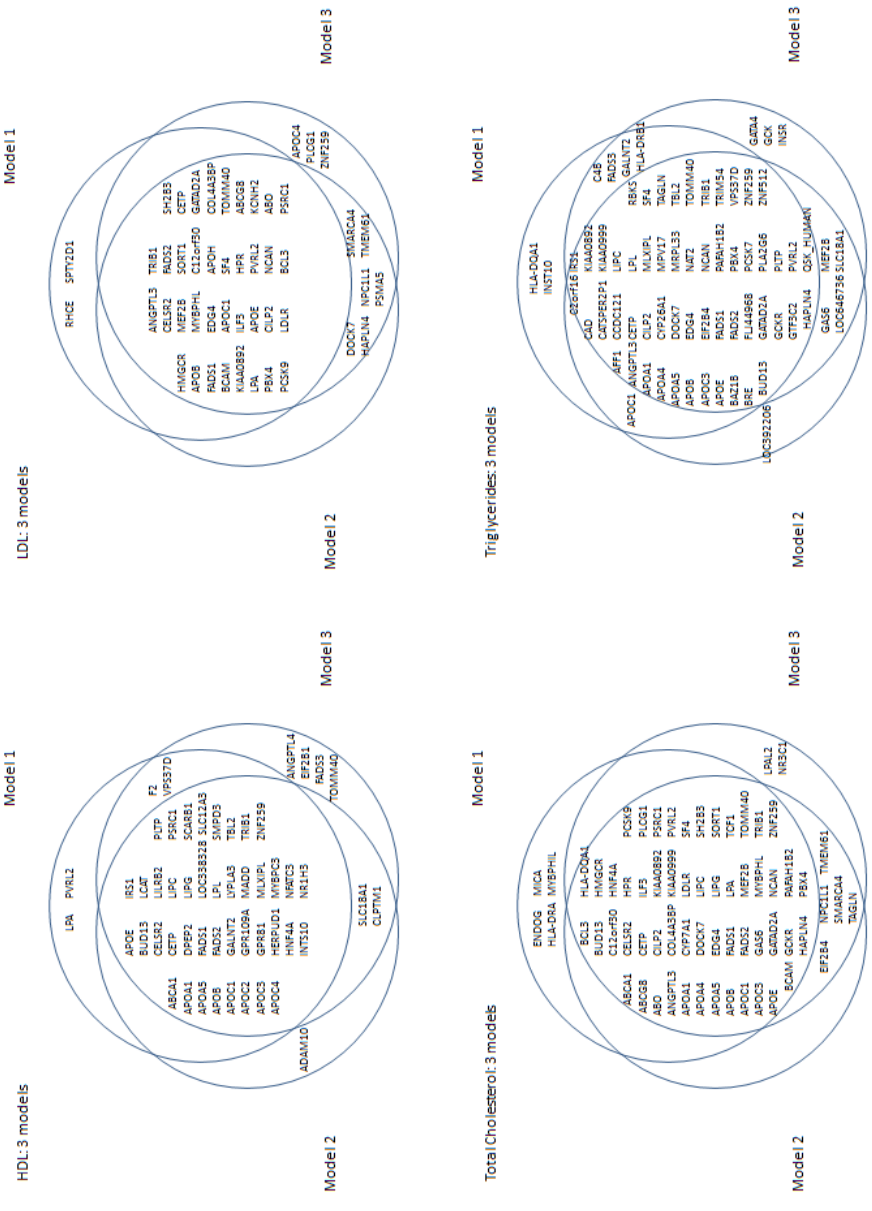

Figure 2.3: Venn Diagram per Phenotype for the Comparison of the Three Models Used.

SNPs were forced into the model including: 41 SNPs for HDL-C, 25 for LDL-C, 22 for TC, and 35 for TG. Based on the AIC evaluation, 71 SNPs in HDL-C, 79 in LDL-C, 120 in TC, and 75 in TG appear to carry additional information beyond the previously reported SNPs. All SNPs retained after variable selection, including the previously reported SNPs that were forced into the model, are described in Supplementary Table 6.

Using the list of the SNPs identified by variable selection, we estimated the percentage of phenotypic variance explained in the subset of studies providing individual level data. After adjustment for age and gender, the SNPs identified explained 9.9% (using 112 SNPs) of the variance in HDL-C, 9.5% of the variance in LDL-C (using 104 SNPs), 10.3% (using 142 SNPs) of the variance in TC and 8.0% (using 110 SNPs) of the variance in TG. Using data derived from previously reported lipid-associated SNPs available in the IBC array, we observed much lower percentages: 6.3% for HDL-C, 4.8% for LDL-C, 4.1% for TC and 5.5% for TG. For comparison, using the common approach of including only the top signal from each locus plus any independent SNP after conditional analysis, we were able to explain 7.9% of the HDL-C variance, 8.4% of the LDL-C variance, 8.2% of the TC variance and 6.3% of the TG phenotypic variance.

To avoid overestimation resulting from using the same datasets for SNP selection and testing, we also estimated the variance explained in the WHII and BWHHS studies that did not contribute individual level data used in the variable selection. For the WHII study, the AIC selected SNPs explained 11.5% of the variations in HDL-C, 15.6% of the variations in LDL-C, 13.2% of the variations in TC and 9.8% of the variations in TG variance, while the previously reported SNPs explained 7.9%, 8.2%, 6.7% and 7.4% of the phenotypic variance in each of these phenotypes, respectively. The corresponding estimates for the BWHHS were 8.2% for HDL-C, 10.7% for LDL-C, 8.1% for TC, and 8.2% for TG when all the selected SNPs were included in the analysis and 6.0% for HDL-C, 4.2% for LDL-C, 2.6% for TC and 5.7 for TG when only the previously reported SNPs were considered. The estimated variance explained, approximate to the heritability due to additive genetic effects, separately for males and females was 10.9% and 12.2% for HDL-C, 12.8% and 11.53% for LDL-C, 12.7% and 12.6% for TC, and 10.7% and 8.6% for TG respectively (Supplementary Table 7).

Confirmation of previously reported signals

The IBC array covered 57 of the 95 loci reported in GLGC[15], and did not include two of the top 20 HDL-C loci (KLF14 (MIM 609393) and LILRA3 (MIM 604818)), two of the top 20 LDL-C loci (TOP1 (MIM 126420) and ST3GAL4 (MIM 104240)),

none of the top 20 TC loci, and two of the top 20 TG loci (KLHL8 (MIM 611967) and FRMD5). Among the directly genotyped SNPs in the IBC array, we were able to replicate the association of 13 out of 18 SNPs with HDL-C, 11 of 21 with LDL-C, 16 of 26 with TC and 9 of 18 with TG, each at a threshold of $p < 10^{-4}$. Similarly, for the previously reported loci, the lowest p-value SNP in our results replicated 23 of 32 available loci for HDL-C, 23 of 32 available loci for LDL-C, 30 of 43 available TC loci, and 21 of 29 loci for TG. Out of the 57 loci cited above, 31 had specifically the same SNP genotyped by both the GLGC and the IBC array. For those, thus, there was information on directions of effect for both GLGC and IBC array. In total, these represent 49 signals (given that one SNP can be significant for more than one trait); 13 of those were significant for HDL-C trait, 11 for LDL-C, 16 for TC and 9 for TG. Only one SNP, rs12027135, from gene LDLRAP1 (MIM 605747), was found with an opposite direction and non-statistically significant effect than that of GLGC, for two traits (LDL-C and TC). We identified additional significant associations for LCAT (MIM 606967), LRP1 (MIM 107770), LPA (MIM 152200), IRS1 (MIM 147545), PCSK9 (MIM 607786) loci at the GWAS p-value cut-off of $p < 5.0x10^{-8}$ in addition to the association reported in previous studies (Supplementary Table 8).

Signals not previously reported

We identified 48 significant SNPs associations in novel and previously reported genes for HDL-C of which 17 are in genes not previously reported, with $p < 1.0x10^{-4}$. For LDL-C, we identified 38 significantly associated SNPs that were not previously reported in either the NHGRI GWAS database or in GLGC[15]. Of these, 18 were located within genes without any previously annotated effect on LDL-C. Similarly, for TC we observed 47 SNPs not previously reported in established genes and 15 SNPs in previously undescribed genes. Finally, for TG, we observed 49 associations, including signals in the 18 genes not previously reported. Assuming the array-wide significance level of $p < 2.4x10^{-6}$, there were 11, 5, 12 and 6 novel SNPs for HDL-C, LDL-C, TC and TG respectively. Several loci not previously reported were observed with an array-wide significance of $p < 2.4x10^{-6}$ (Supplementary Table 9).

Replication

SNPs showing a significant association with a $p < 1.0x10^{-4}$ that were not previously reported to be associated with lipids were considered for replication in additional studies.

These were examined either in our own replication sample of the 25K cohort or using the GLGC data. In total, 23 of the total 69 putative novel gene signals identified in Stage 1 were found to be significantly associated in the replication stage. Three of these SNPs reached a GWAS level of significance ($p < 5x10^{-8}$) in the discovery phase and two replicated (67%), nine more reached our array-wide significance and five replicated (56%), while a further 57 surpassed the permissive 10^{-4} cut-off of which sixteen signals were replicated (28%) Of the all signals tested, 11 associations were replicated for HDL-C, 11 for LDL-C, 17 for TC and 12 for TG levels. These replicated signals were in 21 gene regions not previously reported as associated with the lipids phenotypes considered here. A total of 23 signals were replicated in the 25K cohort and/or GLGC, 5 signals for HDL-C, 5 for LDL-C, 7 for TC and 6 for TG (Supplementary Table 9). Details of the lead SNPs replicated in each of the novel genes are provided in Table 2.1 together with the results of the overall meta-analysis. Additional, not-previously reported, SNPs in known loci were also identified. Four SNPs were associated with HDL-C, 6 with LDL-C, 10 with TC and 4 TG. The results for all SNPs and loci tested for replication are presented in Supplementary Table 9.

DISCUSSION

We used a large scale locus-centric approach, testing 49,227 SNPs carefully prioritized for CVD-related loci, in 32 studies with a combined discovery sample size of up to 66,240 individuals of European ancestry, to explore association with HDL-C, LDL-C, TC and TG levels. Using an additional sample of 25,282 individuals and the available data derived from the GLGC study[15], we identified 21 additional loci that have not been associated with lipid levels before and were able to confirm a number of the previously reported associations. We also observed gender specific differences in multiple loci that were identifiable at the level of variance explained. Finally, although the array covers a smaller proportion of the genome, our heritability estimates were comparable to current GWAS estimates.

Recently, candidate loci and gene-based arrays such as the IBC array, and the "Metabochip", "Immunochip" and "exome-chip" with content derived from GWAS, next generation sequencing and other plausible sources such as functional studies, are becoming increasingly popular and offer significant value to individual investigators and consortia [75, 76, 18]. They allow flexibility to incorporate index SNPs, as well denser probe coverage for finer mapping, across a large number of loci allowing selective coverage for a range of prioritized findings. The GLGC study[15] provides the most current analysis of the entire genome for common polymorphisms that underpin circulating concentrations of HDL-C, LDL-C, TC and TG. Working under the hypothesis that any individual SNP tested is unlikely to have a true effect on the particular phenotype of interest, GLGC was able to identify SNPs explaining 0.05% of the phenotypic variance with 94.75% power, while in our data the same SNPs had 85.07% power to explain the same proportion of the phenotypic variance. This may in part explain our inability to replicate the entire set of SNPs identified in the GLGC study. Nevertheless, these differences in power do not take into account the fact that, compared to the GWAS, the SNPs tested here were more likely to be associated with the phenotypes tested due to the selection of SNPs based on available information concerning a putative role in lipid metabolism. This is apparent from the extreme quantile-quantile QQ plots seen in hypothesis-driven arrays compared to typical GWAS. Such candidate loci arrays also allow "cosmopolitan tagging" approaches to ensure sufficient markers across loci of interest for multiple ancestries are achieved. Much of the lipid content of the 2,000 loci on the IBC array was derived from pathway based candidates[17]. Fifty-seven loci present on the IBC array were associated with lipid traits in the GLGC, although importantly at the time of array design very few loci were shown to be robustly associated with lipid phenotypes, showing the clear utility of such candidate loci approaches to generate putative candidates for validation in large numbers of individuals. One shortcoming of hypothesis-driven genotyping arrays is limited cover-

Table 2.1: Replicated Previously Unreported Genes

						Discovery							25K replication					GLGC replication			Overall meta-analysis
Gene	Trait	Model	SNP	Chr	Position	Allele 1	Allele 2	Freq 1	Weight	P-value	Direction	Het P-Value	Weight	P-value	Direction	Het P-Value	Mult.adj P-value	GLGC P-value	Direction	GLGC Mult.adj.P-value	
PPARG	HDL	3	rs12631819	3	12317861	T	G	0.0285	53187	1.72E-05	-	0.2150	10567	2.42E-01	+	0.6780	1.00E+00	2.27E-04	-	1.64E-02	8.35E-06
GP1HBP1	HDL	2	rs7388248	8	144376728	C	G	0.2878	54144	3.30E-07	+	0.7494	14374	1.42E-02	+	0.5845	1.00E+00	4.03E-04	+	2.90E-02	2.93E-09
DGAT2	HDL	1	rs11236530	11	75167052	A	C	0.4246	61617	2.44E-06	-	0.6566	24732	3.80E-05	-	0.3281	2.74E-03	4.92E-03	-	3.54E-01	1.40E-09
HCAR2	HDL	3	rs4759361	12	121744233	A	T	0.1584	53194	2.05E-09	+	0.8481	13902	6.24E-01	+	0.2468	1.00E+00	1.00E-06	+	7.23E-05	2.50E-12
FTO	HDL	2	rs1421085	16	52358455	T	C	0.6095	54286	5.76E-05	+	0.1952	14373	1.08E-01	+	0.8410	1.00E+00	3.31E-05	+	2.38E-03	2.33E-08
VLDLR	LDL	2	rs7024888	9	2626992	T	C	0.9523	48970	4.38E-05	+	0.8693	2998	1.66E-01	+	0.4035	1.00E+00	4.95E-05	+	3.02E-03	7.51E-08
SPTY2D1	LDL	1	rs11024739	11	18602419	A	C	0.5514	55610	3.09E-07	+	0.3205	24393	7.49E-01	+	0.5482	1.00E+00	4.14E-06	+	2.53E-04	6.04E-10
BRCA2	LDL	3	rs9534275	13	31838345	A	C	0.5149	47864	4.61E-06	-	0.4505	13886	8.26E-01	+	0.4184	1.00E+00	2.48E-07	-	1.51E-05	5.67E-09
SOCS3	LDL	2	rs4082919	17	73889077	T	G	0.5148	48687	2.33E-05	+	0.5148	14347	1.17E-03	+	0.1111	4.45E-02	3.58E-01	+	1.00E+00	4.46E-04
APOH	LDL	1	rs1801689	17	61641042	A	C	0.9646	55363	2.80E-11	-	0.6439	2999	5.56E-01	-	0.3226	1.00E+00	2.10E-05	-	1.28E-03	9.41E-14
C4B	TG	1	rs389883	6	32055439	T	G	0.6323	57442	8.63E-07	+	0.9931	14638	3.81E-02	+	0.9599	1.00E+00	3.95E-15	+	2.85E-13	9.90E-19
LPAL2	TG	3	rs3123629	6	160826076	A	G	0.3445	47279	3.76E-05	+	0.5835	13935	5.23E-01	+	0.5929	1.00E+00	1.20E-06	+	8.60E-05	4.40E-08
GCK	TG	3	rs2070971	7	44164108	T	G	0.132	49207	2.23E-06	+	0.2751	13900	1.18E-02	+	0.8145	8.46E-01	3.08E-04	+	2.22E-02	2.25E-07
GATA4	TG	3	rs6983129	8	11628545	A	C	0.4918	49262	2.30E-06	+	0.9860	13940	2.02E-01	+	0.2548	1.00E+00	7.54E-06	+	5.43E-04	2.62E-08
SERPINF2	TG	3	rs2070863	17	1595252	T	C	0.2109	49243	3.73E-05	+	0.6577	13903	4.41E-04	+	0.3132	3.18E-02	5.95E-03	+	4.28E-01	2.41E-05
INSR	TG	1	rs8112883	19	7130320	T	G	0.3269	57525	8.55E-06	-	0.2601	14635	4.61E-01	-	0.8751	1.00E+00	2.85E-05	-	2.05E-03	4.76E-06
FCGR2A	TC	3	rs1801274	1	159746369	A	G	0.5016	53200	2.25E-05	+	0.9812	13976	2.54E-03	+	0.5643	2.04E-01	1.58E-04	+	1.26E-02	3.67E-08
INSIG2	TC	3	rs12464355	2	118566320	A	G	0.9263	53171	6.26E-05	+	0.6811	13851	7.66E-03	+	0.4101	6.13E-01	8.69E-05	+	6.96E-03	1.19E-08
UGT1A1	TC	1	rs11563251	2	234344123	T	C	0.0934	65731	3.46E-06	+	0.0242	14678	2.56E-01	+	0.0790	1.00E+00	1.50E-06	+	1.20E-04	2.50E-09
CHUK	TC	3	rs11597086	10	101943695	A	C	0.5885	52928	3.54E-05	-	0.4075	13998	2.05E-01	-	0.3244	1.00E+00	5.81E-05	-	4.65E-03	2.99E-09
UBE3B	TC	3	rs7298565	12	108421917	A	G	0.5254	48938	1.47E-05	+	0.8893	13997	7.93E-01	+	0.3195	1.00E+00	1.25E-04	+	1.00E-02	1.98E-06
BRCA2	TC	2	rs9534275	13	31838345	A	C	0.5157	54094	4.16E-06	-	0.6331	14407	3.59E-01	+	0.3510	1.00E+00	2.47E-05	-	1.98E-03	5.39E-07
SOCS3	TC	2	rs4082919	17	73889077	T	G	0.5159	54065	1.22E-05	+	0.7658	14415	1.61E-04	+	0.0394	1.29E-02	9.56E-03	+	7.65E-01	8.91E-07

The following abbreviations are used: chr, chromosome; freq, frequency; het, heterogeneity; mult adj, multiple-variate-adjusted; GLGC, Global Lipids Genetics Consortium; HDL, high-density lipoprotein; LDL, low-density lipoprotein; TGs, triglycerides; and TC, total cholesterol.

age, with 38 out of the reported 95 previously reported GLGC loci not represented on the IBC array. Despite this, the great majority, 74 out of 80, of the strongest associations were covered with greater density than before, highlighting the utility of such approaches. Furthermore, aggregation of datasets such as those presented here, have clear utility for conditional analyses, as we show that 27 loci have more than one independent signal for the examined lipid traits.

Our most significantly associated locus for HDL-C was CETP (MIM 118470). CETP is a hydrophobic glycoprotein, which, upon secretion by the liver, is bound mainly to HDL particles in plasma[77], CETP inhibitors have been shown to significantly increase plasma HDL-C levels, thereby mimicking the hyperalphalipoproteinemia encountered in patients with CETP deficiency[78]. For both LDL-C and TC, LDLR (MIM 606945) (Low density lipoprotein receptor) had the lowest p-value. LDLR encodes the cell surface LDL receptor which removes circulating LDL via receptor-mediated endocytosis. More than 1600 rare, lose-of-function mutations in the LDLR have been shown to cause familial hypercholesterolemia[79, 80, 81]. Finally, the locus with the strongest association with TG levels was BUD13 (functional spliceosome-associated protein 71), which is located at the same chromosome 11 locus that contains the APOA1-C3-A4-A5-ZNF259 (MIM 107680) cluster. In the GLGC GWAS meta-analysis, the top hit for TG, APOA1(MIM 107680) rs964184, lies within the BUD13 promoter. BUD13 is a yeast homolog of an active spliceosome, but little is known about its function in humans. Two of the encoded apolipoproteins within the cluster, apoA-V and apoC-III, influence the activity of lipoprotein lipase (LPL) activity, which is central to hydrolysis of circulating TG-rich lipoproteins. Variants in these genes have long been associated with clinical hypertriglyceridemia[82, 83].

Two of our top signals, CETP and BUD13 show evidence of gender-specific effects. A wide variety of phenotypes, including CHD, demonstrate sexual dimorphism[84]. Thus some of the strongest signals we found might be important in one gender alone. An illustrative example is CETP, for which SNPs rs17231506 and rs12720922 were both differentially associated with HDL-C levels in men and women. This relationship has been previously suggested, and gender-specific differences in expression levels of the gene product were hypothesized[85]. Other previously reported SNPs, also shown here to have gender-interactions, include the three APOB (MIM 107730) SNPs rs531819, rs17398765, rs1367117[86], rs4953023 in ABCG8 (MIM 605460)[87, 88, 89], rs157580 in TOMM40 (MIM 608061) which is close to APOE (MIM 107741)[90], and the APOC4 (MIM 600745) SNPs rs12721109 [91]. In addition to gender differences in associations with individual SNPs, we observed between-sex differences in trait heritability. Of the four lipid phenotypes examined LDL-C and TC had minimal between-gender differences in heritability of 1.06% and 0.2% respectively, while females showed higher heritability

for HDL-C (1.5% difference) and males showed higher heritability for TG (1.9% difference). Our results, except those for TG, are similar to those reported by Weiss et al[42], with LDL-C showing small narrow sense heritability differences while females had higher narrow sense heritability compared to males in HDL-C. In contrast to our findings, Weiss and colleagues[42] showed a stronger but non-significant heritability in females compared to men for TG.

Of the 49,227 SNPs in the array ~21% had a MAF $< 1\%$, while in the 2.273 unfiltered significant associations observed 48% had a MAF $< 1\%$. After filtering, 0.06% of the significant associations were SNPs of MAF $< 1\%$. The higher proportion of rare SNPs passing the array wide p-value threshold compared to their proportion on the array can be attributed, in part, to their high heterogeneity. In the majority of cases, SNPs with genotyping errors show high levels of heterogeneity between studies. This might suggest that uncommon SNPs are more difficult to successfully genotype or call with current methods. A technical note from Illumina[92] reported that accurate calling of rare variants is possible, although there is an increase in the error rate for rare allele homozygotes. It is also possible that carriers of rare functional SNPs will have an extreme phenotype, leading, in some cases, to exclusion from the study or to a greater measurement bias in some studies compared to others. At least some of the rare SNPs in our results are known to have functional mutations with large effects. APOB (MIM 107730) SNP rs5742904 has a p-value of $1.039x10^{-46}$ with LDL-C in our meta-analysis but an I^2 of 96.6%. The rs5742904 SNP is a known rare mutation (R3527Q), involved in hypercholesterolaemia and early CHD[93, 94]. The mutation, which has been shown to reduce the affinity for the LDL cholesterol particle, where ApoB is the single protein component for the receptor, is present in 5% of patients with familial hypercholesterolaemia FH [MIM 143890]) in UK[95]. The identification of rare SNP associations is a substantial challenge and although we observed a number of strong probable associations, high heterogeneity precludes any firm conclusions.

Our results point towards the existence of multiple independent lipid-associated SNPs in several different loci. One example is the LPL (lipoprotein lipase) gene, in which the classical view of the primary functional importance of the S447X variant (rs328), which causes a premature stop codon has been modified by the findings that several different polymorphisms at this locus concurrently affect LPL expression[96, 97]. Interestingly, all of our top signals CETP, LDLR and the BUD13 cluster, show evidence for the existence of more than one functional SNP. Especially for the cluster around BUD13, the risk allele rs9804646-T (MAF 0.08) is on the same haplotype as the protective allele of the top SNP in the region, rs10750097, making the former identifiable only after conditional analysis. If this turns out to be the rule for the genetic architecture of lipid loci, any single identified signal at a locus will underestimate the variance explained. Future clinical use

in prediction of lipid levels will require more sophisticated approaches to fully capture information, irrespective of the levels of significance in discovery and replication studies. A number of statistical and computational criteria to select the most relevant and informative SNPs are available. Here we used the AIC criterion as a balance between being inclusive of the SNPs used while avoiding over-fitting. It is possible that the exclusive use of only the most significant, and not the most informative, SNPs is partly responsible for much of the "missing" heritability that cannot be explained by additional modest-effect common variants[98]. The truly causal polymorphisms are not always included within the genotyped SNPs making heritability estimates dependent on the LD between causal and observed SNPs[99]. Methods accounting for the total information in the area, such as selection with AIC or the approach used by Yang et al[100], will recover some of this missing information as our results suggest. In addition, use of stringent thresholds of statistical significance will exclude polymorphisms explaining a very small percentage of the variation, despite the potential impact of a great number of such SNPs. Our own results and the work by Yang and colleagues[99] suggest that common SNPs, that do not reach generally acceptable significant levels, are likely to hold additional information. Rare variants, yet undiscovered might explain some of the "missing" heritability of plasma lipid phenotypes[101], but it is not clear how much extremely rare changes can contribute towards a population measure such as heritability. Gene X Gene and gene X environment interaction can also play an important role but statistical constraints hinder their identification[101]. Transgenerational epigenetic alterations have also been suggested as possible source of heritability[102], but if they persist for many generations it is likely that they acquire LD with SNP already in the analysis[103].

For these same reasons, we were also less stringent with our criteria in pursuing potentially novel signals for downstream replication, using a $p < 1.0x10^{-4}$ threshold instead of our array-wide significance level of $p < 2.4x10^{-6}$. To avoid any increase of false positive signals the stringent Bonferroni correction was applied in our replication p-values. A very specific definition was used to declare novelty of a signal. Previously unreported loci were defined as those not described in either the NHGRI GWAS database or by the initial GLGC publication[15], and previously unreported SNPs were defined as those that were not reported in the GLGC[15] and having an $r^2 < 0.3$ with lead GLGC SNP. This led to some previously characterized SNPs and loci from candidate gene studies, which had not been replicated in any GWAS, being considered novel in our analysis. GATA4 (MIM 600576) is such an example. Although its association with TG was missed by the GWAS, evidence in mice, and recently humans, reveals that the coded protein is involved in TG absorption from the intestine and underpins plasma TG levels[104].

The most challenging aspect of evaluating large datasets is that an ideal sample for replication, which is larger than the discovery set, is extremely difficult to find in such large

meta-analysis setting were most large studies have been exhausted. We used five previously reported GWAS with and without imputed genotype data. This resulted in an uneven replication in which very few SNPs could be genotyped across the entire replication sample and most were available only in a fraction of the studies. Considering that we likely overestimated the true effect size of each SNP, in accordance with the winners curse, the power to replicate our signals in a smaller sample is markedly reduced. Nevertheless, a small number of signals were replicated, notably DGAT2 (MIM 606983) for HDL-C, SOCS3 (MIM 604176) for TC and LDL and SERPINF2 (MIM 613168) for TG. The published GLGC data evaluated $\geq$ 2.5 million SNPs (both directly genotyped and imputed), in more than 100.000 individuals and as such provided a much more reliable replication set. Based on this replication a further 21 previously unreported loci were confirmed. Moreover, we were also able to identify three SNPs which added additional information to what was previously published. The rs753381 variant is a coding non-synonymous SNP in PLCG1 (MIM 172420), considered in GLGC as part of the LDL-C association with TOP1 (MIM 126420) (rs6029526). The LD of rs753381 and the rs6029526 SNP previously reported is r^2 =0.82 but Phospholipase C, gamma 1 (PLCG1) was reported to effect cholesterol solubility in bile[105]. Therefore, we speculate that this variant may influence serum cholesterol levels through interference with the cholesterol cycle and the relevant locus for the association with LDL-C is PLCG1 rather than TOP1. SNP rs389883, in an intron of C4B (MIM 120820), is significantly associated in both our data and in the GLGC results with TG but it is not included in the GLGC reported signals. Similarly, SNP rs2244608 in TCF1 associated with LDL is only included in the ethnic analysis but not in the main results of the GWAS meta-analysis.

Well-known genes for other metabolic phenotypes were included in the replicated, previously unreported, signals such as FTO (MIM 610966) for BMI. FTO is believed to be involved in the regulation of food intake and to affect lipolysis in adipose tissue[106] while in our data FTO is also associated with HDL-C, probably through its association with BMI, as the loss of significance in Model 3 suggests (p = 0.8805). BRCA2 (MIM 600185), here associated with LDL-C, together with BRCA1 are two of the best known genes in which mutations are associated with breast and ovarian cancers[107]. The precise function of BRCA2 (MIM 600185) is unclear but the protein encoded has been implicated in a variety of processes including DNA repair and recombination, cell cycle control, and transcription[108]. Some of our other signals are already clinically significant. For example, HCAR2 (MIM 609163), also known as niacin receptor 1, is an important biomolecular target of niacin, a widely prescribed drug for the treatment of dyslipidemia, that acts primarily by inhibiting hepatic DGAT-2 (MIM 606983), lowering secretion of TG-rich lipoproteins and so increasing HDL-C levels[109, 110].

The evidence from the cumulative meta-analysis of our-data, the replication studies and the published GLGC results suggest that further ”true” signals might be found with less stringent p-values threshold. Given the recent deluge of available genetic data, we propose that a more careful examination is required of common variants of moderate and small-effects. This might help to explain portions of missing heritability, elucidating the pathways and mechanisms involved in lipid metabolism and CHD, and identifying potential loci in which rare SNPs with large effects on the phenotype can be discovered.

WEB RESOURCES:

The URLs for data presented herein are as follows:

Catalog of Published Genome-Wide Association Studies:
`http://www.genome.gov/gwastudies`
GLGC Meta-analysis Data:
`http://www.sph.umich.edu/csg/abecasis/public/lipids2010/`
Online Mendelian Inheritance in Man (OMIM):
`http://www.omim.org`
SNP Annotation and Proxy Search (SNAP):
`http://www.broadinstitute.org/mpg/snap/`

SUPPLEMENTARY INFORMATION:

Supplemental data is available electronically and includes one figure and nine tables.
DOI: `http://dx.doi.org/10.1016/j.ajhg.2012.08.032`

3 GENE-CENTRIC META-ANALYSIS OF LIPID TRAITS IN AFRICAN, EAST ASIAN AND HISPANIC POPULATIONS

Meta-analyses of European populations has successfully identified genetic variants in over 100 loci associated with lipid levels, but our knowledge in other ethnicities remains limited. To address this, we performed dense genotyping of ~2,000 candidate genes in 7,657 African Americans, 1,315 Hispanics and 841 East Asians, using the IBC array, a custom ~50,000 SNP genotyping array. Meta-analyses confirmed 16 lipid loci previously established in European populations at genome-wide significance level, and found multiple independent association signals within these lipid loci. Initial discovery and in silico follow-up in 7,000 additional African American samples, confirmed two novel loci: rs5030359 within ICAM1 is associated with total cholesterol (TC) and low-density lipoprotein cholesterol (LDL-C) (p = $8.8x10^{-7}$ and p = $1.5x10^{-6}$ respectively) and a nonsense mutation rs3211938 within CD36 is associated with high-density lipoprotein cholesterol (HDL-C) levels (p = $13.5x10^{-12}$). The rs3211938-G allele, which is nearly absent in European and Asian populations, has been previously found to be associated with CD36 deficiency and shows a signature of selection in Africans and African Americans. Finally, we have evaluated the effect of SNPs established in European populations on lipid levels in multi-ethnic populations and show that most known lipid association signals span across ethnicities. However, differences between populations, especially differences in allele frequency, can be leveraged to identify novel signals, as shown by the discovery of ICAM1 and CD36 in the current report.

INTRODUCTION

Plasma levels of circulating total cholesterol (TC), low-density lipoprotein (LDL-C), high-density lipoprotein (HDL-C) and triglycerides (TG) are associated with coronary artery disease (CAD) and are targets for therapeutic intervention[40]. Multiple environmental and genetic factors influence these plasma lipid levels, with heritability estimated to range from 0.28 to 0.78 in twin and family studies[111]. To date, $>$100 lipid-associated loci have been described, using studies mainly based on individuals of European ancestry[112]. Together, known variants affecting plasma lipid levels explain 10-12% of the total variance and 25-30% of the genetic variance[112] indicating that other loci and independent signals in established loci are likely to additionally contribute to the trait.

Lipid levels have been demonstrated to vary between ethnic groups[113]. Africans and East Asians have higher levels of HDL-C and lower levels of TG compared to Europeans[114] though the underlying mechanisms of these ethnic differences remain unknown. Genetic contributors to lipid concentrations are less well understood in non-European populations partly due to less well-powered genetic studies being attempted to date and most genotyping platforms are designed to have optimal coverage in European studies. An important first step towards understanding genetic risk across populations is to establish whether plasma lipid associated loci, identified in Europeans, span across multiple ethnicities or are population-specific. In a recent analysis, most of these known lipid loci had the same direction of association in different ethnic groups as in Europeans, despite presumed differences in linkage disequilibrium (LD) between marker and causal variants in each population[45]. Using regional LD in different ethnicities can help to refine association signals and to distinguish causal variants from correlated markers[115]. Furthermore, independent association signals in established lipid loci in one ethnicity may be useful to highlight causal signal(s) in other ethnicities.

The ITMAT-Broad-CARe (IBC) array (also referred to as the CardioChip or HumanCVD Beadchip [Illumina]) was specifically designed to densely tag ~2000 genes with known or potential roles in lipid and cardiovascular traits using ~50,000 single nucleotide polymorphisms (SNPs)[17]. Sequencing data from European, African American and Yoruba individuals was included for SNP selection in IBC array development. The IBC array drew upon knowledge of lipid metabolism and cardiovascular physiology, as well as early GWAS and sequencing studies to target efforts towards regions with higher a priori evidence of association, reducing cost per sample, and improving efficiency of replication studies. The IBC array has been successfully used for multiple cardiovascular-related phenotypes[43, 47, 44, 50]. Results are reported elsewhere for the association of lipid phenotypes in European-derived cohorts with variants on the IBC array[24].

In this study we set out to discover novel lipid loci, fine map signals to identify causal genes at implicated loci, and gain a greater understanding of the genetic architecture of lipid traits across ethnicities. Here, we have used the IBC array to examine association results for TC, LDL-C, HDL-C and TG across seven non-European study populations, including African Americans (n = 7,657), Hispanics (n = 1,315) and East Asians (n = 841). Using conditional analyses, we sought to identify independent signals from within associated loci. Finally, we assessed the direction of effect in non-Europeans of new and established loci found in European-derived populations, and tested a composite risk score of known loci across ethnicities.

MATERIALS AND METHODS

Ethics statement

All participants in each of the cohorts gave informed written consent. The Institutional Review Boards (IRBs) of each CARe cohort (i.e., the IRBs for each cohort's field centers, coordinating center, and laboratory center) have reviewed and approved the cohort's interaction with CARe. The study described in this manuscript was approved by the Committee on the Use of Humans as Experimental Subjects (COUHES) of the Massachusetts Institute of Technology.

Participating studies

Data from African-American, Hispanic and East Asian participants from seven cohorts were included for this study (Figure 3.1). Participants were $\geq$ 21 years of age. All seven studies contributed individual-level genotypes and phenotypes. Features of the included cohorts are presented in Table S1 and summary statistics are listed in Table S2. Six replication studies were used comprising African American individuals.

Phenotype definitions

Lipid phenotypes were taken from baseline or first measurements for all fasting individuals. All measurements were converted to mmol/L, with TC and HDL-C measurements converted from mg/dL by dividing by 38.67, and TG measurements converted

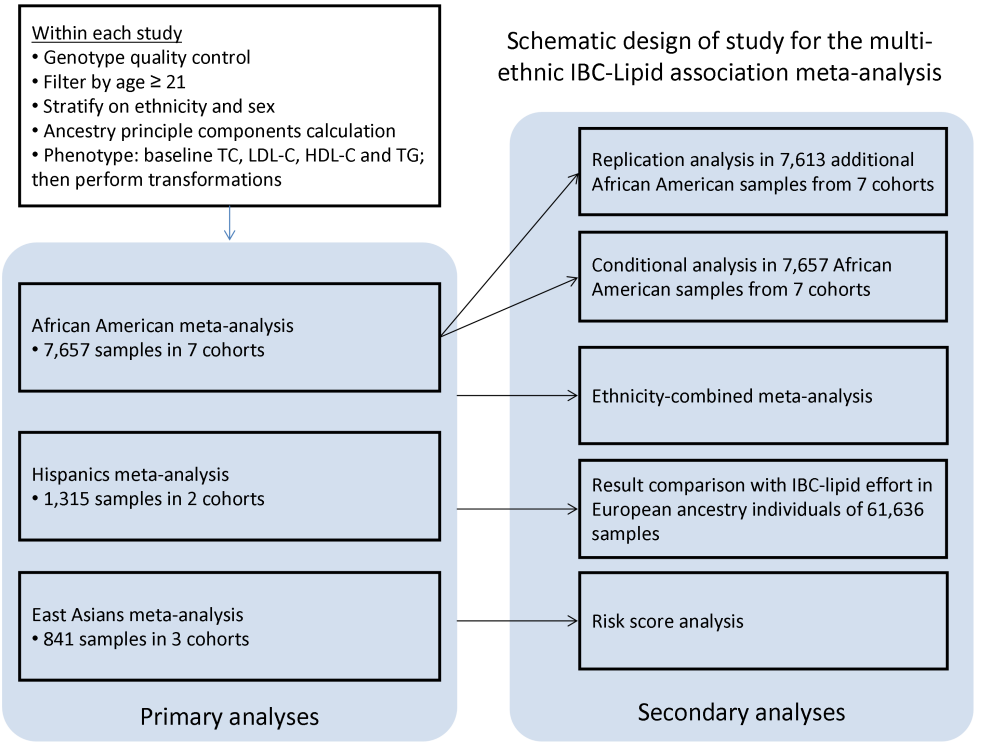

Figure 3.1: Schematic design of study for the multi-ethnic IBC-Lipid association meta-analysis. The workflow includes primary analyses and secondary analyses. Details can be found in the text.

from mg/dL by dividing by 88.57. TG values were log(10)-transformed as TG values were not normally distributed. LDL-C was calculated according to Friedewald's formula L~C - H - kT where C is total cholesterol, H is HDL-C, L is LDL-C, T is TG and k is 0.45 for mmol/L (or 0.20 if measured in mg/dl)[55]. If TG values were >4.51 mmol/L (>400 mg/dL), then LDL-C was treated as a missing value.

Genotyping and quality control

Genotyping in each participating cohort was performed using the IBC array[17]. SNPs were clustered into genotypes using the Illumina Genomestudio software and were subjected to quality control filters at the sample and SNP level, separately within each cohort. Samples were excluded for individual call rates <90%, gender mismatch, and duplicate discordance. SNPs were removed for call rates <95% or Hardy-Weinberg equilibrium (HWE) p $<10^7$. Due to low frequency SNPs included in the design, and the aim to capture low frequency variants of large effect across the combined dataset, we filtered only on minor allele frequency (MAF)<0.005.

Statistical analyses

Evaluation of population stratification

Self-reported ethnicity was verified by multidimensional scaling analysis of identity-by-state distances as implemented in PLINK[58], including HapMap panels as reference standards. After pruning of SNPs in linkage disequilibrium (r^2<0.3), Eigenstrat was used to compute principal components within each ethnic group separately for use as covariates in the regression analyses[116].

Association testing

Association analysis was performed in each study using an additive genetic model with one degree of freedom. Gender stratified analyses were performed using three multivariate models: Model 1, including 10 principal components (PCs); Model 2, including 10 PCs, age, and lipid medication; and Model 3, including 10 PCs, age, lipid medication, type 2 diabetes (T2D), smoking and BMI. The genomic control inflation factor, lambda,

was calculated for each cohort and used for within-study correction before meta-analysis. Genomic control inflation factors (λ) ranged from 1.00 to 1.054.

Meta-analyses within each ethnic group were performed by two independent analysts using a fixed-effect inverse-variance approach in two different software packages: MANTEL ($www.broadinstitute.org/\ debakker/mantel.html$) and METAL[66]. Results were highly concordant, reflecting a robust data analyses pipeline. Additionally, the directions of effect of lead SNPs from previously identified loci from the European IBC array meta-analysis[24] were evaluated for consistency in African Americans, Hispanics and Asians. To gauge an appropriate significance threshold, data from the Candidate gene Association Resource (CARe) IBC array studies[117] which is available on dbGAP ($www.ncbi.nlm.nih.gov/gap$) were employed and it was determined that after accounting for LD, the effective number of independent tests was ~26,500 for African Americans, ~23,500 for Hispanics, and ~15,500 for East Asians. This produces experimental or "array-wide" statistical thresholds of p = $1.9x10^{-6}$, p = $2.1x10^{-6}$ and p = $3.2x10^{-6}$, respectively, to maintain a false positive rate of 5% for each of the three ethnic groups. While we have adopted these "array-wide" statistical thresholds for this study, we also highlight loci associated at a more conventional genome-wide significance threshold of p $<5.0x10^{-8}$.
Additionally, the I^2 statistic was calculated to quantify the proportion of total variation due to heterogeneity, as described previously[71].

Conditional Analyses

Loci harboring evidence for association of P$<1x10^{-5}$ in African Americans were examined for the presence of multiple, independent signals via conditional analyses in PLINK[58]. A term was added to the regression model including the lead SNP as a covariate, and SNPs within a ± 500 kb region were evaluated for significance. A locus-specific Bonferroni correction, as employed in previous IBC studies[50], was applied to determine significance of independent signals within candidate genes genotyped at each locus.
On average, the windows contained 195.2 (± 107.0) variants with a range between 12 for ACADL and 359 for PCSK9. Because of limited power due to low sample size, we did not perform conditional analyses in Hispanics and East Asians.

Genetic Risk Score Analyses and direction of effect.

Within each ethnic group, we generated a genetic risk score using 28 SNPs for TC, 20 SNPs for LDL-C, 24 SNPs for HDL-C, and 21 SNPs for TG that had been found to be array-wide significant (p = $2.6x10^{-6}$) in the European-ancestry IBC meta-analysis[24] (Table S3), weighted by the beta as described previously[118, 119]. To account for missing data we adjusted the values for the number of genotyped risk alleles per individual. We evaluated for each ethnic group the contribution of the weighted genetic risk score to TC, HDL-C, LDL-C and TG in linear regression models adjusting for 10 PCs. Additionally, we compared the relative betas across quartiles of risk by linear regression. These loci were additionally investigated to study direction of effect across ethnicities.

Replication

In order to confirm putative novel loci, we replicated previously undetected lipid signals (p $<1.0x10^{-5}$) in 7,000 African American individuals from six replication cohorts and in 61,636 samples from the European-ancestry IBC meta-analysis[24]. Recent power analyses suggest that large-scale multi-ethnic association studies may have greater statistical power to detect causal alleles because of random genetic drift elevating global risk variants to higher allele frequency in some populations[120]. All but one replication studie provided summary results of SNPs that were genotyped on platforms other than the IBC array, or imputed using 1000 Genomes data. Features of the replication datasets included in this meta-analysis are described in Table S1.

RESULTS

Meta-analyses of African, Hispanic and East Asian populations

Meta-analyses of IBC array association results for plasma TC, LDL-C, HDL-C and TG levels in five African American studies (n = 7,657), two Hispanic studies (n = 1,315) and three East Asian studies (n = 841) were performed independently. Results of different association models did not differ substantially. Therefore, results of model 1, an additive model with 10 PCs as covariates, are presented in the main text (Table 3.1) and results of other models are presented in the supplements (Table S4). After fixed-effect

inverse-variance meta-analysis, we found that 23, five and two loci in African Americans, Hispanics and East Asian samples respectively, were significantly associated with a lipid trait at their respective array-wide significance thresholds, with twelve, three and one loci respectively surpassing the traditional genome-wide significance threshold (see Table 3.1; (Figure 3.1)). Two of these loci, intercellular adhesion molecule 1 (ICAM1) and CD36 molecule thrombospondin receptor (CD36), have not previously been reported to be associated with a lipid trait in a large-scale genomic study (Figure 3.2).

We found five independent loci that were associated with TC at the genome-wide significance threshold. Four of these signals were SNPs lying within previously described loci: LDLR (rs6511720, p = $1.4x10^{13}$); CELSR2 (rs12740374, p = $4.4x10^{-13}$); APOE (rs389261, p = $2.1x10^{-11}$) and PCSK9 (rs11806638, p = $2.00x10^{-9}$), while one signal was a novel SNP within ICAM1 (rs5030359, p = $5.2x10-9$). Three SNPs in the previously known loci, CELSR2 (rs12743074, p = $1.9x10^{-17}$), APOE (rs389261, p = $1.0x10^{-12}$) and PCSK9 (rs11800231, p = $1.0x10^{-10}$) reached genome-wide significance for association with LDL-C. We also identified a novel signal within ICAM1 (rs5030359, p = $1.1x10^{-7}$) that is associated with LDL-C in African Americans at array-wide significance. Genome-wide significant association with HDL-C was observed for three SNPs in previously identified loci within CETP (rs17231520 p = $2.0x10^{-46}$), LPL (rs13702 p = $1.3x10^{-9}$) and LIPC (rs2070895 p = $4.2x10^{-8}$). Of the array-wide significant loci, rs3211938 within CD36 (p = $3.1x10^{-7}$) has been previously described to be associated with HDL-C in a candidate gene study of 2.020 African Americans[121] but had not previously been identified in a large-scale genomic study. For TG, we identified one association signal, rs12721054, within the previously reported APOE locus with TG with at genome-wide significance (p = $1.0x10^{-21}$).

Hispanics

Genome-wide significant association with HDL-C was observed for two SNPs in previously identified loci within CETP (rs3764261, p = $3.4x10^{-11}$) and LIPC (rs8034802, p = $1.8x10^{-8}$). For TG, we identified one genome-wide signal within the previously reported APOA5 locus (rs10750097, p = $2.1x10^{-12}$). Genome-wide significant association for TC and LDL-C was not observed in our Hispanic populations.

East Asians

In East Asians, the rs662799 variant within ZNF259/APOA5 was significantly associated with TG (p = $1.6x10^{-13}$). The opposite allele of the same SNP was study-wide

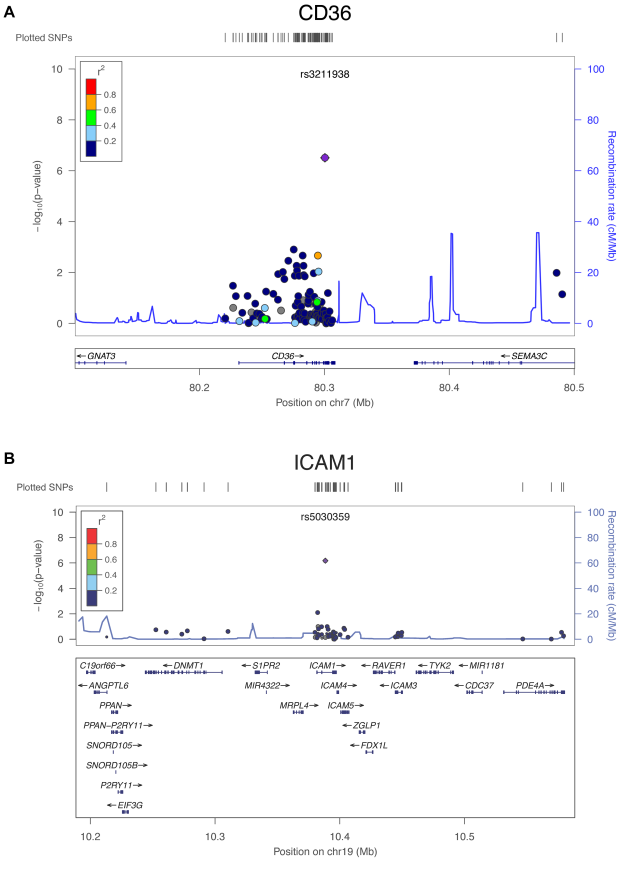

Figure 3.2: Regional plots for novel lipid loci with array-wide significant regions in IBC meta-analysis of African ancestry. A. CD36 region, B. ICAM1 region. Loci are shown as the lead SNP with a flanking region depicting the candidate gene and nearby genes included on the array. The purple diamond represents the lead SNP in the IBC meta-analysis and the dots represent the surrounding SNPs, with the different colors showing the LD relationship with the lead SNP based on YRI HapMap II information. -log10 p-values for association with HDL-C (for CD36) and TC (for ICAM1) are shown for each SNP (left-hand axis). Recombination rates in YRI HapMap II is shown in blue traces (right-hand axis).

Table 3.1: Associated loci with lipid traits in individuals of African, East Asian and Hispanic ancestry.

African-Americans										Results IBC Europeans			
Trait	Chr	BP	Candidate Gene	SNP	Risk allele	RAF	Beta (SE)	P	% I2	RAF	Beta (SE)	P	% I2
TC	1	55290748	PCSK9	rs11806638	C	0.68	0.11 (0.018)	1.99E-09	8	0.94	0.01 (0.013)	2.83E-01	0
TC	1	109619113	CELSR2	rs12740374	G	0.75	0.14 (0.019)	4.43E-13	0	0.76	0.13 (0.008)	1.79E-63	34.5
TC	2	21119200	APOB	rs12720826	T	0.87	0.13 (0.025)	7.84E-08	0	0.999	NA	NA	NA
TC	2	210769915	ACADL	rs6739874	T	0.05	0.17 (0.038)	4.65E-06	0	0.18	NA	NA	NA
TC	19	10249462	ICAM1	rs5030359	G	0.99	0.57 (0.098)	5.22E-09	0	0.998	NA	NA	NA
TC	19	11063306	LDLR	rs6511720	G	0.86	0.18 (0.024)	1.39E-13	26.2	0.88	0.17 (0.009)	1.81E-73	24.3
TC	19	50112183	APOE	rs389261	A	0.25	0.13 (0.020)	2.07E-11	0	0.001	NA	NA	NA
LDL-C	1	55290528	PCSK9	rs11800231	G	0.83	0.14 (0.022)	1.02E-10	37.3	0.96	0.009 (0.015)	6.04E-01	0
LDL-C	1	109619113	CELSR2	rs12740374	G	0.75	0.16 (0.019)	1.92E-17	0.8	0.75	0.12 (0.007)	2.62E-66	41.6
LDL-C	2	21141826	APOB	rs562338	G	0.4	0.09 (0.017)	6.54E-08	0	0.82	0.12 (0.008)	7.38E-51	0
LDL-C	17	61641042	APOH	rs1801689	C	0.007	0.49 (0.098)	4.70E-07	25.5	0.03	0.11 (0.016)	8.57E-12	0
LDL-C	19	10249462	ICAM1	rs5030359	G	0.99	0.51 (0.096)	1.09E-07	0	0.998	NA	NA	NA
LDL-C	19	11059187	LDLR	rs17248720	C	0.73	0.11 (0.019)	7.13E-10	0	0.88	0.16 (0.009)	8.81E-72	37.5
LDL-C	19	50112183	APOE	rs389261	A	0.25	0.14 (0.020)	1.03E-12	0	0.001	NA	NA	NA
HDL-C	7	80138385	CD36	rs3211938	G	0.09	0.06 (0.012)	3.09E-07	28.3	0.0005	NA	NA	NA
HDL-C	8	19868772	LPL	rs13702	C	0.51	0.04 (0.007)	1.34E-09	46.8	0.3	0.04 (0.002)	3.57E-74	24
HDL-C	15	56511231	LIPC	rs2070895	A	0.51	0.04 (0.007)	4.16E-08	0	0.21	0.04 (0.002)	9.76E-58	41.2
HDL-C	16	55553328	CETP	rs17231520	A	0.07	0.19 (0.013)	2.03E-46	14.5	0.002	0.16 (0.044)	3.28E-04	0
HDL-C	16	66534352	LCAT	rs35673026	T	0.004	0.28 (0.059)	2.40E-06	0	0.001	NA	NA	NA
TG	2	27584444	GCKR	rs1260326	T	0.15	0.06 (0.012)	5.54E-07	13.7	0.41	0.06 (0.003)	1.56E-83	32.6
TG	8	19864004	LPL	rs328	C	0.93	0.08 (0.016)	1.74E-07	19.9	0.9	0.08 (0.005)	5.73E-16	38.1
TG	11	116170289	APOA5	rs9804646	C	0.65	0.04 (0.009)	2.55E-06	2.3	0.92	0.02 (0.005)	1.57E-07	11.4
TG	19	50114427	APOE	rs12721054	A	0.89	0.12 (0.013)	1.01E-21	2.7	0.998	NA	NA	NA

East Asians										Results IBC Europeans			
Trait	Chr	BP	Candidate Gene	SNP	Risk allele	RAF	Beta (SE)	P	% I2	RAF	Beta (SE)	P	% I2
HDL-C	11	116168917	ZNF259/APOA5	rs662799	A	0.71	0.09 (0.018)	8.91E-07	0	0.93	0.03 (0.004)	1.42E-18	28.3
TG	11	116168917	ZNF259/APOA5	rs662799	G	0.29	0.21 (0.029)	1.57E-13	28.4	0.07	0.1102 (0.0058)	5.93E-89	58

Hispanics										Results IBC Europeans			
Trait	Chr	BP	Candidate Gene	SNP	Risk allele	RAF	Beta (SE)	P	% I2	RAF	Beta (SE)	P	% I2
TC	1	109619113	CELSR2	rs12740374	G	0.78	0.22 (0.045)	7.88E-07	41.5	0.76	0.13 (0.008)	1.79E-63	34.5
TC	12	130896484	MMP17	rs10902456	G	0.73	0.19 (0.043)	6.96E-06	0	0.69	-0.005 (0.006)	3.97E-01	0
LDL-C	1	109619113	CELSR2	rs12740374	G	0.78	0.21 (0.040)	1.10E-07	25.8	0.76	0.13 (0.008)	2.62E-66	34.5
LDL-C	19	47603609	LIPE	rs34052647	G	0.96	0.43 (0.098)	9.65E-06	0	0.999	NA	NA	NA
HDL-C	1	202392868	REN	rs11571087	T	0.01	0.28 (0.060)	2.58E-06	53	0.03	0.006 (0.006)	8.92E-01	13.1
HDL-C	3	149905771	AGTR1	rs12721308	G	0.01	0.27 (0.060)	4.95E-06	0	0.001	NA	NA	NA
HDL-C	15	56512084	LIPC	rs8034802	A	0.48	0.07 (0.012)	1.82E-08	2.5	0.28	0.03 (0.002)	4.86E-43	18.2
HDL-C	16	55550825	CETP	rs3764261	A	0.31	0.08 (0.012)	3.42E-11	63.2	0.31	0.07 (0.002)	8.45E-270	65.3
TG	11	116169250	APOA5	rs10750097	G	0.4	0.14 (0.020)	2.14E-12	0	0.21	0.07 (0.004)	4.16E-93	52.9

Chr chromosome, BP base pair, RAF risk allele frequency, SE standard error.

significantly associated with HDL-C. Genome-wide or study-wide significant genetic association was not observed for LDL-C or TC in our East Asian populations.

Independent signals within single genetic loci in African Americans

The current investigation using the IBC array included rare SNPs at candidate loci collected in sequencing data from Europeans and Africans and dense genotyping, which can potentially be used to identify independent signals for lipids within genes at known or novel loci. We repeated association studies conditioning on the lead SNP in 23 loci with P $<1.0x10^{-5}$. After Bonferroni correction for the number of SNPs at each candidate gene locus, we found independent lipids signals at the LDLR, APOE, PCSK9 and APOB loci for TC, at the APOE, PCSK9, LDLR, and APOB loci for LDL-C, at the APOC1/APOE, and LPL loci for TG and at the CETP, LPL, CD36 and the TRADD/LCAT for HDL-C (Table 3.2).
Three loci harbored two independent signals at genome-wide significance. The alleles rs6511720-G (risk allele frequency (RAF) = 0.86) and rs17242787-T (RAF = 0.98) within the LDLR gene showed association with TC with a p-value of $1.04x10^{-13}$ and $4.7x10^{-9}$ respectively in the original analyses. After conditioning on rs6511720-G, the p value for rs17242787-T remained significant (p = $2.4x10^{-10}$). Also for LDL-C, we found two independent genome-wide significant signals within the APOE locus: rs389261-A (RAF = 0.25) and rs283813-T (RAF = 0.67). Furthermore, the SNPs rs17231520-A (RAF = 0.07) and rs4783961-A (RAF = 0.44) within the CETP gene were both strongly associated with HDL-C and after conditioning on the lead signal, the secondary signal remained significant with p = $2.8x10^{-20}$. Interestingly, the newly identified CD36 locus also harbored two independent signals, with the second signal showing association with locus-wide significance. The r^2 between the two SNPs in HapMap-YRI was 0.118.

Replication

In order to confirm putative novel signals, we carried out in silico follow-up of ten SNPs within novel loci and previously unreported SNPs within known lipid-associated loci (P$<1.0x10^{-5}$) in six African American studies, comprising together 7,000 samples. Only HeartSCORE was genotyped using the IBC array and provided association results for all SNPs. All other replication studies contributed association results for up to seven genotyped and imputed SNPs. Imputed SNPs were only included in the study when passing the 95% confidence threshold. Combined meta-analysis of the discovery and replication studies led to genome-wide significant signals at the CD36 locus (p = $13.5x10^{-12}$;

Table 3.2: Loci with significant evidence of independent lipid association signals

									SECOND HIT		THIRD HIT		
Trait	Gene	SNP	Chr	Position	Risk Allele	RAF	Beta (SD)	P	Beta (SD)	P	Beta (SD)	P	r^2 with lead SNP*
TC	LDLR	rs6511720	19	11.063.306	G	0.86	0.18 (0.024)	1.39E-13					
		rs17242787	19	11.063.460	T	0.98	0.35 (0.059)	4.67E-09	0.37 (0.059)	2.44E-10			0.004
	APOE	rs389261	19	50.112.183	A	0.25	0.13 (0.0120)	2.07E-11					
		rs283813	19	50.081.014	T	0.67	0.09 (0.018)	3.60E-07	0.09 (0.018)	1.30E-06			0.001
		rs12721054	19	50.114.427	A	0.88	0.15 (0.027)	4.75E-08			0.10 (0.027)	1.85E-04	0.025
	PCSK9	rs11806638	1	55.290.748	C	0.68	0.11 (0.018)	1.99E-09					
		rs505151	1	55.301.775	G	0.24	0.11 (0.019)	7.50E-09	0.09 (0.0120)	9.78E-06			0.085
	APOB	rs12720826	2	21.119.200	T	0.87	0.13 (0.025)	7.84E-08					
		rs562338	2	21.141.826	G	0.4	0.09 (0.017)	2.00E-07	0.07 (0.018)	1.20E-04			0.054
LDL-C	APOE	rs389261	19	50.112.183	A	0.25	0.14 (0.020)	1.03E-12					
		rs283813	19	50.081.014	T	0.67	0.12 (0.018)	7.27E-12	0.12 (0.018)	3.30E-11			0.001
		rs166907	19	50.078.695	G	0.12	0.08 (0.026)	2.92E-03			0.16 (0.031)	1.96E-07	0.001
	PCSK9	rs11800231	1	55.290.528	G	0.82	0.14 (0.022)	1.02E-10					
		rs505151	1	55.301.775	G	0.24	0.12 (0.019)	1.91E-10	0.10 (0.020)	2.15E-07			0.12
		rs1165287	1	55.292.800	A	0.2501	0.09 (0.02)	7.42E-06			0.08 (0.021)	1.17E-04	0.08
	LDLR	rs17248720	19	11.059.187	C	0.73	0.11 (0.019)	7.13E-10					
		rs6511720	19	11.063.306	T	0.86	0.19 (0.024)	7.91E-15	0.15 (0.030)	2.93E-07			0.251
		rs17242787	19	11.063.460	T	0.98	0.31 (0.06)	1.76E-07			0.33 (0.062)	8.12E-08	0.071
	APOB	rs562338	2	21.119.200	G	0.4	0.09 (0.017)	6.54E-08					
		rs12720826	2	21.119.200	T	0.87	0.13 (0.025)	7.84E-08	0.11 (0.026)	0.0001301			0.054
HDL-C	CETP	rs17231520	16	55.553.328	A	0.07	0.19 (0.013)	2.03E-46					
		rs4783961	16	55.552.395	A	0.44	0.09 (0.007)	6.08E-40	0.06 (0.007)	2.83E-20			0.165
		rs7499892	16	55.564.091	C	0.62	0.07 (0.007)	7.27E-24			0.04 (0.006)	6.40E-09	0.078
	LPL	rs13702	8	19.868.772	C	0.51	0.04 (0.007)	1.34E-09					
		rs3289	8	19.867.472	T	0.93	0.07 (0.013)	5.07E-08	0.06 (0.013)	2.70E-05			0.047
	CD36	rs3211938	7	80.138.385	G	0.09	0.06 (0.012)	3.09E-07					
		rs3211849	7	80.121.259	G	0.54	0.02 (0.007)	5.50E-03	0.03 (0.007)	1.33E-05			0.118
	TRADD/LCAT	rs35673026	16	66.534.352	T	0.004	0.28 (0.059)	2.40E-06					
		rs2233455	16	65.765.434	T	0.29	0.03 (0.007)	3.72E-06	0.03 (0.007)	2.78E-06			
TG	APOC1/APOE	rs12721054	19	50.114.427	A	0.89	0.12 (0.013)	1.01E-21					
		rs7258987	19	50.124.360	T	0.03	0.11 (0.024)	2.14E-06	0.11 (0.024)	3.42E-06			0.003
	LPL	rs328	8	19.864.004	C	0.93	0.08 (0.016)	1.74E-07					
		rs3289	8	19.867.472	G	0.07	0.08 (0.016)	2.82E-06	0.07 (0.016)	1.52E-05			0.003

Chr chromosome, BP base pair, RAF risk allele frequency, SE standard error.

Table 3.3) for association with HDL-C. A signal within ACADL was not significant after meta-analysis of the discovery and replication studies. However, the direction of effect was consistent with our discovery dataset in three of six studies, so it is possible that the signal has a weak effect and the locus is undetectable due to limited statistical power. Also, previously unidentified signals in known lipid loci showed genome-wide significant association in the combined discovery and replication meta-analysis: rs11806638 within PCSK9 was found to be associated with TC; rs389261 within APOE was associated with LDL-C levels; rs17231520 within the CETP locus and rs35673026 within the LCAT locus were found to be associated with HDL-C; and rs12721054 within APOE was associated with TG levels (Table 3.3).

Comparison of lipid loci in African Americans to Europeans

Utilizing the results of each of the meta-analyses from the three available ethnicities, we sought to refine localization of known lipid signals or reveal novel independent signals within known loci based upon differential LD (see Table 3.1). The dense genotyping within each locus on the IBC array enabled detailed comparisons of loci that harbored array-wide significant SNPs in Africans Americans, Hispanics and East Asians as well as in the IBC meta-analysis of up to 61,636 individuals of Europeans ancestry[24] (see Table 3.1 and Table S3).

The strongest signal for HDL-C in African Americans is rs17231520 within CETP (p = $2.0x10-46$; Table 3.1). This SNP is associated with HDL-C in the same direction in Europeans with p = $3.3x10^{-4}$. However, in Europeans there is less power to detect this signal at array-wide significance, as the MAF in Europeans is only 0.2% (versus 7% in African Americans) and was screened out in many European studies for the IBC meta-analysis. Furthermore, rarer variants are often not correctly clustered optimally during QC, making them less likely to pass the standard quality control (including genotyping threshold or HWE check). This is also observed for the most strongly associated SNPs within CD36 (rs3211938) and LCAT (rs35673026) for HDL-C in African-Americans, as they show the same direction of effect in Europeans, but do not reach significance, given low MAF and absence in the majority of European studies for IBC meta-analysis. For two loci, LIPC and LPL, the strongest associated SNP in African Americans for HDL-C was the same or among the most highly associated SNPs in Europeans. Also, for the LDL-C-associated loci CELSR2, APOB, APOH and LDLR, the strongest signals in African Americans did overlap or represented similar signals that were highly associated with LDL-C in Europeans. The newly identified SNP for LDL-C, rs5030359 within ICAM1, has an observed MAF of 0.8% in African Americans and 0.2% in Europeans.

Table 3.3: Replication results of nine signals in 7,000 African Americans.

Novel loci					Discovery Set (N=7.657)		Replication Set (N=7.000)		Combined Set (N=14.657)	
Trait	Candidate Gene	SNP	Risk allele	RAF	Beta (SE)	P	Beta (SE)	P	Beta (SE)	P
TC	ACADL	rs6739874	T	0.05	0.17 (0.038)	4.65E-06	0.01 (0.043)	8.60E-01	0.11 (0.029)	1.72E-04
TC	ICAM1	rs5030359	G	0.99	0.57 (0.098)	5.22E-09	0.48 (0.295)	1.04E-01	0.44 (0.095)	4.33E-06
LDL-C	ICAM1	rs5030359	G	0.99	0.51 (0.096)	1.09E-07	0.47 (0.264)	7.74E-02	0.40 (0.089)	8.74E-06
HDL-C	CD36	rs3211938	G	0.09	0.06 (0.012)	3.09E-07	0.06 (0.013)	3.12E-06	0.06 (0.009)	3.49E-12

Unidentified signals in previously known loci					Discovery Set (N=7.657)		Replication Set (N=7.000)		Combined Set (N=14.657)	
Trait	Candidate Gene	SNP	Risk allele	RAF	Beta (SE)	P	Beta (SE)	P	Beta (SE)	P
TC	PCSK9	rs11806638	C	0.68	0.11 (0.018)	1.99E-09	0.15 (0.029)	3.78E-07	0.12 (0.016)	2.38E-14
LDL-C	APOE	rs389261	A	0.25	0.14 (0.020)	1.03E-12	0.13 (0.028)	4.43E-06	0.14 (0.016)	1.70E-17
HDL-C	CETP	rs17231520	A	0.07	0.19 (0.013)	2.03E-46	0.18 (0.021)	2.35E-17	0.18 (0.011)	1.70E-62
HDL-C	LCAT	rs35673026	T	0.004	0.28 (0.059)	2.40E-06	0.05 (0.12)	7.07E-01	0.24 (0.053)	9.06E-06
TG	APOE	rs12721054	A	0.89	0.12 (0.013)	1.01E-21	0.05 (0.008)	4.71E-13	0.07 (0.007)	1.31E-28

RAF risk allele frequency, SE standard error.

In Europeans, this SNP is not associated with LDL-C (p = 0.3231), but the SNP is only present in very few European studies that are included in the IBC meta-analysis. The most associated signals within PCSK9 and APOE in African Americans are different, independent signals compared to the most associated SNPs within these loci in Europeans. Again, both signals are common in African Americans and have very low frequencies in Europeans: MAF for SNPs in PCSK9 and APOE are 17% and 25% in African Americans and 0.5% and 0.1% in Europeans respectively.

Among the array-wide statistically significant loci that were associated with TG in African Americans, three SNPs within GCKR, LPL and APOA5 were the same as or amongst the most highly associated SNPs in Europeans. SNP rs12721054 in APOE appeared to be a novel independent signal for TG in African Americans. This SNP showed an opposite effect in European-derived cohorts, although it was observed rarely in the meta-analysis of European populations (MAF = 0.2%)[24].

For TC, we observed the same pattern as for other lipid traits. The strongest associated SNPs within loci associated with TC overlapped with the same signals in Europeans (SNPs within CELSR2, APOB, LDLR and APOE), or were independent signals in African Americans that could not be replicated in Europeans because of low frequency (PCSK9, ACADL and ICAM1).

Direction of effect concordance with lead SNPs identified in European populations

Direction of effect across different ethnicities was studied for 28 previously established TC risk loci, 20 LDL-C loci, 24 HDL-C loci, and 21 TG associated loci. Not all SNPs passed the initial quality control, so number of investigated SNPs differed by trait and ethnicity (Table S3).

Concordance in direction of effect was observed for 21/27 (p = 0.033), 15/20 (p = 0.102), 16/23 (p = 0.176) and 19/21 (p = 0.004) association signals for TC, LDL-C, HDL-C and TG, respectively, between Europeans and African Americans; 23/28 (p = 0.011), 16/20 (p = 0.047), 21/23 (p = 0.002) and 19/21 (p = 0.004) SNPs were concordant in direction of effect for TC, LDL-C, HDL-C and TG respectively between Europeans and Hispanics. Finally, 17/24 SNPs for TC (p = 0.140), 11/16 SNPs for LDL-C (p = 0.279), 16/29 SNPs for HDL-C (p = 0.196) and 17/21 (p = 0.035) SNPs for TG were concordant between Europeans and East Asians (Table S3).

Genetic risk score analysis

To study whether we could find elevated lipid levels in multi-ethnic samples with cumulative numbers of risk alleles that were previously found to be associated in Europeans, we evaluated the contribution of the weighted genetic risk score for lipids in linear regression models adjusting for 10 PCs and compared the relative beta's ratios across quartiles of risk. We demonstrated a significant per quartile risk effect in African-Americans (ranging from p$<10^{-10}$ for TG to p$<10^{-33}$ for HDL-C), Hispanics (ranging from p$<10^{-l}$ for LDL-C to p$<10^{-23}$ for TC) and East Asians (ranging from p$<$0.02 for HDL-C to p$<10^{-6}$ for TG) (see Table 3.4). Quartiles based on weighted risk alleles and lipid level distribution for each ethnicity is shown in Figure S1.

DISCUSSION

The current study reports a meta-analysis of lipid association studies in African Americans, Hispanics and East Asians using the IBC array, and has identified two novel loci associated with TC and LDL-C levels (rs5030359 in ICAM1) and HDL-C levels (rs3211938 in CD36) in African Americans. Additionally, we have uncovered multiple independent association signals within established lipid loci, demonstrating the value of dense SNP genotyping to uncover genetic variation associated with lipid levels. Furthermore, we have evaluated the impact of established SNPs, previously associated with lipids in Europeans populations, on lipid levels in three additional populations, showing that many known association signals for lipids span across ethnicities.

CD36

This study shows association between the nonsense coding variant rs3211938-G in CD36 and HDL-C levels at conventional genome-wide significance for African Americans (p$<5x10^{-9}$). This SNP has previously been reported to be associated with increased HDL-C levels (p = 0.00018), decreased TG levels (p = 0.0059) and protection against metabolic syndrome (p = 0.0012) in a candidate gene study including 2,020 African Americans that did not overlap with samples in our meta-analyses[121]. Also, a variant within CD36 was associated with LDL levels in two small studies[122, 123]. The CD36 finding is present in an accompanying paper [52] from the wider NHLBI CARe lipid studies which essentially uses the same discovery cohorts for African Americans that we present here although our analysis differs in that (a) it screened out related individuals

Table 3.4: Risk score analysis of lipid profile in multiethnic populations, using weighted score of known lipid SNPs.

Trait	TC	LDL-C	HDL-C	TG
Beta (SE)	0.12 (0.011)	0.08 (0.01)	0.05 (0.004)	0.02 (0.004)
P	7.57E-23	5.45E-15	3.13E-33	4.83E-10
Quartiles of Risk Alleles				
Q1 Beta (SE)	ref	ref	ref	ref
Q2 BETA (SE)	0.12 (0.034)	0.12 (0.033)	0.05 (0.014)	0.06 (0.017)
P	2.99E-04	3.07E-04	1.23E-04	8.04E-04
Q3 BETA (SE)	0.25 (0.035)	0.16 (0.034)	0.10 (0.014)	0.05 (0.017)
P	1.88E-12	1.88E-06	2.45E-12	0.002619
Q4 BETA (SE)	0.33 (0.036)	0.26 (0.034)	0.16 (0.013)	0.12 (0.017)
P	1.98E-20	7.87E-15	6.73E-32	3.37E-13
Trait	TC	LDL-C	HDL-C	TG
Beta (SE)	0.17 (0.017)	0.08 (0.015)	0.07 (0.009)	0.07 (0.009)
P	2.16E-23	2.07E-07	4.43E-14	4.43E-14
Quartiles of Risk Alleles				
Q1 Beta (SE)	ref	ref	ref	ref
Q2 BETA (SE)	0.09 (0.063)	0.06 (0.046)	0.06 (0.019)	0.05 (0.036)
P	0.146	0.19	0.005	1.45E-01
Q3 BETA (SE)	0.26 (0.063)	0.09 (0.048)	0.08 (0.02)	0.02 (0.033)
P	2.78E-05	0.059	9.65E-05	0.556
Q4 BETA (SE)	0.47 (0.056)	0.24 (0.047)	0.166	0.18 (0.034)
P	6.42E-17	2.65E-07	3.42E-16	2.68E-07
Trait	TC	LDL-C	HDL-C	TG
Beta (SE)	0.0642 (0.0288)	0.0137 (0.0272)	0.105 (0.044)	0.088 (0.0186)
P	0.02615	0.6153	0.01702	2.29E-06
Quartiles of Risk Alleles				
Q1 Beta (SE)	ref	ref	ref	ref
Q2 BETA (SE)	0.32 (0.097)	-0.03 (0.084)	0.09 (0.039)	0.08 (0.059)
P	8.37E-04	0.74	0.02	0.17
Q3 BETA (SE)	0.22 (0.094)	-0.08 (0.092)	0.13 (0.038)	0.10 (0.063)
P	0.017	0.37	6.82E-04	0.1
Q4 BETA (SE)	0.25 (0.092)	0.05 (0.085)	0.19 (0.039)	0.27 (0.058)
P	0.007	0.53	2.27E-06	2.39E-06

ref reference group, SE standard error.

(b) it takes additional covariates into account through the use of the three multivariate models and (c) our analysis filtered more stringently on I^2 and (d) we replicated these findings in additional studies.

CD36, which is present on gustatory, olfactory and intestinal epithelial cells, is involved in the orosensory perception of fatty acids[124, 125]. Also, lipid ingestion affects lingual CD36 expression in mice[126]. Therefore, CD36 may influence fat intake, and hence, serum lipid levels. SNPs within CD36, other than the one we found in this study, were linked to obesity in a case-control study[127]. However, this finding could not be replicated in a larger cohort[128]. In mouse models, CD36 deficiency impairs intestinal lipid secretion and results in hypertriglyceridemia[129] and others show that CD36 deficiency rescues lipotoxic cardiomyopathy[130].

CD36 is an integral membrane protein found on the surface of many cell types and binds many ligands including oxidized lipid proteins[131, 132], long-chain fatty acids[133] and erythrocytes that are parasitized with the malaria parasite Plasmodium falciparum[134]. The rs3211938-G variant is nearly absent in Europeans and Asians and shows a signature of selection in African Americans and some African populations[135, 136]. Additionally, rs3211938-G has been shown in previous studies to be associated with CD36 deficiency and with susceptibility to malaria, although this has not been confirmed in other studies[137, 138].

ICAM1

The rs5030359 variant in ICAM1, is observed in this study to be associated with TC and LDL-C at conventional genome-wide significance. ICAM1 encodes a cell surface glycoprotein that is typically expressed on endothelial cells and cells of the immune system[139]. However, rs5030359 maps to a gene-dense region (Figure 2b), so it cannot be excluded that there is another gene underlying the signal. The rs5030359 variant is ~800 kb downstream of a previously identified lipids signal within the LDLR region, but conditional analyses showed that the two loci are independent. Using fine-mapping in non-African populations to point to the most likely gene underlying the signal, is not possible as the SNP is very rare in Europeans, with a MAF of 0.002, and absent in our Hispanic and East Asian populations. Previously, common variants within ICAM1 were found to be associated with soluble ICAM1 (sICAM1) concentrations in Europeans[140, 141]. sICAM1 has been associated with several common diseases such as diabetes, heart disease, stroke, and malaria[142, 143]. sICAM1 levels were associated with progression of carotid intima media thickness in young adults[144, 145] and in asymptomatic

dyslipidaemia subjects[146]. Additionally, sICAM1 levels were found to be higher in Europeans than in Africans[145].

Differences in signals within lipid loci in multiple ethnicities

We were able to use the dense SNP genotyping in loci on the IBC array to analyze and compare lipid-associated loci, particularly between African Americans and Europeans. Our analyses showed multiple examples of signals that were associated with lipid levels in one ethnicity but not another (Table 3.1).

First, some of the strongest associated SNPs in one ethnicity may be rare or absent in other ethnicities. This is a well-established phenomenon, e.g., truncation mutations in PCSK9 that are of low frequency in African Americans and absent in individuals of European origin, that result in a robust reduction in LDL-C levels and coronary heart disease risk[147, 148]. In this study we find that the majority of the observed discrepancies across ethnicities in association of SNPs with lipid traits can be attributed to differences in allele frequency. For example, rs3211938 in CD36 is much more highly associated with HDL-C in African Americans (p = $1.8x10^{-11}$) than in Europeans (p = 0.08) with a large discrepancy in RAFs (7% vs. 0.2%).

In other loci, the strongest associated polymorphisms varied across populations, for example in the BUD13/ZNF259/APOA5 region (Table S3, Figure S2). In theory these regions could be excellent candidates for fine-mapping, but our efforts and association results could not narrow down the loci. When conducting meta-analyses across multiple ethnicities we observed that the stronger p-value association typically tracked with the higher heterogeneity I^2 values (Figure S3). This high I^2 suggests high heterogeneity, but it could also be the effect of low sample sizes of the combined cohorts (especially for Hispanics and East Asians).

One limitation of this study is the sample size available particularly the Hispanic and the East Asian available samples and this obviously limited our ability to find new signals in these populations and to replicate many previously established lipid signals. Also, not all previously described signals for lipids were present on the IBC array, as the array was designed to densely cover genes regions, rather than the whole genome. However, using this approach we did find signals for lipids that remained uncovered using the genome-wide association approach, as both rs5030359 within ICAM1 and rs3211938 within CD36 were not present on conventional genome-wide arrays.

In conclusion, we performed dense genotyping of ~2,000 candidate genes in 7,657 African Americans, 1,315 Hispanics and 841 East Asians using IBC 50K SNP genotyping array and we found and confirmed two novel signals for lipids by replication in 7,000 African Americans. Additionally we evaluated the effect of SNPs established in European populations on lipid levels in multi-ethnic populations and show that most known lipid association signals span across ethnicities. However, differences between populations, especially differences in allele frequency, can be leveraged to identify novel signals.

SUPPLEMENTARY INFORMATION:

Supplemental data is available electronically and includes three figures, four tables and supplementary acknowledgments. DOI: https://doi.org/10.1371/journal.pone.0050198

Part II

Coronary Artery Disease (CAD)

4 COMMON GENETIC VARIANTS DO NOT ASSOCIATE WITH CAD IN FAMILIAL HYPERCHOLESTROLEMIA

In recent years, multiple loci dispersed on the genome have been shown to be associated with coronary artery disease. We investigated whether these common genetic variants also hold value for coronary artery disease prediction in a large cohort of patients with familial hypercholesterolemia.

We genotyped a total of 41 single nucleotide polymorphisms in 1701 familial hypercholesterolemia patients, of whom 482 patients (28.3%) had at least one coronary event during an average follow up of 66 years. The association of each single nucleotide polymorphism with event-free survival time was calculated with a Cox proportional hazard model.

In the cardiovascular disease risk factor adjusted analysis, the most significant single nucleotide polymorphism was rs1122608:G>T in the SMARCA4 gene near the LDLR gene, with a hazard ratio for coronary artery disease risk of 0.74 (95% CI $0.49 - 0.99$; p-value 0.021). However none of the single nucleotide polymorphisms reached the Bonferroni threshold.

Of all the known coronary artery disease loci analysed, the SMARCA4 locus near the LDLR had the strongest negative association with coronary artery disease in this high-risk familial hypercholesterolemia cohort. The effect is contrary to what was expected. None of the other loci showed association with coronary artery disease.

INTRODUCTION

Familial hypercholesterolemia (FH) is an autosomal dominant disorder characterized by increased low-density lipoprotein cholesterol (LDL-C) levels and preponderance to coronary artery disease (CAD). The diagnosis is based on stringent clinical criteria or on the identification of mutations in the LDL-receptor (LDLR), apolipoprotein B or proprotein convertase subtilisin/kexin type 9 (PCSK9) gene. The frequency of heterozygosity is at least 1/500 in most European countries[149]. By virtue of the elevated LDL-C levels, FH results in lipid accumulation in the arterial wall and as a consequence accelerated atherosclerosis[150, 151]. If left untreated, 50% of male and 30% of female heterozygous FH patients will develop CAD before 60 years of age[152]. The age of onset and severity of CAD varies considerably between FH patients, even among individuals who share an identical gene defect[10].

Previously, we performed a retrospective multi-centre cohort study of 2400 FH patients, of whom 782 patients (32.6%) had at least one cardiovascular event during an average follow up of 66 years[153]. In this cohort we demonstrated that LDL-C levels are more important than the LDLR mutation type in determining the age of onset of CAD[154]. We also showed that classical risk factors including male gender, smoking, hypertension, Type 2 diabetes, low levels of high-density lipoprotein cholesterol (HDL-C) and elevated lipoprotein(a) levels were independent risk factors for the development of CAD. However, these factors combined explained only 18.7% of the variation in CVD occurrence[153]. Thus, a considerable part of the variability in CVD occurrence remains to be disentangled and common genetic variation might provide one of the explanations.

Recent large-scale genome-wide association (GWA) studies have revealed common genetic variations at 45 loci which moderately affect (Hazard Ratios (HR) varying between 1.1 and 2.0) the incidence of coronary artery disease (CAD) in the general Caucasian population[155, 156, 157, 14]. We set out to address whether common variations within the 45 previously identified loci by recent GWA studies are modifiers of CAD risk in a high-risk population of heterozygous FH cases.

MATERIAL AND METHODS

Ethics Statement

Written informed consent was obtained from all patients. The Medical Ethics Review Board of each participating hospital approved the protocol, which complies with the Declaration of Helsinki.

Study Design and Study Population

The Genetic Identification of Risk Factors in Familial Hypercholesterolemia (GIRaFH) is a retrospective multicenter cohort study. The study design and study population have been described elsewhere[153]. Briefly, DNA samples from patient who, based on clinically oriented algorithms are anticipated to suffer from FH are being sent in to the central core molecular diagnostic laboratory by physicians working at one of the nationwide lipid clinics. LDLR gene variation was genotyped according to previously published methods[158]. DNA of a total of 9,300 hypercholesterolemic patients was stored in the DNA database at time of initiation of the cohort. Only those cases from larger lipid clinics were selected (9,188) for further analysis, as smaller clinics normally only send DNA samples of the rare, usually very serious FH cases. Of this set, 4,000 cases were randomly selected. After review of medical records, a group of 2,400 patients fulfilled the FH diagnostic criteria based on internationally established criteria and were included in the study[159]. Phenotypic, CVD event and cause of death data were acquired from medical charts. None of the study population received primary prevention in the form of beta-blockers or aspirin. CAD was defined as angina pectoris (AP), acute coronary syndrome (ACS), percutaneous coronary interventions (PCI) or coronary artery bypass grafting (CABG).

SNP selection

Based on the latest published meta-analysis, a total of 45 SNPs associated with CAD were identified[14]. We did not include SNPs which were only associated with an intermediate trait such as lipid levels, type 2 diabetes or hypertension. If these SNPs were not directly genotyped, imputed data, using MACH and the HapMap phase 2 data sets

(build 36 release 22)[160, 161, 162] was used. Finally if this failed proxies were looked up within a window of 500kb ($r^2 \geq 0.8$)[160, 161, 162], see Figure 4.1.

Genotyping and imputation

For 1,701 of the 2,400 GIRaFH cases DNA was available for additional genotyping. Genotyping was performed using the 50K gene-centric Human CVD BeadChip[17] and genotypes were called using the BeadStudio software (Illumina, San Diego, California, USA) and subjected to quality control filters at the sample and SNP level. After genotyping, PLINK v1.07 ($http://pngu.mgh.harvard.edu/purcell/plink/$) was used to test the SNPs for population substructure which could introduce false-positive associations. This was done by means of multidimensional scaling implementation[58]. Also the SNPs were subjected to additional quality control filters based on sample size and minor allele frequencies (MAF). Samples with a call rate of $<95\%$ were excluded from further analysis. Genetic markers with a MAF $<1\%$ were excluded from further analysis. An identity-by-state analysis was performed to ensure that only Caucasian individuals were included in the final association analyses.

Statistical analysis

With an effective sample size of 1,701 cases and 483 events the GIRaFH sample has 80% power to identify statistically significant associations for SNPs conferring a relative risk > 2.2 and MAF > 0.10. Differences between subgroups were tested with Chi square statistics or an independent sample t-test where appropriate. Triglycerides had a skewed distribution and therefore statistical analyses were performed on log-transformed data. The association of each SNP with event-free survival time was calculated with a Cox proportional hazard model in the R package ProbABEL[163] The event-free survival time was defined as time from birth to date of CAD event, or when no event had occurred as time from birth to date of inclusion in the study. An additive genetic model was applied in the Cox model and classical cardiovascular risks were used as co-variates[159]. We corrected for factors that had previously been shown to be associated with CAD risk in this population: age, gender, smoking, Type 2 diabetes, hypertension and body mass index (BMI). Analyses were performed for the loci previously reported to be associated with CAD. Of the 45 reported SNPs, data was available of 41 SNPs after imputation. Significance was defined as a p-value <0.05 divided by the number of SNPs tested, yielding a significance level of $1.22x10^{-3}$ (41 SNPs based on the literature). All analyses were

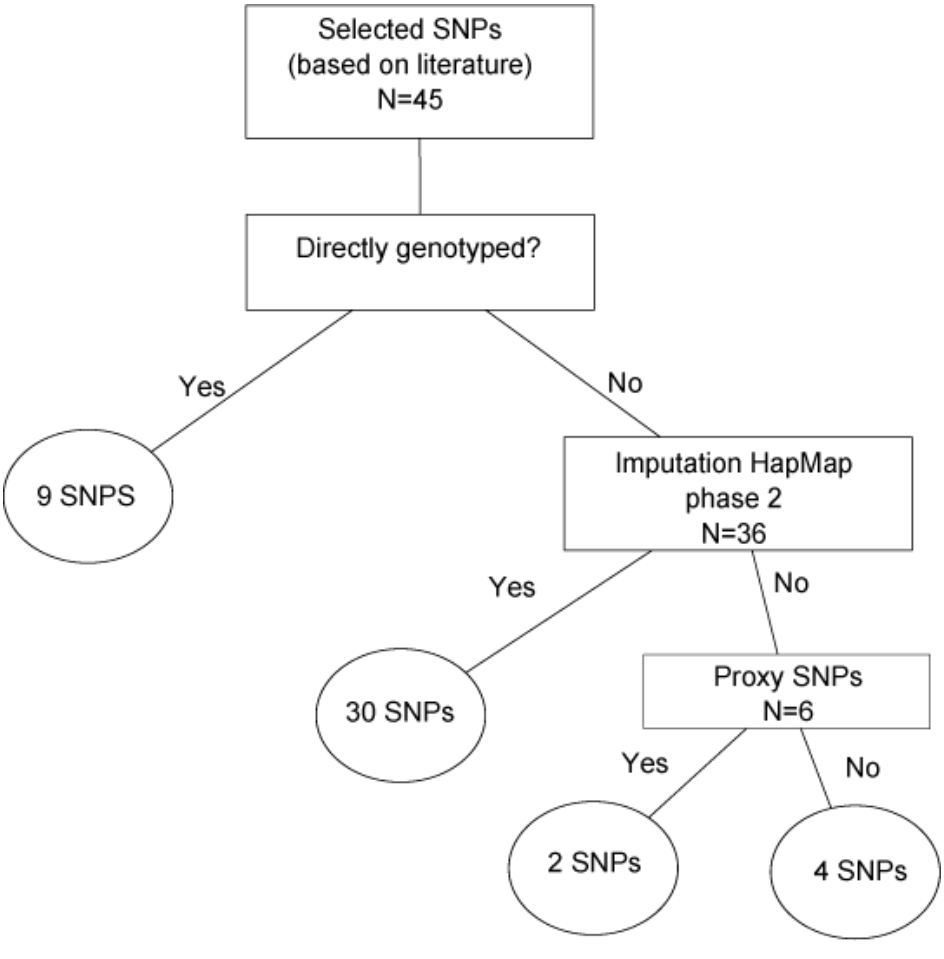

Figure 4.1: SNP selection procedure

also performed separately for males and females. Statistical analyses were performed using SPSS software (version 17; Chicago, Illinois, USA). All described variants will be submitted to the following public variant database LOVD3: Leiden Open Variation Database ($http://databases.lovd.nl/shared/genes/$).

RESULTS

Genotyping and imputation

Of the 1,701 DNA samples, 7 individuals did not cluster appropriately in the IBS, reflecting non-Caucasoid origin, and were consecutively excluded, leaving a total of 1,694 DNA samples for analysis. A total of 38,978 SNPs met our quality control steps. No subjects were excluded because of low call rate. The genomic inflation factor was close to 1 (λ = 1.07), indicating that the influence of population substructure and genotyping errors was negligible. Using HapMap we were able to impute up to 2.5 million SNPs for all individuals. Out of the 45 SNPs, 9 were directly genotyped and 30 were imputed. A proxy of 2 of the remaining 6 SNPs (rs12205331:C>T -> rs12197124:C>T (R^2= 1.0) and rs9369640:C>A -> rs7751826:C>T (R^2= 1.0) could be found in LD with the lead SNP. So a total of 9 SNPs was genotyped directly, 30 SNPs were imputed, for two SNPs, proxy SNPs were found and 4 SNPs could not be found within the imputed data, because they were not available in the reference panel or they were poorly imputed. We only analysed the 41 SNPs which were available after imputation.

Demographic data

Demographic data of the 1,694 study cases are listed in Table 4.1. The average age at inclusion was higher in the CAD group. The mean age of onset of CAD was 49.1 (standard deviation; SD 10.7) years and the mean event-free survival in individuals without CAD was 47.3 (SD 12.6) years. During follow-up, 28% of our cohort developed CAD. Cardiovascular risk factors were significantly more prevalent in CAD cases than in controls. Treatment-naive LDL-C levels at the time of inclusion in the cohort did not differ between patients with and without CAD. All first visits to the lipid clinic took place between March 1969 and November 2002.

Table 4.1: Baseline characteristics

	CAD+ n=482	CAD- n = 1212	p-value
Age first visit lipid clinic (yrs)	50.5 (11.1)	42.2 (12.3)	<0.0001
Age last visit lipid clinic (yrs)	57.1 (11.2)	46.6 (12.5)	<0.0001
Male gender	301 (62)	516 (42)	<0.0001
Diabetes mellitus	54 (11)	45 (4)	<0.0001
Hypertension	79 (16)	77 (6)	<0.0001
History of Smoking	365 (75)	766 (63)	<0.0001
Body mass index, kg/m2	25.7 $\pm$ 3.2	24.9 $\pm$ 3.6	<0.0001
Family history CVD	204 (42)	607 (50)	0.005
Treatment-naive lipids levels:			
Total cholesterol, mmol/L	9.6 (2.1)	9.4 (1.9)	0.146
LDL-cholesterol, mmol/L	7.0 (1.9)	7.2 (1.8)	0.152
HDL-cholesterol, mmol/L	1.1 (0.3)	1.2 (0.4)	<0.0001
Triglycerides*, mmol/L	1.8 (1.3-2.5)	1.5 (1.0-2.0)	<0.0001
LDL-Receptor mutation proven	207 (43)	655 (54)	0.017
Age of CAD onset, years	49.1 $\pm$ 10.7	NA	NA
Age of start statin use	48.8 $\pm$ 11.5	40.9 $\pm$ 12.6	<0.0001

Abbreviations: CAD, coronary artery disease; CVD, cardiovascular disease; LDL, low-density lipoprotein; HDL, high-density lipoprotein.
Values are given as number (percentage) or mean$\pm$SD unless indicated otherwise.

SNPs and risk of CAD

None of the 41 SNPs reached a significant p-value after Bonferroni correction ($p < 1.11x10^{-3}$) (see Table 4.2). The best performing SNP in the cardiovascular risk factors adjusted analysis was rs1122608:G>T, in the SMARCA4 gene near the LDLR gene, with HR 0.74 (CI 0.49-0.99) and p-value 0.021 (Figure 4.2). No differences were observed in gender specific analysis (data not shown).

DISCUSSION

We tested the hypothesis that common genetic variants which were previously shown in GWA studies to be associated with CAD risk in the general population, might affect the risk of CAD in a high-risk cohort of FH patients. As previously reported, established risk factors do associate with risk of CAD in FH patients[153, 34]. However, none of the tested CAD-associated SNPs significantly modified the risk of CAD in our FH cohort in analyses unadjusted or adjusted for established cardiovascular risk factors. The lowest observed p-value of association was for a SNP in the SMARCA44 gene, near the LDLR gene. (in adjusted analysis; p=0.021), however it showed a paradoxically protective effect.

FH patients are known to be at high CAD risk. Other patient cohorts with high risk are those with established cardiovascular disease and those with Type 2 diabetes. Of these three patient categories, the effect of 9p21 variants on survival had been tested only in those with established CAD. In a prospective observational study including 846 Caucasoid cases who underwent CABG, the 9p21 SNP rs10116277:G>T was independently associated with all-cause mortality during 5 years follow-up after surgery[164]. Homozygotes for the minor allele of this SNP had an increased risk of all-cause mortality (HR 1.7; CI 1.1-2.7). The SNP even remained associated with outcome after adjustment for the Euroscore, a score commonly used to predict CVD outcome after CABG. In contrast, in a larger cohort of patients with established CAD (>8000 patients), a haplotype block with 8 of the strongest 9p21 SNPs was associated with better prognosis in whites but not in blacks or Hispanics[165]. Moreover, in line with our findings, the HRs for prognosis among the risk alleles were in opposite direction compared to the published HRs for CAD/MI risk in the Caucasoid population for both the two most widely reported 9p21 SNPs; rs2383207:A>G and rs10757278:A>G, G alleles were 0.75 (0.60-0.93) and p=0.0083 and 0.81 (0.66-1.0) p=0.0523. However, a less commonly reported linkage disequilibrium consisting of six 9p21 SNPs was associated with worse prognosis[165]. Compared to these two studies in high-risk populations, our study adds a considerable number of events in a high-risk populations; a total of 482 CAD events occurred during follow-up in our cohort whereas the studies by Muehlschlegel and Gong reported analyses on 38 and 134 CVD events, respectively. In summary, the only two other studies that addressed the effect of 9p21 SNPs showed conflicting results in high risk populations.

Previous work conducted by our group to determine the genetic modifiers of CVD risk among FH patients showed that the G20210A polymorphism in the protrombin gene was strongly associated with significantly increased CVD risk[159]. However in that publication, the threshold to reach statistical significance was rather lenient. In this paper

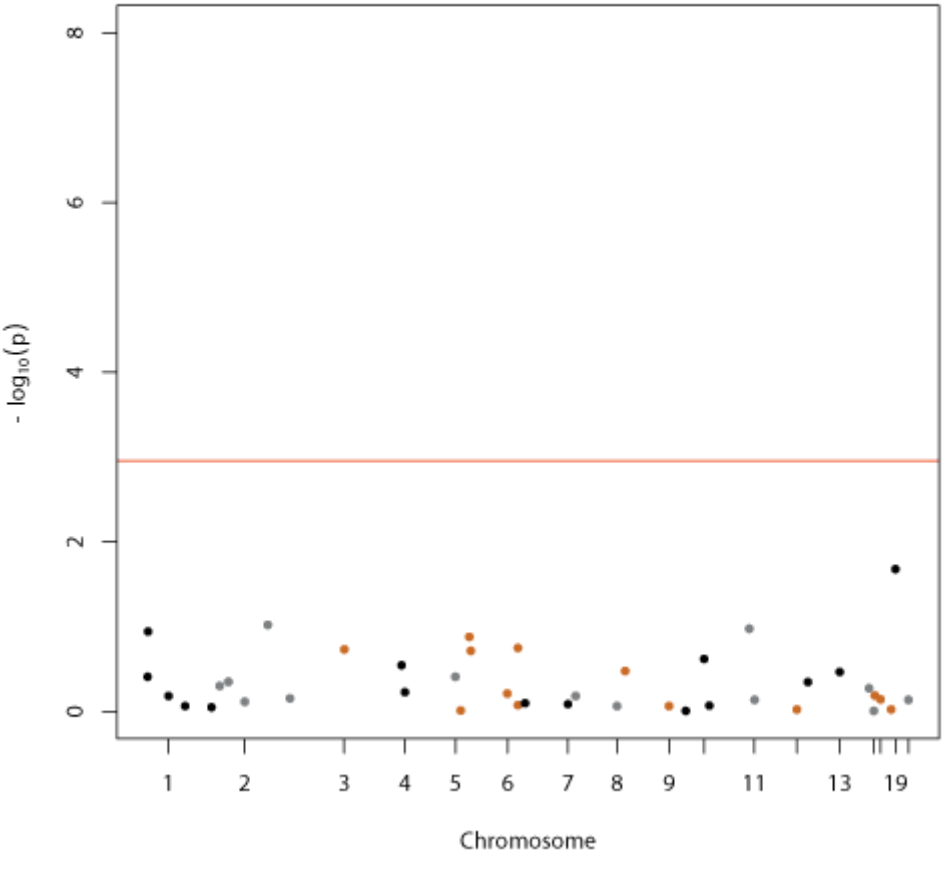

Figure 4.2: Manhattan plot containing the entire set of analyzed SNPs (41) associated with CAD. The Manhattan plot presents the -log10(p-value) of all the SNPs analysed in the current analysis adjusted for cardiovascular risk factors (age, gender, smoking, Type 2 diabetes, hypertension and BMI). Four of the 45 SNPs associated with CAD were not available for analysis in our data after imputation. The red line represents the p-value indicating statistical significance, taking into account Bonferroni correction for multiple testing.

a p-value <0.001 was considered statistically significant; however applying Bonferroni, which is common practice nowadays would suggest a p-value < -0.00076. None of the SNPs reached the a priori determined value for statistical significance. Because the reported protrombin variant is not considered to be a CAD risk SNP its effect on survival was not calculated in our current analysis. In current analysis only 1701 samples of the original 1940 were available. A selection bias has not taken place. This is merely a reflection of the usage of the DNA. Baseline characteristic are similar in both studies.

Our study did not address the underlying explanation for the rather counterintuitive result. The top SNP in our analysis, rs1122608:G>T, had a protective effect, contrary to the latest papers. However, smaller reports have reported a protective effect of this variant for CAD and PAD[166, 167]. Martinelli et all. suggest that the effect they observed is due to the lipid effects of the variants[166, 167].

The effect of common CAD-associated genetic variants on general population samples has also yielded conflicting results. In a study among approximately 3000 cases and 3000 controls, a gene score comprising nine SNPs was associated with CAD risk[168]. People in the top quintile for this gene score had a two-fold increased risk of MI compared to those in the bottom quintile corrected for age, sex and ancestry. In line, SNP based risk score designed in a Finnish cohort of 30,725 participants free of CVD was associated with the risk of a first CAD event, with a relative risk of 1.7 between the highest and lowest quintiles of the score adjusted for traditional risk factors; sex, LDL-C, HDL-C, current smoking, BMI, systolic and diastolic blood pressure, blood pressure treatment an prevalent Type 2 diabetes[169]. In contrast, Paynter and co-workers did not observe an independent association between genetic risk factors and CAD risk in a cohort of 19,313 initially healthy women during 12.3 years follow-up. A risk score based on 12 SNPs was clearly associated with CAD risk after adjustment for age. However, after additional adjustment for other traditional risk factors this association disappeared[170].

We have previously shown that the type of LDLR mutations underlying FH, variation in LDL-C levels, and established classical risk factors explain 21.3% of the variation in CAD risk. Our current analysis suggests that environmental and/or unknown genetic factors may play a role. Since there was no standardized information available on lifestyle factors such as dietary habits and physical activity, the effect of environmental factors and their potential interaction with genetic variants could not be studied, but it is unlikely that this could explain significant proportions of the remaining 80% of CAD risk prediction. At maximum, the common genetic variants we tested explained only 10% of the heritability of the trait, if we consider the heritability estimates of 40% for CAD to be correct[156]. Much of the missing heritability is expected to be explained by common variants not yet identified, rare variants, structural variants and copy number variations.

Several aspects of the design of our study need to be considered. The major strength of this study is the unparalleled cohort size, the detailed information on CAD events during follow-up, and the high CAD event rate. However, our study is hampered by several limitations. First, power calculations suggest that our study was at the limit of the power needed to detect statistically significant associations with CAD, since the studied SNPs were previously shown to have a moderate-effect (RR 1.1-1.7). Also the MAF of some of the SNPs were lower then used in the power calculation. Secondly, the majority of the CAD SNPs analysed were not directly genotyped, but imputed using HapMap and we can not rule out misclassifications. However, Southam et al have shown that imputation of common variants is generally very accurate[162]. Finally, the patients who were included in this study were referred to a Lipid Clinic. In theory, patients with the most detrimental genetic profiles might have died before referral. Therefore the effect of genetic variants associated with a more severe CAD phenotype or early death could have been underestimated or missed.

CONCLUSION

In this high-risk cohort of patients with FH, common SNPs shown to be associated with CAD risk in the general population could not be associated with the disease.

Table 4.2: The effect of the 41 SNPs analyzed in our study on CAD in FH

SNP	proxy	Chr	Pos	RA	RAF	HR	95% CI lower	95% CI upper	P-value	Nearest Gene	Genotyped / Imputed
rs17464857		1	220829332	G	0.11	1.0339	0.57	1.50	0.89	MIA3	Imputed
rs17114036		1	56735409	G	0.11	1.4222	0.98	1.86	0.11	PPAP2B	Imputed
rs11206510		1	55268627	C	0.17	1.0886	0.89	1.28	0.39	PCSK9	Genotyped
rs602633		1	109623034	T	0.19	1.0458	0.85	1.24	0.66	SORT1	Genotyped
rs4845625		1	152688691	T	0.44	1.0139	0.86	1.17	0.86	IL6R	Genotyped
rs6725887		2	203454130	T	0.14	1.2882	0.02	2.56	0.70	WDR12	Imputed
rs515135		2	21139562	T	0.15	1.0793	0.86	1.30	0.50	APOB	Imputed
rs6544713		2	43927385	T	0.33	1.0643	0.90	1.22	0.45	ABCG5-ABCG8	Imputed
rs1561198		2	85663500	C	0.46	0.9765	0.82	1.13	0.77	VAMP5-VAMP8-GGCX	Imputed
rs2252641		2	145517931	T	0.48	1.5015	1.03	1.97	0.10	ZEB2-AC074093.1	Imputed
rs9818870		3	139604812	T	0.16	0.2871	-1.55	2.13	0.18	MRAS	Imputed
rs1878406		4	148613114	T	0.15	0.8840	0.66	1.11	0.29	EDNRA	Imputed
rs7692387		4	156854759	A	0.2	0.0230	-13.48	13.53	0.59	GUCY1A3	Imputed
rs273909		5	131695252	G	0.15	0.8902	0.63	1.15	0.39	SLC22A4-SLC22A5	Imputed
rs12205331	rs12197124	6	35029419	T	0.21	0.6550	0.10	1.21	0.13	ANKS1A	Imputed
rs10947789		6	39282900	C	0.22	0.0001	-13.46	13.46	0.19	KCNK5	Imputed
rs4252120		6	161063598	C	0.3	0.9838	0.83	1.14	0.84	PLG	Imputed
rs9369640	rs7751826	6	12900977	C	0.37	0.9236	-3.59	5.44	0.97	PHACTR1	Imputed
rs2048327		6	160783522	T	0.37	0.8940	0.73	1.06	0.18	SLC22A3-LPAL2-LPA	Imputed
rs12190287		6	134256218	C	0.38	1.0986	0.73	1.46	0.61	TCF21	Imputed
rs2023938		7	19003300	C	0.12	1.1024	0.37	1.83	0.79	HDAC9	Imputed
rs11556924		7	129450732	C	0.4	0.9822	0.83	1.14	0.82	ZC3HC1	Genotyped
rs264		8	19857460	G	0.14	0.9500	0.72	1.18	0.66	LPL	Genotyped
rs2954029		8	126560154	T	0.45	0.9869	0.84	1.14	0.86	TRIB1	Genotyped
rs579459		9	135143989	C	0.25	0.9838	0.81	1.16	0.85	ABO	Imputed
rs3217992		9	21993223	C	0.37	1.0781	0.93	1.23	0.33	CDKN2BAS1	Imputed
rs12413409		10	104709086	A	0.07	1.0449	0.60	1.49	0.85	CYP17A1-CNNM2-NT5C2	Imputed
rs2047009		10	43859919	G	0.44	0.9702	-1.63	3.57	0.98	CXCL12	Imputed
rs11203042		10	90979089	T	0.44	0.9130	0.76	1.06	0.24	LIPA	Genotyped
rs9326246		11	116116943	C	0.08	0.9456	0.64	1.25	0.72	ZNF259-APOA5-APOA1	Imputed
rs974819		11	103165777	T	0.18	0.4280	-0.58	1.44	0.11	PDGFD	Imputed
rs3184504		12	110368991	C	0.44	0.9943	0.84	1.15	0.94	SH2B3	Genotyped
rs9515203		13	109847624	C	0.31	2.8507	0.67	5.03	0.34	COL4A1-COL4A2	Imputed
rs9319428		13	27871621	G	0.33	1.0639	0.90	1.22	0.45	FLT1	Imputed
rs7173743		15	76928839	T	0.45	0.9068	0.60	1.21	0.53	ADAMTS7	Imputed
rs17514846		15	89217554	C	0.46	0.9977	0.85	1.15	0.98	FURIN-FES	Genotyped
rs2281727		17	2064695	G	0.32	1.0620	0.80	1.32	0.65	SMG6	Imputed
rs12936587		17	17484447	A	0.47	1.0324	0.86	1.20	0.72	RAI1-PEMT-RASD1	Imputed
rs15563		17	44360192	G	0.48	0.9946	0.85	1.14	0.94	UBE2Z	Imputed
rs1122608		19	11024601	T	0.21	0.7415	0.49	0.99	0.02	LDLR	Imputed
rs9982601		21	34520998	C	0.21	0.6858	-1.43	2.80	0.73	Gene desert (KCNE2)	Imputed

Abbreviations: CAD, coronary artery disease; FH, familial hypercholesterolemia; SNPs, single-nucleotide polymorphisms.

5 PREDICTIVE VALUE OF A GENETIC RISK SCORE ON CARDIOVASCULAR RISK IN STATIN-TREATED, CORONARY PATIENTS

A large number of SNPs have been shown to be associated with risk for coronary artery disease (CAD) in genome wide association studies. Our objective was to determine whether these SNPs also have predictive value for a second cardiovascular event in atorvastatin-treated patients with established CAD.

We analyzed the genotype data of 1877 patients enrolled in the Treating to New Targets trial (TNT). Of these, 24.5% suffered from a second vascular event during follow-up, cases and controls were matched 1:3 on the basis. We investigated the predictive power of two different genetic risk scores (GRS), based on either the number of risk alleles or on the weighted effect-sizes of these 45 GWAS CAD SNPs. The cohort was divided in quartiles of GRS, in order to study the effect of an increasing GRS on developing a second event. The risk for a second vascular event was not statistically different in subjects in the highest versus the lowest GRS quartile (unweighted OR in 4th quartile=1.005 (0.99-1.01), p-value=0.06 and weighted OR in 4th quartile 1.004 (0.99 - 1.01), p-value=0.07). The association between the CAD SNPs with event-free survival time was analyzed using a Cox proportional hazard model, none of the SNPs was significantly associated with a second vascular event after Bonferroni correction.

These findings suggest that common CAD SNPs are not major determinants of the risk for a second vascular event in patients treated with atorvastatin. This holds true for SNPs tested individually and combined in a GRS.

INTRODUCTION

Coronary artery disease (CAD) remains a leading cause of morbidity and mortality[171] and the identification of additional risk factors and risk markers is therefore urgently warranted. In recent years, genome wide association studies (GWAS) have been successful in identifying genetic markers for CAD. It is, however, largely unknown whether these are also associated with CAD risk in secondary prevention settings. The latter is highly relevant because patients who suffered from a CAD event are at the highest risk of a recurrent event, even while being managed according to current guidelines[172]. Additional differentiation of risk stratification is warranted in these patients, who, in general, are treated with risk modifying agents such as statins, anti-platelet therapy and anti-hypertensive medications. A recent large-scale GWAS identified common genetic variations at 45 loci that moderately affected the incidence of a first CAD event in Caucasians (odds ratios [OR] varying between 1.1 and 2.0)[157, 14]. We set out to address the following questions: 1) Do genetic risk scores based on those 45 CAD risk single nucleotide polymorphisms (SNPs) have any predictive value in secondary care prevention settings? and 2) Are the 45 SNPs also independently associated with the risk of a vascular event to occur in patients who already suffered a CAD event? For this purpose, we analyzed the data of a large randomized clinical trial that compared the effect of atorvastatin 80 mg vs. atorvastatin 10 mg on cardiovascular events in patients with coronary heart disease, the Treating to New Target (TNT) study.

MATERIAL AND METHODS

Study design, genotyping and imputation

The design and outcome of the TNT study as well as the genotyping study have been previously described[173]. In short, 10,001 patients aged 25 to 70 years with clinical evident coronary heart disease (defined as previous myocardial infarction, previous or current angina with objective evidence of atherosclerosis or a history of coronary revascularization) were randomized to either a daily dose of 10 or 80 mg atorvastatin. Patients were followed for 5 years. A total of 2092 individuals were selected based on cardiovascular events during the study for genotyping on the Perlegen 322K platform, as previously described[174]. The genotype data from 1,984 patients was analysed for the current analysis; 489 of these patients suffered an adverse vascular event during the study and 1,495 did not suffer from such event and were thus considered a control. Cases and controls

were matched 1:3 on the basis of age, gender, treatment arm, smoking, diabetes, hypertension, baseline lipid values, baseline glucose levels and screening plasma low density lipoprotein-cholesterol levels (LDL-C). A vascular event was defined as death from coronary heart disease, nonfatal myocardial infarction, and resuscitation after cardiac arrest, and fatal or nonfatal stroke. After genotyping, the GenABEL package in R was used to test the SNPs for population substructure that could introduce false-positive associations. An identity-by-state analysis was performed to ensure that only Caucasians were included in the association analyses. SNPs were subjected to quality control (QC) filters based on sample-size, minor allele frequency (MAF) and Hardy-Weinberg equilibrium. Samples with a call-rate of $< 95\%$ were excluded from further analysis. Thirty-seven of the 45 CAD SNPs were not available on the Perlegen 322K platform and were imputed using MACH[161] and the HapMap phase 2 datasets (build 36 release 22). Imputed SNPs with a $R^2 < 0.3$ were removed during post-imputation QC. After imputation all 45 CAD SNPs were available for analysis.

Statistical Analysis

We combined the 45 CAD SNPs into two different genetic risk scores (GRS); the first GRS was based on the total number of risk alleles (RA) in each subject, according to a previously described method[175]. The second GRS was a weighted GRS, where the risk allele was weighted with the reported effect-size in the original published article[14] (odds ratios ranged from 1.04 to 1.28). We ran two different logistic regression models for both GRS, one while including age and sex as covariates, and a second including age, sex, body mass index (BMI), LDL-C, hypertension, type 2 diabetes and smoking). For the GRS analyses a p-value of <0.05 was considered statistically significant. In order to illustrate the effect of an increasing GRS on developing a secondary event, we separated the TNT cohort into four quartiles, on the basis of the calculated GRS, using the lowest GRS quartile as reference. Statistical analyses were performed using R (version 3.0.2).

In a second analysis we tested the association for each of the 45 CAD SNP with occurrence of an event using Cox proportional hazards-models, adjusted for classical risk factors of CAD: age, sex, smoking, type 2 diabetes, hypertension and BMI in ProbABEL[163]. Significance was defined as a p-value below $1.11x10^{-3}$ (which is the result of a Bonferroni corrected p value, i.e. 0.05 divided by 45; the number of SNPs tested).

RESULTS

Demographic data

Upon quality control, genotyping data from 1,877 of the original 1,984 individuals and a total of 259,580 of the 322,185 SNPs were available for analysis. A total of 107 individuals were excluded after QC because of low call-rate and population stratification. 459 (24.5%) patients suffered a primary endpoint during follow-up (median 4.9 years), while the remainder (n=1,418) were free from a primary endpoint at the end of the study. The characteristics of these 1,877 individuals are shown in Table 5.1.

GRS and the risk of a second vascular event

The average GRS was not significantly higher in CAD cases compared to controls (47.0 $\pm$ 4.15 versus 46.6 $\pm$ 4.04, p=0.07). Moreover, the unadjusted GRS was not significant in the risk of developing a secondary CAD event while comparing patients in the lowest GRS quartile versus the highest GRS quartile (OR= 1.005 95% CI (0.99-1.01), p-value=0.06). A similar result was obtained while comparing the weighted unadjusted GRS in these quartiles (OR= 1.004 95% CI 0.99 - 1.1, p=0.07) (Table 5.2).

SNPs and risk of a secondary CAD event

Eight of the selected CAD SNPs were directly available on the genotyping platform and upon imputation, genotype data was available for all 45 SNPs (mean r^2 = 0.90 and mean quality index 0.96). None of the SNPs were significantly associated with a second CAD event after Bonferroni correction (p$< 1.11x10^{-3}$) (see Table 5.3). rs4252120 in the plasminogen (PLG) gene (MAF 0.30) showed the strongest association with CAD risk (HR 0.85 (0.71-1.00) uncorrected p= 0.03) followed by rs9515203 in the COL4A1-COL4A2 gene cluster (MAF 0.28, HR=0.85 (0.68-1.03, p-value= 0.07) and rs15563 in the UBE2Z gene (MAF 0.45, HR=0.89 (0.76-1.02), p-value= 0.09 (see Table 5.3).

Table 5.1: Baseline characteristics of study patients with/without a primary event

	Patients with primary endpoint (N=459)	Patients without primary endpoint (N=1418)	p-value
Age	62.38 ± 8.59	62.60 ± 8.26)	0.54
Males (n,%)	377 (82.13%)	1169 (82.44%)	0.88
BMI [kg/m2]	29.62 ± 5.28	28.87± 4.43	0.006
SBP [mmHg]	133.58 ± 18.14	131.84 ± 17.06)	0.07
DBP [mmHg]	78.54 ± 10.02)	77.77 ± 9.46	0.15
Diabetes (n,%)	107 (23.31%)	310 (21.86%)	0.52
Hypertension (n,%)	305 (66.45%)	932 (65.73%)	0.78
Treatment group(number and percentage treated with 80mg atorvastatin) (n,%)	280 (43.79%)	617 (43.51%)	0.85
HDL-C* [mg/dl]	45.12 ± 10.37	45.80 ± 9.84)	0.22
LDL-C* [mg/dl]	98.50 ± 17.49	97.21 ± 15.48	0.16
TC* [mg/dl]	175.95 ± 24.12	173.73 ± 21.54	0.09
TG *[mg/dl]	163.92 ± 83.59	154.31 ± 67.27	0.03
Glucose [mg/dl]	113.81 ± 35.59	111.36 ± 32.60	0.19

Values are given as number (percentage) or mean ± standard deviation unless indicated otherwise;
BMI indicates Body Mass Index; SBP indicates Systolic Blood Pressure; DBP indicates Diastolic Blood Pressure; LDL indicates low-density lipoprotein cholesterol; HDL indicates high-density lipoprotein cholesterol; TC indicates Total Cholesterol; TG indicates Triglycerides.
* Lipid values were measured after a run-in period on atorvastatin 10 mg.

Table 5.2: Association between the 2 genetic risk scores and the primary endpoint.

			Number of risk alleles					
GRS	model	<=44	>44 <=47	>47 <=49	>49	All p-value*	OR	95% CI
CAD GRS	1	ref	0.99 (0.93-1.044)	0.99 (0.93-1.05)	1.07 (1.01 - 1.13)	0.063	100.453	0.9997	10.093
CAD GRS	2	ref	0.99 (0.94-1.05)	1.03 (0.97-1.09)	1.05 (0.995 - 1.10)	0.062	100.455	0.9997	10.093
CAD GRS	3	ref	0.995 (0.94-1.05)	1.03 (0.98-1.09)	1.05 (0.997 - 1.11)	0.054	100.470	0.9999	10.095
N		545	524	365	443			
number cases		125	117	94	123			
% cases		22.9	22.3	25.8	27.8			
				Number of risk alleles				
GRS	model	<=46.96	>46.96 <=50.13	>50.13 <=52.40	>52.40	All p-value*	OR	95% CI
weighted CAD GRS	1	ref	1.00 (0.95 - 1.06)	1.01 (0.95 - 1.06)	1.04 (0.99 - 1.10)	0.0655	10.042	0.9997	10.087
weighted CAD GRS	2	ref	1.00 (0.95 - 1.06)	1.01 (0.95 - 1.06)	1.04 (0.99- 1.10)	0.0643	10.042	0.9997	10.087
weighted CAD GRS	3	ref	1.00 (0.95 - 1.06)	1.01 (0.95 - 1.06)	1.05 (0.99- 1.10)	0.0563	10.044	0.9999	10.089
N		545	524	365	443			
number cases		125	117	94	123			
% cases		22.9	22.3	25.8	27.8			

*comparison between the highest and lowest quartile

Model 1: logistic regression model unadjusted

Model 2: logistic regression model adjusted for age and sex

Model 3: logistic regression model adjusted for age,sex and known CAD Riskfactors (BMI,LDL-C,Hypertension,Diabetes and smoking)

Table 5.3: Association results of the 45 CAD SNPs with event free survival time

SNP	Chr	Pos	RA	RAF	HR	SE	95% CI lower	95% CI upper	p-value	Nearest Gene	RAF Cardiogram
rs4252120	6	161063598	T	0,7	0,85	0,07	0,71	1	0,03	PLG	0,73
rs9515203	13	109847624	T	0,72	0,85	0,09	0,68	1,03	0,07	COL4A1-COL4A2	0,74
rs15563	17	44360192	G	0,55	0,89	0,07	0,76	1,02	0,09	UBE2Z	0,52
rs2048327	6	160783522	C	0,38	1,12	0,07	0,99	1,26	0,1	SLC22A3-LPAL2-LPA	0,35
rs2047009	10	43859919	G	0,47	0,91	0,07	0,78	1,04	0,16	CXCL12	0,48
rs2895811	14	99203695	C	0,42	0,91	0,07	0,78	1,04	0,16	HHIPL1	0,43
rs11556924	7	129450732	T	0,6	0,9	0,08	0,74	1,06	0,2	ZC3HC1	0,65
rs602633	1	109623034	T	0,81	1,13	0,1	0,94	1,32	0,2	SORT1	0,77
rs17464857	1	220829332	T	0,9	0,83	0,15	0,55	1,12	0,21	MIA3	0,87
rs12190287	6	134256218	C	0,66	1,08	0,07	0,95	1,22	0,25	TCF21	0,59
rs273909	5	131695252	A	0,13	1,13	0,11	0,92	1,34	0,25	SLC22A4-SLC22A5	0,14
rs12936587	17	17484447	G	0,65	0,93	0,07	0,79	1,06	0,29	RAI1-PEMT-RASD1	0,59
rs6725887	2	203454130	C	0,13	1,1	0,09	0,92	1,29	0,3	WDR12	0,11
rs12413409	10	104709086	G	0,92	0,9	0,12	0,66	1,14	0,39	CYP17A1-CNNM2-NT5C2	0,89
rs2954029	8	126560154	A	0,56	1,05	0,07	0,92	1,19	0,44	TRIB1	0,55
rs3217992	9	21993223	A	0,41	1,05	0,07	0,92	1,19	0,44	CDKN2BAS1	0,38
rs10947789	6	39282900	T	0,79	1,07	0,08	0,9	1,23	0,45	KCNK5	0,76
rs7692387	4	156854759	G	0,82	0,94	0,09	0,77	1,11	0,46	GUCY1A3	0,81
rs17114036	1	56735409	A	0.91	1,09	0,12	0,85	1,33	0,49	PPAP2B	0,91
rs12539895	7	106879085	A	0,23	1,06	0,08	0,9	1,21	0,5	7q22	0,19
rs2023938	7	19003300	C	0,11	1,08	0,11	0,86	1,29	0,5	HDAC9	0,1
rs1561198	2	85663500	C	0,49	0,96	0,07	0,82	1,09	0,51	VAMP5-VAMP8-GGCX	0,45
rs9319428	13	27871621	A	0,31	1,05	0,07	0,91	1,19	0,52	FLT1	0,32
rs9818870	3	139604812	C	0,17	1,06	0,09	0,88	1,24	0,52	MRAS	0,14
rs17514846	15	89217554	A	0,48	0,95	0,08	0,79	1,12	0,56	FURIN-FES	0,44
rs9369640	6	13009427	A	0,66	1,04	0,07	0,9	1,18	0,6	PHACTR1	0,65
rs11203042	10	90979089	T	0,47	0,97	0,07	0,83	1,1	0,62	LIPA	0,38
rs2281727	17	2064695	A	0,36	1,03	0,07	0,9	1,17	0,63	SMG6	0,36
rs12205331	6	35006433	C	0,81	0,96	0,09	0,8	1,13	0,65	ANKS1A	0,81
rs445925	19	50107480	G	0,88	1,08	0,18	0,73	1,43	0,68	ApoE-ApoC1	0,9
rs974819	11	103165777	C	0,28	0,97	0,08	0,82	1,12	0,68	PDGFD	0,29
rs2252641	2	145517931	C	0,47	1,03	0,07	0,89	1,16	0,71	ZEB2-AC074093.1	0,46
rs2505083	10	30375128	C	0,42	0,97	0,08	0,82	1,13	0,74	KIAA1462	0,42
rs4845625	1	152688691	T	0,44	1,02	0,07	0,89	1,15	0,76	IL6R	0,47
rs515135	2	21139562	T	0,83	1,02	0,1	0,84	1,21	0,82	APOB	0,83
rs9326246	11	116116943	C	0,07	1,03	0,13	0,77	1,29	0,83	ZNF259-APOA5-APOA1	0,1
rs1878406	4	148613114	C	0,14	1,02	0,1	0,83	1,21	0,86	EDNRA	0,15
rs6544713	2	43927385	C	0,33	1,01	0,07	0,88	1,15	0,86	ABCG5-ABCG8	0,3
rs3184504	12	110368991	T	0,49	0,99	0,07	0,86	1,12	0,93	SH2B3	0,4
rs579459	9	135143989	C	0.23	1,01	0,08	0,85	1,17	0,94	ABO	0,21
rs11206510	1	55268627	T	0.83	1	0,12	0,77	1,22	0,97	PCSK9	0,84
rs1122608	19	11024601	G	0,78	1	0,08	0,84	1,15	0,97	LDLR	0,76
rs9982601	21	34520998	T	0.16	1	0,09	0,82	1,17	0,97	Gene desert (KCNE2)	0,13
rs7173743	15	76928839	T	0.46	1	0,07	0,87	1,13	0,98	ADAMTS7	0,38
rs264	8	19857460	G	0.86	1	0,1	0,8	1,2	0,99	LPL	0,86

DISCUSSION

In the current study we showed that genetic risk scores based on 45 known CAD risk loci have no predictive value in a cohort of atorvastatin-treated patients with coronary heart disease. It should be noted, however, that the p-values for both GRS models almost reached significance (p=0.06 and p=0.07), which is comparable with the finding by Tragante and co-workers, who showed a borderline significant association between a CAD GRS comprising 30 CAD SNPs and the risk of recurrent myocardial infarction (p=0.05)[176]. The latter study was performed in a relatively small number of patients suffering from a second event (n=72), and despite the significant larger number of patients suffering from a second event (n=459) we do not observe a significant association between our GRS and recurrent events either. It might be that the association would become statistical significant upon a further increase of the sample size, but the clinical relevance of the effect size is likely to remain very small. Moreover, almost none of the tested single CAD-associated SNPs significantly modified the risk of CAD in our cohort in analyses unadjusted or adjusted for established cardiovascular risk factors. Even the strongest association (rs4252120 in the plasminogen (PLG) gene p=0.03) did not reach the significance level upon Bonferroni correction.

Several aspects of the design of our study require consideration. The major strength of this study is the relatively large population and the focus on adjudicating endpoints during the conduct of the clinical trial. However, our study has limitations. First, power calculations suggest that our study was at the limit of the power (power of 45 SNPs between 5%-91%) needed to detect statistically significant associations with vascular events during follow-up, since the studied SNPs were previously shown to have a moderate-effect. Unfortunately we were not able to conduct a replication study to confirm the absence of an effect of the GRS on vascular endpoints. Secondly, the majority of the CAD SNPs (n=37) analyzed were not directly genotyped, but imputed, and we cannot rule out misclassifications. However, imputation has been shown to generate accurate classifications[162]. Thirdly, the CAD SNPs might not necessarily be predictive for the endpoint "vascular events" (including stroke, which was the predefined endpoint in the TNT trial) studied in this analysis. However, while only including patients who suffered from a second CAD event, we did not observe an effect of the GRS either (data not shown).

Finally, the patients who were included in this study all survived a first CAD event, so this could be considered a "selected high-risk" population. In theory, patients with the most detrimental genetic profiles might have died before inclusion into this study. We

might therefore be confronted with survival bias, where the effect of genetic variants associated with a more severe CAD phenotype or early death are underestimated or missed. As a contrast one could also hypothesize that we enriched our analysis for high-risk alleles, because all patients had coronary heart disease. The MAF of the 45 known CAD loci were, however, comparable to the MAF in the CARDIoGRAMplusC4D consortium report, which makes the latter effect somewhat unlikely[14].

In conclusion, the 45 SNPs combined in two distinct GRS were not associated with a second vascular event, and none of the 45 CAD SNPs on their own was significantly associated with outcome in this large cohort of patients participating in a secondary prevention statin trial either. These findings suggest that the putative CAD SNPs identified in GWAS have no, or at best a minor effect on the risk for a secondary vascular event in patients treated with atorvastatin. This implicates that the secondary event is caused by other factors, and that these factors may not be explained by common genetic variations, that have been shown to be a determinant for the first cardiovascular event.

Part III

Imputation

6 EXTENDING THE USE OF GWAS DATA BY COMBINING DATA FROM DIFFERENT GENETIC PLATFORMS

In the past decade many Genome-wide Association Studies (GWAS) were performed that discovered new associations between single-nucleotide polymorphisms (SNPs) and various phenotypes. Imputation methods are widely used in GWAS. They facilitate the phenotype association with variants that are not directly genotyped. Imputation methods can also be used to combine and analyse data genotyped on different genotyping arrays. In this study we investigated the imputation quality and efficiency of two different approaches of combining GWAS data from different genotyping platforms. We investigated whether combining data from different platforms before the actual imputation performs better than combining the data from different platforms after imputation.

In total 979 unique individuals from the AMC-PAS cohort were genotyped on 3 different platforms. A total of 706 individuals were genotyped on the MetaboChip, a total of 757 individuals were genotyped on the 50K gene-centric Human CVD BeadChip, a total of 955 individuals was genotyped on the HumanExome chip. A total of 397 individuals was genotyped on all 3 individual platforms. After pre-imputation quality control (QC), Minimac in combination with MaCH was used for the imputation of all samples with the 1000genomes reference panel. All imputed markers with an r^2 value of $<$0.3 were excluded in our post-imputation QC.
A total of 397 individuals were genotyped on all three platforms. All three datasets were carefully matched on strand, SNP ID and genomic coordinates. This resulted in a dataset of 979 unique individuals and a total of 258,925 unique markers. A total of 4,117,036 SNPs were available when imputation was performed before merging the three datasets. A total of 3,933,494 SNPs were available when imputation was done on the combined set. Our results suggest that imputation of individual datasets before merging performs slightly better than after combing the different datasets.
Imputation of datasets genotyped by different platforms before merging generates more SNPs than imputation after putting the datasets together.

INTRODUCTION

In the past decade many Genome-wide Association Studies (GWAS) were executed in order to assess the relative contribution of single-nucleotide polymorphisms (SNPs) in various phenotypes. At this moment more than 21,750 SNPs were significantly associated, in more than 2,437 studies ($http://www.ebi.ac.uk/GWAS$ Accessed [May 2016]) with one or more phenotypes by GWAS. These GWAS were performed on a wide range of different genotyping platforms, with a great diversity in the number and density of SNPs, ranging from 50k to 1 million SNPs.

Imputation methods are widely used in GWAS, since this will provide information about variants that are not genotyped directly. Imputation can also be used to combine data genotyped on different genotyping arrays. The quality of imputation is discussed in several papers[162, 177], and imputation methods are used in all large genetic meta-analysis consortia (eg. CARDIoGRAM, MAGIC and GIANT)[14, 178, 179]. The standard method used in these large consortia combining data of studies genotyped on different platforms is to impute the individual cohorts locally and combine them later centrally. The most import reason to analyse data using this method is the computational efficiency. Another imputation method mainly used in case-control studies, is to use only SNPs that are common to all the genotyping platform used in the analysis to remove potential platform specific imputation errors, when cases and controls are genotyped on different platforms.

In this paper we investigated whether there is a difference in the imputation quality (number of imputed SNPs with $r^2 > 0.3$) and efficiency between imputing different platforms by themselves and combine them afterwards versus combining platforms before the imputation. We hypothesize that combining the individual platforms before the imputation will lead to more imputed good quality SNPs, because we had more genotyped SNPs. We used the AMC-PAS[180] cohort which is genotyped on 3 different platforms: MetaboChip, 50K gene-centric Human CVD BeadChip and the HumanExome BeadChip.

MATERIALS AND METHODS:

We included a total of 979 unique patients from the previously described prospective cohort AMC-PAS[180], with symptomatic Coronary Artery Disease (CAD) before the age of 51 years, defined as Myocardial Infarction (MI), coronary revascularization, or

evidence of at least 70% stenosis in a major epicardial artery. The samples were genotyped on at least one of the three different platforms, all manufactured by Illumina: The MetaboChip[16], The Human cardiovascular disease (HumanCVD) BeadChip (Illumina, San Diego, CA, USA), also known as the ITMAT-Broad-CARe (IBC)(IBCv2 array)[17] and the HumanExome BeadChip(version 24 v1.0)[181].
The MetaboChip consists of approximately 200,000 SNPs chosen based on GWAS meta-analyses of 23 metabolic traits[16].
The 50K gene-centric Human CVD BeadChip has approximately 50,000 SNPs on the array in about 2000 genes in relevant loci across a range of cardiovascular, metabolic and inflammatory syndromes[17].
The HumanExome BeadChip contains about 250,000 variants based on the data of 12,000 sequenced genomes and exomes. Each variant on the chip has been seen at least 3+ times across at least 2 different data sets.
A total of 706 individuals were genotyped on the MetaboChip, a total of 757 individuals were genotyped on the 50K gene-centric Human CVD BeadChip and a total of 955 individuals were genotyped on the HumanExome chip. A total of 397 individuals were genotyped on all three platforms. We only included autosomal chromosomes in our analysis.

Ethics:

The Institutional Review Board of the Academic Medical Center in approved the protocol. All patients gave informed consent.

Pre-imputation quality control:

After genotyping, PLINK v1.07 $(http://pngu.mgh.harvard.edu/purcell/plink/)$ was used in the genotype data generated by all 3 platforms to test the SNPs for population substructure which could introduce false-positive associations. This was done by means of multidimensional scaling[58], individuals identified as population outliers were removed from the genotype data from all 3 platforms. Also SNPs were subjected to quality control filters based on sample size and minor allele frequencies (MAF). Samples with a call rate of $<95\%$ were excluded from further analysis. Genetic markers with a MAF $<1\%$ and Hardy-Weinberg equilibrium p $< 10^{-4}$ were excluded from further analysis. An identity-by-state (IBS) analysis was performed to remove related samples from the analyses. We checked if the genotypes for SNPs available on multiple platforms were

concordant, by comparing the genotypes of the SNPs available on all 3 platforms using the *-merge-mode 7* option in PLINK.

Imputation:

Minimac[182] in combination with MaCH[161] were used for the imputation of the combined set of individuals and markers with the 1000 Genomes reference panel (Phase 1 Version march 2012) including all ethnicities. We used a two-step imputation approach. First, the haplotypes of the entire sample were estimated using MaCH followed by haplotype-to-haplotype imputation by Minimac. Minimac generates the allele dosages for each of the variants. Minimac also generates the SNP-level quality metric Rsq (r^2). The r^2 value is the SNP-specific estimated squared correlation between the allele dosages and the unknown true genotype. r^2 is an efficient post-imputation quality control value. We used an r^2 value of $>$ 0.3 as our post-imputation QC[182, 183, 184].
To validate our method, we performed the same analysis with the other widely used imputation software IMPUTE2[185] (chromosome 22 only) and we performed the same analysis for only 2 platforms (chromosome 22 only). We imputed all individuals that were genotyped on the Exomechip (N=853) and the individuals that were only genotyped on the Metabochip (N=78). First we performed imputation analyses on these two datasets separately, and combined the results afterwards. In our second approach we merged the two data sets before we performed the imputation analyses.

Association Analysis:

Finally, we performed association analysis of both imputed data sets for the Low-Density lipoprotein concentration (LDL-C) phenotype relevant for premature cardio atherosclerotic disease. All analyses were adjusted for age and sex. The results were filtered based on bonferroni correction (p-values $< 5x10^{-8}$).

RESULTS:

The number of SNPs available after QC is shown in Table 6.1 The concordance for SNPs genotyped on all 3 genotyping platforms was perfect. All SNPs genotyped on more than one platform had the same genotype call. The call rate of the individual arrays was 99.8%, when we combine the different arrays the call rate drops to 74%, this is

Table 6.1: Number of genotyped SNPs available after QC on the different genotyping platforms

Array	#Individuals (unique)	#SNPs	Call rate
IBC CardioChip	718	35,092	0.998
MetaboChip	585	113,685	0.997
Exomechip	853	114,967	0.999
Combined	979	258,925	0.739
Combined (only overlapping individuals	397	258,896	0.998455

because we introduce more missing SNP data, because not all SNPs are available on all 3 platforms.

Combining different platforms together:

All three datasets were carefully matched on strand, SNP ID and genomic coordinates. This resulted in a dataset of 979 unique individuals and a total of 258,925 SNPs. 397 individuals were genotyped on all three platforms. A total of 36 individuals were only genotyped on the IBC Cardiochip, 24 individuals were only genotyped on the Metabochip and 115 individuals were uniquely genotyped on the Exomechip. The Venn diagram in Figure 6.1 illustrates the overlap of individuals on the different genotyping platforms after pre-imputation quality control.

Imputation:

The computational time needed to impute a chromosome varied between 153 hours (chr1) and 23 hours (chr22). There was no significant difference between combining the datasets before imputation or combining the datasets after imputation.
After imputation and post-imputation quality control there were a total of 70,7871 SNPs available for the IBC Cardiochip array, a total of 2,853,265 SNPs available for the Metabochip, a total of 1,586,399 SNPs for the Exomechip and a total of 3,933,494 SNPs for the combined dataset of those 3 platforms. The Venn-diagram in Figure 6.2 gives an overview of the overlap of the total number of SNPs available after imputation between the different data sets. As an example: 305,373 SNPs were uniquely available after imputation of the MetaboChip. A total of 109,229 unique SNPs were available from each

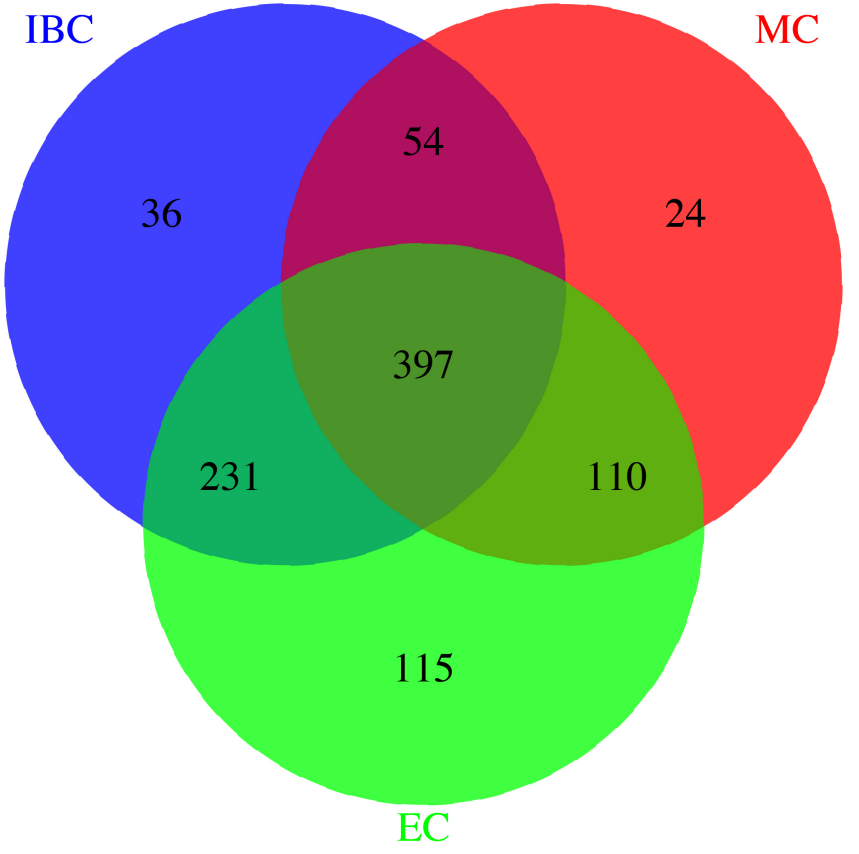

Figure 6.1: Overlap of individuals genotyped with the three platforms (after QC).

Method	<1%	>1% <5%	>5% <10%	>10%
Combined before imputation	512,579	591,523	486,281	2,343,111
Combined after imputation	615,779	579,945	497,244	2,715,283
% Difference between methods	17%	2%	2%	14%

Table 6.2: MAF distribution of the two imputation methods

of the 3 different platforms and the combined dataset. To validate our analyses, we did perform the same analysis with IMPUTE2 for both methods and we did find the same result, combining after imputation leads to more good quality SNPs.

We observed a difference in the number of available SNPs after imputation between the two imputation approaches. Combining the individual datasets after imputation, resulted in more SNPs than combining the 3 datasets before imputation. A total of 4,117,036 unique SNPs was available after combining the three different sets after separate imputation versus a total of 3,933,494 SNPs after the imputation of the combined set, a 5% difference (ranging between 1% (chr15) and 10% (chr22)) in the number of available SNPs after imputation, see Figure 6.3.
The overlap of SNPs of the combined data sets and the union of the independently imputed datasets was 3,428,387 SNPs. This meant that 1,242,341 SNPs were only imputed by one of the two different imputation approaches.
We also observed that more low frequency variants were imputed with the method when we combined the different data sets after imputation than imputation after combining the datasets, shown in Table 6.2. Figure 6.4 shows the $r^2 >$ distribution for all overlapping SNPs for both methods and the MAF distribution for both methods.

The analysis where we combined data of 2 platforms after imputation analyses (chromosome 22 only) resulted in a total of 31,241 available SNPs of good quality ($r^2 > 0.3$) after merging the 2 separate imputed datasets. In our second approach we merged the two data sets before we performed imputation analyses, this analyses resulted in a total of 13,516 available SNPs of good quality ($r^2 > 0.3$).

We repeated the imputation analysis for the subgroup of individuals that were genotyped on all three platforms (n=397). The number of SNPs available after combining the three separately imputed sets was 3,245,164 versus a total of 3,528,322 SNPs after the imputation of the combined set, a 8% difference between the 2 imputation methods in favour of the method combining the 3 platforms before imputation.

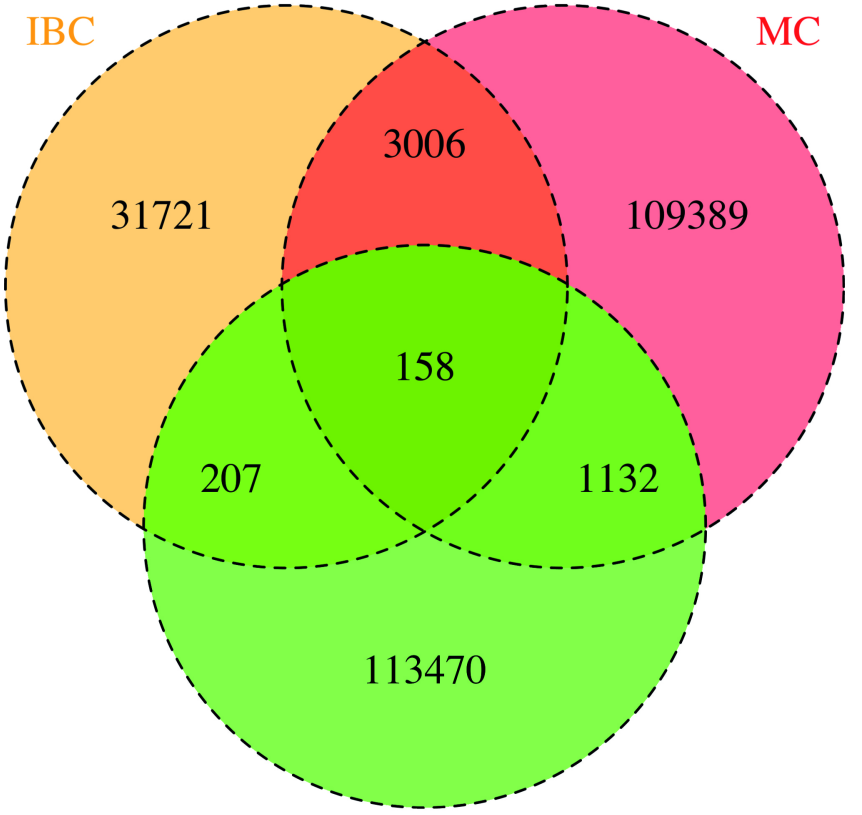

Figure 6.2: Overlap SNPs after imputation on the different platforms and the combined imputation.

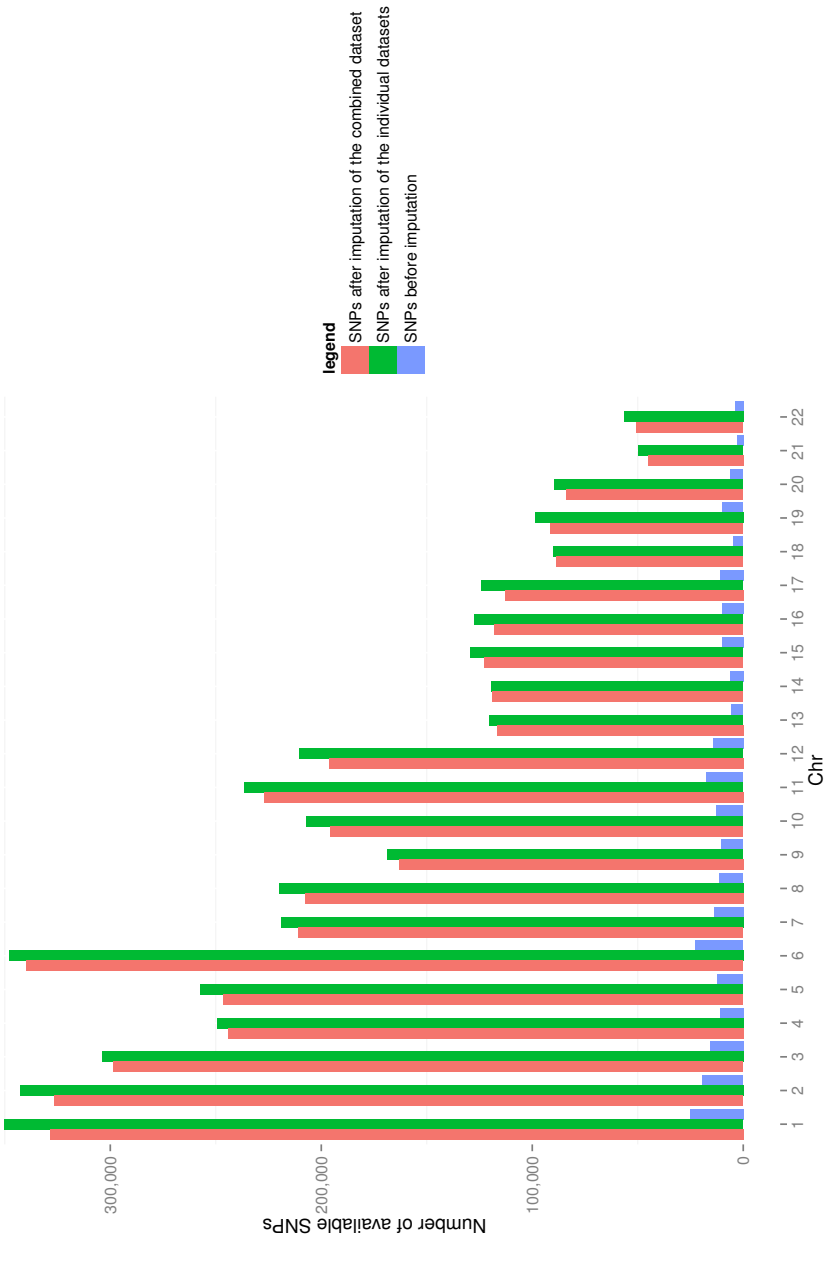

Figure 6.3: Histogram of number of SNPs available after both imputation methods

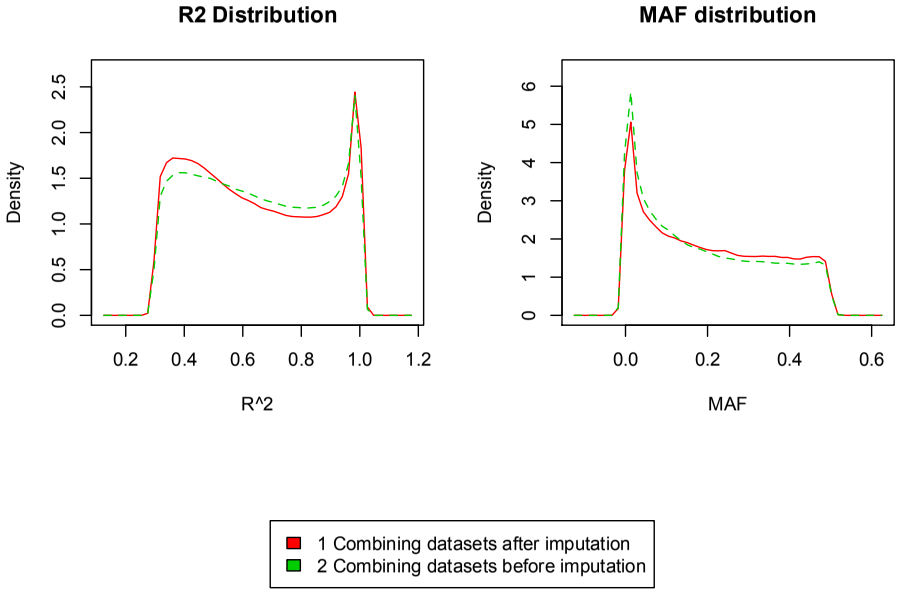

Figure 6.4: r^2 and MAF distribution for the two imputation methods

Association analysis comparison:

Association analysis was performed in both imputed datasets with the LDL-C phenotype in the AMC-PAS cohort. No significant association was found with the LDL-C phenotype in both methods. Manhattan plots and QQ plots for both imputed datasets are shown in Figure 6.5.
We have highlighted known associated SNPs reported in the NHGRI GWAS Catalogue (available at: http://www.ebi.ac.uk/gwas/) for the LDL cholesterol trait. After removing duplicate SNPs, a total of 118 SNPs were available. Of these, 113 SNPs were available in the dataset combined before imputation analyses and 105 SNPs in the dataset combined after imputation. None of the p-values of these known LDL-C associated SNPs was associated with LDL-c in our analyses. In both datasets a variant in the HMGCR gene was the most significant known LDL-c associated SNP. In the dataset combined after imputation rs3846662 p=$6.62x10^{-4}$ and in the dataset combined before imputation rs7703051 p=$7.97x10^{-4}$.

DISCUSSION:

Combining different imputed GWAS datasets is a very powerful method to identify new loci using data from different genotyping platforms[186].
Our study illustrates that combining imputed data from different platforms before and after imputation does result in differential numbers and quality of SNPs to be analysed. When imputation of different data sets is performed prior to combining resulted in 5% more SNPs that satisfied the post-imputation QC, this was not what we hypothesized and we deem those to explained by the r^2 selection during the imputation process. When a SNP is genotyped on only one of the platforms, a lot more uncertainty will be introduced in the set when we combined the 3 platforms before imputation, because this SNP is only available on one of the three platforms. When we apply a call-rate filter of 95% or higher, we found the opposite result, the method where we combine the datasets before imputation results in more good quality SNPs. To validate our method, we performed the same analysis with IMPUTE2[185] (chromosome 22) and we found the same result, combined after imputation leads to more good quality SNPs. To validate our chosen r^2 threshold we did run the analysis with different r^2 thresholds ($r^2 > 0.5$ and $r^2 > 0.8$), resulting in the same difference of good quality SNPs between the different imputation methods.
We found the opposite result when we analysed only data from individuals (N=397) genotyped on all 3 platforms, where we were able to impute 8% more SNPs that satisfied our

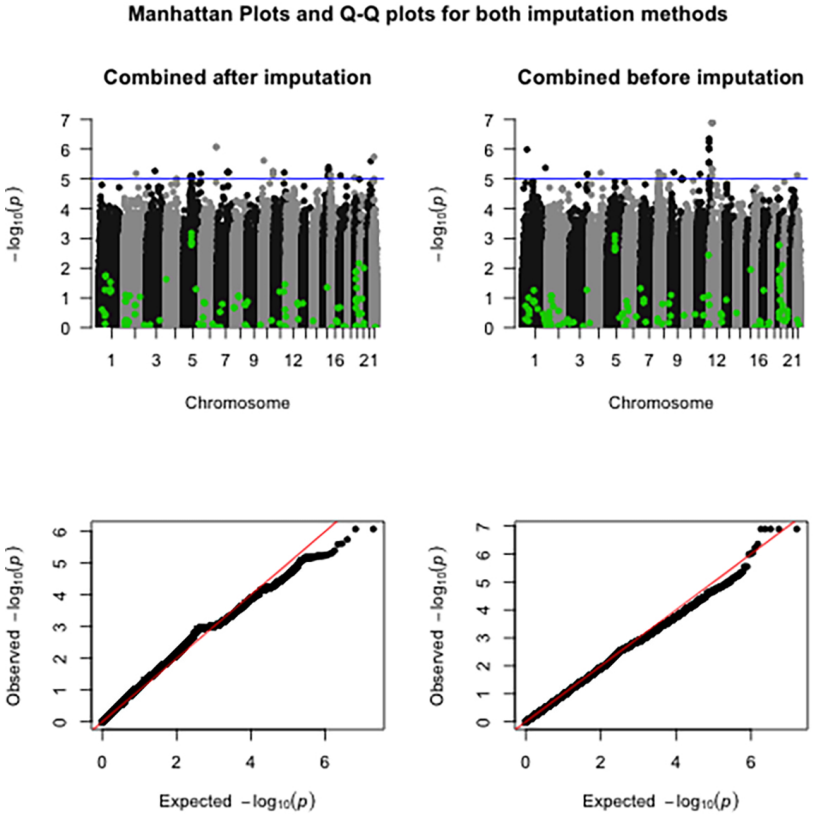

Figure 6.5: Manhattan plots and Q-Q plots for the association analyses for both imputation methods

post-imputation QC when we first combine before performing imputation. We think this is because we do not introduce extra missing data.
To validate our findings we performed the same analysis for only 2 platforms (chromosome 22 only). This result is in favor of combining after imputation and is in line with our previous results.
Imputation of the individual datasets seems to result in more good quality SNPs ($r^2 > 0.3$). However, several limitations apply to our study. Firstly, the results presented here were based on 3 different gene-centric genotyping platforms. This meant that the variant density around known and interesting genes is higher than on a normal GWAS platforms. Therefore, the results of our analyses can be different if applied to non-gene-centric GWAS datasets. Secondly, the relatively small sample size of 979 unique individuals in de AMC-PAS cohort could have influenced the total number of variants with good quality available after the imputation. Thirdly, due to the relative small sample size we did not find a significant difference in computational time needed for both imputation methods. In datasets with more individuals and SNPs, the computation time needed for the combined imputation method will increase exponentially, this is the main reason why the combined after imputation is seen as the golden-standard within all large consortia at the moment. On the other hand we have a unique data collection, there are few cohorts genotyped on 2 or more different platforms. Other studies have analysed the quality of imputed SNPs after combining the genotyping data of different cohorts on different genotyping platforms[187].

CONCLUSION:

In conclusion, our results indicate that combining the data from three different platforms together after imputation performs better than combining the data of the 3 platforms before imputation.

Part IV

(Genetic) Risk Factors

7 GENETIC ANALYSIS OF EMERGING RISK FACTORS IN CORONARY ARTERY DISEASE

Type 2 diabetes (T2D), low-density lipoprotein-cholesterol (LDL-c), body mass index (BMI), blood pressure and smoking are established risk factors that play a causal role in coronary artery disease (CAD). Numerous common genetic variants associating with these and other risk factors have been identified, but their association with CAD has not been comprehensively examined in a single study. Our goal was to comprehensively evaluate the associations of established and emerging risk factors with CAD using genetic variants identified from Genome-wide Association Studies (GWAS).

We tested the effect of 60 traditional and putative risk factors with CAD, using summary statistics obtained in GWAS. We approximated the regression of a response variable onto an additive multi-SNP genetic risk score in the Coronary Artery DIsease Genomewide Replication And Meta-analysis (CARDIoGRAM) consortium dataset weighted by the effect of the SNP on the risk factors.

The strongest association with risk of CAD was for LDL-c SNPs (p $= 3.96x10^{-34}$). For non-established CAD risk factors, we found significant CAD associations for coronary artery calcification (CAC), Lp(a), LP-PLA2 activity, plaque, vWF and FVIII. In an attempt to identify independent associations between risk factors and CAD, only SNPs with an effect on the target trait were included. This identified CAD associations for Lp(a)(p$= 1.77x10^{-21}$), LDL-c (p$= 4.16x10^{-06}$), triglycerides (TG) (p$= 1.94x10^{-05}$), Height (p$= 2.06x10^{-05}$), CAC (p$= 3.13x10^{-23}$) and Carotid plaque (p$= 2.08x10^{-05}$). We identified SNPs associated with the emerging risk factors Lp(a), TG, plaque, Height and CAC to be independently associated with risk of CAD. This provides further support for-ongoing clinical trials of Lp(a) and TG, and suggests that CAC and plaque could be used as surrogate markers for CAD in clinical trials.

INTRODUCTION

In recent years many genome wide association studies (GWAS) have been conducted for established and non-established risk factors for CAD (Supplementary Table 1)[188, 24, 189, 190, 191, 44, 192, 193, 194, 195, 196, 197, 198, 199, 200, 201, 202, 48, 203, 204, 205, 206, 207, 208, 179, 209, 15, 210, 211, 212, 213, 214, 215] with the aim to discovering genetic determinants of risk factors. For each of the 60 a priori-selected risk factors, Supplementary Table 1 lists the studies and their characteristics used in this analysis. Even if tested, in most cases the relative contribution of risk factors to CAD has not been comprehensively investigated, since each study typically tests for the association of only one or two traits with CAD at any one time.

Our goal was to comprehensively evaluate the associations of established and emerging risk factors with CAD using genetic variants identified from GWAS. The rationale is that this could identify risk factors for follow up studies, for example in much larger and more comprehensive Mendelian randomization (MR) studies, and/or provide support for ongoing clinical trials targeting selected biomarkers. Furthermore this may increase our understanding of the aetiological and genetic landscape of CAD.

METHODS

Trait selection

To identify GWAS of established and non-established risk factors for CAD we queried the NHGRI GWAS Catalogue (available at: $http://www.ebi.ac.uk/gwas/$) in May 2015. GWAS with summary level data on SNP, effect size, standard error of effect size, risk allele and risk allele frequency publicly available in the GWAS catalog or in the original paper were included in the analysis.

We considered T2D, LDL-c, BMI, blood pressure and smoking to be established risk factors for CAD. The non-established risk factors were selected from the GWAS Catalogue by reviewing the traits in the GWAS catalog, and select potential traits of interest. Each trait in the GWAS Catalogue which has been linked to CAD based on its pathophysiology in the literature by PubMed searches was included in the analysis; i.e. Height, Psoriasis, TG, HDL-c, LP-PLA2[191, 44], adiponectin[194], MMP-1[190], homocysteine[207], white blood cell count(WBC)[216], glycated hemoglobin levels[217, 218] hsCRP[195], coagulation disorders (vWF)[208], mean platelet volume(MPV)[215], platelet count[215], platelet aggregation[219], FVIII[208], Protein C[220], PAI-I[202], fibrinogen[193, 195, 213]); chronic inflammation (systemic lupus erythematosis (SLE) [199, 200, 201]1, inflammatory bowel disease (IBD)[203], rheumatoid arthritis[198, 209]) and cardiac imaging (Coronary Artery Calcification (CAC)[206], cIMT[189]) or reported possible risk factors for CAD (vitamin D1, glomerular filtration rate (GFR)[221], homoarginine levels[222],serum uric acid levels[223], serum dimethylarginine levels (symmetric)[224], serum dimethylarginine levels (asymmetric/symmetric rate)[224], fat body mass(FBM)[225], and chronic kidney disease(CKD)[226]).

SNPs selection for association

For all analyses, we selected SNPs that reached a significance level of at least p$< 5.0x10^{-08}$ in the original GWAS of European individuals.
The goal was to test the association between the SNPs associated with risk factors and risk of CAD. For each trait we only selected independent SNPs (not in linkage disequilibrium (LD) with other SNPs for the same trait at r^2 cut-off<0.5). If two SNP were in LD this means that the alleles of both SNPs are inherited together more often than would be expected by chance. When we encountered SNPs in LD, the variant with the most significant p-value for the trait of interest was included in the analysis for that trait.

Traits with a significant association with CAD after the first analysis were selected for a secondary analysis to investigate independence of association by removing SNPs with pleiotropic effects. To do this, we categorized traits into two groups; 'upstream' and 'downstream' risk factors. This was in order to avoid removing SNPs that associated with several traits along a causal pathway from risk factor to disease. The upstream risk factors consist of the immediately modifiable risk factors BMI, lipids, lipoproteins, blood pressure, inflammation markers, coagulation traits and chronic inflammation traits. In contrast, downstream markers were much closer to the CAD phenotype and included CAC, cIMT and plaque. For upstream markers, we removed pleiotropic SNPs (by performing a pairwise LD-analysis using SNAP ($https://www.broadinstitute.org/mpg/snap/$) and removing those at r^2 >0.5) that were within the upstream group, but retained SNPs in LD with downstream markers (as this association could reflect a causal pathway). For downstream traits, we removed only those SNPs pleiotropic with other downstream traits.

Association of SNPs with CAD

To test the association between the selected risk factors and CAD we used the grs.summary function from R package Genetics ToolboX. This package implements a summary statistic method for approximating the regression of a response variable onto an additive multi-SNP genetic risk score in a given testing dataset weighting the association statistic by the effect of the SNP on the risk factors. This method uses single SNP association summary statistics: effect size, standard error of effect size and risk allele. Odds ratios (ORs) and confidence intervals (CIs) were transformed to effect sizes and standard error of effect sizes with an inverse natural logarithm, when necessary. We used the CARDIoGRAM GWAS publicly available data as our validation dataset[13]. In brief, CARDIoGRAM includes 22,233 cases of CAD and 64,762 controls[13]. An overly conservative Bonferroni corrected p-value significance of $8.33x10^{-04}$ was set (0.05/60 traits), to account for the number of traits assessed. As a negative control, we tested SNPs associated with eye color for their association with risk of CAD. We limited our study to GWAS performed in Caucasians since the CARDIoGRAM validation dataset is of the same ethnicity.

RESULTS

Trait selection

We included 69 studies in which a total of 60 risk factors were described. Supplementary Table 1 provides an overview of the papers incorporated in the analysis, listing the unit of exposure per increased risk allele, the variance explained by the reported SNPs, the number of SNPs discovered and the sample size of the study.

SNPs selection for association

The number of SNPs remaining after excluding duplicates and correcting for LD can be found in Table 7.1. Height was the trait with most SNPs (173) whereas some traits only had one SNP (e.g. smoking cessation). With these SNPs we performed a genetic association analysis with CAD using the CARDIoGRAM data.

Association of risk factors with CAD using all independent SNPs

In our analysis 15 out of 60 risk factors (Table 7.1) were significantly associated with CAD outcome in the CARDIoGRAM consortium. All established risk factors for CAD (LDL-c, SBP, DBP, BMI, T2D, smoking, HDL and TG) were identified to associate with CAD. The genetic risk score of LDL-c associated SNPs had the most significant p-value, with an OR_{CAD} 1.54, 95% CI: 1.44-1.65 p= $3.96x10^{-34}$ per 1-SD increase in LDL-C. CAC had the most significant p-value for the non-traditional risk factors, OR_{CAD} 1.91, 95% CI: 1.68-2.16 p= $3.13x10^{-23}$.
As a negative control, we tested eye color as a risk factor: no association between eye color and CAD risk was found OR_{CAD}=1.00, 95% CI: 0.99-1.02 p=0.93.

Association of risk factors with CAD using non-pleiotropic SNPs

Table 7.2 provides the number of SNPs that remained after pleiotropic SNPs were excluded. We analyzed 15 risk factors in our secondary analysis (Table 7.2) and found that six of the 15 traits had a significant association; the lowest p-value and highest effect size for CAD was for CAC OR_{CAD} 1.91,95% CI: 1.68-2.16 p= $3.13x10^{-23}$. In contrast to

Table 7.1: Association results of the primary analysis including all SNPs associated with exposures.

Trait	#SNPS	OR_{CAD}	95% CI	P-value
LDL-c	54	1.542	1.438-1.653	$3.96x10^{-34}$
CAC	2	1.906	1.678-2.164	$3.13x10^{-23}$
TG	48	1.399	1.305-1.499	$3.01x10^{-21}$
Lp(a)	5	1.249	1.193-1.308	$3.92x10^{-21}$
Diastolic Blood pressure	27	1.486	1.355-1.631	$5.19x10^{-17}$
Systolic Blood pressure	25	1.492	1.359-1.639	$5.47x10^{-17}$
LP-PLA2 (activity)	9	1.377	1.257-1.510	$8.09x10^{-12}$
HDL-c	74	0.789	0.737-0.845	$1.34x10^{-11}$
T2D	92	1.221	1.132-1.317	$2.19x10^{-07}$
Plaque	2	1.348	1.175-1.547	$2.08x10^{-05}$
Height	173	0.866	0.800-0.932	$2.06x10^{-05}$
BMI	69	1.082	1.042-1.123	$3.85x10^{-05}$
Factor VIII	5	2.249	1.504-3.363	$7.83x10^{-05}$
von Willebrand factor	8	0.786	0.696-0.888	$1.09x10^{-04}$
Mean arterial pressure	22	1.342	1.152-1.563	$1.54x10^{-04}$
hsCRP	20	0.893	0.833-0.958	$1.63x10^{-03}$
Hypertension	11	1.300	1.101-1.535	$2.00x10^{-03}$
GFR	3	1.031	1.010-1.053	$3.86x10^{-03}$
cIMT	3	1.521	1.226-1.816	$5.33x10^{-03}$
Homoarginine levels	1	0.856	0.757-0.967	$1.26x10^{-02}$
Glycated hemoglobin levels	12	1.124	1.024-1.235	$1.43x10^{-02}$
ADMA/SDMA	1	0.879	0.780-0.991	$3.49x10^{-02}$
SDMA	1	1.119	1.008-1.242	$3.49x10^{-02}$
Ulcerative colitis	19	1.071	1.004-1.143	$3.89x10^{-02}$
White blood cell count	9	0.675	0.456-1.000	$5.01x10^{-02}$
Psoriasis	15	1.041	0.997-1.084	$7.35x10^{-02}$
Inflammatory Bowel Disease	96	1.037	0.996-1.081	$7.62x10^{-02}$
Smoking Cessation	1	0.491	0.203-1.185	$1.14x10^{-01}$
Body Fat Mass	1	1.311	0.921-1.867	$1.33x10^{-01}$
Stroke	6	1.097	0.968-1.243	$1.48x10^{-01}$
Platelet count	55	1.040	0.984-1.099	$1.60x10^{-01}$
Protein C	4	1.025	0.990-1.061	$1.62x10^{-01}$

Continued on next page

Table 7.1 – *Continued from previous page*

Trait	#SNPS	OR_{CAD}	95% CI	P-value
Mean Platelet Volume	27	0.977	0.943-1.012	$1.92x10^{-01}$
Serum Uric Acid	3	1.073	0.947-1.217	$2.70x10^{-01}$
Fibrinogen	28	1.015	0.987-1.044	$2.95x10^{-01}$
Vitamin D	10	1.006	0.993-1.019	$3.76x10^{-01}$
SLE	23	0.983	0.945-1.022	$3.86x10^{-01}$
PA: ADP aggregation (GS)	3	0.940	0.806-1.097	$4.34x10^{-01}$
Chronic kidney disease	3	1.018	0.974-1.064	$4.38x10^{-01}$
Smoking Initiation	2	1.135	0.818-1.574	$4.50x10^{-01}$
PA: epinephrine aggregation(FHS)	3	0.956	0.847-1.079	$4.68x10^{-01}$
PA: epinephrine aggregation(GS)	3	1.062	0.886-1.273	$5.14x10^{-01}$
Pulse Pressure	11	0.949	0.788-1.143	$5.83x10^{-01}$
PA: ADP aggregation (FHS)	3	0.967	0.848-1.101	$6.11x10^{-01}$
Adiponectin	13	1.003	0.991-1.014	$6.57x10^{-01}$
Alcohol consumption	2	1.001	0.997-1.005	$6.64x10^{-01}$
COPD	9	1.001	0.998-1.003	$6.73x10^{-01}$
Crohn's disease	25	0.992	0.940-1.047	$7.83x10^{-01}$
Rheumatoid Arthritis	40	1.006	0.960-1.055	$7.98x10^{-01}$
Matrix metalloproteinase-1	2	1.007	0.952-1.065	$8.17x10^{-01}$
NT-proBNP	3	1.010	0.900-1.134	$8.69x10^{-01}$
Smoking	3	1.025	0.754-1.393	$8.74x10^{-01}$
Homocysteine	20	1.004	0.943-1.070	$8.92x10^{-01}$
Caffeine	2	0.969	0.548-1.711	$9.13x10^{-01}$
PAI-1	8	1.007	0.870-1.166	$9.24x10^{-01}$
Eye color	7	1.001	0.986-1.016	$9.31x10^{-01}$
PA: collagen lag time (FHS)	1	0.996	0.892-1.114	$9.49x10^{-01}$
PA: collagen lag time (GS)	1	0.987	0.667-1.462	$9.49x10^{-01}$

the primary analysis, no significant association was identified for LP-PLA2, T2D, HDL-c, SBP, DBP, MAP, vWF and FVIII when SNPs were thinned out by pleiotropy (Figure 7.1).

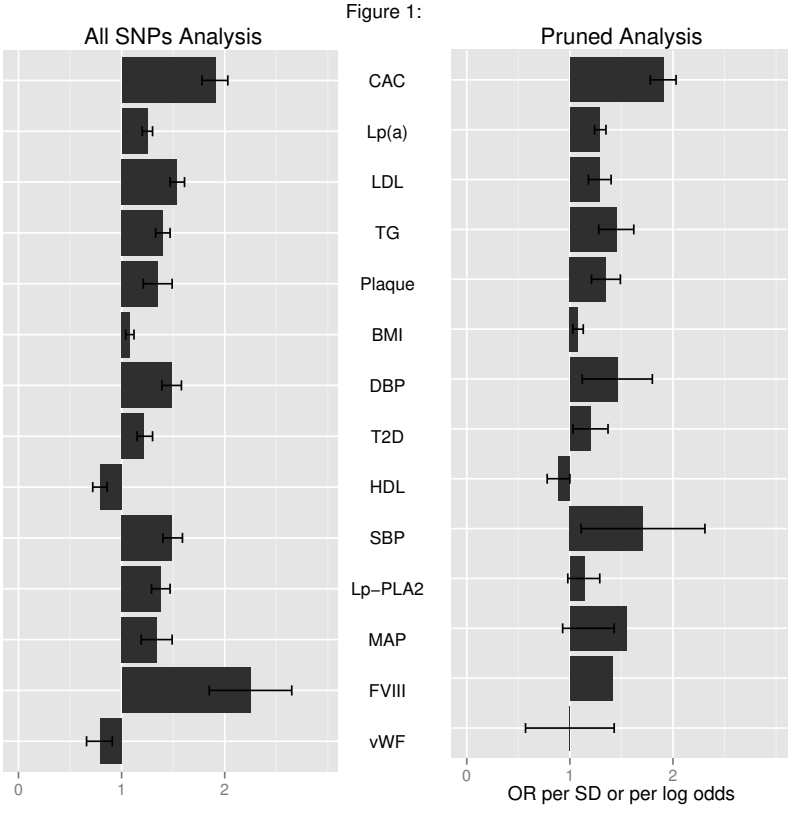

Figure 7.1: Association results of the 15 significant risk factors in the primary analysis and their associations in the pruned analysis

Table 7.2: Association results of secondary analysis using only SNPs that are specific for the exposure of interest.

Trait	#SNPS	OR_{CAD}	95%CI	P-value
CAC	2	1.906	1.678-2.164	$3.13x10^{-23}$
Lp(a)	3	1.293	1.226-1.363	$1.77x10^{-21}$
LDL-c	31	1.293	1.159-1.443	$4.16x10^{-06}$
TG	27	1.448	1.222-1.716	$1.94x10^{-05}$
Plaque	2	1.348	1.175-1.547	$2.08x10^{-05}$
Height	172	0.867	0.801-0.934	$2.54x10^{-05}$
BMI	63	1.080	1.029-1.134	$1.75x10^{-03}$
Diastolic Blood pressure	3	1.456	1.038-2.044	$2.96x10^{-02}$
T2D	47	1.198	1.012-1.418	$3.55x10^{-02}$
HDL-c	47	0.890	0.799-0.992	$3.56x10^{-02}$
Systolic Blood pressure	1	1.708	0.935-3.120	$8.18x10^{-02}$
LP-PLA2 (activity)	3	1.135	0.976-1.320	$1.01x10^{-01}$
Mean Arterial Pressure	7	1.210	0.913-1.604	$1.85x10^{-01}$
Factor VIII	2	1.548	0.249-9.618	$6.39x10^{-01}$
Von Willebrand factor	5	0.997	0.648-1.534	$9.89x10^{-01}$

DISCUSSION

We conducted a comprehensive study to investigate the association of risk factors for their association with CAD using summary-level genetic data. Using all available SNPs, we found a significant association with CAD for the following traditional risk factors: LDL-C, HDL-C, SBP, DBP, BMI, T2D and TG. In addition, we found the following traits to associate with CAD: Height, CAC, Lp(a), LP-PLA2, plaque, factor VIII, von willebrand factor and mean arterial pressure. Of these emerging risk factors, when we removed potentially pleiotropic SNPs, associations with CAD persisted for Height, CAC, Lp(a), and plaque.

After removing potentially pleiotropic SNPs for FVIII and vWF, the association with CAD no longer persisted. Furthermore, the association between LP-PLA2 and CAD also diminished when limited to nominally non-pleiotropic SNPs, which is in keeping with a recent MR study[227]. This arises from an overlap in genetic variation between LP-PLA2 activity and lipid phenotypes in our data: 6 of 9 LP-PLA2 SNPs in our primary analysis where excluded in our secondary analysis because they were in LD with one of the lipid (LDL-C, HDL-C or TG) phenotypes. An explanation for this phenomena

might be that in the bloodstream, two-thirds of LP-PLA2 circulates primarily bound to LDL; the remaining third is distributed between HDL and VLDL[228]. Measures of LP-PLA2 might thus partially reflect the concentration of proatherogenic lipoproteins. Recent clinical trials with the LP-PLA2 inhibitor darapladib in coronary heart disease patients yield similar results to our analysis[229, 230]. In patients with stable CAD after optimal treatment for dyslipidemia, there was no added benefit of reducing LP-PLA2. Regarding vWF, other studies identified a weak association between vWF levels and CAD but it disappeared after adjusting for coexisting riskfactors[231, 232].

For IMT and PAI-1, we do not find an association with CAD in our primary analysis. This is in contrast with other studies[233, 202]. This may be due to low variance of the exposure explained by the SNPs for IMT (1.10%). Bis et al report two SNPs that are associated with IMT phenotype and CAD. Of the two SNPs, only rs17045031 near LRIG1 was significantly associated with IMT phenotype and therefore included in our analysis, the other SNP was not included in our analysis because of our threshold for significance, this might explain the discrepancy in results[233]. For PAI-1, out of the 10 SNPs identified by Huang et al, only two SNPs in ARNTL were nominally associated with CAD (rs6486122: OR 1.04; 95% CI: 1.01-1.07; rs3816360: OR 1.03; 95% CI: 1.01-1.06)[202]. In contrast, our chosen method incorporated all SNPs associated with PAI-1, which did not associate with CAD.

We did identify a clear association between Lp(a) and CAD (OR_{CAD} 1.25, 95% CI: 1.19-1.31 p= $3.92x10^{-21}$), and the association became stronger when we removed pleiotropic SNPs. This is in keeping with recent genetic studies including that by Clarke et al, which identified two common SNPs in the LPA gene that had an association with both the Lp(a) lipoprotein level and the risk of CAD[44].

The association between height and CAD (OR_{CAD} 0.87, 95% CI: 0.80-0.93 p= $2.06x10^{-05}$) remained significant after removing pleiotropic SNPs. This is inline with recently published MR-studies by Nuesch et al. They conclude that taller individuals have a lower risk of developing CAD[234].

The clear association between TG and CAD remained significant after removing pleiotropic SNPs: OR_{CAD}=1.45, 95% CI: 1.22-1.72 p= $1.94x10^{-05}$. This is in line with the findings by Do et al[235], They used 185 common SNPs to examine the role of TG on risk for CAD. They show that loci with only a strong magnitude of association with TG are associated with CAD. Furthermore, Holmes et al. performed a Mendelian randomization analysis based on individual participant level data from 62,000 individuals with >12,000 CHD events. They found a causal role for TG in the analysis restricted to SNPs only associated with TG and no association with any other lipid trait[236]. More

recently, White et al used 140 SNPs and identified a consistent causal association of TG with risk of CAD using different MR approaches[237].

Although HDL-C associated with CAD on initial analysis, when we limited the genetic instrument to only non-pleiotropic SNPs, this association diminished. This is in keeping with prior reports[236, 237, 238, 239]. To take one example, in the paper by Voight et al., a Mendelian randomization analysis was conducted using a genetic risk score consisting of 14 common SNPs that associated predominantly with HDL cholesterol and tested this score in up to 12,482 cases of myocardial infarction and 41,331 controls. They found that a 1 SD increase in HDL cholesterol due to genetic score was not associated with risk of myocardial infarction (OR 0.93, 95% CI 0.68-1.26, p=0.63)[238].

Interestingly, in our analysis restricted to non-pleiotropic SNPs, the established causal risk factors SBP and DBP were no longer associated with CAD. This arose because restricting to non-pleiotropic SNPs resulted in the removal of the majority of SNPs with only 3 SNPs remaining for DBP and 1 SNP for SBP. SNPs were removed because they were also associated with MAP, PP, HTN. While the causal role of blood pressure in CAD is well-established, this suggests that our approach overly-penalized some traits that have a highly pleiotropic genetic architecture.

The association of CAC and plaque with CAD retained after removing pleiotropic SNPs and is worthy of further comment. Over the past few years there has been a move to use surrogate markers of CHD in clinical trials as a marker of ”hard” clinical outcomes, but at a lower cost than conducting a full outcome-based clinical trial. Given that these downstream traits are proximal to the CAD phenotype, they are less likely to be confounded compared to a trait that is more distant (or upstream trait, such as a blood lipid profile). Our data therefore suggest that, given the relationship of these traits with risk of CAD, that they may represent a proxy for CAD, however further investigations are needed, for example relating SNPs such as LDL-C and SBP on plaque and CAC[240] to see whether they associate with these traits.

Our study has several strengths and weaknesses that merit discussion. To the best of our knowledge we are the first to report on the association between SNPs associated with multiple non-established risk factors and risk of CAD. We are, however restricted to published studies, which might have resulted in selection bias. Furthermore, the data was derived from the publications, and no source data quality assessment could be performed, we were dependent on the quality of the published papers. As an example, traits where the genetic landscape is not fully captured by current GWAS might bias our results. Therefore, at the moment we cannot rule out any association between these risk factors and CAD. Because summary level data were used for analyses, we were not able to perform age and gender corrections (however these would have been conducted in

the original GWAS); we are therefore dependent on the corrections performed in the published papers. Finally, this study represents a rather broad but crude genetic analysis; more sophisticated analyses such as multivariate MR[241] and/or MR-Egger[242] would represent next steps to more comprehensively assessing and accounting for pleiotropy of genetic instruments and testing for independence of causal estimates across multiple risk factors.

In conclusion, our multiple trait genetic analysis of established and emerging risk factors for CAD provides further evidence that TG and Lp(a) should be prioritized as potential therapeutic targets for CAD prevention, and suggests that CAC and plaque could be potential surrogate markers for CAD.

8 GENERAL DISCUSSION AND FUTURE PERSPECTIVES

CVD is still the most important cause of death in Western societies; a total of 17.7 million people die each year due to CVD (2015) according to the World Health Organisation (WHO), this represents 31% of all global deaths[243]. In numerous large prospective studies, male gender, a positive family history of CVD, high plasma cholesterol levels, high blood pressure, diabetes and smoking have invariably been shown to be independent risk factors for cardiovascular disease[244]. These risk factors are therefore implemented in the CVD risk assessment tools used in the clinical setting. While notifying the relevance of these factors, it is important to state that risk prediction is far from precise with the use of these risk factors and that these factors do not fully explain the heritability of CVD. Discovery of additional risk factors are therefore needed. Identification of these factors will not only improve CVD risk estimation but may also identify novel drug targets to treat CVD. Genetics can help in the identification of novel underlying biological mechanism using an agnostic approach. Twin studies have indicated that the heritability of CVD is 30-60%[245]. With heritability, we mean the inter-individual differences resulting from genetic factors. The heritability of CVD is considered to be the result of genetic variants (common and rare) with small and large effects on the expression of CVD [246]. Recent genetic findings explain approximately 15% of the heritability suggesting that still undiscovered pathways are involved in the development of CVD[247]. Upon the notion that family history of CVD is a major risk factor for CVD, several linkage studies were performed in distinct FH families and a number of mutations in several genes were found in these studies[10]. In recent years, these linkage studies were followed by candidate-gene studies and GWAS.

The advances in molecular biology and the decrease in cost made it possible to perform more and larger studies. The GWAS studies in CVD discovered 65 variants up till now, all however with small effects associated with CVD[248].

In our meta-analysis on 4 lipid traits consisting of 66,240 individuals of European ancestry we used a large-scale locus-centric approach, testing 49,227 SNPs carefully prioritized for CVD-related loci in 32 studies to explore association with HDL-C, LDL-C, TC and TG levels. Using an additional sample of 25,282 individuals and the available data derived from the GLGC study[24], we identified 21 additional loci that have not been associated with lipid levels before and were able to confirm a number of the previously reported associations. The evidence from the cumulative meta-analysis of our data, the replication studies, and the published GLGC results suggest that further ”true” signals might be found with less stringent p value thresholds. Given the recent deluge of available genetic data, we propose that a more careful examination is required of common variants of moderate and small effects. This might help explain portions of missing heritability, elucidate the pathways and mechanisms involved in lipid metabolism and CHD, and identify potential loci in which rare SNPs with large effects on the phenotype can be discovered.
In our multi-ethnic meta-analysis of lipid association studies in African Americans, Hispanics and East Asians we used the same large-scale locus-centric approach and we identified two novel loci associated with TC and LDL-C levels[61].
Furthermore, we evaluated SNPs previously associated with the 4 lipid traits in European populations, on lipid levels in three populations of other ethnicities, showing that many known association signals for lipids span across ethnicities[61].

Established variants for CVD are often combined in a genetic risk score (GRS) to predict the risk of CVD, but they have not been used in a clinical setting yet for prediction. We tested whether a GRS of 46 variants associated with CVD predict events in a cohort of FH patients and in a cohort of patients treated with statins.
We did not find a significant association between the GRS of the 46 variants and the number of events in the cohort of 2400 FH patients (GiraFH)[36]. The lack of association found in this study may very well be due to the fact that our study was at the limit of the power needed to detect statistically significant associations with CVD, since the studied SNPs were previously shown to have a moderate to small effect[14].
In the TNT-cohort[249] of 1877 patients treated with statins we did not find an association either [NOT PUBLISHED], which may also be due to a lack of power or the fact that the patients in this cohort all survived a first CAD event, which might have led to survival bias or index event bias. In theory, it could be possible thatpatients with the most harmful genetic profiles might have died before inclusion into our study. But both results illustrate that the reservations of many against implementing GRSs in the clinic may be justified.

More variants associated with CVD should be discovered to make the GRS a robust tool to predict the risk of CVD for a specific patient. Besides GWAS, whole-exome and whole-genome sequencing are new methods that became affordable in the last few years to apply in large cohorts of patients with CVD. By using these sequencing methods on a large scale, we may be able to discover rare variants with large effects that explain the missing heritability. The combination of these rare variants with large effects and the common variants can then be used to generate a GRS that could then be validated in clinic in order to identify patients at risk for CVD.

Due to the large amount of available genetic data, Mendelian randomization (MR) studies can be performed. A MR study is a study in which genetic variants are used to investigate the causality of a biomarker on risk the risk of CVD. MR studies have been very successful in the field of CVD demonstrating strong evidence of causality for established and novel biomarkers (such as LP(a) and drug targets (such as PCSK9)[250]. Due to developments in genotyping and availability of publicly available genetic data, MR analyses using existing genetic data, have identified potentially modifiable exposures that, if shown to be causal, may be tested in future intervention studies[251].
We aimed to identify new risk factors of CVD by using publicly available summary-level genetic data available in the GWAS catalog and meta-analysis data from the CARDIOGRAM consortium[14] to investigate the association of risk factors with CAD. We found the following traits to associate with CAD: Height, coronary artery calcification (CAC), Lipoprotein(a) (Lp(a)) and carotid plaque after correction for pleiotropy. In our analysis, we were restricted by the fact that we used studies in the GWAS catalog, which might however have resulted in selection bias. Traits that are not fully captured by current GWAS might bias our results. Therefore, we cannot rule out any association between risk factors, where the genetic landscape is not fully captured by published GWAS studies, and CVD. Because we used summary level data in our analyses, we were not able to correct for age and gender ourselves, so we were dependent on the methods performed in the published papers. Furthermore, the data we used in our analysis was derived from the publications, and no source data quality assessment could be performed[37].

A method to increase the power is to combine data from different data sources. We investigated if combining different (imputed) GWAS datasets is a powerful method to identify new variants using data from different genotyping platforms. We found that combining imputed data from different platforms before and after imputation does result in differential numbers and quality of SNPs to be analyzed. When imputation of different data sets is performed prior to combining resulted in 5% more SNPs that satisfied the post-imputation QC[252].

Future perspectives

In this thesis, we used various methods to discover new common genetic variants associated with CVD related traits. We have identified new variants with small effects on lipid levels, that need to be validated in other cohorts. These variants only explain a limited proportion of the genetic heritability of CVD, and identification of additional (rare) variants will require larger patient cohorts. This could only be achieved by setting up large biobanking initiatives like the UK biobank[253], GONL[27], UK10K[28], Lifelines[254], Parelsnoer Institute[255] and the HELIUS project[256] and collaborating in large consortia bringing together multiple initiatives/studies. These initiatives have only value if the participants also have detailed phenotype data available. For this matter, linkage of biobanks to standardized Electronic Patients Records (EPR) can play a key role; the EPR should provide automatic access to and extraction of data of the clinical, imaging and laboratory data of the participants in the different biobanks. This is not yet available in most of the university medical centers in the Netherlands. When duplicate data entry is not needed anymore and data is registered at the source (in the EPR), we will be able to recruit larger cohorts of patients and controls to explain the genetic heritability of CVD. The analysis of a cohort of ~200.000 patients with CVD on the Exomechip did not result in any genome-wide significant associations that were not described yet[257]. Therefore, larger cohorts are needed. This can be achieved by collecting the bio specimens (DNA, blood, urine, feces and tissue) of a huge number of people in the general population. Then genotype and/or sequence these participants and collect extensive phenotype data of these participants (eg. clinical data, imaging data, ECG data, survey data). By doing this, we hopefully are able to explain more of the missing heritability of CVD and other diseases.

Another way to unravel more of the genetic background of CVD is using published GWAS data. GWAS have been instrumental in our understanding of the genetic background of many diseases in the last 10 years. Most, but not all of this data is publicly available in the GWAS catalog[258] maintained by The National Human Genome Research Institute (NHGRI). As of 10-04-2018, the GWAS Catalog contains 3349 publications and 59,967unique SNP-trait associations. Figure 8.1 shows the increase of the number of studies included in the GWAS catalog in the last 12 years. The data in the GWAS catalog can be used in new research unraveling the genetic background of CVD. The GWAS catalog contains however only summary level data.

If a researcher wants to make use of individual level data published in the GWAS catalog by combining different cohorts he needs to contact the authors of the published papers and ask for access to the dataset. Fortunately, more and more datasets, for example

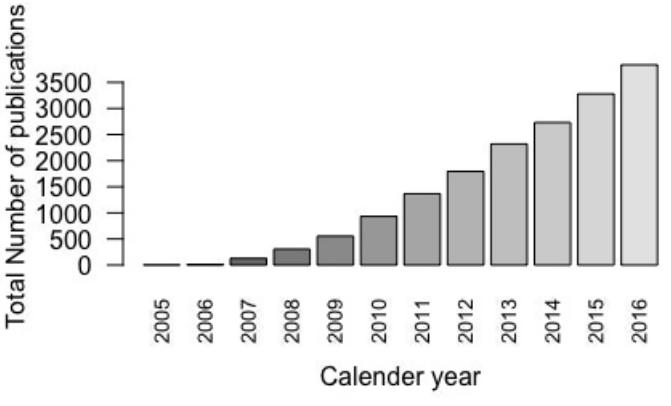

Figure 8.1: Number of published GWAS studies in GWAS catalog in the last 12 years.

GWAS results, are shared publicly, but a large part of (GWAS) datasets is not yet shared by the authors, because authors do not ”Dare to Share” their datasets. Fortunately, more and more scientific journals and funding agencies demand authors to publish their data according to the FAIR principles (Findable, Accessible, Interoperable, Reusable) in data repositories. In short this means that datasets should have at least a unique persistent identifier, rich metadata, this metadata should be retrievable, the data should make use of standardized vocabularies and the data set should have a data usage license and the dataset should be registered in a searchable resource[259]. Using these principles will hopefully lead to more datasets of good quality that can be used for new research projects, within legal and ethical boundaries.

In conclusion, in this thesis we have identified new variants with small effects associated with lipid levels. Unfortunately, we did not find an association between GRS of common variants with small effects on CVD and the number of events in a cohort of FH patients and a cohort of patients treated with statins. By using publicly available GWAS data we found that TG, LP(a), CAC and plaque are associated with CVD. We feel that sharing and combining datasets by making (summary-level) data publicly available for further research will contribute to the discovery of new (genetic) risk factors of CVD.

SUMMARY

Cardiovascular Disease

Cardiovascular disease (CVD) is the leading cause of mortality in the Western Societies[1]. CVD encompasses a number of different disease entities, many of which are related to atherosclerosis. Atherosclerosis is the process of development of plaque formation in the sub endothelium of arterial walls. These plaques narrow the arterial lumen, thereby disabling blood flow. Upon rupture of the fibrous cap, a blood clot forms, which results in total obstruction of blood flow, which, depending of the site, may lead to myocardial infarction (MI), stroke or peripheral artery disease (PAD). Atherosclerosis has a multifactorial origin involving abnormalities[2] in lipid metabolism, hypertension, obesity, diabetes mellitus, smoking, inflammation, coagulation, and fibrinolysis, amongst others. At present, we lack a complete understanding of the relevance of these individual risk factors and their interplay in the disease process[3]. It has been suggested that genetic factors contribute to the risk of CVD[4].
The main aim of this thesis was to find new genetic risk factors for CVD, by using different analysis methods using new and existing publicly available data.

Part I:

First, a meta-analysis in **chapter 2** of this thesis found 21 new common variants associated with one or more lipid traits: PPARG, GP1HBP1, DGAT2, HCAR2, FTO, VLDLR, SPTY2D1, BRCA2, SOCS3, APOH, C4B, LPAL2, GCK, GATA4, SERPINF2, INSR, FCGR2A, INSIG2, UGT1A1, CHUK, UBE3B[24] using the data from the IBC cardiochip consortium on 4 different lipid traits (LDL-C, HDL-C, TG, TC) in 66,240 individuals from 32 different studies.

We performed another meta-analysis in **chapter 3** in 7,657 African Americans, 1,315 Hispanics and 841 East Asians using the IBC 50K SNP genotyping array and we found and confirmed two novel signals for lipids by replication in 7,000 African Americans.

Additionally we evaluated the effect of SNPs established in European populations on lipid levels in multi-ethnic populations and show that most known lipid association signals span across ethnicities.

Part II:

In **chapter 4** of this thesis we investigated whether common genetic variants associated with CVD also hold value for CVD prediction in a large cohort of patients with Familial Hypercholesterolemia (FH). A total of 46 SNPs in 1701 FH patients were genotyped, of whom 482 patients (28.3%) had at least one cardiovascular event during 112.943 person-years follow-up. The association of each SNP with event-free survival time was calculated with a Cox proportional hazard model. In CVD risk adjusted analysis, the lead SNP at the well-known 9p21 locus rs1333049 near CDKN2B-AS1 had a HR for CAD risk of 0.82 (95% CI 0.77-0.87; p-value 0.000945). None of the other tested CAD-associated SNPs were significantly associated with CAD risk. Of all the loci analyzed, the 9p21 locus had the strongest negative association with CAD in this high-risk FH cohort. None of the SNPs at neither this 9p21 locus, however, nor any of the other tested CAD-associated SNPs were significantly associated with risk of CAD according to a priori defined significance threshold that took into account multiple testing.

In **chapter 5** our objective was to determine whether common genetic variants associated with CVD have also a predictive value for a second cardiovascular event in a cohort of 1877 atorvastatin-treated patients with established CAD. Of these, 24.5% suffered from a second vascular event during follow-up. Using two different genetic risk scores (GRS) we investigated wether common genetic variants associated CVD are determinants of the risk for a second vascular event in patients treated with atorvastatin. The cohort was divided in quartiles of GRS, in order to study the effect of an increasing GRS on developing a second event. The risk for a second vascular event was not statistically different in subjects in the highest versus the lowest GRS quartile. These findings suggest that the putative CVD SNPs identified in GWAS have no, or minor effect on the risk for a secondary vascular event in patients treated with atorvastatin. This implicates that the secondary event is caused by other factors, and that these factors may not be explained by common genetic variations, that have been shown to be a determinant for the first cardiovascular event.

Part III:

In **chapter 6** we investigated the imputation quality and efficiency of two different approaches of combining GWAS data from different genotyping platforms. We investigated whether combining data from different platforms before the actual imputation performs better than combining the data from different platforms after imputation. In total 979 unique individuals from the AMC-PAS cohort were genotyped on the MetaboChip, 50K gene-centric Human CVD BeadChip and the HumanExome chip. A total of 397 individuals was genotyped on all 3 individual platforms. Using Minimac in combination with MaCH we imputed all samples using the the 1000 genomes reference panel. All imputed markers with an r^2 value of $<$0.3 were excluded in our post-imputation QC. This resulted in a total of 4,117,036 SNPs of good quality when imputation was performed before merging the three datasets. A total of 3,933,494 SNPs were available when imputation was done on the combined set. This suggests that imputation of individual datasets before merging performs slightly better than after combing the different datasets.

In **chapter 7** we investigated the effect of 60 traditional (Type 2 diabetes (T2D), low-density lipoprotein-cholesterol (LDL-c), body mass index (BMI), blood pressure and smoking) and emerging risk factors with CAD, using summary statistics obtained in GWAS. Our goal was to comprehensively evaluate the associations of these traditional and putative risk factors with CAD using genetic variants identified from Genome-wide Association Studies (GWAS). Type 2 diabetes (T2D), low-density lipoprotein-cholesterol (LDL-c), body mass index (BMI), blood pressure and smoking are established risk factors that play a causal role in coronary artery disease (CAD). Numerous common genetic variants associating with these and other risk factors have been identified, but their association with CAD has not been comprehensively examined in a single study. Our goal was to comprehensively evaluate the associations of established and emerging risk factors with CAD using genetic variants identified from Genome-wide Association Studies (GWAS). The strongest association with risk of CAD in our analysis was for LDL-c SNPs ($p = 3.96x10^{-34}$). For non-established CAD risk factors, we found significant CAD associations for coronary artery calcification (CAC), Lp(a), LP-PLA2 activity, plaque, vWF and FVIII. In an attempt to identify independent associations between risk factors and CAD, only SNPs with an effect on the target trait were included. We identified CAD associations for Lp(a)($p= 1.77x10^{-21}$), LDL-c ($p= 4.16x10^{-06}$), triglycerides (TG) ($p= 1.94x10^{-05}$), Height ($p= 2.06x10^{-05}$), CAC ($p= 3.13x10^{-23}$) and Carotid plaque ($p= 2.08x10^{-05}$).

This provides further evidence that TG and Lp(a) should be prioritized as potential therapeutic targets for CAD prevention. And our results suggests also that CAC and plaque could be used as potential surrogate markers for CAD in clinical trials.

SAMENVATTING

Hart- en vaatziekten

Hart- en vaatziekten (HVZ) zijn nog steeds doodsoorzaak nummer één in de westerse wereld[1]. HVZ omvatten een aantal verschillende ziektebeelden, hiervan zijn er veel gerelateerd aan atherosclerose.
Atherosclerose is de medische term voor slagaderverkalking. Atherosclerose is het proces waarbij plaque zich ontwikkeld in het sub-endotheel van de arteriële vaatwanden. Deze plaques vernauwen het arteriële lumen, waardoor de bloedstroom (gedeeltelijk of helemaal) wordt beperkt. Bij een breuk van deze vezelachtige kap vormt zich een bloedstolsel, dit resulteert dan vaak in een totale afsluiting van de bloedstroom, die afhankelijk van de locatie, kan leiden tot een hartinfarct of beroerte. Atherosclerose heeft een multifactoriële oorsprong[2]; afwijkingen in lipiden waarden, hoge bloeddruk, diabetes mellitus, roken, ontsteking en afwijkingen in de coagulatie en fibrinolyse kunnen allemaal ten grondslag liggen aan atherosclerose. Op dit moment weten we nog niet precies wat de relevantie van deze individuele risicofactoren en hun wisselwerking op het ziekteproces is[3]. Er wordt steeds meer bewijs gevonden dat genetische factoren een rol spelen bij het krijgen van HVZ[4].

Het belangrijkste doel van dit proefschrift was om nieuwe genetische risicofactoren voor HVZ te vinden, door verschillende analysemethoden toe te passen op nieuwe en reeds bestaande (openbaar) beschikbare data.

Deel I

In **hoofdstuk 2** hebben we met behulp van een meta-analyse 21 nieuwe varianten geïdentificeerd die geassocieerd zijn met een of meer lipiden fenotypes: PPARG, GP1HBP1, DGAT2, HCAR2, FTO, VLDLR, SPTY2D1, BRCA2, SOCS3, APOH, C4B, LPAL2, GCK, GATA4, SERPINF2, INSR, FCGR2A, INSIG2, UGT1A1, CHUK, UBE3B[24].

Hiervoor hebben we de lipiden data (LDL-C, HDL-C, TG, TC) van het IBC-cardiochip-consortium gebruikt van in totaal 66,240 patiënten die geïncludeerd waren in 32 verschillende studies.

In **hoofdstuk 3** hebben we een meta-analyse uitgevoerd in een cohort van 7657 Afro-Amerikanen, 1315 Latijns-Amerikaanse Amerikanen en 841 Oost-Aziaten. Voor deze meta-analyse hebben we ook de data van de IBC 50K SNP genotyperingsarray gebruikt. In deze analyse hebben we gekeken naar de etnische verschillen, we hebben twee nieuwe varianten gevonden die geassocieerd zijn met een van de lipiden fenotypes en we hebben deze resultaten gerepliceerd in een cohort van 7000 Afro-Amerikanen. Daarnaast hebben we gekeken of de effecten van SNPs die in de Europese populatie een associatie hebben met een van de lipiden fenotypes ook een associatie laten zien in het multi-etnische cohort. Onze analyse laat zien dat deze varianten ook in cohorten van een andere etniciteit geassocieerd zijn met de verschillende lipiden fenotypes.

Deel II

In **hoofdstuk 4** van dit proefschrift hebben we onderzocht of bekende genetische varianten die geassocieerd zijn met HVZ ook een voorspellende waarde hebben op de kans van het krijgen van een event in een groot cohort van patiënten met Familiaire Hypercholesterolemie (FH). We hebben 46 SNPs bij 1701 FH-patiënten gegenotypeerd. In 112,943 persoonsjaren kregen 482 patiënten ten minste één cardiovasculair event (28,3%). De associatie van elke SNP met event vrije overlevingstijd werd berekend met een Cox proportioneel risico model. In de analyse waarbij we corrigeerden voor HVZ was de meest significante SNP, een SNP op de bekende 9p21-locus, rs1333049 vlakbij het CDKN2B-AS1 gen. Deze SNP had een hazard ratio voor het risico op HVZ van 0,82 (95% CI 0,77-0,87; p-waarde 0,000945). Geen van de andere bekende CAD-geassocieerde SNPs waren significant geassocieerd met CAD-risico in dit cohort van patiënten met FH. Van alle geanalyseerde SNPs waren de SNPs in de 9p21-locus het meest geassocieerd met CAD. Geen van de geanalyseerde SNPs had echter een statistisch significante associatie met het risico op CAD volgens het vooraf gedefinieerd significantie niveau, waarbij werd gecorrigeerd voor ”multiple testing”.

In **hoofdstuk 5** was ons doel om te bepalen of bekende genetische varianten waarvan een eerdere associatie met HVZ al is aangetoond, ook een voorspellende waarde hebben voor het krijgen van een tweede cardiovasculair event in een cohort van 1877 patiënten met

CAD, die behandeld werden met atorvastatine. In dit cohort kreeg 24,5% een tweede cardiovasculair event tijdens de follow-up. Door gebruik te maken van twee verschillende genetische risicoscores (GRS) hebben we onderzocht of de bekende genetische varianten mogelijke voorspellers zijn voor het risico op het krijgen van een tweede cardiovasculair event We hebben het cohort verdeeld in kwartielen op basis van de genetische risicoscore om zo het effect te kunnen bepalen van een hogere GRS op de ontwikkeling van een tweede cardiovasculair event. Het risico op een tweede cardiovasculair event was niet statistisch significant tussen het hoogste en het laagste GRS-kwartiel. Deze bevindingen suggereren dat de bekende HVZ-SNPs die in verschillende GWAS zijn geïdentificeerd, geen of een heel klein effect hebben op het risico voor het krijgen van een secundair cardiovasculair event bij patiënten die behandeld worden met atorvastatine. Dit suggereert dat een secundair event wordt veroorzaakt door andere (genetisch) factoren en dat deze factoren mogelijk niet worden verklaard door genetische variaties, waarvan reeds is aangetoond dat ze geassocieerd zijn met HVZ.

In **hoofdstuk 6** hebben we de imputatie kwaliteit en de imputatie efficiëntie onderzocht van twee verschillende benaderingen om GWAS-data van verschillende genotyperingsplatformen te combineren. We hebben onderzocht of het combineren van data van verschillende platformen vóór de daadwerkelijke imputatie beter presteert ten opzichte van het combineren van de data van verschillende platformen na de imputatie.
In totaal werden 979 unieke patiënten uit het AMC-PAS-cohort gegenotypeerd op de MetaboChip, 50K Human CVD BeadChip en de HumanExome-chip. 397 individuen waren op alle 3 de platformen gegenotypeerd. Met behulp van Minimac in combinatie met MaCH hebben we alle samples geïmputeerd door gebruik te maken van het 1000 genomes-referentie panel. Alle geïmputeerde SNPs met een r^2 waarde van $< 0{,}3$ werden in onze kwaliteitscontrole na de imputatie geëxcludeerd. Dit resulteerde in een totaal van 4.117.036 SNPs van goede kwaliteit wanneer we de drie data sets los van elkaar imputeerden. Een totaal van 3.933.494 SNPs van goede kwaliteit was beschikbaar nadat de imputatie werd uitgevoerd op de gecombineerde dataset. Dit suggereert dat de imputatie van afzonderlijke datasets vóór het samenvoegen iets beter presteert dan na het combineren van de verschillende datasets.

In **hoofdstuk 7** hebben we het effect van 60 traditionele (Type 2 diabetes (T2D), low-density lipoprotïene-cholesterol (LDL-c), body mass index (BMI), bloeddruk en roken) en nieuwe risicofactoren met CAD onderzocht, met behulp van bestaande publiekelijk toegankelijke GWAS data. Ons doel was om de associaties van deze traditionele en potentiële risicofactoren met CAD uitgebreid te analyseren met behulp van genetische varianten die al eerder geïdentificeerd waren in eerder uitgevoerde GWAS.
T2D, LDL-c, BMI, bloeddruk en roken zijn bekende risicofactoren die een oorzakelijke

rol spelen bij HVZ. Veel bekende genetische varianten die met deze en andere risicofactoren zijn geassocieerd, zijn geïdentificeerd, maar hun associatie met CAD is nog niet uitgebreid onderzocht in een studie. De meest significante risicofactor in onze analyse voor het risico op CAD was LDL-c ($p = 3,96x10^{-34}$). Voor niet-traditionele CAD-risicofactoren vonden we significante associaties met CAD voor verkalking van de kransslagader (CAC), Lp(a), LP-PLA2-activiteit, stenose, vWF en FVIII.
In een poging om onafhankelijke associaties tussen risicofactoren en CAD te identificeren, hebben we in een tweede analyse, alleen SNPs meegenomen die in eerdere GWAS geen associatie hadden met een andere risicofactor. In deze analyse vonden we de volgende significante associaties van de risicofactor met CAD: Lp(a) ($p = 1,77x10^{-21}$), LDL-c ($p = 4,16x10^{-06}$), triglyceriden (TG) ($p = 1,94x10^{-05}$), Lengte ($p = 2.06x10^{-05}$), CAC ($p = 3.13x10^{-23}$) en carotid stenose ($p = 2.08x10^{-05}$).

De resultaten van deze analyse leveren verder bewijs dat TG en Lp(a) prioriteit zouden moeten krijgen als potentiële therapeutische doelen om CAD preventief te behandelen. Ook suggereren onze bevindingen dat CAC en plaque kunnen worden gebruikt als potentiële surrogaatmarkers voor CAD in klinische trials.

DANKWOORD

Ik wil graag iedereen bedanken die op welke manier dan ook heeft bijgedragen aan de totstandkoming van dit proefschrift. Een aantal van deze mensen wil ik graag apart bedanken voor hun bijdrage aan dit proefschrift.

Als eerste mijn promotoren en co-promotor: prof. dr. A. H. Zwinderman, beste Koos, bedankt voor alle begeleiding bij het schrijven van de verschillende stukken uit mijn proefschrift. Ik ben je dankbaar voor de begeleiding en de vrijheid die je mij hebt gegeven tijdens het schrijven van mijn proefschrift. En ik heb de afgelopen jaren enorm veel van je geleerd.

prof. dr. F.W Asselbergs, beste Folkert, we hebben de afgelopen jaren in veel projecten samengewerkt, zowel voor dit proefschrift als bij een heleboel verschillende projecten van het Durrer Center. Door jouw altijd positieve en enthousiaste aanpak was het altijd fijn met jou samenwerken. Bedankt voor al je goede feedback op mijn stukken en mijn proefschrift.

prof. dr. G.K Hovingh, beste Kees, ook jou wil ik bedanken voor de plezierige samenwerking en je altijd (positieve) kritische feedback op de verschillende stukken in mijn proefschrift.

Dan wil ik de leden van de promotiecommissie: prof. dr. E. J. Meijers-Heijboer, prof. dr. J. W. Jukema, prof. dr. M. A. Swertz, prof. dr. O. H. Franco, dr. J. Hamann, dr. P. Henneman graag bedanken voor het lezen en beoordelen van dit proefschrift en dat zij zitting hebben willen nemen in mijn promotiecommissie.

Ook wil ik alle co-auteurs die op de verschillende stukken staan (dat zijn er nogal veel) bedanken voor hun bijdrage en samenwerking.

Ook wil ik graag alle collega's op de afdeling Klinische Epidemiologie, Biostatistiek en Bioinformatica (KEBB) in het AMC bedanken. Ik heb me door de jaren heen altijd thuis gevoeld op de KEBB en genoten van de verschillende activiteiten die elk jaar georganiseerd worden. Ik wil Iris, Raha, Michel, Wouter en Marit bedanken voor de altijd leuke en gezellige sfeer in 207 de afgelopen jaren. Michel en Wouter dank jullie wel dat jullie als paranimf willen optreden.

Suthesh bedankt voor de samenwerking en de gastvrijheid in Boston! (al is dat alweer even een tijd geleden). Ik denk dat we inderdaad echt alles uit de cardiochip data hebben gehaald wat erin zat.

Dan wil ik mijn huidige collega's (en oud-collega's) bij het Netherlands Heart-institute/ Durrer Center en AMC biobank bedanken. Wanda, Jörg, John, René, Astrid, Susan, Peter, Pim, Folkert, Jan, Carina, Alex en Peter bedankt voor de plezierige samenwerking in het biobank-team, de wekelijkse werkbesprekingen, de kwartaaloverleggen, de congres bezoeken en de leuke en gezellige uitjes en BBQ's die we hebben gehad. En natuurlijk ook de weekenden skiën in Winterberg, hopelijk volgen er hier in de toekomst nog meer van.

Ook wil ik mijn collega's bij TraIT/BBMRI: David, Tieneke, Jeroen, Morris, Jan-Willem, Rita en Erna bedanken, voor de samenwerking de afgelopen jaren in de verschillende projecten. Ik heb altijd (en nog steeds) met veel plezier aan de verschillende projecten gewerkt.

Familie en vrienden ook jullie bedankt, jullie hebben altijd gezorgd voor afleiding en ontspanning. Teamgenoten bedankt! Volleyballen was (en is nog steeds) een goede manier voor mij om te ontspannen, mede dankzij jullie.

Dan wil ik mijn ouders bedanken voor alle leuke dingen die we in het verleden hebben gedaan. Helaas zijn jullie er beiden niet meer. Wat was jij trots Mama, gelukkig heb ik je nog wel kunnen vertellen dat ik mijn proefschrift op 10 oktober 2018 ga verdedigen. Dank jullie wel voor alles!

Lief zusje, wat hebben wij een rot tijd achter de rug, ik wil jou ook heel erg bedanken voor de 31 jaar die jij al deel uitmaakt van mijn leven. Hopelijk gaan we nu alleen nog maar leuke dingen meemaken!

Als laatste wil ik mijn "gezinnetje" bedanken, Lieve Kim, wat ben ik blij dat ik jou heb leren kennen en wat hebben wij een heleboel leuke dingen samen meegemaakt de afgelopen 7 jaar (helaas ook een heleboel minder leuke dingen...). Gelukkig hadden wij elkaar in deze moeilijke tijden. Dank je wel voor alles!
En als allerlaatste wil ik Fleur en Tijn bedanken! Mijn (ons) leven is zo veel rijker geworden sinds jullie geboortes. Wat ben ik blij dat jullie er zijn. Ik hoop dat we vanaf nu alleen nog maar leuke dingen mee gaan maken met z'n vieren.

CURRICULUM VITAE

Erik Pieter Adriaan van Iperen was born on May 26, 1985 in Schiedam and grew up in Hilversum. After finishing the gymnasium at the Alberdingk Thijm College in Hilversum, he studied Medical Informatics at the University of Amsterdam. He finished his bachelor in Medical Informatics in 2006 and his master Medical Informatics in 2008.

After finishing his master he started as a PhD student at the Durrer Center for cardiovascular research and at the clinical epidemiology, biostatistics and bioinformatics department (KEBB) at the Academic Medical Center (AMC) in Amsterdam.

Under the supervision of prof. dr. A.H. Zwinderman, prof. dr. F.W. Asselbergs and prof. dr. G.K. Hovingh he analyzed large genetic datasets with a primary focus on CVD and lipids metabolism.

During his PhD he was also closely involved in the setup of the ICT infrastructure and multiple Research Datamanagement and biobanking (infrastructure) projects within the Durrer Center for cardiovascular research and AMC Biobank, he was also involved in the biobanking work-package of the CTMM Translational research IT (TraIT) project, ICT coordinator for the Parelsnoer Institute (PSI) project in the AMC and taskleader within the BBMRI 2.0 work package 5 (Access to data and samples). Within the Academic medical center, Erik is involved in the newly formed ICT for research initiative.

As of October 2018 he will continue working for the Netherlands Heart Institute: Durrer Center, AMC biobank, PSI, BBMRI-NL and the AMC ICT for research initiative.

LIST OF PUBLICATIONS

1. Kroner, A, **Erik P A van Iperen**, J Horn, J M Binnekade, P E Spronk, J Stoker, and M J Schultz. The low therapeutic efficacy of postoperative chest radiographs for surgical intensive care unit patients.*Minerva anestesiologica 77.2 (Feb. 2011), pp. 14753*

2. Lanktree, Matthew B, Yiran Guo, Muhammed Murtaza, et al. Meta-analysis of Dense Gene centric Association Studies Reveals Common and Uncommon Variants Associated with Height. American journal of human genetics 88.1 (Jan. 2011), pp. 618.

3. **Erik P A van Iperen** et al. Large-scale gene-centric meta-analysis across 32 studies identifies multiple lipid loci. *American journal of human genetics 91.5 (Nov. 2012), pp. 823 38. issn: 1537-6605. doi: 10 . 1016 / j . ajhg . 2012 . 08 . 032.*

4. C. C. Elbers, Y. Guo, V. Tragante, **Erik P A van Iperen**, M. B. Lanktree, B. A. Castillo, F. Chen, L. R. Yanek, M. K. Wo jczynski, Y. R. Li, B. Ferwerda, C. M. Ballantyne, S. G. Buxbaum, Y.-D. I. Chen, W.-M. Chen, L. A. Cupples, M. Cushman, Y. Duan, D. Duggan, M. K. Evans, J. K. Fernandes, M. Fornage, M. Garcia, W. T. Garvey, N. Glazer, F. Gomez, T. B. Harris, I. Halder, V. J. Howard, M. F. Keller, M. I. Kamb oh, C. Kooperberg, S. B. Kritchevsky, A. LaCroix, K. Liu, Y. Liu, K. Musunuru, A. B. Newman, N. C. Onland-Moret, J. Ordovas, I. Peter, W. Post, S. Redline, S. E. Reis, R. Saxena, P. J. Schreiner, K. a. Volcik, X. Wang, S. Yusuf, A. B. Zonderland, S. S. Anand, D. M. Becker, B. Psaty, D. J. Rader, A. P. Reiner, S. S. Rich, J. I. Rotter, M. M. Sale, M. Y. Tsai, I. B. Borecki, R. a. Hegele, S. Kathiresan, M. a. Nalls, H. a. Taylor, H. Hakonarson, S. Sivapalaratnam, F. W. Asselbergs, F. Drenos, J. G. Wilson, and B. J. Keating. Gene-centric meta-analysis of lipid traits in African, East Asian and Hispanic populations. *PloS one 7.12 (Jan. 2012), e50198. issn: 1932-6203. doi: 10 . 1371 / journal . pone . 0050198.*

5. I. M. Purmer, **Erik P A van Iperen**, L. F. M. Beenen, M. J. Kuip er, J. M. Binnekade, P. W. van der Top, M. J. Schultz, and J. Horn. Brain computer tomography in critically ill patients a prospective cohort study. *BMC medical imaging 12.1 (Jan. 2012), p. 34. issn: 1471-2342. doi: 10 . 1186 / 1471 - 2342 - 12 - 34.*

6. R. Saxena et al. Large-scale gene-centric meta-analysis across 39 studies identifies type 2 diabetes loci.*American journal of human genetics 90.3 (Mar. 2012), pp. 410 25. issn: 1537-6605. doi: 10 . 1016 / j . ajhg . 2011 . 12 . 022.*

7. F. H. van der Baan, M. J. Knol, A. H. Maitland-van der Zee, J. J. Regieli, **Erik P A van Iperen**, A. C. G. Egberts, O. H. Klungel, D. E. Grobbee, and J. W. Jukema. Added value of pharmacogenetic testing in predicting statin response: results from the REGRESS trial. 2012. doi: 10 . 1038 / tpj.2012.12 8. K. Ganesh et al. Loci influencing blood pressure

identified using a cardiovascular gene-centric array.*Human molecular genetics 22.8 (Apr. 2013), pp. 1663 78. issn: 1460-2083. doi: 10 . 1093 / hmg / dds555*

9. Y. Guo, M. B. Lanktree, K. C. Taylor, **Erik P A van Iperen**, H. Hakonarson, L. a. Lange, and B. J. Keating. Gene-centric meta-analyses of 108 912 individuals confirm known body mass index loci and reveal three novel signals. *Human molecular genetics 22.1 (Jan. 2013), pp. 184 201. issn: 1460-2083. doi: 10 . 1093 / hmg / dds396.*

10. M. V. Holmes et al. Secretory phospholipase A2-I IA and cardiovascular disease: A mendelian randomization study. *Journal of the American College of Cardiology 62.21 (2013), pp. 1966 1976. issn: 07351097. doi: 10.1016/j.jacc.2013.06.044*

11. L. Jansen, A. de Niet, F. Stelma, **E. P. A. van Iperen**, K. A. van Dort, M. J. T. Plat-Sinnige, R. B. Takkenberg, D. J. Chin, A. H. Zwinderman, U. Lopatin, N. A. Kootstra, and H. W. Reesink. HBsAg loss in patients treated with peginterferon alfa-2a and adefovir is associated with SLC16A9 gene variation and lower plasma carnitine levels. 2013. doi: 10 . 1016 / j . jhep.2014.05.004

12. M. V. Holmes, F. W. Asselbergs, T. M. Palmer, F. Drenos, M. B. Lanktree, C. P. Nelson, C. E. Dale, S. Padmanabhan, C. Finan, D. I. Swerdlow, V. Tragante, **E. P. A. van Iperen**, S. Sivapalaratnam, S. Shah, C. C. Elbers, T. Shah, J. Engmann, C. Giambartolomei, J. White, D. Zabaneh, R. Sofat, S. McLachlan, P. A. Doevendans, A. J. Balmforth, A. S. Hall, K. E. North, B. Almoguera, R. C. Hoogeveen, M. Cushman, M. Fornage, S. R. Patel, S. Redline, D. S. Siscovick, M. Y. Tsai, K. J. Karczewski, M. H. Hofker,W. M. Verschuren, M. L. Bots, Y. T. van der Schouw, O. Melander, A. F. Dominiczak, R. Morris, Y. Ben-Shlomo, J. Price, M. Kumari, J. Baumert, A. Peters, B. Thorand, W. Koenig, T. R. Gaunt, S. E. Humphries, R. Clarke, H. Watkins, M. Farrall, J. G. Wilson, S. S. Rich, P. I. W. de Bakker, L. a. Lange, G. Davey Smith, A. P. Reiner, P. J. Talmud, M. Kivimaki, D. a. Lawlor, F. Dudbridge, N. J. Samani, B. J. Keating, A. D. Hingorani, and J. P. Casas. Mendelian randomization of blood lipids for coronary heart disease. *European heart journal (2014), pp. 1 27. issn: 1522-9645. doi: 10 . 1093 / eurheartj / eht571.*

13. V. Tragante et al. Gene-centric meta-analysis in 87,736 individuals of European ancestry identifies multiple blood-pressure-related loci. *American Journal of Human Genetics 94.3 (2014), pp. 349 360. issn: 15376605. doi: 10.1016/j.ajhg.2013.12.016*

14. **E.P. A. van Iperen**, S. Sivapalaratnam, S. M. Boekholdt, G. K. Hovingh, S. Maiwald, M. W. Tanck, N. Soranzo, J. C. Stephens, J. G. Sambrook, M. Levi, W. H. Ouwehand, J. J. Kastelein, M. D. Trip, and A. H. Zwinderman. Common genetic variants do not associate with CAD in familial hypercholesterolemia. *Eur J Hum Genet 22.6 (2014), pp. 809 813. issn: 1476-5438. doi: 10.1038/ejhg.2013.242.*

15. E. Nuesch, C. Dale, T. M. Palmer, J. White, B. J. Keating, **E. P. A. van Iperen**, A. Goel, S. Padmanabhan, F. W. Asselbergs, E.-N. Investigators, W. M. Verschuren, C. Wijmenga, Y. T. Van der Schouw, N. C. Onland-Moret, L. A. Lange, G. K. Hovingh, S. Sivapalaratnam, R. W. Morris, P. H. Whincup, G. S. Wannamethe, T. R. Gaunt, S. Ebrahim, L. Steel, N. Nair, A. P. Reiner, C. Ko op erb erg, J. F. Wilson, J. L. Bolton, S. McLachlan, J. F. Price, M. W. Strachan, C. M. Robertson, M. E. Kleb er, G. Delgado, W. Marz, O. Melander, A. F. Dominiczak, M. Farrall, H. Watkins, M. Leusink, A. H. Maitland-van der Zee, M. C. de Groot, F. Dudbridge, A. Hingorani, Y. Ben-Shlomo, D. A. Lawlor, U. Investigators, A. Amuzu, M. Caueld, A. Cavadino, J. Cooper, T. L. Davies, I. Day, F. Drenos, J. Engmann, C. Finan, C. Giambartolomei, R. Hardy, S. E. Humphries, E. Hypponen, M. Kivimaki, D. Kuh, M. Kumari, K. Ong, V. Plagnol, C. Power, M. Richards, S. Shah, T. Shah, R. Sofat, P. J. Talmud, N. Wareham, H. Warren, J. C. Whittaker, A. Wong, D. Zabaneh, G. Davey Smith, J. C. Wells, D. A. Leon, M. V. Holmes, and J. P. Casas. Adult height, coronary heart disease and stroke: a multi-lo cus Mendelian randomization meta-analysis. *International journal of epidemiology (2015). issn: 1464-3685. doi: 10 . 1093 / ije / dyv074.*

16. G. B. Ehret et al. The genetics of blood pressure regulation and its target organs from association studies in 342,415 individuals. *Nature genetics 48.10 (2016), pp. 1171-1184. issn: 1546-1718. doi: 10 . 1038 / ng . 3667.*

17. M. Leusink, A. H. Maitland-van der Zee, B. Ding, F. Drenos, **E. P. A. van Iperen**, H. R. Warren, M. J. Cauleld, L. A. Cupples, M. Cushman, A. D. Hingorani, R. C. Hoogeveen, G. K. Hovingh, M. Kumari, L. A. Lange, P. B. Munro e, F. Nyberg, P. J. Schreiner, S. Sivapalaratnam, P. I. de Bakker, A. de Boer, B. J. Keating, F. W. Asselbergs, and N. C. Onland- Moret. A genetic risk score is associated with statin-induced low-density lipoprotein cholesterol lowering. *Pharmacogenomics 17.6 (Apr. 2016), pp. 583 91. issn: 1744-8042. doi: 10.2217/pgs.16.8.*

18. Myocardial Infarction Genetics and CARDIoGRAM Exome Consortia Investigators et al. Coding Variation in ANGPTL4, LPL, and SVEP1 and the Risk of Coronary Disease. *The New England journal of medicine 374.12 (Mar. 2016), pp. 1134 44. issn: 1533-4406. doi: 10 . 1056 / NEJMoa1507652.*

19. **E. P. A. van Iperen**, S. Sivapalaratnam, M. V. Holmes, G. K. Hovingh, A. H. Zwinderman, and F. W. Asselbergs. Genetic analysis of emerging risk factors in coronary artery disease. *Atherosclerosis 254 (2016) pp. 35 41. issn: 1879-1484. doi: 10 . 1016 / j . atherosclerosis . 2016 . 09.008*

20. E. R. Holzinger, S. S. Verma, C. B. Moore, M. Hall, R. De, D. Gilbert-Diamond, M. B. Lanktree, N. Pankratz, A. Amuzu, A. Burt, C. Dale, S. Dudek, C. E. Furlong, T.

R. Gaunt, D. S. Kim, H. Riess, S. Sivapalaratnam, V. Tragante, **E. P. A. van Iperen**, A. Brautbar, D. S. Carrell, D. R. Crosslin, G. P. Jarvik, H. Kuivaniemi, I. J. Kullo, E. B. Larson, L. J. Rasmussen-Torvik, G. Tromp, J. Baumert, K. J. Cruickshanks, M. Farrall, A. D. Hingorani, G. K. Hovingh, M. E. Kleber, B. E. Klein, R. Klein, W Koenig, L. A. Lange, W. Mfdfdrz, K. E. North, N. Charlotte Onland- Moret, A. P. Reiner, P. J. Talmud, Y. T. van der Schouw, J. G. Wilson, M. Kivimaki, M. Kumari, J. H. Moore, F. Drenos, F. W. Asselbergs, B. J. Keating, and M. D. Ritchie. Discovery and replication of SNP-SNP interactions for quantitative lipid traits in over 60,000 individuals. *BioData mining 10 (2017), p. 25. issn: 1756-0381. doi: 10.1186/s13040- 017 - 0145 - 5.*

21. A. F. Schmidt et al. PCSK9 genetic variants and risk of type 2 diabetes: a mendelian randomisation study. *The lancet. Diabetes endocrinology 5.2 (Feb. 2017), pp. 97 105. issn: 2213-8595. doi: 10 . 1016 / S2213 - 8587(16) 30396 - 5.*

22. **E. P. A. van Iperen**, G. K. Hovingh, F. W. Asselbergs, and A. H. Zwinderman. Extending the use of GWAS data by combining data from different genetic platforms. *PloS one 12.2 (2017), e0172082. issn: 1932-6203. doi: 10 . 1371 / journal . pone . 0172082.*

23. T. R. Webb et al. Systematic Evaluation of Pleiotropy Identifies 6 Further Loci Associated With Coronary Artery Disease. *Journal of the American Col lege of Cardiology 69.7 (Feb. 2017), pp. 823 836. issn: 1558- 3597. doi: 10.1016/j.jacc.2016.11.056.*

24. E. Wheeler et al. Impact of common genetic determinants of Hemoglobin A1c on type 2 diabetes risk and diagnosis in ancestrally diverse populations: A transethnic genome-wide meta-analysis. *PLoS medicine 14.9 (Sept. 2017), e1002383. issn: 1549-1676. doi: 10 . 1371 / journal . pmed . 1002383.*

PHD PORTFOLIO AMC GRADUATE SCHOOL

PhD training

Course	*Year*
ISCB Marseille: Pre-conference course 3: "Analysis of high-throughput SNP association studies"	2010
ASHG Workshop: Social Media + Scientists = Success	2012
OpenClinica Training Course	2013
ASHG 2013: High-Throughput Data Analysis and Visualization with Galaxy Workshop	2013
REDCap user training	2014
GATK Best Practices + Building Analysis Pipelines with Queue Workshop	2013
BBMRI-NL's social media workshop	2015
IBS Channel meeting: Splines Course (by Paul Eilers, Rotterdam, NL)	2015
Labvantage User training	2016-2018
OpenSpecimen user meeting (Amsterdam)	2017
Weekly departmental KEBB seminars	2008-2017

Teaching

Name	*Year*
Tutorial bachelor medicine: medical statistics	2011-2013

Presentations (orals and posters)

Conference	*Year*
American Society of Human Genetics Annual Meeting: Common genetic variants do not predict CAD in Familial Hypercholesterolemia	2012
American Society of Human Genetics Annual Meeting: A comparison of imputation quality: combining different GWAS platforms	2013
BBMRI: Connecting biobanks: Durrer Center for Cardiovascular Research	2014
ICIN Themamiddag Biobanking: Durrer Center for Cardiovascular Research and Biobanking eCRFs, catalogs and Virtual Biobanking	2014
IBS Channel Meeting:Enrichment of GWAS data, combining different sources of genetic data.	2015
BBMRI Hands on Biobanks (Milan): Standardized workflow for bio-material requests	2015
Health-RI meeting	2016
Global Biobankweek (GBW) (Sweden): Generic Portal to request access to biobank data and samples	2017
Health-RI meeting: Podium (Pitch)	2017
CTEC: Labvantage User meeting (Lisbon)	2018
European Biobank Week (Antwerp): Making biobank data and samples findable and accessible	2018

Funding

Funding	*Year*
BBMRI Voucher: Project to make OpenSpecimen available as a free Biobank Information Management system (BIMS) for (Dutch) researchers	2016-2017

Other

Activity	*Year*
1 month visit Brendan Keating's lab (Philadelphia): IBC cardiochip consortium	2011

REFERENCES

[1] A MORDENTE, B GUANTARIO, E MEUCCI, et al. “Lycopene and cardiovascular diseases: an update.” In: *Current medicinal chemistry* 18.8 (Jan. 2011), pp. 1146–63.

[2] RAJA B SINGH, SUSHMA A MENGI, YAN-JUN XU, et al. “Pathogenesis of atherosclerosis: A multifactorial process.” In: *Experimental and clinical cardiology* 7.1 (2002), pp. 40–53.

[3] HIMADRI ROY, SHALINI BHARDWAJ, and SEPPO YLA-HERTTUALA. “Molecular genetics of atherosclerosis.” In: *Human genetics* 125.5-6 (June 2009), pp. 467–91.

[4] GEORGE THANASSOULIS and RAMACHANDRAN S VASAN. “Genetic cardiovascular risk prediction: will we get there?” In: *Circulation* 122.22 (Nov. 2010), pp. 2323–34.

[5] PETER M VISSCHER, WILLIAM G HILL, and NAOMI R WRAY. “Heritability in the genomics era concepts and misconceptions”. In: *Nature Reviews Genetics* 9.4 (Apr. 2008), pp. 255–266.

[6] J J NORA, R H LORTSCHER, R D SPANGLER, et al. “Genetic–epidemiologic study of early-onset ischemic heart disease.” In: *Circulation* 61.3 (Mar. 1980), pp. 503–8.

[7] T A MANOLIO, F S COLLINS, N J COX, et al. *Finding the missing heritability of complex diseases*.

[8] H WATKINS, W J MCKENNA, L THIERFELDER, et al. “Mutations in the genes for cardiac troponin T and alpha-tropomyosin in hypertrophic cardiomyopathy.” In: *The New England journal of medicine* 332.16 (Apr. 1995), pp. 1058–64.

[9] M E CURRAN, I SPLAWSKI, K W TIMOTHY, et al. “A molecular basis for cardiac arrhythmia: HERG mutations cause long QT syndrome.” In: *Cell* 80.5 (Mar. 1995), pp. 795–803.

[10] E J SIJBRANDS, R G WESTENDORP, J C DEFESCHE, et al. “Mortality over two centuries in large pedigree with familial hypercholesterolaemia: family tree mortality study”. eng PT - Journal Article SB - AIM SB - IM. In: *BMJ* 322.0959-8138 (Print) (Apr. 2001), pp. 1019–1023.

[11] INTERNATIONAL HAPMAP CONSORTIUM. “The International HapMap Project.” In: *Nature* 426.6968 (Dec. 2003), pp. 789–96.

[12] GLOBAL LIPIDS GENETICS CONSORTIUM, CRISTEN J WILLER, ELLEN M SCHMIDT, et al. “Discovery and refinement of loci associated with lipid levels.” In: *Nature genetics* 45.11 (Nov. 2013), pp. 1274–83.

[13] MICHAEL PREUSS, INKE R KNIG, JOHN R THOMPSON, et al. “Design of the Coronary ARtery DIsease Genome-Wide Replication And Meta-Analysis (CARDIoGRAM) Study: A Genome-wide association meta-analysis involving more than 22 000 cases and 60 000 controls.” In: *Circulation. Cardiovascular genetics* 3.5 (Oct. 2010), pp. 475–83.

[14] PANOS DELOUKAS, STAVROULA KANONI, CHRISTINA WILLENBORG, et al. “Large-scale association analysis identifies new risk loci for coronary artery disease.” In: *Nature genetics* 45.1 (Jan. 2013), pp. 25–33.

[15] T M TESLOVICH, K MUSUNURU, A V SMITH, et al. *Biological, clinical and population relevance of 95 loci for blood lipids*.

[16] BENJAMIN F VOIGHT, HYUN MIN KANG, JUN DING, et al. "The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits." In: *PLoS genetics* 8.8 (Jan. 2012), e1002793.

[17] BRENDAN J KEATING, SAM TISCHFIELD, SARAH S MURRAY, et al. "Concept, design and implementation of a cardiovascular gene-centric 50 k SNP array for large-scale genomic association studies." In: *PloS one* 3.10 (Jan. 2008), e3583.

[18] *Exome Chip Design*. 2012.

[19] 1000 GENOMES PROJECT CONSORTIUM, CORRESPONDING AUTHORS, STEERING COMMITTEE, et al. "A global reference for human genetic variation." In: *Nature* 526.7571 (Sept. 2015), pp. 68–74.

[20] ELEFTHERIA ZEGGINI, LAURA J SCOTT, RICHA SAXENA, et al. "Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes." In: *Nature genetics* 40.5 (2008), pp. 638–45.

[21] CRISTEN J WILLER, ELIZABETH K SPELIOTES, RUTH J F LOOS, et al. "Six new loci associated with body mass index highlight a neuronal influence on body weight regulation." In: *Nature genetics* 41.1 (Jan. 2009), pp. 25–34.

[22] ANDREW D PATERSON, DARYL WAGGOTT, ANDREW P BORIGHT, et al. "A genome-wide association study identifies a novel major locus for glycemic control in type 1 diabetes, as measured by both A1C and glucose." In: *Diabetes* 59.2 (Feb. 2010), pp. 539–49.

[23] INTERNATIONAL CONSORTIUM FOR BLOOD PRESSURE GENOME-WIDE ASSOCIATION STUDIES, GEORG B EHRET, PATRICIA B MUNROE, et al. "Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk." In: *Nature* 478.7367 (Sept. 2011), pp. 103–9.

[24] FOLKERT W ASSELBERGS, YIRAN GUO, ERIK P A VAN IPEREN, et al. "Large-scale gene-centric meta-analysis across 32 studies identifies multiple lipid loci." In: *American journal of human genetics* 91.5 (Nov. 2012), pp. 823–38.

[25] YUN R LI and BRENDAN J KEATING. "Trans-ethnic genome-wide association studies: advantages and challenges of mapping in diverse populations." In: *Genome medicine* 6.10 (2014), p. 91.

[26] DANIELLE WELTER, JACQUELINE MACARTHUR, JOANNELLA MORALES, et al. "The NHGRI GWAS Catalog, a curated resource of SNP-trait associations." In: *Nucleic acids research* 42.Database issue (Jan. 2014), pp. D1001–6.

[27] DORRET I BOOMSMA, CISCA WIJMENGA, ELINE P SLAGBOOM, et al. "The Genome of the Netherlands: design, and project goals." In: *European journal of human genetics : EJHG* 22.2 (Feb. 2014), pp. 221–7.

[28] UK10K CONSORTIUM, WRITING GROUP, PRODUCTION GROUP, et al. "The UK10K project identifies rare variants in health and disease." In: *Nature* 526.7571 (Oct. 2015), pp. 82–90.

[29] W B KANNEL, T R DAWBER, A KAGAN, et al. "Factors of risk in the development of coronary heart disease–six year follow-up experience. The Framingham Study." In: *Annals of internal medicine* 55 (July 1961), pp. 33–50.

[30] JULIA HIPPISLEY-COX, CAROL COUPLAND, YANA VINOGRADOVA, et al. "Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study." In: *BMJ (Clinical research ed.)* 335.7611 (July 2007), p. 136.

[31] MICHAEL J PENCINA, RALPH B D'AGOSTINO, MARTIN G LARSON, et al. "Predicting the 30-year risk of cardiovascular disease: the framingham heart study." In: *Circulation* 119.24 (June 2009), pp. 3078–84.

[32] PAUL M RIDKER, NINA P PAYNTER, NADER RIFAI, et al. "C-reactive protein and parental history improve global cardiovascular risk prediction: the Reynolds Risk Score for men." In: *Circulation* 118.22 (Nov. 2008), 2243–51, 4p following 2251.

[33] P W WILSON, R B D'AGOSTINO, D LEVY, et al. "Prediction of coronary heart disease using risk factor categories." In: *Circulation* 97.18 (May 1998), pp. 1837–47.

[34] UMESH N KHOT, MONICA B KHOT, CHRISTOPHER T BAJZER, et al. "Prevalence of conventional risk factors in patients with coronary heart disease." In: *JAMA : the journal of the American Medical Association* 290.7 (Aug. 2003), pp. 898–904.

[35] DANIEL W BELSKY, TERRIE E MOFFITT, KAREN SUGDEN, et al. "Development and evaluation of a genetic risk score for obesity." In: *Biodemography and social biology* 59.1 (2013), pp. 85–100.

[36] ERIK P A VAN IPEREN, SUTHESH SIVAPALARATNAM, S MATTHIJS BOEKHOLDT, et al. "Common genetic variants do not associate with CAD in familial hypercholesterolemia". In: *Eur J Hum Genet* 22.6 (2014), pp. 809–813.

[37] ERIK P A VAN IPEREN, SUTHESH SIVAPALARATNAM, MICHAEL V HOLMES, et al. "Genetic analysis of emerging risk factors in coronary artery disease." In: *Atherosclerosis* 254 (Nov. 2016), pp. 35–41.

[38] A D LOPEZ, C D MATHERS, M EZZATI, et al. *Global and regional burden of disease and risk factors, 2001: systematic analysis of population health data*.

[39] G S BERENSON, S R SRINIVASAN, W BAO, et al. *Association between multiple cardiovascular risk factors and atherosclerosis in children and young adults. The Bogalusa Heart Study*.

[40] B J ARSENAULT, S M BOEKHOLDT, and J J KASTELEIN. *Lipid parameters for measuring risk of cardiovascular disease*.

[41] ANGELANTONIO E DI, N SARWAR, P PERRY, et al. *Major lipids, apolipoproteins, and risk of vascular disease*.

[42] L A WEISS, L PAN, M ABNEY, et al. *The sex-specific genetic architecture of quantitative traits in humans*.

[43] THE IBC and C A D CONSORTIUM. *Large-scale gene-centric analysis identifies novel variants for coronary artery disease.* Sept. 2011.

[44] R CLARKE, J F PEDEN, J C HOPEWELL, et al. *Genetic variants associated with Lp(a) lipoprotein level and coronary disease*.

[45] MATTHEW B LANKTREE, SONIA S ANAND, SALIM YUSUF, et al. "Replication of genetic associations with plasma lipoprotein traits in a multiethnic sample." In: *Journal of lipid research* 50.7 (July 2009), pp. 1487–96.

[46] P J TALMUD, F DRENOS, S SHAH, et al. *Gene-centric association signals for lipids and apolipoproteins identified via the HumanCVD BeadChip*.

[47] E R FOX, J H YOUNG, Y LI, et al. *Association of genetic variation with systolic and diastolic blood pressure among African Americans: the Candidate Gene Association Resource study*.

[48] A D JOHNSON, R E HANDSAKER, S L PULIT, et al. *SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap*.

[49] T P CAPPOLA, M LI, J HE, et al. *Common variants in HSPB7 and FRMD4B associated with advanced heart failure*.

[50] RICHA SAXENA, CLARA C ELBERS, YIRAN GUO, et al. "Large-scale gene-centric meta-analysis across 39 studies identifies type 2 diabetes loci." In: *American journal of human genetics* 90.3 (Mar. 2012), pp. 410–25.

[51] G A WALFORD, T GREEN, B NEALE, et al. *Common genetic variants differentially influence the transition from clinically defined states of fasting glucose metabolism*.

[52] MATTHEW B. LANKTREE, YIRAN GUO, MUHAMMED MURTAZA, et al. "Meta-analysis of dense genecentric association studies reveals common and uncommon variants associated with height". In: *American Journal of Human Genetics* 88.1 (Jan. 2011), pp. 6–18.

[53] S D DE FERRANTI. *Childhood cholesterol disorders: the iceberg base or nondisease?*

[54] B M KAESS, M TOMASZEWSKI, P S BRAUND, et al. *Large-scale candidate gene analysis of HDL particle features*. 2011.

[55] W T FRIEDEWALD, R I LEVY, and D S FREDRICKSON. *Estimation of the concentration of low-density lipoprotein cholesterol in plasma, without use of the preparative ultracentrifuge*.

[56] M MARMOT and E BRUNNER. *Cohort Profile: the Whitehall II study*.

[57] M R LAW, N J WALD, and A R RUDNICKA. *Quantifying effect of statins on low density lipoprotein cholesterol, ischaemic heart disease, and stroke: systematic review and meta-analysis*.

[58] SHAUN PURCELL, BENJAMIN NEALE, KATHE TODD-BROWN, et al. "PLINK: a tool set for whole-genome association and population-based linkage analyses." In: *American journal of human genetics* 81.3 (Sept. 2007), pp. 559–75.

[59] A L PRICE, N J PATTERSON, R M PLENGE, et al. *Principal components analysis corrects for stratification in genome-wide association studies*.

[60] A L PRICE, J BUTLER, N PATTERSON, et al. *Discerning the ancestry of European Americans in genetic association studies*.

[61] C C ELBERS. *Gene-centric meta-analysis of lipid traits in African, East Asian and Hispanic populations*. 2012.

[62] K S LO, J G WILSON, L A LANGE, et al. *Genetic association analysis highlights new loci that modulate hematological trait variation in Caucasians and African Americans*.

[63] M B LANKTREE, Y GUO, M MURTAZA, et al. *Meta-analysis of Dense Genecentric Association Studies Reveals Common and Uncommon Variants Associated with Height*.

[64] S A BACANU, B DEVLIN, and K ROEDER. *Association studies for quantitative traits in structured populations*.

[65] T A PEARSON and T A MANOLIO. *How to interpret a genome-wide association study*.

[66] C J WILLER, Y LI, and G R ABECASIS. *METAL: fast and efficient meta-analysis of genomewide association scans*.

[67] P I DE BAKKER, M A FERREIRA, X JIA, et al. *Practical aspects of imputation-driven meta-analysis of genome-wide association studies*.

[68] W VIECHTBAUER. *Conducting meta-analyses in R with the metafor package*. 2010.

[69] J.E.& SCHMIDT HUNTER F.L. *Methods of meta-analysis: Correcting error and bias in research findings*. 1990.

[70] J R THOMPSON, J ATTIA, and C MINELLI. *The meta-analysis of genome-wide association studies*.

[71] J P HIGGINS and S G THOMPSON. *Quantifying heterogeneity in a meta-analysis*.

[72] P SCHEET and M STEPHENS. *A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase*.

[73] L A HINDORFF, P SETHUPATHY, H A JUNKINS, et al. *Potential etiologic and functional implications of genome-wide association loci for human diseases and traits*.

[74] H AKAIKE. "A new look at the statistical model identification". In: *IEEE Trans.Automat.Contr.* 19 (1974), pp. 716–723.

[75] ADRIAN CORTES and MATTHEW BROWN. *Promise and pitfalls of the Immunochip*. 2011.

[76] S BUYSKE, Y WU, C L CARTY, et al. *Evaluation of the Metabochip Genotyping Array in African Americans and Implications for Fine Mapping of GWAS-Identified Loci: The PAGE Study*. 2012.

[77] P BARTER. *CETP and atherosclerosis*.

[78] P J BARTER, H B BREWER JR., M J CHAPMAN, et al. *Cholesteryl ester transfer protein: a novel target for raising HDL and inhibiting atherosclerosis*.

[79] M A AUSTIN, C M HUTTER, R L ZIMMERN, et al. *Familial hypercholesterolemia and coronary heart disease: a HuGE association review*.

[80] R A HEGELE. *Plasma lipoproteins: genetic influences and clinical implications*.

[81] S E LEIGH, A H FOSTER, R A WHITTALL, et al. *Update and analysis of the University College London low density lipoprotein receptor familial hypercholesterolemia database*.

[82] M V HOLMES, S HARRISON, P J TALMUD, et al. *Utility of genetic determinants of lipids and cardiovascular events in assessing risk*.

[83] K W VAN DIJK, P C RENSEN, P J VOSHOL, et al. *The role and mode of action of apolipoproteins CIII and AV: synergistic actors in triglyceride metabolism?*

[84] A STROMBERG and J MARTENSSON. *Gender differences in patients with heart failure*.

[85] K K ANAGNOSTOPOULOU, G D KOLOVOU, P M KOSTAKOU, et al. *Sex-associated effect of CETP and LPL polymorphisms on postprandial lipids in familial hypercholesterolaemia.*

[86] N R MATTHAN, S M JALBERT, P H BARRETT, et al. *Gender-specific differences in the kinetics of nonfasting TRL, IDL, and LDL apolipoprotein B-100 in men and premenopausal women*.

[87] M Z DIETER, J M MAHER, X CHENG, et al. *Expression and regulation of the sterol half-transporter genes ABCG5 and ABCG8 in rats*.

[88] L P DUAN, H H WANG, A OHASHI, et al. *Role of intestinal sterol transporters Abcg5, Abcg8, and Npc1l1 in cholesterol absorption in mice: gender and age effects*.

[89] S E HAZARD and S B PATEL. *Sterolins ABCG5 and ABCG8: regulators of whole body dietary sterols*.

[90] O H KANTARCI, D D HEBRINK, S J ACHENBACH, et al. *Association of APOE polymorphisms with disease severity in MS is limited to women*.

[91] M I KAMBOH, C E ASTON, and R F HAMMAN. *DNA sequence variation in human apolipoprotein C4 gene and its effect on plasma lipid profile*.

[92] ILLUMINA INC. *Genotyping rare variants. A simulated analysis achieves high cell rates and low error rates from loci containing rare variants. Pub. No. 370-2010-008.* Tech. rep.,

[93] L F SORIA, E H LUDWIG, H R CLARKE, et al. *Association between a specific apolipoprotein B mutation and familial defective apolipoprotein B-100.*

[94] A TYBJAERG-HANSEN and S E HUMPHRIES. *Familial defective apolipoprotein B-100: a single mutation that causes hypercholesterolemia and premature coronary artery disease*.

[95] N B MYANT. *Familial defective apolipoprotein B-100: a review, including some comparisons with familial hypercholesterolaemia*.

[96] A J SMITH, J PALMEN, W PUTT, et al. *Application of statistical and functional methodologies for the investigation of genetic determinants of coronary heart disease biomarkers: lipoprotein lipase genotype and plasma triglycerides as an exemplar*.

[97] R C DEO, D REICH, A TANDON, et al. *Genetic differences between the determinants of lipid profile phenotypes in African and European Americans: the Jackson Heart Study*.

[98] A R SHULDINER and T I POLLIN. *Genomics: Variations in blood lipids*.

[99] J YANG, T A MANOLIO, L R PASQUALE, et al. *Genome partitioning of genetic variation for complex traits using common SNPs*.

[100] J YANG, B BENYAMIN, B P MCEVOY, et al. *Common SNPs explain a large proportion of the heritability for human height*.

[101] K A FRAZER, S S MURRAY, N J SCHORK, et al. *Human genetic variation and its contribution to complex traits*.

[102] F JOHANNES, E PORCHER, F K TEIXEIRA, et al. *Assessing the impact of transgenerational epigenetic variation on complex traits*.

[103] M SLATKIN. *Epigenetic inheritance and the missing heritability problem*.

[104] C LAMINA, S COASSIN, T ILLIG, et al. *Look beyond one's own nose: combination of information from publicly available sources reveals an association of GATA4 polymorphisms with plasma triglycerides*.

[105] N R PATTINSON and K E WILLIS. *Effect of phospholipase C on cholesterol solubilization in model bile. A concanavalin A-binding nucleation-promoting factor from human gallbladder bile*.

[106] R J LOOS and C BOUCHARD. *FTO: the first gene contributing to common forms of human obesity*.

[107] J M LANCASTER, R WOOSTER, J MANGION, et al. *BRCA2 mutations in primary breast and ovarian cancers*.

[108] A R VENKITARAMAN. *Cancer susceptibility and the functions of BRCA1 and BRCA2*.

[109] A VILJOEN and A S WIERZBICKI. *Safety and efficacy of laropiprant and extended-release niacin combination in the management of mixed dyslipidemias and primary hypercholesterolemia*. 2010.

[110] A WISE, S M FOORD, N J FRASER, et al. *Molecular identification of high and low affinity receptors for nicotinic acid*.

[111] D A HELLER, U DE FAIRE, N L PEDERSEN, et al. "Genetic and environmental influences on serum lipid levels in twins." In: *The New England journal of medicine* 328.16 (Apr. 1993), pp. 1150–6.

[112] TANYA M TESLOVICH, KIRAN MUSUNURU, ALBERT V SMITH, et al. "Biological, clinical and population relevance of 95 loci for blood lipids." In: *Nature* 466.7307 (Aug. 2010), pp. 707–13.

[113] EARL S FORD, WAYNE H GILES, and WILLIAM H DIETZ. "Prevalence of the metabolic syndrome among US adults: findings from the third National Health and Nutrition Examination Survey." In: *JAMA* 287.3 (Jan. 2002), pp. 356–9.

[114] YONG-WOO PARK, SHANKUAN ZHU, LATHA PALANIAPPAN, et al. "The metabolic syndrome: prevalence and associated risk factor findings in the US population from the Third National Health and Nutrition Examination Survey, 1988-1994." In: *Archives of internal medicine* 163.4 (Feb. 2003), pp. 427–36.

[115] GUILLAUME LETTRE, CAMERON D PALMER, TAYLOR YOUNG, et al. "Genome-wide association study of coronary heart disease and its risk factors in 8,090 African Americans: the NHLBI CARe Project." In: *PLoS genetics* 7.2 (2011), e1001300.

[116] ALKES L PRICE, NICK J PATTERSON, ROBERT M PLENGE, et al. "Principal components analysis corrects for stratification in genome-wide association studies." In: *Nature genetics* 38.8 (Aug. 2006), pp. 904–9.

[117] K MUSUNURU, G LETTRE, T YOUNG, et al. *Candidate gene association resource (CARe): design, methods, and proof of concept*.

[118] MARILYN C CORNELIS, LU QI, CUILIN ZHANG, et al. "Joint effects of common genetic variants on the risk for type 2 diabetes in U.S. men and women of European ancestry." In: *Annals of internal medicine* 150.8 (Apr. 2009), pp. 541–50.

[119] KEVIN M WATERS, DANIEL O STRAM, MOHAMED T HASSANEIN, et al. "Consistent association of type 2 diabetes risk variants found in europeans in diverse racial and ethnic groups." In: *PLoS genetics* 6.8 (Aug. 2010).

[120] SARA L PULIT, BENJAMIN F VOIGHT, and PAUL I W DE BAKKER. "Multiethnic genetic association studies improve power for locus discovery." In: *PloS one* 5.9 (2010), e12600.

[121] LATISHA LOVE-GREGORY, RICHARD SHERVA, LINGWEI SUN, et al. "Variants in the CD36 gene associate with the metabolic syndrome and high-density lipoprotein cholesterol." In: *Human molecular genetics* 17.11 (June 2008), pp. 1695–704.

[122] TOSHIYUKI MORII, YOICHI OHNO, NORIHIRO KATO, et al. "CD36 single nucleotide polymorphism is associated with variation in low-density lipoprotein-cholesterol in young Japanese men." In: *Biomarkers : biochemical indicators of exposure, response, and susceptibility to chemicals* 14.4 (June 2009), pp. 207–12.

[123] ESTIBALIZ GOYENECHEA, LAURA J COLLINS, DOLORES PARRA, et al. "CD36 gene promoter polymorphisms are associated with low density lipoprotein-cholesterol in normal twins and after a low-calorie diet in obese subjects." In: *Twin research and human genetics : the official journal of the International Society for Twin Studies* 11.6 (Dec. 2008), pp. 621–8.

[124] DANY GAILLARD, FABIENNE LAUGERETTE, NICOLAS DARCEL, et al. "The gustatory pathway is involved in CD36-mediated orosensory perception of long-chain fatty acids in the mouse." In: *FASEB journal : official publication of the Federation of American Societies for Experimental Biology* 22.5 (May 2008), pp. 1458–68.

[125] FABIENNE LAUGERETTE, PATRICIA PASSILLY-DEGRACE, BRUNO PATRIS, et al. "CD36 involvement in orosensory detection of dietary lipids, spontaneous fat preference, and digestive secretions." In: *The Journal of clinical investigation* 115.11 (Nov. 2005), pp. 3177–84.

[126] CLINE MARTIN, PATRICIA PASSILLY-DEGRACE, DANY GAILLARD, et al. "The lipid-sensor candidates CD36 and GPR120 are differentially regulated by dietary lipids in mouse taste buds: impact on spontaneous fat preference." In: *PloS one* 6.8 (2011), e24014.

[127] SZILVIA BOKOR, VANESSA LEGRY, ALINE MEIRHAEGHE, et al. "Single-nucleotide polymorphism of CD36 locus and obesity in European adolescents." In: *Obesity (Silver Spring, Md.)* 18.7 (July 2010), pp. 1398–403.

[128] HLNE CHOQUET, YANN LABRUNE, FRANCK DE GRAEVE, et al. “Lack of association of CD36 SNPs with early onset obesity: a meta-analysis in 9,973 European subjects.” In: *Obesity (Silver Spring, Md.)* 19.4 (Apr. 2011), pp. 833–9.

[129] VICTOR A DROVER, MOHAMMAD AJMAL, FATIHA NASSIR, et al. “CD36 deficiency impairs intestinal lipid secretion and clearance of chylomicrons from the blood.” In: *The Journal of clinical investigation* 115.5 (May 2005), pp. 1290–7.

[130] JOHN YANG, NANDAKUMAR SAMBANDAM, XIANLIN HAN, et al. “CD36 deficiency rescues lipotoxic cardiomyopathy.” In: *Circulation research* 100.8 (Apr. 2007), pp. 1208–17.

[131] G ENDEMANN, L W STANTON, K S MADDEN, et al. “CD36 is a receptor for oxidized low density lipoprotein.” In: *The Journal of biological chemistry* 268.16 (June 1993), pp. 11811–6.

[132] D CALVO, D GMEZ-CORONADO, Y SUREZ, et al. “Human CD36 is a high affinity receptor for the native lipoproteins HDL, LDL, and VLDL.” In: *Journal of lipid research* 39.4 (Apr. 1998), pp. 777–88.

[133] N A ABUMRAD, M R EL-MAGHRABI, E Z AMRI, et al. “Cloning of a rat adipocyte membrane protein implicated in binding or transport of long-chain fatty acids that is induced during preadipocyte differentiation. Homology with human CD36.” In: *The Journal of biological chemistry* 268.24 (Aug. 1993), pp. 17665–8.

[134] P OQUENDO, E HUNDT, J LAWLER, et al. “CD36 directly mediates cytoadherence of Plasmodium falciparum parasitized erythrocytes.” In: *Cell* 58.1 (July 1989), pp. 95–101.

[135] GAURAV BHATIA, NICK PATTERSON, BOGDAN PASANIUC, et al. “Genome-wide comparison of African-ancestry populations from CARe and other cohorts reveals signals of natural selection.” In: *American journal of human genetics* 89.3 (Sept. 2011), pp. 368–81.

[136] GEORGE AYODO, ALKES L PRICE, ALON KEINAN, et al. “Combining evidence of natural selection with association analysis increases power to detect malaria-resistance variants.” In: *American journal of human genetics* 81.2 (Aug. 2007), pp. 234–42.

[137] T J AITMAN, L D COOPER, P J NORSWORTHY, et al. “Malaria susceptibility and CD36 mutation.” In: *Nature* 405.6790 (June 2000), pp. 1015–6.

[138] ANDREW E FRY, ANITA GHANSA, KERRIN S SMALL, et al. “Positive selection of a CD36 nonsense variant in sub-Saharan Africa, but no association with severe malaria phenotypes.” In: *Human molecular genetics* 18.14 (July 2009), pp. 2683–92.

[139] A VAN DE STOLPE and P T VAN DER SAAG. “Intercellular adhesion molecule-1.” In: *Journal of molecular medicine (Berlin, Germany)* 74.1 (Jan. 1996), pp. 13–33.

[140] GUILLAUME PAR, DANIEL I CHASMAN, MARK KELLOGG, et al. “Novel association of ABO histoblood group antigen with soluble ICAM-1: results of a genome-wide association study of 6,578 women.” In: *PLoS genetics* 4.7 (July 2008), e1000118.

[141] SUZETTE J BIELINSKI, ALEX P REINER, DEBORAH NICKERSON, et al. “Polymorphisms in the ICAM1 gene predict circulating soluble intercellular adhesion molecule-1(sICAM-1).” In: *Atherosclerosis* 216.2 (June 2011), pp. 390–4.

[142] P M RIDKER, C H HENNEKENS, B ROITMAN-JOHNSON, et al. “Plasma concentration of soluble intercellular adhesion molecule 1 and risks of future myocardial infarction in apparently healthy men.” In: *Lancet (London, England)* 351.9096 (Jan. 1998), pp. 88–92.

[143] YIQING SONG, JOANN E MANSON, LESLEY TINKER, et al. “Circulating levels of endothelial adhesion molecules and risk of diabetes in an ethnically diverse cohort of women.” In: *Diabetes* 56.7 (July 2007), pp. 1898–904.

[144] MYRON D GROSS, SUZETTE J BIELINSKI, JOSE R SUAREZ-LOPEZ, et al. "Circulating soluble intercellular adhesion molecule 1 and subclinical atherosclerosis: the Coronary Artery Risk Development in Young Adults Study." In: *Clinical chemistry* 58.2 (Feb. 2012), pp. 411–20.

[145] QUOC MANH NGUYEN, SATHANUR R SRINIVASAN, JI-HUA XU, et al. "Distribution and cardiovascular risk correlates of plasma soluble intercellular adhesion molecule-1 levels in asymptomatic young adults from a biracial community: the Bogalusa Heart Study." In: *Annals of epidemiology* 20.1 (Jan. 2010), pp. 53–9.

[146] D KARASEK, H VAVERKOVA, Z FRYSAK, et al. "Soluble intercellular cell adhesion molecule-1 and vascular cell adhesion molecule-1 in asymptomatic dyslipidemic subjects." In: *International angiology : a journal of the International Union of Angiology* 30.5 (Oct. 2011), pp. 441–50.

[147] JONATHAN COHEN, ALEXANDER PERTSEMLIDIS, INGRID K KOTOWSKI, et al. "Low LDL cholesterol in individuals of African descent resulting from frequent nonsense mutations in PCSK9." In: *Nature genetics* 37.2 (Feb. 2005), pp. 161–5.

[148] JONATHAN C COHEN, ERIC BOERWINKLE, THOMAS H MOSLEY, et al. "Sequence variations in PCSK9, low LDL, and protection against coronary heart disease." In: *The New England journal of medicine* 354.12 (Mar. 2006), pp. 1264–72.

[149] J L GOLDSTEIN, M K SOBHANI, J R FAUST, et al. "Heterozygous familial hypercholesterolemia: failure of normal allele to compensate for mutant allele at a regulated genetic locus". eng PT - Journal Article PT - Research Support, U.S. Gov't, P.H.S. In: *Cell* 9.0092-8674 (Print) (Oct. 1976), pp. 195–203.

[150] R HUIJGEN, M N VISSERS, J C DEFESCHE, et al. "Familial hypercholesterolemia: current treatment and advances in management". eng PT - Journal Article PT - Review. In: *Expert.Rev.Cardiovasc.Ther.* 6.1744-8344 (Electronic) (Apr. 2008), pp. 567–581.

[151] E S VAN AALST-COHEN, A C JANSEN, M W TANCK, et al. "Diagnosing familial hypercholesterolaemia: the relevance of genetic testing". eng PT - Journal Article PT - Multicenter Study PT - Research Support, Non-U.S. Gov't. In: *Eur.Heart J.* 27.0195-668X (Print) (Sept. 2006), pp. 2240–2246.

[152] N J STONE, R I LEVY, D S FREDRICKSON, et al. "Coronary artery disease in 116 kindred with familial type II hyperlipoproteinemia". eng PT - Journal Article. In: *Circulation* 49.0009-7322 (Print) (Mar. 1974), pp. 476–488.

[153] A C JANSEN, E S VAN AALST-COHEN, M W TANCK, et al. "The contribution of classical risk factors to cardiovascular disease in familial hypercholesterolaemia: data in 2400 patients". eng PT - Journal Article PT - Multicenter Study PT - Research Support, Non-U.S. Gov't. In: *J.Intern.Med.* 256.0954-6820 (Print) (Dec. 2004), pp. 482–490.

[154] O W SOUVEREIN, J C DEFESCHE, A H ZWINDERMAN, et al. "Influence of LDL-receptor mutation type on age at first cardiovascular event in patients with familial hypercholesterolaemia". eng PT - Journal Article PT - Multicenter Study PT - Research Support, Non-U.S. Gov't. In: *Eur.Heart J.* 28.0195-668X (Print) (Feb. 2007), pp. 299–304.

[155] J F PEDEN, J C HOPEWELL, D SALEHEEN, et al. "A genome-wide association study in Europeans and South Asians identifies five new loci for coronary artery disease". ENG PT - JOURNAL ARTICLE. In: *Nat.Genet.* 43.1546-1718 (Electronic) (2011), pp. 339–344.

[156] H SCHUNKERT, I R KONIG, S KATHIRESAN, et al. "Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease". ENG PT - JOURNAL ARTICLE. In: *Nat.Genet.* 43.1546-1718 (Electronic) (2011), pp. 333–338.

[157] S SIVAPALARATNAM, M M MOTAZACKER, S MAIWALD, et al. “Genome-wide association studies in atherosclerosis”. eng PT - Journal Article PT - Research Support, Non-U.S. Gov’t SB - IM. In: *Curr.Atheroscler.Rep.* 13.1534-6242 (Electronic) (June 2011), pp. 225–232.

[158] M A UMANS-ECKENHAUSEN, J C DEFESCHE, E J SIJBRANDS, et al. “Review of first 5 years of screening for familial hypercholesterolaemia in the Netherlands”. eng PT - Journal Article PT - Research Support, Non-U.S. Gov’t. In: *Lancet* 357.0140-6736 (Print) (Jan. 2001), pp. 165–168.

[159] A C JANSEN, E S VAN AALST-COHEN, M W TANCK, et al. “Genetic determinants of cardiovascular disease risk in familial hypercholesterolemia”. eng PT - Journal Article PT - Research Support, Non-U.S. Gov’t SB - IM. In: *Arterioscler.Thromb.Vasc.Biol.* 25.1524-4636 (Electronic) (July 2005), pp. 1475–1481.

[160] K HAO, E CHUDIN, J MCELWEE, et al. “Accuracy of genome-wide imputation of untyped markers and impacts on statistical power for association studies”. eng PT - Journal Article PT - Research Support, N.I.H., Extramural PT - Research Support, Non-U.S. Gov’t. In: *BMC.Genet.* 10.1471-2156 (Electronic) (2009), p. 27.

[161] YUN LI, CRISTEN J WILLER, JUN DING, et al. “MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes.” In: *Genetic epidemiology* 34.8 (Dec. 2010), pp. 816–34.

[162] L SOUTHAM, K PANOUTSOPOULOU, N W RAYNER, et al. “The effect of genome-wide association scan quality control on imputation outcome for common variants”. ENG PT - JOURNAL ARTICLE. In: *Eur.J.Hum.Genet.* 19.1476-5438 (Electronic) (May 2011), pp. 610–614.

[163] YURII S AULCHENKO, MAKSIM V STRUCHALIN, and CORNELIA M VAN DUIJN. “ProbABEL package for genome-wide association analysis of imputed data.” In: *BMC bioinformatics* 11 (Jan. 2010), p. 134.

[164] J D MUEHLSCHLEGEL, K Y LIU, T E PERRY, et al. “Chromosome 9p21 variant predicts mortality after coronary artery bypass graft surgery”. eng PT - Clinical Trial PT - Journal Article PT - Multicenter Study PT - Research Support, N.I.H., Extramural PT - Research Support, Non-U.S. Gov’t SB - AIM SB - IM. In: *Circulation* 122.1524-4539 (Electronic) (Sept. 2010), S60–S65.

[165] Y GONG, A L BEITELSHEES, R M COOPER-DEHOFF, et al. “Chromosome 9p21 haplotypes and prognosis in white and black patients with coronary artery disease”. eng PT - Journal Article PT - Research Support, N.I.H., Extramural PT - Research Support, Non-U.S. Gov’t SB - IM. In: *Circ.Cardiovasc.Genet.* 4.1942-3268 (Electronic) (Apr. 2011), pp. 169–178.

[166] NICOLA MARTINELLI, DOMENICO GIRELLI, BARBARA LUNGHI, et al. “Polymorphisms at LDLR locus may be associated with coronary artery disease through modulation of coagulation factor VIII activity and independently from lipid profile.” In: *Blood* 116.25 (Dec. 2010), pp. 5688–97.

[167] JOANNE M MURABITO, CHARLES C WHITE, MARYAM KAVOUSI, et al. “Association between chromosome 9p21 variants and the ankle-brachial index identified by a meta-analysis of 21 genome-wide association studies.” In: *Circulation. Cardiovascular genetics* 5.1 (Feb. 2012), pp. 100–12.

[168] S KATHIRESAN, B F VOIGHT, S PURCELL, et al. “Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants”. eng PT - Journal Article PT - Research Support, N.I.H., Extramural PT - Research Support, Non-U.S. Gov’t SB - IM. In: *Nat.Genet.* 41.1546-1718 (Electronic) (Mar. 2009), pp. 334–341.

[169] SAMULI RIPATTI, EMMI TIKKANEN, MARJU ORHO-MELANDER, et al. “A multilocus genetic risk score for coronary heart disease: case-control and prospective cohort analyses.” In: *Lancet* 376.9750 (Oct. 2010), pp. 1393–400.

[170] N P PAYNTER, D I CHASMAN, J E BURING, et al. “Cardiovascular disease risk prediction with and without knowledge of genetic variation at chromosome 9p21.3”. eng PT - Journal Article PT - Research Support, N.I.H., Extramural PT - Research Support, Non-U.S. Gov’t SB - AIM SB - IM. In: *Ann.Intern.Med.* 150.1539-3704 (Electronic) (Jan. 2009), pp. 65–72.

[171] RAFAEL LOZANO, MOHSEN NAGHAVI, KYLE FOREMAN, et al. “Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010.” In: *Lancet* 380.9859 (Dec. 2012), pp. 2095–128.

[172] ANNA KOTSIA, EMMANOUIL S BRILAKIS, CLAES HELD, et al. “Extent of coronary artery disease and outcomes after ticagrelor administration in patients with an acute coronary syndrome: Insights from the PLATelet inhibition and patient Outcomes (PLATO) trial.” In: *American heart journal* 168.1 (July 2014), 68–75.e2.

[173] JOHN C LAROSA, SCOTT M GRUNDY, DAVID D WATERS, et al. “Intensive lipid lowering with atorvastatin in patients with stable coronary disease.” In: *The New England journal of medicine* 352.14 (Apr. 2005), pp. 1425–35.

[174] JOHN F THOMPSON, CRAIG L HYDE, LINDA S WOOD, et al. “Comprehensive whole-genome and candidate gene analysis for response to statin therapy in the Treating to New Targets (TNT) cohort.” In: *Circulation. Cardiovascular genetics* 2.2 (Apr. 2009), pp. 173–81.

[175] ANIKA A M VAARHORST, YINGCHANG LU, BASTIAAN T HEIJMANS, et al. “Literature-based genetic risk scores for coronary heart disease: the Cardiovascular Registry Maastricht (CAREMA) prospective cohort study.” In: *Circulation. Cardiovascular genetics* 5.2 (Apr. 2012), pp. 202–9.

[176] VINICIUS TRAGANTE, PIETER A F M DOEVENDANS, HENDRIK M NATHOE, et al. “The impact of susceptibility loci for coronary artery disease on other vascular domains and recurrence risk.” In: *European heart journal* 34.37 (Oct. 2013), pp. 2896–904.

[177] BOSHAO ZHANG, DEGUI ZHI, KUI ZHANG, et al. “Practical Consideration of Genotype Imputation: Sample Size, Window Size, Reference Choice, and Untyped Rate.” In: *Statistics and its interface* 4.3 (Jan. 2011), pp. 339–352.

[178] ROBERT A SCOTT, VASILIKI LAGOU, RYAN P WELCH, et al. “Large-scale association analyses identify new loci influencing glycemic traits and provide insight into the underlying biological pathways.” In: *Nature genetics* 44.9 (Sept. 2012), pp. 991–1005.

[179] ELIZABETH K SPELIOTES, CRISTEN J WILLER, SONJA I BERNDT, et al. “Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index.” In: *Nature genetics* 42.11 (Nov. 2010), pp. 937–48.

[180] M. D. TRIP. “Frequent Mutation in the ABCC6 Gene (R1141X) Is Associated With a Strong Increase in the Prevalence of Coronary Artery Disease”. In: *Circulation* 106.7 (Aug. 2002), pp. 773–775.

[181] MEGAN L. GROVE, BING YU, BARBARA J. COCHRAN, et al. “Best Practices and Joint Calling of the HumanExome BeadChip: The CHARGE Consortium”. In: *PLoS ONE* 8.7 (2013).

[182] BRYAN HOWIE, CHRISTIAN FUCHSBERGER, MATTHEW STEPHENS, et al. “Fast and accurate genotype imputation in genome-wide association studies through pre-phasing.” In: *Nature genetics* 44.8 (Aug. 2012), pp. 955–9.

[183] YUN JU SUNG, LIHUA WANG, TUOMO RANKINEN, et al. “Performance of genotype imputations using data from the 1000 Genomes Project.” In: *Human heredity* 73.1 (2012), pp. 18–25.

[184] LAURA J SCOTT, KAREN L MOHLKE, LORI L BONNYCASTLE, et al. “A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants.” In: *Science (New York, N.Y.)* 316.5829 (June 2007), pp. 1341–5.

[185] BRYAN N HOWIE, PETER DONNELLY, and JONATHAN MARCHINI. "A flexible and accurate genotype imputation method for the next generation of genome-wide association studies." In: *PLoS genetics* 5.6 (June 2009), e1000529.

[186] PHILIP E STUART, RAJAN P NAIR, EVA ELLINGHAUS, et al. "Genome-wide association analysis identifies three psoriasis susceptibility loci." In: *Nature genetics* 42.11 (Nov. 2010), pp. 1000–4.

[187] HAE-WON UH, JORIS DEELEN, MARIAN BEEKMAN, et al. "How to deal with the early GWAS data when imputing and combining different arrays is necessary." In: *European journal of human genetics : EJHG* 20.5 (May 2012), pp. 572–6.

[188] JIYOUNG AHN, KAI YU, RACHAEL STOLZENBERG-SOLOMON, et al. "Genome-wide association study of circulating vitamin D levels". In: *Human Molecular Genetics* 19.13 (2010), pp. 2739–2745.

[189] JOSHUA C BIS, MARYAM KAVOUSI, NORA FRANCESCHINI, et al. "Meta-analysis of genome-wide association studies from the CHARGE consortium identifies common variants associated with carotid intima media thickness and plaque". In: *Nature Genetics* 43.10 (2011), pp. 940–947.

[190] YU CHING CHENG, WEN HONG L KAO, BRAXTON D. MITCHELL, et al. "Genome-wide association Scan Identifies variants near matrix metalloproteinase (MMP) genes on chromosome 11q21-22 strongly associated with serum MMP-1 levels". In: *Circulation: Cardiovascular Genetics* 2.4 (2009), pp. 329–337.

[191] AUDREY Y CHU, FRANCO GUILIANINI, HARALD GRALLERT, et al. "Genome-wide association study evaluating lipoprotein-associated phospholipase A2 mass and activity at baseline and after rosuvastatin therapy." In: *Circulation. Cardiovascular genetics* 5.6 (Dec. 2012), pp. 676–85.

[192] MARILYN C. CORNELIS, KERI L. MONDA, KAI YU, et al. "Genome-wide meta-analysis identifies regions on 7p21 (AHR) and 15q24 (CYP1A2) as determinants of habitual caffeine consumption". In: *PLoS Genetics* 7.4 (2011).

[193] JACQUELINE S DANIK, GUILLAUME PAR, DANIEL I CHASMAN, et al. "Novel loci, including those related to Crohn disease, psoriasis, and inflammation, identified in a genome-wide association study of fibrinogen in 17 686 women: the Women's Genome Health Study." In: *Circulation. Cardiovascular genetics* 2.2 (Apr. 2009), pp. 134–41.

[194] ZARI DASTANI, MARIE-FRANCE HIVERT, NICHOLAS TIMPSON, et al. "Novel loci for adiponectin levels and their influence on type 2 diabetes and metabolic traits: a multi-ethnic meta-analysis of 45,891 individuals." In: *PLoS genetics* 8.3 (Jan. 2012), e1002607.

[195] ABBAS DEHGHAN, JOSE DUPUIS, MAJA BARBALIC, et al. "Meta-analysis of genome-wide association studies in ¿80 000 subjects identifies multiple loci for C-reactive protein levels." In: *Circulation* 123.7 (Feb. 2011), pp. 731–8.

[196] M. FABIOLA DEL GRECO, CRISTIAN PATTARO, ANDREAS LUCHNER, et al. "Genome-wide association analysis and fine mapping of NT-proBNP level provide novel insight into the role of the MTHFR-CLCN6-NPPA-NPPB gene cluster". In: *Human Molecular Genetics* 20.8 (2011), pp. 1660–1671.

[197] GEORG B EHRET, PATRICIA B MUNROE, KENNETH M RICE, et al. "Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk." In: *Nature* 478.7367 (Oct. 2011), pp. 103–9.

[198] STEVE EYRE, JOHN BOWES, DOROTHE DIOGO, et al. "High-density genetic mapping identifies new susceptibility loci for rheumatoid arthritis." In: *Nature genetics* 44.12 (2012), pp. 1336–40.

[199] ROBERT R GRAHAM, CHRIS COTSAPAS, LEELA DAVIES, et al. "Genetic variants near TNFAIP3 on 6q23 are associated with systemic lupus erythematosus." In: *Nature genetics* 40.9 (2008), pp. 1059–1061.

[200] JOHN B HARLEY, MARTA E ALARCN-RIQUELME, LINDSEY A CRISWELL, et al. "Genome-wide association scan in women with systemic lupus erythematosus identifies susceptibility variants in ITGAM, PXK, KIAA1542 and other loci." In: *Nature genetics* 40.2 (2008), pp. 204–210.

[201] GEOFFREY HOM, D PH, ROBERT R GRAHAM, et al. "Association of Systemic Lupus Erythematosus with". In: (2008).

[202] JIE HUANG, MARIA SABATER-LLEAL, FOLKERT W ASSELBERGS, et al. "Genome-wide association study for circulating levels of PAI-1 provides novel insights into its regulation." In: *Blood* 120.24 (Dec. 2012), pp. 4873–81.

[203] LUKE JOSTINS, STEPHAN RIPKE, RINSE K WEERSMA, et al. "of Inflammatory Bowel Disease". In: 491.7422 (2013), pp. 119–124.

[204] JASON Z LIU, FEDERICA TOZZI, DAWN M WATERWORTH, et al. "Meta-analysis and imputation refines the association of 15q25 with smoking quantity." In: *Nature genetics* 42.5 (2010), pp. 436–440.

[205] ANDREW P MORRIS, BENJAMIN F VOIGHT, TANYA M TESLOVICH, et al. "Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes." In: *Nature genetics* 44.9 (Sept. 2012), pp. 981–90.

[206] CHRISTOPHER J. ODONNELL, MARYAM KAVOUSI, ALBERT V. SMITH, et al. "Genome-wide association study for coronary artery calcification with follow-up in myocardial infarction". In: *Circulation* 124.25 (2011), pp. 2855–2864.

[207] GUILLAUME PAR, DANIEL I. CHASMAN, ALEXANDER N. PARKER, et al. "Novel associations of CPS1, MUT, NOX4, and DPEP1 with plasma Homocysteine in a healthy population a genome-wide evaluation of 13 974 participants in the women's genome health study". In: *Circulation: Cardiovascular Genetics* 2.2 (2009), pp. 142–150.

[208] NICHOLAS L SMITH, MING-HUEI CHEN, ABBAS DEHGHAN, et al. "Novel associations of multiple genetic loci with plasma levels of factor VII, factor VIII, and von Willebrand factor: The CHARGE (Cohorts for Heart and Aging Research in Genome Epidemiology) Consortium." In: *Circulation* 121.12 (Mar. 2010), pp. 1382–92.

[209] ELI A STAHL, SOUMYA RAYCHAUDHURI, ELAINE F REMMERS, et al. "Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci." In: *Nature genetics* 42.6 (2010), pp. 508–514.

[210] THORGEIR E THORGEIRSSON, DANIEL F GUDBJARTSSON, IDA SURAKKA, et al. "Sequence variants at CHRNB3-CHRNA6 and CYP2A6 affect smoking behavior." In: *Nature genetics* 42.5 (2010), pp. 448–453.

[211] MATTHEW TRAYLOR, MARTIN FARRALL, ELIZABETH G. HOLLIDAY, et al. "Genetic risk factors for ischaemic stroke and its subtypes (the METASTROKE Collaboration): A meta-analysis of genome-wide association studies". In: *The Lancet Neurology* 11.11 (2012), pp. 951–962.

[212] LOUISE V WAIN, GERMAINE C VERWOERT, PAUL F O'REILLY, et al. "Genome-wide association study identifies six new loci influencing pulse pressure and mean arterial pressure". In: *Nature Genetics* 43.10 (2011), pp. 1005–1011.

[213] CHRISTINA L WASSEL, LESLIE A LANGE, BRENDAN J KEATING, et al. "Association of genomic loci from a cardiovascular gene SNP array with fibrinogen levels in European Americans and African-Americans from six cohort studies: the Candidate Gene Association Resource (CARe)." In: *Blood* 117.1 (Jan. 2011), pp. 268–75.

[214] JIAN YANG, RUTH J F LOOS, JOSEPH E POWELL, et al. "FTO genotype is associated with phenotypic variability of body mass index." In: *Nature* 490.7419 (Oct. 2012), pp. 267–72.

[215] C GIEGER, B KHNEL, A RADHAKRISHNAN, et al. "New gene functions in megakaryopoiesis and platelet formation". In: *Nature* 480.7376 (2011), pp. 201–208.

[216] MICHAEL A NALLS, DAVID J COUPER, TOSHIKO TANAKA, et al. "Multiple loci are associated with white blood cell phenotypes." In: *PLoS genetics* 7.6 (June 2011), e1002113.

[217] PING AN, IVA MILJKOVIC, BHARAT THYAGARAJAN, et al. "Genome-wide association study identifies common loci influencing circulating glycated hemoglobin (HbA1c) levels in non-diabetic subjects: the Long Life Family Study (LLFS)." In: *Metabolism: clinical and experimental* 63.4 (Apr. 2014), pp. 461–8.

[218] NICOLE SORANZO, SERENA SANNA, ELEANOR WHEELER, et al. "Common variants at 10 genomic loci influence hemoglobin A(C) levels via glycemic and nonglycemic pathways." In: *Diabetes* 59.12 (Dec. 2010), pp. 3229–39.

[219] ANDREW D JOHNSON, LISA R YANEK, MING-HUEI CHEN, et al. "Genome-wide meta-analyses identifies seven loci associated with platelet aggregation in response to agonists." In: *Nature genetics* 42.7 (July 2010), pp. 608–13.

[220] WEIHONG TANG, SAONLI BASU, XIAOXIAO KONG, et al. "Genome-wide association study identifies novel loci for plasma levels of protein C: the ARIC study." In: *Blood* 116.23 (Dec. 2010), pp. 5032–6.

[221] ADRIENNE TIN, ELIZABETH COLANTUONI, ERIC BOERWINKLE, et al. "Using multiple measures for quantitative trait association analyses: application to estimated glomerular filtration rate." In: *Journal of human genetics* 58.7 (July 2013), pp. 461–6.

[222] MARCUS E KLEBER, ILKKA SEPPL, STEFAN PILZ, et al. "Genome-wide association study identifies 3 genomic loci significantly associated with serum levels of homoarginine: the AtheroRemo Consortium." In: *Circulation. Cardiovascular genetics* 6.5 (Oct. 2013), pp. 505–13.

[223] PATRICK SULEM, DANIEL F GUDBJARTSSON, G BRAGI WALTERS, et al. "Identification of low-frequency variants associated with gout and serum uric acid levels." In: *Nature genetics* 43.11 (Nov. 2011), pp. 1127–30.

[224] ILKKA SEPPL, MARCUS E. KLEBER, LEO PEKKA LYYTIKINEN, et al. "Genome-wide association study on dimethylarginines reveals novel AGXT2 variants associated with heart rate variability but not with overall mortality". In: *European Heart Journal* 35.8 (2014), pp. 524–530.

[225] YU FANG PEI, LEI ZHANG, YONGJUN LIU, et al. "Meta-analysis of genome-wide association data identifies novel susceptibility loci for obesity". In: *Human Molecular Genetics* 23.3 (2014), pp. 820–830.

[226] ANNA KTTGEN, CRISTIAN PATTARO, CARSTEN A BGER, et al. "New loci associated with kidney function and chronic kidney disease." In: *Nature genetics* 42.5 (2010), pp. 376–384.

[227] LINDA M POLFUS, RICHARD A GIBBS, and ERIC BOERWINKLE. "Coronary heart disease and genetic variants with low phospholipase A2 activity." In: *The New England journal of medicine* 372.3 (Jan. 2015), pp. 295–6.

[228] WOLFGANG KOENIG, DOROTHEE TWARDELLA, HERMANN BRENNER, et al. "Lipoprotein-associated phospholipase A2 predicts future cardiovascular events in patients with coronary heart disease independently of traditional risk factors, markers of inflammation, renal function, and hemodynamic stress." In: *Arteriosclerosis, thrombosis, and vascular biology* 26.7 (July 2006), pp. 1586–93.

[229] MICHELLE L O'DONOGHUE, EUGENE BRAUNWALD, HARVEY D WHITE, et al. "Effect of darapladib on major coronary events after an acute coronary syndrome: the SOLID-TIMI 52 randomized clinical trial." In: *JAMA* 312.10 (Sept. 2014), pp. 1006–15.

[230] HARVEY D WHITE. "Darapladib for Preventing Ischemic Events in Stable Coronary Heart Disease". In: *New England Journal of Medicine* (2014), p. 140330050005008.

[231] ALEXANDER O SPIEL, JAMES C GILBERT, and BERND JILMA. "von Willebrand factor in cardiovascular disease: focus on acute coronary syndromes." In: *Circulation* 117.11 (Mar. 2008), pp. 1449–59.

[232] A R FOLSOM, K K WU, W D ROSAMOND, et al. "Prospective study of hemostatic factors and incidence of coronary heart disease: the Atherosclerosis Risk in Communities (ARIC) Study." In: *Circulation* 96.4 (Aug. 1997), pp. 1102–8.

[233] JOSHUA C BIS, MARYAM KAVOUSI, NORA FRANCESCHINI, et al. "Meta-analysis of genome-wide association studies from the CHARGE consortium identifies common variants associated with carotid intima media thickness and plaque". In: *Nature Genetics* 43.10 (2011), pp. 940–947.

[234] EVELINE NESCH, CAROLINE DALE, TOM M PALMER, et al. "Adult height, coronary heart disease and stroke: a multi-locus Mendelian randomization meta-analysis." In: *International journal of epidemiology* 45.6 (Dec. 2016), pp. 1927–1937.

[235] RON DO, CRISTEN J WILLER, ELLEN M SCHMIDT, et al. "Common variants associated with plasma triglycerides and risk for coronary artery disease." In: *Nature genetics* 45.11 (2013), pp. 1345–52.

[236] MICHAEL V HOLMES, FOLKERT W ASSELBERGS, TOM M PALMER, et al. "Mendelian randomization of blood lipids for coronary heart disease." In: *European heart journal* (2014), pp. 1–27.

[237] JON WHITE, DANIEL I SWERDLOW, DAVID PREISS, et al. "Association of Lipid Fractions With Risks for Coronary Artery Disease and Diabetes." In: *JAMA cardiology* (Aug. 2016).

[238] BENJAMIN F VOIGHT, GINA M PELOSO, MARJU ORHO-MELANDER, et al. "Plasma HDL cholesterol and risk of myocardial infarction: a mendelian randomisation study." In: *Lancet* 380.9841 (Aug. 2012), pp. 572–80.

[239] STEPHEN BURGESS, DANIEL F FREITAG, HASSAN KHAN, et al. "Using Multivariable Mendelian Randomization to Disentangle the Causal Effects of Lipid Fractions". In: 9.10 (2014).

[240] JESSICA VAN SETTEN, IVANA IGUM, SONALI PECHLIVANIS, et al. "Serum lipid levels, body mass index, and their role in coronary artery calcification: a polygenic analysis." In: *Circulation. Cardiovascular genetics* 8.2 (Apr. 2015), pp. 327–33.

[241] STEPHEN BURGESS and SIMON G THOMPSON. "Multivariable Mendelian randomization: the use of pleiotropic genetic variants to estimate causal effects." In: *American journal of epidemiology* 181.4 (Feb. 2015), pp. 251–60.

[242] JACK BOWDEN, GEORGE DAVEY SMITH, and STEPHEN BURGESS. "Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression." In: *International journal of epidemiology* 44.2 (Apr. 2015), pp. 512–25.

[243] *World Health Organization. Media centre, Fact Sheets: Cardiovascular diseases (CVDs).*

[244] CHRISTOPHER J O'DONNELL and ROBERTO ELOSUA. "[Cardiovascular risk factors. Insights from Framingham Heart Study]." In: *Revista espaola de cardiologa* 61.3 (Mar. 2008), pp. 299–310.

[245] M E MARENBERG, N RISCH, L F BERKMAN, et al. "Genetic susceptibility to death from coronary heart disease in a study of twins." In: *The New England journal of medicine* 330.15 (Apr. 1994), pp. 1041–6.

[246] TERI A MANOLIO, FRANCIS S COLLINS, NANCY J COX, et al. "Finding the missing heritability of complex diseases." In: *Nature* 461.7265 (Oct. 2009), pp. 747–53.

[247] THOMAS R WEBB, JEANETTE ERDMANN, KATHLEEN E STIRRUPS, et al. "Systematic Evaluation of Pleiotropy Identifies 6 Further Loci Associated With Coronary Artery Disease." In: *Journal of the American College of Cardiology* 69.7 (Feb. 2017), pp. 823–836.

[248] RAJAT M GUPTA, JOSEPH HADAYA, ADITI TREHAN, et al. "A Genetic Variant Associated with Five Vascular Diseases Is a Distal Regulator of Endothelin-1 Gene Expression." In: *Cell* 170.3 (July 2017), 522–533.e15.

[249] DAVID D WATERS, JOHN R GUYTON, DAVID M HERRINGTON, et al. "Treating to New Targets (TNT) Study: does lowering low-density lipoprotein cholesterol levels below currently recommended guidelines yield incremental clinical benefit?" In: *The American Journal of Cardiology* 93.2 (Jan. 2004), pp. 154–158.

[250] HENNING JANSEN, NILESH J SAMANI, and HERIBERT SCHUNKERT. "Mendelian randomization studies in coronary artery disease." In: *European heart journal* 35.29 (Aug. 2014), pp. 1917–24.

[251] DERRICK A BENNETT and MICHAEL V HOLMES. "Mendelian randomisation in cardiovascular research: an introduction for clinicians." In: *Heart (British Cardiac Society)* 103.18 (Sept. 2017), pp. 1400–1407.

[252] E P A VAN IPEREN, G K HOVINGH, F W ASSELBERGS, et al. "Extending the use of GWAS data by combining data from different genetic platforms." In: *PloS one* 12.2 (2017), e0172082.

[253] *UK Biobank (2006). Protocol for a large-scale prospective epidemiological resource.*

[254] SALOME SCHOLTENS, NYNKE SMIDT, MORRIS A SWERTZ, et al. "Cohort Profile: LifeLines, a three-generation cohort study and biobank." In: *International journal of epidemiology* 44.4 (Aug. 2015), pp. 1172–80.

[255] JAN L TALMON, MAURITS G ROS', and DINK A LEGEMATE. "PSI: The Dutch Academic Infrastructure for shared biobanks for translational research." In: *Summit on translational bioinformatics* 2008 (Mar. 2008), pp. 110–4.

[256] KARIEN STRONKS, MARIEKE B SNIJDER, RON J G PETERS, et al. "Unravelling the impact of ethnicity on health in Europe: the HELIUS study." In: *BMC public health* 13 (Apr. 2013), p. 402.

[257] MYOCARDIAL INFARCTION GENETICS AND CARDIOGRAM EXOME CONSORTIA INVESTIGATORS, NATHAN O STITZIEL, KATHLEEN E STIRRUPS, et al. "Coding Variation in ANGPTL4, LPL, and SVEP1 and the Risk of Coronary Disease." In: *The New England journal of medicine* 374.12 (Mar. 2016), pp. 1134–44.

[258] JACQUELINE MACARTHUR, EMILY BOWLER, MARIA CEREZO, et al. "The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog)". In: *Nucleic Acids Research* 45.D1 (Jan. 2017), pp. D896–D901.

[259] MARK D WILKINSON, MICHEL DUMONTIER, I JSBRAND JAN AALBERSBERG, et al. "The FAIR Guiding Principles for scientific data management and stewardship." In: *Scientific data* 3 (Mar. 2016), p. 160018.