January 2014

A Rare View Of Coding Mutations And Plasma Lipid Levels

Aniruddh Pradip Patel
Yale School of Medicine, aniruddh.patel@yale.edu

A Rare View of Coding Mutations and Plasma Lipid Levels

A Thesis Submitted to the
Yale University School of Medicine
in Partial Fulfillment of the Requirements for the
Degree of Doctor of Medicine

by
Aniruddh Pradip Patel
2014

A RARE VIEW OF CODING MUTATIONS AND PLASMA LIPID LEVELS. Aniruddh P. Patel, Sekar Kathiresan. Center for Human Genetics Research, Massachusetts General Hospital, Harvard Medical School, Boston, MA and Program in Medical and Population Genetics, the Broad Institute of Harvard and MIT, Cambridge, MA (Sponsored by Richard P. Lifton, Department of Genetics, Yale University School of Medicine, New Haven, CT).

Plasma low-density lipoprotein cholesterol (LDL-C), high-density lipoprotein cholesterol (HDL-C), and triglycerides (TG) are quantitative, heritable risk factors for coronary heart disease. Genome-wide association screens (GWAS) of common DNA sequence variants have identified many loci associated with plasma lipid levels. Targeted re-sequencing of exons has been proposed as a strategy to pinpoint causal variants and genes based in GWAS loci. Additionally, genotyping of rare and low frequency variants in large cohorts using an exome array has been proposed as a method to assess the contribution of rare variation to plasma lipid levels at the population level.

We tested the hypothesis that each genomic region identified with a significant HDL-C level association by GWA studies contains at least one gene causal for HDL-C metabolism. We performed solution-based hybrid selection of 4,118 exons at 407 genes within 47 loci associated with HDL-C and subsequently sequenced individuals drawn from the extremes of the HDL-C distribution (high HDL-C, n=385, mean=102 mg/dl or low HDL-C, n=334, mean=32 mg/dl) using next-generation sequencing technology. We tested whether rare coding sequence variants, individually or aggregated within a gene, were associated with HDL-C. To replicate findings, we performed follow-up genotyping using the Exome Array (Illumina HumanExome BeadChip) in independent participants with extremely high

HDL-C (n=514, mean=98 mg/dl) or low HDL-C (n=580, mean=32 mg/dl). Through sequencing, we identified 8,138 rare (minor allele frequency < 5%) missense, nonsense, or splice site variants. Across discovery sequencing and replication genotyping, we found 3 variants to be significantly associated with HDL-C. Of these, none were novel. In gene-level association analyses where rare variants within each gene are collapsed, only the *CETP* gene was associated with plasma HDL-C ($P=2.0 \times 10^{-6}$). After sequencing genes from GWAS loci in participants with extremely high or low HDL-C, we did not identify any new rare coding sequence variants with a strong effect on HDL-C. These results provide insight regarding the design of similar sequencing studies for cardiovascular traits with respect to sample size, follow-up, and analysis methodology.

We then tested the hypothesis that rare coding and splice-site mutations contribute to inter-individual variability in plasma lipid concentrations in the population. We contributed to the design of a new, rare-variant genotyping array based on the sequences of the protein-coding regions of ~18,500 genes ("the exome") in >12,000 individuals. This genotyping array ("the Exome Chip") includes approximately 250,000 non-synonymous and splice-site mutations and is estimated to capture nearly all such variation with a >1:1000 allele frequency in the European population. We obtained Exome Chip genotype data in >130,000 individuals from 58 studies. Within each study, we tested the association of plasma lipids with individual rare variants. To combine statistical evidence across studies, we performed meta-analysis. Top results for each trait replicated established associations in the genes *APOE*, *CETP*, and *APOA5* for LDL-C, HDL-C, and TG, respectively. We identified 11 new genes associated with plasma lipid levels: *ABCA6* with LDL-C (C1359R, frequency = 1:100, effect=+8.2 mg/dl, P=9.7 x 10^{-32}, *SERPINA* with LDL-C (E366K, frequency = 2:100, effect = +3.1 mg/dl, P=2.3 x 10^{-7}), *REST* with LDL-C (R645W, frequency = 6:10000, effect = +13.7 mg/dl, P=5.0 x 10^{-7}), *FBLN1* with LDL-C (H695R, frequency = 2:100, effect = -2.7 mg/dl, P=5.3 x 10^{-7}), *CCDC117* with LDL-C (T232I, frequency = 9:1000, effect = -4.3 mg/dl, P=7.3 x 10^{-7}), *TMED6* with HDL-C (F6L, frequency = 4:100, effect = -0.8 mg/dl, P=4.4 x 10^{-9}), *CDC25A* with HDL-C (Q24H, frequency = 3:100, effect = -1.0 mg/dl, P=8.4 x 10^{-8}), *MAP1A* (P2349L, frequency = 3:100) with HDL-C (effect= -1.4mg/dl, P=3.9 x 10^{-14}) and TG (effect=+8.4mg/dl, P=3.2 x 10^{-26}), *PRRC2A* with TG (S1219Y, frequency = 2:100, effect = +6.6 mg/dl, P=4.6 x 10^{-17}), *COL18A1* with TG (V125I, frequency = 1:1000, effect = +18.0 mg/dl, P=1.3 x 10^{-7}), and *EDEM3* with TG (P746S, frequency = 1:100, effect = -5.4 mg/dl, P=2.4 x 10^{-7}).

In addition, at some genes previously known to affect lipids, we identified new associations for variants: *APOC3* (R19Stop, frequency = 3:10,000) with HDL (effect=+11mg/dl, P=9.9 x 10^{-12}) and with TG (effect=-65.9mg/dl, P=5.8 x 10^{-23}); (splicesite IVS2+1 G>A, frequency = 2:1000) with HDL (effect=+10.6mg/dl, P=3.5 x 10^{-42}) and with TG (effect=-65.2mg/dl, P=2.0 x 10^{-81}). Using the Exome Chip rare variant genotyping array, we have discovered several new genes and variants associated with plasma lipids.

Acknowledgements

I would like to thank Dr. Sekar Kathiresan for his incredible mentorship and tireless support of all aspects of my academic development. I would also like to thank Gina Peloso for patiently teaching me bioinformatics and guiding me through these projects. I would like to thank the Kathiresan laboratory group members for making my research experience a pleasure.

I would like to thank Dr. Richard Lifton for advising me in this process and supporting my work over the past five years. Along with Dr. Lifton, I would also like to thank my other previous research mentors, Dr. Timothy Graubert and Ute Scholl, for cultivating my interest in genetics.

I would also like to thank Dr. John Forrest, Donna Carranzo and Mae Geter at the Office of Student Research for their guidance and encouragement.

Finally, I would like to thank the Stanley J. Sarnoff Cardiovascular Research Foundation, whose generous funding and support made this year possible. I would particularly like to thank Dr. Lauren Cohn for introducing me to the wonderful Sarnoff family and Dr. Agustin Melian for his advice throughout the research year.

Many individuals from the Kathiresan lab and from collaborating groups contributed to the work presented in this thesis through assistance in patient collection and characterization, exome array design, genotyping, sequencing, analysis, and meta-analysis. Part I of this thesis has been submitted for publication, and the ordered list of authors includes myself, Gina Peloso, James P. Pirruccello,

Daniel B. Larach, Namrata Gupta, Michael Boehnke, Gonçalo R. Abecasis, John J.P. Kastelein, G. Kees Hovingh, Robert A. Hegele, Daniel J. Rader, and Sekar Kathiresan. Part II will be submitted for publication with first authorship by Dajiang J. Liu, second authorship by myself and Hayato Tada, auxilliary authorship by members of the Global Lipids Genetics Consortium Exome Array Working Group, and corresponding authorship shared among Sekar Kahtiresan and other lead investigators.

This thesis was also funded by grants from the American Heart Association, the Donovan Family Foundation, Fondation Leducq, Massachusetts General Hospital, Merck, National Institutes of Health Grant R01HL107816, National Heart, Lung, and Blood Institute Grants 5RC1HL099793 and T32HL007208, and Pfizer.

Table of Contents

INTRODUCTION

Cardiovascular disease (CVD) is the leading cause of death in the United States and in the world.[1,2] In addition to being the cause of death for almost 600,000 individuals in the United States each year, CVD accounts for much of the nation's morbidity and health care system expenditure. In the US, 11.5% of non-institutionalized adults have been diagnosed with heart disease, and it is the primary diagnosis for over 12.4 million annual physician office visits and over 3.7 million hospital discharges, with the average length of inpatient stay being 4.6 days. Much of the mortality and morbidity from CVD stems from ischemic coronary heart disease (CHD), which includes angina pectoris, myocardial infarction, silent myocardial ischemia, and mortality resulting from coronary heart disease.[3] A deeper understanding of the causes of CHD will help guide prevention and treatment measures to decrease its burden of morbidty and mortality.

Risk Factors for Coronary Heart Disease

The Framingham Heart Study first identified common risk factors contributing to CHD after following the development of disease in over 10,000 individuals across two generations through their lifetimes. In the 1960's, the Study linked cigarette smoking, elevated cholesterol level, hypertension, electrocardiogram abnormalities, obesity, and physical inactivity with increased risk of heart disease.[4] In the 1970's the Study reported heart disease associations with diabetes, menopause, and psychosocial factors.

These risk factors were incorporated into the calculation of a Framingham Risk score adjusted for sex and age, which has been used by physicians to predict CHD risk in patients without disease.[5] Using the Framingham data, the lifetime risk of developing CHD calculated at age 40 was one in two for men and one in three for women in the US, with the risk decreasing to one in three for men and one in four for women when calculated at 70 years.[6] A recent meta-analysis of over 250,000 individuals confirmed that optimization of the burden of modifiable risk factors such as total cholesterol level, blood pressure, smoking, and diabetes resulted in a significant decrease in the lifetime risk of developing cardiovascular disease.[7]

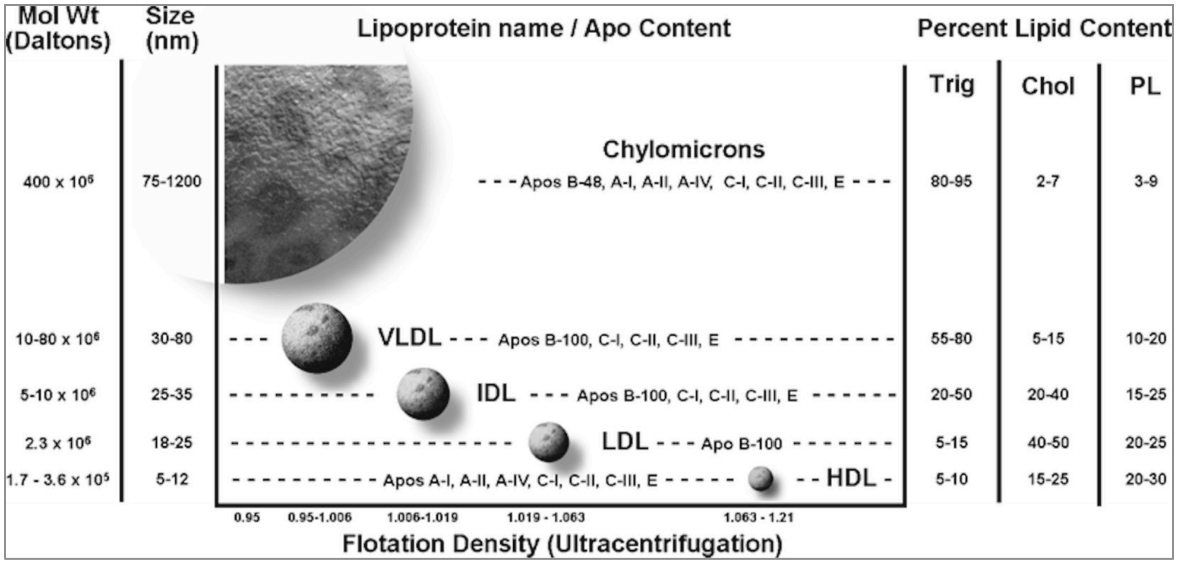
Of the largely lifestyle modifiable risk factors, diabetes, hypertension, and cholesterol levels are complex traits with a significant unmodifiable genetic component predisposing individuals to disease. In the past few decades, hundreds of genes and loci have been associated with these complex traits. Linkage and sequencing studies have identified numerous genes with rare variants involved in the pathogenesis of type 2 diabetes,[8-16] and a series of GWA studies have reported common variants influencing risk of developing the disease.[17-22] Consortia have also reported common variants with significant associations with hypertension.[23-25] Our group (laboratory of R.P. Lifton) has investigated and reported a number of genes involved in the pathogenesis of Mendelian hypertension syndromes.[26-35] Furthermore, our group (laboratory of S. Kathiresan) has reported several genes and variants contributing to variation of plasma lipid levels in families and in the population.

Role of Plasma Lipids in Coronary Heart Disease

Human plasma contains five major lipid subgroups differentiated by their density and apoprotein content: high density lipoprotein (HDL), low density lipoprotein (LDL), intermediate density lipoprotein (IDL), very low density lipoprotein (VLDL), and chylomicrons. Naturally hydrophobic plasma lipids are made soluble in plasma through encapsulation by lipoproteins that carry lipids to tissues to serve as fuel, structural components, and building blocks for steroid hormones and bile acids. The lipoprotein particle is composed of a shell of phospholipids, cholesterol, and apoproteins filled with triglycerides and cholesterol esters. Lipoprotein particles are classified by their density and apoprotein content, with LDL carrying mostly cholesterol and VLDL and chylomicrons carrying mostly triglycerides (**Figure 1**). Triglyceride measurements capture mainly chylomicrons and VLDL particles, LDL-C measurements capture LDL particles, and HDL-C measurements capture HDL particles.

Total cholesterol, LDL-C, HDL-C, and triglyceride levels are lab tests commonly ordered as part of a lipid panel for the screening and monitoring of coronary heart disease. Total cholesterol and HDL-C are measured using direct methods and are reliable regardless of whether the individual is fasting or not fasting. Triglyceride levels must be measured in the non-fasting state for baseline uniformity. The LDL-C level is calculated using the Friedewald equation where HDL-C and triglyceride components are removed from the total cholesterol value: $(LDL = TC - HDL - (TG/5))$.[36] This method of obtaining LDL-C is only reliable in individuals who are fasting and without other presentations of hypertriglyceridemia (TG>400 mg/dl).

Figure 1: Classification and Composition of Plasma Lipid Particles



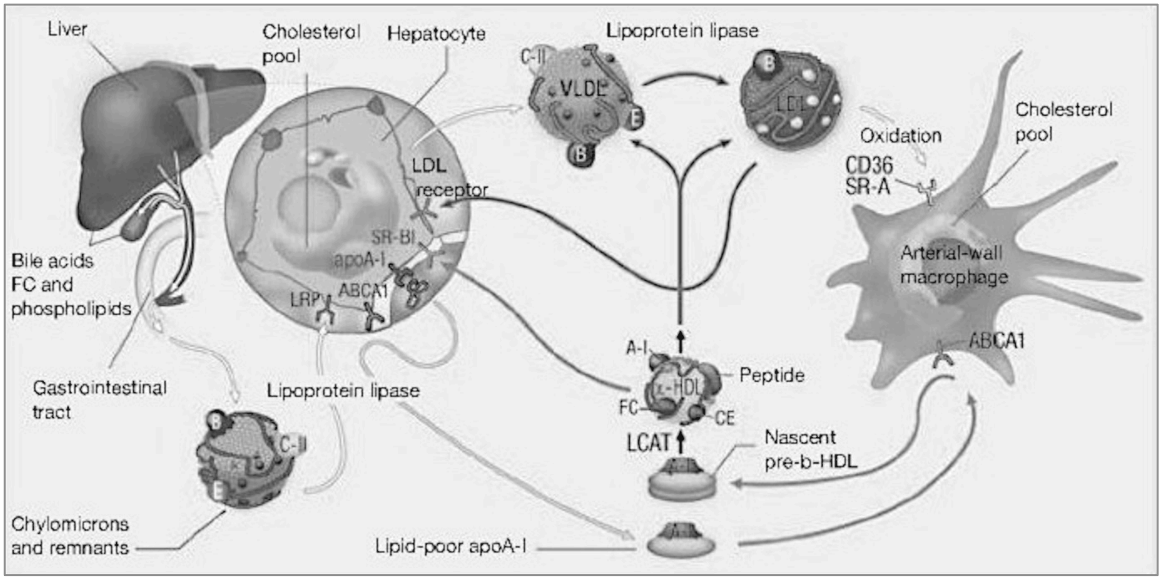
Mol Wt (Daltons)	Size (nm)	Lipoprotein name / Apo Content	Percent Lipid Content		
			Trig	Chol	PL
400×10^6	75-1200	Chylomicrons Apos B-48, A-I, A-II, A-IV, C-I, C-II, C-III, E	80-95	2-7	3-9
$10\text{-}80 \times 10^6$	30-80	VLDL Apos B-100, C-I, C-II, C-III, E	55-80	5-15	10-20
$5\text{-}10 \times 10^6$	25-35	IDL Apos B-100, C-I, C-II, C-III, E	20-50	20-40	15-25
2.3×10^6	18-25	LDL Apo B-100	5-15	40-50	20-25
$1.7 - 3.6 \times 10^5$	5-12	Apos A-I, A-II, A-IV, C-I, C-II, C-III, E HDL	5-10	15-25	20-30

Plasma lipid particles are classified based on their flotation density/ultracentrifugation. The physical properties, apolipoprotein content, and plasma lipid composition of the different particles are shown. Figure from Saland and Ginsberg, 2007.[37]

Lifestyle factors play a large role in determining plasma lipid levels. A diet high in cholesterol and saturated fats from animal products or hydrogenated oils raises plasma cholesterol and triglyceride levels, but diets with higher proportions of polyunsaturated fats lower total cholesterol levels. Physical activity and moderate amounts of meat and ethanol intake help raise HDL-C levels, but inactivity, smoking, and obesity are associated with decreased HDL-C levels.[38]

Dietary cholesterol and fatty acids are absorbed through the intestinal epithelium, where fatty acids are combined with glycerol to form triglycerides and cholesterol is esterified. These lipids are packaged into chylomicrons and transported to the tissues via the circulation. The liver also synthesizes triglycerides and cholesterol esters and assembles them into VLDL particles for secretion into the blood stream and delivery to the tissues. As VLDL particles are depleted of their triglyceride content, their density increases. These remnants are then cleared by the liver along with the chylomicron remnants, or they are remodeled into LDL particles.[39] Lipid metabolism plays a key role in the pathogenesis of CHD and atheroma formation and modification. (**Figure 2**)

Figure 2: Overview of Lipid Metabolism and Role in Atherosclerosis



Chylomicron particles generated in the gastrointestinal tract and VLDL and LDL particles generated in the liver contribute to atheroma formation in the arterial wall. HDL particles help transport cholesterol from the tissues to the liver. Figure adapted from Badimón and Ibáñez, 2010.[40]

Low-density lipoprotein cholesterol is primarily responsible for atherosclerosis. LDL-C is first transported into the artery wall though a concentration-dependent process which is accelerated in the setting of endothelial injury and hypercholesterolemia. The LDL-C retained in the vessel intima is then oxidized by free radical species generated by nearby endothelial cells and macrophages. Local macrophages phagocytize the oxidized LDL-C through the scavenger receptor and transform into foam cells storing massive amounts of oxidized lipids. Oxidized LDL-C stimulates an inflammatory response through the release of cytokines and chemokines by local cells that leads to monocyte and lymphocyte recruitment and vascular smooth muscle cell proliferation. This results in the development of a raised atheromatous plaque with a lipid core and fibrous cap which can obstruct coronary blood flow, weaken the vessel wall leading to aneurysm formation, and rupture to lead to thrombosis and myocardial infarction.[38]

In contrast to LDL-C, HDL-C has been shown to have a cardioprotective function. HDL is the primary mediator of reverse cholesterol transport, which removes cholesterol from peripheral tissues and brings it to the liver for biliary excretion. Furthermore, HDL has a number of non-cholesterol-mediated functions that may contribute to endothelial integrity and atheroprotection, such as anti-inflammatory effects, anti-oxidant effects, and anti-thrombotic effects.[41]

The role of triglyceride levels in coronary heart disease remains unclear because hypertriglyceridemia usually occurs in the setting of low HDL levels and increased LDL levels.[42] VLDL triglyceride particles enriched in apolipoprotein E or B have been shown to increase cholesterol uptake and oxidation in atheromas.[43] Furthermore, hypertriglyceridemia is conducive to hypercoagulability due to increased blood viscosity.[44]

Although the initial epidemiological studies have shown that plasma lipid levels have an association with coronary heart disease and functional studies have postulated the general role of lipids in metabolism, these studies alone cannot distinguish if lipids play a pathologically causal role or are just markers of underlying disease. The gold standard for assessing such causality is through large-scale randomized trials. For LDL-C, results from several, large, randomized controlled trials of statin medications that lower plasma LDL-C levels and rates of myocardial infarction suggest that LDL-C plays a causal role in coronary heart disease.[45-49] Large randomized controlled trials using niacin and torcetrapib to raise HDL-C levels did not result in a significant reduction in coronary heart disease, suggesting that HDL-C does not play a causal role in the underlying pathophysiology. Randomized controlled trials using fish oils and fibrates to lower triglyceride levels had mixed results in altering the risk of coronary heart disease, and the causal role of triglycerides is difficult to interpret because the administered medications have significant effects on the other lipid fractions.[50,51]

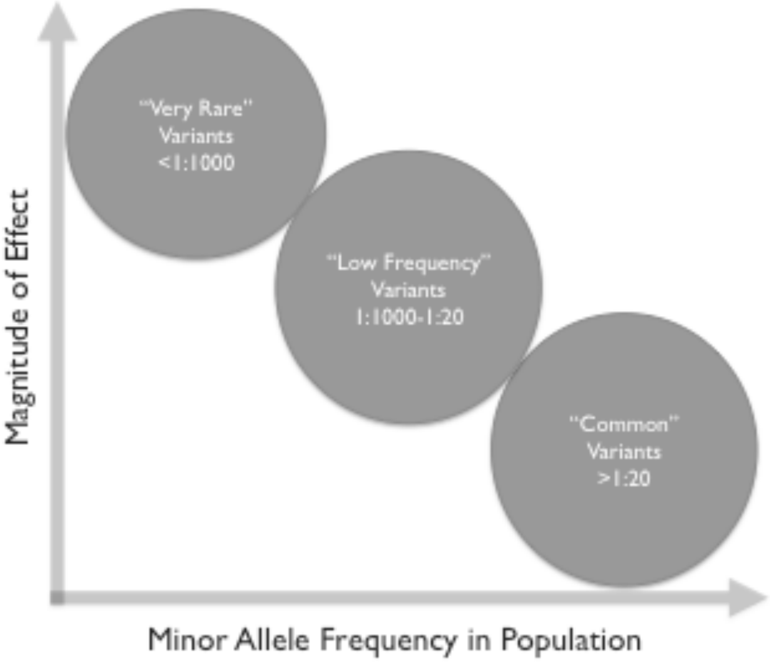
Analyzing lipid levels in the context inherited DNA variation may also be employed to distinguish causality of a biomarker using the theory of Mendelian randomization, which relies on the fact that genotypes are randomly assigned at meiosis and remain independent of non-genetic confounders or other disease processes.[52] In this sense, the assignment of genotypes at birth is analogous to the randomized, double-blinded administration of a medication in a clinical trial. Based on a representative set of significant SNPs tested using Mendelian randomization, our group recently reported evidence suggesting independent, causal roles of TG, in addition to LDL-C, in the pathogenesis of CHD.[53] We also reported similar evidence suggesting that some common genetic mechanisms that raise HDL-C do not contribute to CHD risk.[54]

Genetics of Plasma Lipid Levels

Many groups have determined that a large portion of the variation that we see in plasma lipid levels at the population level is heritable, or explained by inherited genetic differences.[55] Heritability estimates of lipid levels based on the Framingham data were calculated to be 66% for LDL-C, 69% for HDL-C, and 58% for TG.[56] The systematic study of these genetic differences can help reveal causal biologic mechanisms of lipid metabolism. Studies of linkage analysis and candidate genes were first used to identify genes implicated in lipid metabolism. The development of DNA sequencing technologies has helped fine map larger and larger genetic regions in individuals to discover variants influencing plasma lipid levels.[57] Furthermore, the completion of the Human Genome Project and the International Haplotype Map Project has further expanded investigative possibilities by allowing genome-wide screening of common variants to discover novel associations with plasma lipids.[58,59] These insights from genetics can be applied to the development novel therapeutics to treat at risk patients.[60]

Plasma lipid levels are quantitative traits with complex inheritance patterns in which multiple genes and non-genetic factors collude to influence the final phenotype. The genetic architecture underlying plasma lipids can be divided and examined by the minor allele frequency and effect sizes of significant variant associations (**Figure 3**). Very rare variants are those with a frequency of less than 0.1%, reflecting that fewer than 1 in 1000 people have the variant. These tend to have a larger effect on plasma lipid levels and may need to be grouped together to test their associations with lipid levels in aggregate. Low frequency variants are defined as having allele frequencies between 0.1% and 5%, or 1:1000-1:20 in the population. Common variants are those with a frequency of greater than 5%, or >1:20 in the population, and they tend to have smaller effects on lipid levels.

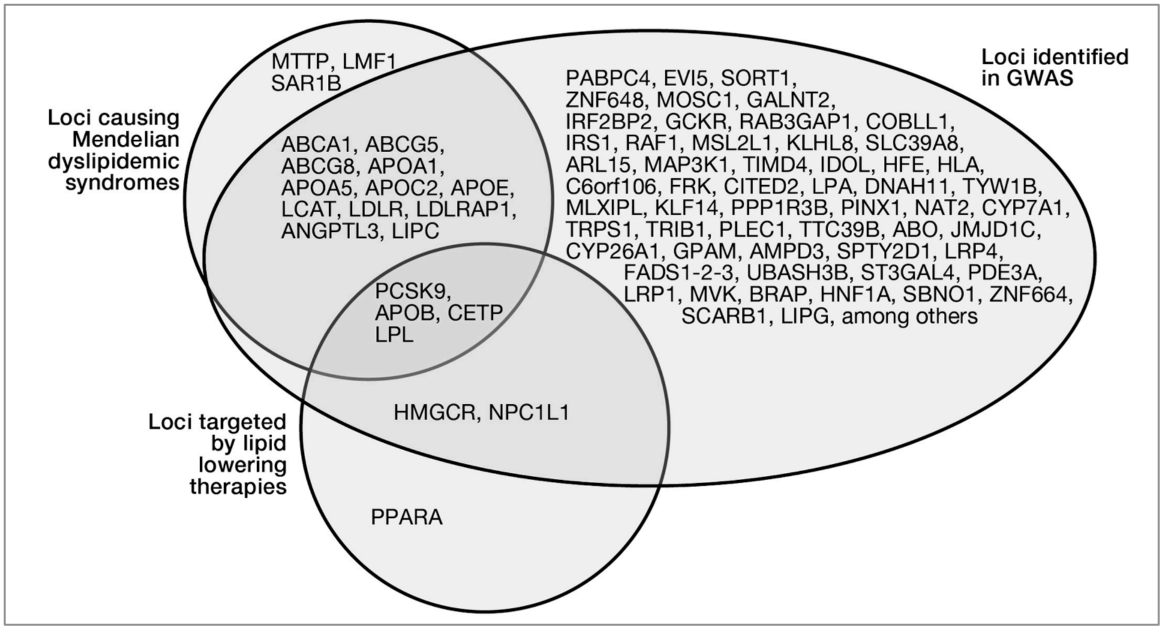
Figure 3: Genetic Architecture of Complex Traits



Variants are grouped as very rare (<1:1000), low frequency (1:1000-1:20), and common (>1:20) based on allele frequency, which tends to be inversely correlated with magnitude of effect. Adapted from Kathiresan and Srivastava[60]

Genome-wide linkage mapping in families with extreme lipid phenotypes has identified over a dozen genes with very rare mutations responsible for Mendelian dyslipidemias. Linkage and DNA sequencing analyses were first used to identify a large deletion in the LDL receptor in a patient with familial hypercholesterolemia.[61] Further study of families with hypercholesterolemia with linkage analysis and sequencing led to the identification of *PCSK9* (inducer of LDL receptor degredation in liver) and *APOB* (an LDL receptor ligand for uptake in tissues) as having casual roles in dyslipidemias.[62,63] Other genes discovered through similar methods include *ARH* (an adaptor protein for LDLR) in autosomal recessive hypercholesterolemia and *ABCG5/ABCG8* (ATP-binding cassette transporters involved in biliary sterol excretion) in sitosterolemia.[64,65] Recently, our lab used exome sequencing to discover that rare mutations in *ANGPTL3* (an inhibitor of lipoprotein lipase and hepatic lipase) are responsible for combined hypolipidemia in a family with markedly low levels of LDL-C, HDL-C, and TG in their plasma.[66] Many of the genes implicated in Mendelian dyslipidemias are in the vicinity of common variant associations, some of which are targets of lipid lowering drugs. (**Figure 4**).

Figure 4: Overlap of Genetic Loci Causing Mendelian Dyslipidemias, Loci Targeted by Lipid-Lowering Drugs, and Loci Identified in GWAS


MTTP, LMF1
SAR1B
Loci causing
Mendelian
dyslipidemic
syndromes
ABCA1, ABCG5,
ABCG8, APOA1,
APOA5, APOC2, APOE,
LCAT, LDLR, LDLRAP1,
ANGPTL3, LIPC
PCSK9,
APOB, CETP
LPL
Loci identified
in GWAS
PABPC4, EVI5, SORT1,
ZNF648, MOSC1, GALNT2,
IRF2BP2, GCKR, RAB3GAP1, COBLL1,
IRS1, RAF1, MSL2L1, KLHL8, SLC39A8,
ARL15, MAP3K1, TIMD4, IDOL, HFE, HLA,
C6orf106, FRK, CITED2, LPA, DNAH11, TYW1B,
MLXIPL, KLF14, PPP1R3B, PINX1, NAT2, CYP7A1,
TRPS1, TRIB1, PLEC1, TTC39B, ABO, JMJD1C,
CYP26A1, GPAM, AMPD3, SPTY2D1, LRP4,
FADS1-2-3, UBASH3B, ST3GAL4, PDE3A,
LRP1, MVK, BRAP, HNF1A, SBNO1, ZNF664,
SCARB1, LIPG, among others
Loci targeted
by lipid
lowering
therapies
HMGCR, NPC1L1
PPARA

GWAS loci are named for a plausible candidate gene in the proximity of the variant with the top association. Many genes identified as causal in Mendelian dyslipidemias are near GWAS loci. Medications currently target several of these genes. Figure from Kathiresan and Srivastava, 2012.[60]

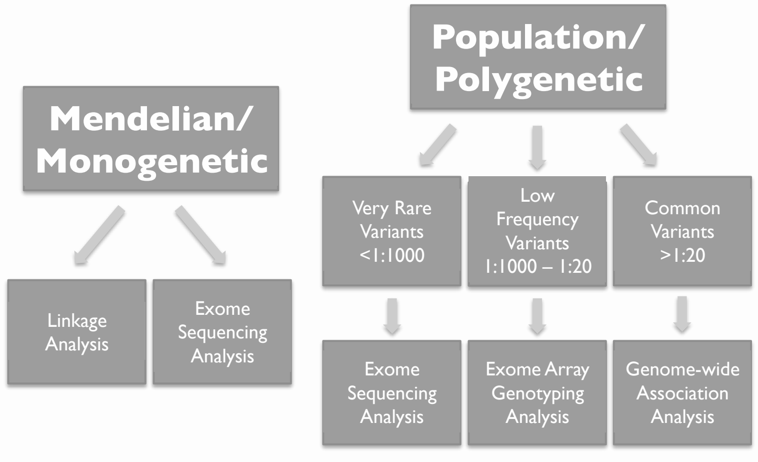
Genome-wide association (GWA) studies have been used to identify common variants associated with the plasma lipid levels through the efficient genotyping of hundreds of thousands of common single nucleotide polymorphisms (SNPs) distributed throughout the genome. Significant associations detected through GWA studies implicate not just the variant with the strongest signal, but also several variants in the surrounding locus found to be in linkage disequilibrium with the index signal. Our lab has helped lead a concerted effort in using genome-wide association (GWA) mapping in populations to define many new loci related to plasma lipids, with the most recent iteration cataloging common variants at 157 loci with genome-wide significant associations.[67-71] Furthermore, our group has reported that one of these common, noncoding GWAS-implicated variants at the chromosome 1p13 locus is involved in a novel regulatory pathway where altered transcription factor binding leads to altered expression of *SORT1* (sortilin 1, involved in pre-secretory degradation of VLDL-C) and altered risk for CHD.[72]

Despite this progress in linkage analysis, sequencing technologies, and genotyping capabilities, the identified alleles with significant associations explain only a modest fraction of the overall variance in plasma lipid concentrations. Moreover the common variant associations identified by GWA studies only implicate a collection of variants in a general locus, and the true causes of the GWA signals are unknown. For loci mapped using GWAS, we need to move from locus to gene and pinpoint the specific causal gene and causal variant responsible for the association. Furthermore, although family studies aid in the investigation of Mendelian syndromes and GWA studies help uncover common variants with strong associations, low frequency variants with allele frequencies of 1:1000-1:20 in population are difficult to discover with the methods mentioned so far. The contribution of rare and low frequency genetic variation at the population level is unknown. Within the bounds of the exome, we need to interrogate the significance of low frequency and rare variants in large cohorts to discover new genes implicated in lipid metabolism and coronary heart disease.

STRUCTURE OF THESIS

This thesis focuses on the role of rare variants and plasma lipid levels. Genetic architecture generally determines the method of investigation. Mendelian syndromes lend themselves to linkage and exome sequencing analysis in the context of pedigrees with well-developed phenotypes. Common diseases and traits such as plasma lipid levels tend to have polygenic contributions, and the minor allele frequency range of interest determines method of inquiry (**Figure 5**).

Figure 5: Investigative Genetic Method by Trait Type and Frequency Range


Population/ Polygenetic
Mendelian/ Monogenetic
Very Rare Variants <1:1000
Low Frequency Variants 1:1000 – 1:20
Common Variants >1:20
Linkage Analysis
Exome Sequencing Analysis
Exome Sequencing Analysis
Exome Array Genotyping Analysis
Genome-wide Association Analysis

The investigative genetic method is determined by trait type and allele frequency of interest.

This thesis is divided into two chapters that cover two main research projects completed over the past 20 months. The first chapter focuses on targeted exome sequencing of previously identified HDL-C GWAS loci to identify rare coding variants contributing to the initial association signals. The second chapter focuses on an exome array meta-analysis of 130,000 individuals and reports the discovery of novel genes with rare and low frequency variants associated with plasma lipids.

Part I: Targeted Sequencing of GWAS Loci in Extremes of the High-density Lipoprotein Cholesterol Distribution

BACKGROUND

High-density lipoprotein cholesterol (HDL-C) is a highly heritable risk factor for coronary heart disease.[4] Genome-wide association (GWA) studies have identified many new single nucleotide polymorphisms (SNP) related to HDL-C in the population.[17,67,68,70,71,73-75] Most associated SNPs are non-coding (intergenic or intronic) and fall in regions of linkage disequilibrium extending tens of thousands of bases and containing many protein-coding genes. Thus, a major challenge at each GWAS locus is to identify the culprit gene and variant responsible for the association signal.

One approach to pinpoint causal genes and variants at GWAS loci is to perform fine mapping through targeted sequencing. Sequencing may identify a protein-altering variant in a gene, which if associated with HDL-C would suggest that the gene is influencing HDL-C variation. The discovery of rare nonsense alleles that affect a trait may be particularly informative. Targeted sequencing of GWAS loci has been used to pinpoint independent rare variants and causal genes for diabetes mellitus,[76] fetal hemoglobin,[77] age-related macular degeneration,[78-82] and Crohn's disease.[83]

HYPOTHESIS AND SPECIFIC AIMS

We hypothesized that each genomic region identified with a significant HDL-C level association by GWA studies contains at least one gene causal for HDL-C metabolism. Our specific aims were as follows:

1). Identified 47 GWAS loci for HDL-C, targeted all exons at 407 genes within these genomic regions for sequencing, and sequenced targeted regions in individuals with extremely high or low HDL-C.

2). Attempted replication in an independent sample through array-based genotyping.

3). Assessed for novel, coding variants and genes with large effect in GWAS loci associated with HDL-C to try to determine the functional HDL-C gene at each locus.

METHODS

Discovery Cohort Selection

Individuals of European ancestry who had an abnormally high or low HDL-C level (<35 mg/dl for women and <28 mg/dl for men or > 100mg/dl for women and >80 mg/dl men) were recruited at the University of Pennsylvania to participate in the study. Individuals with no history of liver disease or HIV and who were not pregnant, nursing, or taking hormone replacement therapy or niacin had ~40cc of blood drawn. Plasma lipid levels were measured, and whole genomic DNA was extracted from the blood of these individuals. Individuals with HDL-C levels greater than the 95th percentile were selected for targeted sequencing (n=389, mean HDL-C = 102 mg/dl). Healthy age and sex matched controls with plasma HDL-C levels < 25th percentile were also sequenced (n=387, mean HDL=32mg/dl) (**Table 1**).

Targeted Sequencing

We studied 47 loci previously mapped for HDL-C ($P < 5 \times 10^{-8}$) in a genome-wide association study (GWAS) meta-analysis involving >100,000 individuals by the Global Lipids Genetics Consortium.[70,71] For each locus, we selected all of the exons of genes within 300 kb from the lead GWAS SNP identified in that locus. This represented 4,118 exons at 407 genes. Solution-based hybrid selection was used to select exons.[84]

To amplify exons, target-specific oligonucleotides 170 bases in length were designed to cover the entire coding sequence (hybrid selection bait size: 262,873 bases). These 170-mers were flanked on both sides with universal primer sequence to allow for PCR amplification. A T7 promoter was added in a second round of PCR, and in vitro transcription in the presence of biotin-UTP was performed to generate single-stranded hybridization bait to capture targets of interest from the DNA sample. Genomic DNA from individuals was randomly sheared and ligated to Illumina sequencing adapters. The fragments of this sheared and ligated genomic DNA were PCR amplified for 12 cycles and hybridized with biotinylated RNA bait. The hybridized DNA was extracted and PCR amplified to generate 36-base sequencing reads off of the Illumina adaptor sequence at the ends of each fragment.

Next generation sequencing reactions were performed using Illumina Genome Analyzers. Base pairs were called and sequencing reads were aligned to the human genome reference GRCh37 (hg19). Sequencing metrics were calculated using the Picard data-processing pipeline with an output of Binary Alignment Map (BAM) files. The Genome Analysis Toolkit suite was used to genotype all variant sites, calculate initial quality control metrics, and filter based on these values to result in an output of Variant Call Format (VCF) files, which were used for further quality control and analysis.[85] Variants were annotated using SnpEFF.[86]

Discovery Cohort Quality Control

Samples that failed in any step of the solution hybrid selection component of the targeted sequencing process were excluded. Population clustering was assessed through multidimensional scaling using pruned common variants (>5% minor allele frequency) with high call rates and that were not in linkage disequilibrium. Outliers on a plot of the first two principal components generated from multidimensional scaling were excluded. Samples with high heterozygosity rates (number of heterozygote sites/number of variants per sample) were excluded as presumptively contaminated, and those with high singleton counts (> three interquartile range above the median) were excluded due to presumptive sequencing error. Variants with low mean depth (<8) and low call rate (<95%) were excluded.

Replication Cohort Selection

To replicate our findings, we selected samples from an additional unrelated 1,245 individuals with historically high and low HDL-C who had participated in other genetic studies related to lipids. Plasma lipid levels were measured and individuals with HDL-C levels greater than the 95th percentile (n = 580, mean HDL = 98 mg/dl) or below the 25th percentile (n = 514, mean HDL-C = 32 mg/dl) were selected for follow up (**Table 2**).

Exome Array Genotyping

Exome array is a genotyping chip designed to query rare variation in the European population. We contributed a small number of variants to the design of this array by depositing 128 variants from preliminary HDL-C targeted sequencing analyses with marginally significant (P<0.05) results for variants with a MAF <5% and low-frequency nonsynonymous variants contributing to significant (P<0.05) gene burden test results.

Exome array genotyping was performed at the Center for Applied Genomics at the Children's Hospital of Pennsylvania. Genotyping calls were generated using

Genome Studio and all samples had a call rate >98%. To improve genotype calling for rare variants on the exome array, uncalled sites were recalled using zCall, which has been described in detail elsewhere.[87] Briefly, zCall is a rare-variant caller that models the relationship between intensity profiles of common allele homozygote and heterozygote clusters at common sites, and then uses these models to assign rare genotypes marked as missing in an initial Genome Studio analysis.

Replication Cohort Quality Control

All samples had a high call rate (>98%). Six samples had heterozygosity rates three interquartile ranges above the median and were excluded. Population clustering was assessed through multidimensional scaling using pruned common variants (>5% minor allele frequency) that were not in linkage disequilibrium. No samples were excluded based on this metric. Individuals that were determined to have a high degree of relatedness through identity-by descent calculation (Pi-Hat > 0.4) were also removed from the analysis.

Statistical Analysis

Single variant association results were computed using adaptive permutations on a dichotomous phenotype of high and low levels HDL-C using Fisher's exact test. For the exome array genotyping cohort, clustering algorithms based on pruned genotypes were used to account for ancestry differences and to serve as a sensitivity analysis. Using a minor allele frequency cutoff of 5%, the C-alpha[88] and variable threshold[89] gene burden tests was used to identify significantly associated genes with a Bonferroni corrected P value based on the total number of genes sequenced at the same locus. The C-alpha test is a gene burden test that aggregates variants within a gene to identify if a mixture of non-neutral alleles (risk and/or protective allele) are present that result in a deviation from variance expected under binomial model.[90] With the variable threshold test, the allele frequency threshold on which variants are pooled within a gene is optimized and the pooled variants are assigned equal weight and directionality in burden to calculate their collective burden.[91] All single variant associations and gene-based associations with a P value < 0.05 were compared with respective association results in the genotyping replication population. All analyses were performed using R,[92] GATK,[93] PLINK,[94] PLINK/SEQ.[95]

RESULTS

Discovery Sequencing

Of the 776 individuals of European descent that underwent targeted sequencing, 731 individuals remained after quality control measures. Of this group, 719 individuals had HDL-C values distinctly either above the 95th percentile or below the 25th percentile for their age and sex. The final targeted sequencing association analysis was performed on 334 individuals with low HDL-C levels (mean HDL-C = 31.6 mg/dL) and 385 individuals with very high HDL-C levels (mean HDL-C = 101.8 mg/dL) (**Table 1**).

Of the 262,873 targeted bases, 76% were covered at greater than 30-fold coverage whereas 81% were covered at greater than 20-fold coverage. Across the 719 individuals, we identified 8,714 missense, nonsense, or splice site DNA sequence variants. Of these, 8,138 had a minor allele frequency <5%.

Table 1: Characteristics of participants who underwent targeted sequencing

Targeted Sequencing Cohort	**Low HDL-C (n=334)**	**High HDL-C (n=385)**
HDL-C (mg/dl)	31.6	101.8
LDL-C (mg/dl)	103.1	123.3
TG (mg/dl)	155.2	75.4
Age (years)	63	60
Female (%)	58%	60%
Body mass index (kg/m^2)	28.8	23.4
Type II Diabetes (%)	6.4%	5.2%

Mean phenotypic characteristics of individuals with low HDL-C (<25th percentile) and high HDL-C (>95th percentile) who underwent targeted sequencing. All individuals who underwent targeted sequencing were of European ancestry.

Single Variant Association Analysis from Sequence Data

We first tested the association of individual variants with plasma HDL-C. Quantile-quantile plots of the single variant association results show that most of the variants fall along the expected null distribution, indicating that the study is well calibrated. A small fraction of variants (n=122 coding variants) displayed nominal evidence for association (P<0.05). Of these, 9 were loss-of-function mutations (stop gained, frameshift, splicesite) and 113 were missense mutations. The variants with the lowest P values were in genes with well-characterized roles in HDL metabolism including *CETP*, *ABCA1*, and *APOA1*. The rare, nonsense variant with the strongest association evidence was in the *PPP1R15A* gene (E118X, 0.4% frequency, OR for high HDL-C of 9.81, P= 0.04). The rare, missense or splice-site variant with the strongest association evidence was in the *CETP* gene (A330P, 3.3% frequency, OR for high HDL-C of 0.23, P=1.0 x 10^{-5}).

Gentoyping-based Replication of Single Variant Results

One proposed method for testing low-frequency and rare DNA variation is to first sequence to discover variation and then, subsequently, to genotype the discovered variants in a larger number of individuals to test for association with phenotype. Towards this end, we had contributed rare variants identified to be marginally significant in preliminary single variant and gene burden analyses from targeted sequencing to the design of the exome genotyping array. The final designed content of the array captures 38% of the variants with MAF < 5% from the HDL-targeted sequencing. We also evaluated the extent to which the array captured all of the low-frequency and rare variants discovered from the targeted sequencing. Of the 8,714 missense, nonsense, and splice-site variants discovered from the targeted sequencing of 719 individuals, 43% were present on the final designed content of the exome genotyping array.

Of the 1,250 individuals who underwent exome array genotyping, 1,228 remained after sample quality control measures. Of this group, 1,094 individuals had HDL-C values at above the 95th percentile or below the 25th percentile adjusted for age and sex. The final exome array genotyping association analysis was performed on 580 individuals with low HDL-C levels (mean HDL-C = 32.0 mg/dL) and 514 individuals with very high HDL-C level (mean HDL-C = 97.9 mg/dL) (**Table 2**).

In the single variant analysis of the exome array genotyping data, we found 3,638 coding variants to be marginally significant ($P<0.05$). Of these, 167 were loss of function mutations and 3,472 were missense mutations. Of the significant coding variants, 116 were in the vicinity of the 47 HDL-C loci. Quantile-quantile plots were well calibrated.

Of the 122 coding variants associated with HDL-C at nominal significance in the targeted sequencing single variant analysis, 31 were available for replication on the exome array data. Of these, 14 variants in the genes *ZNF259, APOA5, CCDC92, CETP, FBN3, SNX21, APOB, TBL2,* and *LPL* also showed a $P<0.05$ in the exome array association analysis.

Across discovery and replication, 3 variants associated with HDL-C after accounting for the 31 variants tested (threshold $P<1.6 \times 10^{-3}$): the A390P variant in *CETP* (4% frequency, OR for high HDL-C of 0.33, $P=2.0 \times 10^{-6}$, *CETP* locus), the S19W variant in *APOA5* (7% frequency, OR for high HDL-C of 1.78, $P=8.1 \times 10^{-4}$, *APOA1* locus), and the S474X variant in *LPL* (9% frequency, OR for high HDL-C of 1.66, $P=1.5 \times 10^{-3}$, *LPL* locus). All 3 variants have been previously studied.[96-98] (**Table 3**)

Table 2: Characteristics of Participants Who Underwent Exome Array Genotyping

Exome Array Genotyping Cohort	**Low HDL-C (n=580)**	**High HDL-C (n=514)**
HDL-C (mg/dl)	31.7	98.1
LDL-C (mg/dl)	103.4	121.3
TG (mg/dl)	343.6	78.6
Age (years)	49	57
Female (%)	58%	65%
Body mass index (kg/m^2)	32.6	23.5
Type II Diabetes (%)	26.1%	3.9%

Mean phenotypic characteristics of individuals with low HDL-C (<25th percentile) and high HDL-C (>95th percentile) who underwent exome array genotyping.

Table 3: Top Single Variant Association Results Including Discovery and Replication

Gene	Position	Protein	MAF	OR	P Value Targeted	P Value Exome Array	Locus
CETP	16:57015091	ALA390PRO	0.04	0.33	6.8 x 10^{-5}	2.0 x 10^{-6*}	*CETP*
APOA5	11:116662407	SER19TRP	0.07	1.78	4.6 x 10^{-2}	8.1 x 10^{-4*}	*APOA1*
LPL	8:19819724	SER474stop	0.09	1.66	3.0 x 10^{-3}	1.5 x 10^{-3*}	*LPL*
ZNF259	11:116655600	ALA264VAL	0.06	0.58	3.6 x 10^{-2}	2.9 x 10^{-3}	*APOA1*
CETP	16:57016092	VAL422ILE	0.36	0.76	3.3 x 10^{-4}	5.0 x 10^{-3}	*CETP*
APOB	2:21231524	PRO2739LEU	0.22	1.31	2.6 x 10^{-2}	9.0 x 10^{-3}	*APOB*
CCDC92	12:124427306	SER70CYS	0.33	0.79	8.0 x 10^{-3}	1.2 x 10^{-2}	*ZNF664*
SNX21	20:44469698	ALA290THR	0.003	5.66	3.8 x 10^{-2}	1.4 x 10^{-2}	*PLTP*
TBL2	7:72985148	VAL345ILE	0.05	0.61	2.7 x 10^{-2}	1.8 x 10^{-2}	*MLXIPL*
FBN3	19:8176945	SER1293GLY	0.18	1.33	3.8 x 10^{-2}	2.1 x 10^{-2}	*ANGPTL4*

Association results from exome array replication for variants were found to have P value < 0.05 in targeted sequencing single variant analysis and exome array genotyping single variant analysis; *indicates association significant after Bonferroni correction for number of variants tested; REF=Reference allele, ALT=Alternate allele, MAF=minor allele frequency, OR=odds ratio, Locus=gene name assigned to HDL-C GWA SNP[70]

Gene-level Association Analysis Using Sequence and Genotype Data

As the majority of the queried variants are rare, we collapsed the missense, nonsense, and splice site mutations in each gene and performed gene burden testing. The analysis of the targeted sequencing data using the variable threshold test found 12 genes at nominal significance ($P<0.05$) and one gene, *SPDEF* ($P=8.0 \times 10^{-4}$), to achieve significance after Bonferroni correction for number of genes tested within respective HDL-C GWAS locus. Performing the variable threshold test in the exome array genotyping data found 19 genes to be marginally significant ($P<0.05$) and three genes to achieve significance after Bonferroni correction for the number of genes tested within their respective HDL-C loci (*LILRB2* ($P=8.6 \times 10^{-5}$), *LIPG* ($P=3.0 \times 10^{-3}$), and *SCARB1* ($P=1.0 \times 10^{-3}$)). Marginally significant associations in *CETP* (15 pooled variants, $P = 1.1 \times 10^{-2}$), *LIPG* (14 pooled variants, $P=3.0 \times 10^{-3}$), and *NCOR2* (52 pooled variants, $P=3.4 \times 10^{-2}$) identified using the variable threshold test in the targeted sequencing data were replicated in the exome array genotyping cohorts (**Table 4**).

To accommodate genes where variants could have both gain-of-function and loss-of-function variants affects HDL-C, we performed the c-alpha test using nonsense, missense, and splice site variants with a minor allele frequency lower than 5%. The analysis of the targeted sequence data using the c-alpha test showed that 11 genes were marginally significant ($P<0.05$), and 2 genes reached significance after Bonferroni correction for the number of genes tested within their respective HDL-C loci: *ABCA1* (1.4×10^{-4}) and *CETP* ($P=1.3 \times 10^{-5}$). Performing the c-alpha test on the exome array genotyping data using the same inclusion criteria showed that

19 genes were marginally significant and five genes reached significance after Bonferroni correction for the number of genes tested within their respective HDL-C loci: *CD40* (P=3.4 x 10^{-4}), *CETP* (P=2.0 x 10^{-6}), *LILRB2* (P=2.1 x 10^{-5}), *PLG* (P=8.9 x 10^{-5}), and *TCAP* (P=1.0 x 10^{-3}). The marginally significant association identified in *CETP* (P=2.0 x 10^{-6}) using the c-alpha test in targeted sequencing was replicated in the exome array genotyping analysis (**Table 4**).

Table 4: Top gene-level association results after discovery and replication

Position	Gene	Test	P value from targeted sequencing	P value from exome array
chr16:56995908..57017292	*CETP*	CALPHA/VT	1.3 x 10^{-5}/ 2.0 x 10^{-2}	2.0 x 10^{-6} /1.1 x 10^{-2}
chr18:47088754..47109955	*LIPG*	VT	1.6 x 10^{-2}	3.0 x 10^{-3}
chr12:124810093..124979749	*NCOR2*	VT	1.9 x 10^{-2}	3.4 x 10^{-2}

Association results from exome array replication for pooling of variants led to gene-based associations with P <0.05 in targeted sequencing burden analysis and exome array genotyping burden analysis using C-alpha and variable threshold tests (VT).

DISCUSSION

We identified 47 GWAS loci for HDL-C, targeted all exons at 407 genes within these genomic regions, sequenced individuals with extremely high or low HDL-C, and attempted replication in an independent sample through array-based genotyping. After performing single variant and gene-burden analyses across discovery and replication cohorts, we did not identify any new rare coding sequence variants or genes with a large effect on plasma HDL-C levels.

This study was successful in replicating known genes with previously defined associations with plasma lipid levels. The functions of *CETP, APOA5*, *LPL, APOB,* and *LIPG* in HDL-C metabolism have been well established.[63,97-100] Although the functional role of the CCD92 and ZNF259 genes remain unclear, these loci have previously been associated with plasma lipoprotein size, concentration, and cholesterol content.[73]

The study permits several conclusions. First, since we were unable pinpoint specific coding variants responsible for the genome-wide association signals for HDL-C, we can speculate that the GWAS association signals may be truly due to non-coding, regulatory variants. Of the 47 total HDL GWAS loci that were fine mapped, 36 loci (77%) remain without any marginally significant single coding variant or gene-based association. Only the S70C variant in *CCDC92* and the P2739L variant in *APOB* were found to have identical minor allele frequencies and similar effect size estimates as the non-coding variants in their respective GWAS loci (*APOB* locus: rs1042034, frequency = 22%, effect = +4.16 mg/dL, P = 4.08 x 10^{-96} with *APOB* variant P2739L, frequency = 22%, OR = 1.31, P = 2.6 x 10^{-2} and *ZNF664* locus: rs4765127, frequency = 34%, effect = +0.44 mg/dL, P = 2.89 x 10^{-10} with *CCDC92* variant S70C, frequency = 36%, OR = 0.79, P = 8.0 x 10^{-3}).[70] This suggests that for these two loci, the identified coding variants may be responsible for the initial common noncoding variant GWAS association. For the remaining loci, intronic or intergenic SNPs in vicinity of the HDL-C GWAS loci may be involved in regulation and expression of coding genes involved in lipid metabolism.[101]

Secondly, targeted sequencing may have limited utility in discovering functional causes of GWAS signals. The absence of rare coding variants of large effect in GWAS loci is consistent with reports from other groups who have performed targeted sequencing-based variant discovery and genotyping-based replication studies to investigate variants in GWAS loci for autoimmune diseases with larger sample sizes.[102] Although targeted sequencing has previously been used to identify a few genes implicated in various diseases, hundreds of GWAS loci have collectively been fine mapped in the course of these studies, and the functional significance of the association signal at the vast majority of these loci remains unresolved.[76-78,83] Therefore revisiting and systematically studying the initially discovered non-coding variants in the implicated loci will be necessary to better understand the biologic underpinnings of these associations.

Several limitations of the present study need to be considered. The collective sample size of 719 individuals may be too small to provide sufficient power to detect associations of rare alleles with more modest effect. The targeted sequencing analysis has 80% power to identify 0.7% frequency variants with odds ratio greater than 3.25, 1% frequency variants with odds ratio greater than 2.56, and 5%

frequency variants with an odds ratio of greater than 1.47 in the study population.[103] However, the total number of individuals in the study is similar to the sample sizes studied for the analysis of different traits by other groups who were able to implicate coding mutations to GWAS SNPS using fine mapping and replication.[76-78,83]

Furthermore, the replication study was performed using exome array genotyping rather than additional sequencing. As a result, we were unable to fully test the following: 1) rare variants discovered in the targeted sequencing but not present on the exome array: and 2) a model where a burden of multiple rare alleles exclusively contributes to association signal.

Although this study successfully identified common variants and genes previously implicated in HDL-C metabolism, we did not identify any new rare coding variants or genes with sufficiently significant, replicating associations in the 47 loci with genome-wide associations with HDL-C. Fine mapping of coding regions surrounding GWAS loci may have limited utility in the investigation of the cause of these association signals. Though the study may have been limited by power and its genotying-based replication, it suggests that noncoding variation may be playing a significant role in determining plasma HDL-C levels. These results provide insight regarding the design of similar sequencing studies for cardiovascular traits with respect to sample size, follow-up, and analysis methodology.

Part II: Plasma Lipids Association Study Including ~130,000 Individuals Genotyped Using the Exome Array

BACKGROUND

Plasma low-density lipoprotein cholesterol (LDL-C), high-density lipoprotein cholesterol (HDL-C), and triglycerides (TG) are quantitative, heritable risk factors for coronary heart disease.[3] Linkage and sequencing studies have identified a number of very rare variants in genes implicated in Mendelian dyslipidemias.[61-63,66] Genome-wide association (GWA) studies have identified many common single nucleotide polymorphisms (SNPs) associated with plasma lipid levels.[16,66,67,69,70,72-74] However, the impact of rare and low frequency variation on plasma lipid levels in the population remains largely unstudied due to logistical limitations of current methodology. Querying this range of variation may help provide potential novel targets for therapeutics for preventing and treating coronary heart disease.

A large cohort of samples is necessary to sufficiently power a study to detect rare associations, and exome sequencing on this scale is currently prohibitively expensive. Other groups have designed custom arrays for fine mapping and candidate gene genotyping of large cohorts of individuals with cardiovascular, metabolic, and immunologic diseases.[104-106] A consortium of groups studying a broad range of diseases has taken a similar approach with a broader stroke of attempting to capture all replicable rare and low frequency variants from the entire exome on an array based on the variation seen in exome sequencing of over 12,000 individuals. As this array makes the analysis of variants in tens of thousands of individuals considerably faster and less expensive, it lends itself to interrogating rare and low frequency variants in complex traits via large collaborative efforts.

HYPOTHESIS AND SPECIFIC AIMS

We hypothesized that using a novel genotyping array focusing on low-frequency and rare variants, we can discover new genes and variants associated with plasma lipid levels. Our specific aims were as follows:

1). Contributed to the to the design of the exome array.

2). Obtained plasma lipid level data and performed genotyping and data processing in all participating cohorts.

3). Analyzed data from each cohort using a common analysis plan.

4). Combined test statistics from participating studies and meta-analyzed results in single variant and gene based testing.

METHODS

Cohort Selection

The more than 130,000 samples that were genotyped and analyzed came from 58 cohorts of individuals from around the world. These cohorts consist of cases and controls for early onset myocardial infarction and type 2 diabetes mellitus, as well as members of the general population. The majority of these individuals report having European ancestry while a significant portion report South Asian and African ancestries.

Phenotype Modeling

Plasma lipid trait data was collected and modeled uniformly across all cohorts. Blood was collected, plasma lipid levels were measured, and DNA was extracted from each study participant's blood sample. Only individuals who had blood drawn while fasting were included in the analysis of LDL-C and TG. All individuals were included in the analysis of HDL-C, regardless of fasting status at the time of blood draw. Calculations involving LDL-C, HDL-C values were performed using units of mg/dl, and calculations involving TG were performed using the natural log transform of the TG level in mg/dl. The mean levels for each trait across the cohorts were representative of the general population (LDL-C: 161 mg/dl, HDL-C: 53 mg/dl, TG: 107 mg/dl, and TC: 239 mg/dl). For data collected after 1994, the plasma TC measurements of subjects that were known to be on lipid medication were adjusted (adjusted TC = TC / 0.8) with the assumption that they were prescribed statins following the publication of the 4S trial.[49] For data that was collected prior to 1994, no adjustments were made to account for lipid medication. LDL-C was calculated using the Friedewald formula for individuals with a triglyceride level < 400 mg/dl (LDL = TC – HDL – (TG/5)).[36] For individuals whose TC levels were modified due to lipid medication status, the modified TC was used in the formula. Regardless of medication status, no adjustments were made for HDL-C or TG.

For each lipid trait, residuals were calculated after adjusting for age, age^2, sex, and at least 4 principal components computed through multiple dimensional scaling. The inverse normal transformation of the trait residuals served as the phenotypes in the analysis.

Exome Array Design

Exome array is a genotyping chip designed to query rare variation down to 1:10,000 in the European population. The array includes coding variants seen in multiple existing sequence datasets of ~12,000 sequenced genomes and exomes of individuals of mainly European ancestry. It also contains all of the variants implicated in GWAS found in the NHGRI catalog and a small amount of insertion-deletions, micro RNA target sites, mitochondrial variants, and ancestry markers. A more detailed description of exome array SNP content and selection is available on the design website.[107]

Exome Array Genotyping

Exome array genotyping was performed at academic institutes around the world using Illumina HumanExome BeadChip arrays. All genotypes were called using zCall, a variant caller specifically designed for calling rare SNPs from the exome array.[87] This caller was implemented as a post-processing step after a default autocalling algorithm was applied. Linear regression was used to determine the relationship between the intensity profile of the common allele homozygote clusters for common sites, and an optimal z value was obtained, which was then used to recall the genoypes at each site.

Genotyping Quality Control

After genotyping, each cohort of samples underwent uniform and stringent quality control measures. Samples with an autocall rate <98% or a Z-call rate <99% were removed. Samples with a high heterozygosity rate (a measure of contamination) three interquartile ranges above the median were excluded. Samples exhibiting gender discordance were removed. Population clustering was assessed through multidimensional scaling using pruned common variants (>5% minor allele frequency) that were not in linkage disequilibrium, and outliers based on these metrics were excluded. Individuals that were determined to have a high degree of relatedness through identity-by descent calculation (Pi-Hat > 0.4) were also removed from the analysis.

Association Analysis

Each cohort was analyzed independently. For study groups ascertained on case or control status for early onset myocardial infarction or type 2 diabetes, the two sub-groups were modeled as separate studies. Each self-reported racial subgroup was also modeled as a separate study.

Single variant association was performed using a linear regression-based analysis of missense, nonsense, and splice variants with minor allele frequency >1%. The statistical software programs rvtests and RareMetalWorker were used to run the rare variant analyses and to generate summary statistics. The rvtests program was used for cohorts with all unrelated individuals. RareMetalWorker was used to analyze cohorts that included related individuals and generated an empirical kinship matrix based on the cohort's genotype data to account for hidden relatedness or substructures during the analysis. Additional computation was performed using R[92] and PLINK.[94] The summary level statistics generated by both programs included: allele frequency for each variant, single variant association score test statistics with direction of effect, covariance matrix for each genetic region, and metrics for assessing genotype qualities, including Hardy Weinberg equilibrium and call rate. The output files were uniformly prepared by each study and then aggregated and meta-analyzed at one site.

Gene-based association involving burden analysis of pooled rare variants within a gene was used to assess if rare mutations in aggregate associate with the lipid traits. Using summary statistics for each variant, genes were annotated based on entries in the refFlat/refGene database. Only non-synonymous variants, stop gain mutations, stop loss mutations, and splice site mutations with minor allele frequencies less than 5% were used in separate burden analyses. The variable threshold test assigns equal weight and directionality of effect to each variant,[91] while the SKAT (Sequence Kernel Association Tests) test better assesses the effects of causal variants with opposite effects present in a gene region.[108] Both analyses were performed for each cohort, and the results were meta-analyzed.

Meta-analysis

In order to power the study well, very large sample sizes from a collaboration of groups were needed. We gathered summary statistics generated using a common analysis plan from each of the 58 participating studies and combined single variant test statistics using the Mantel-Haenszel method.[109] We then subsequently meta-analyzed gene-level association tests for a total of >130,000 individuals using RAREMETAL.[110]

Conditional Analysis

As all variants on the exome array were included in the analysis without exclusions based on SNPs being in linkage disequilibrium, conditional analysis was necessary to identify the true, independent association signals. Using the covariance matrices generated during the analysis step for each cohort, each variant was conditioned on the SNP with the lowest P value within a 1 megabase window. No conditional analysis was performed for the SNPs with the strongest associations in a 1 megabase window block.

RESULTS

Single Variant Analysis From Exome Array Genotyping

After stringent quality control measures, over 130,000 samples were included in the final analysis of each lipid trait. The mean plasma lipid levels and phenotypic distributions for the combined cohorts were representative of the general population. All participating studies reported excellent genotyping data quality, and minimal samples were excluded during the quality control phase. Quantile-quantile plots for each lipid trait were well calibrated.

In the single variant analysis of the population exome array genotyping data, we identified over 230,000 polymorphic variants each for LDL, HDL, and TG. This includes 194,470 missense variants, 9,600 splicesite variants, 4,546 nonsense variants, 76 frame shift variants, and 30,894 non-coding variants. Of these, 27,827 or 12.1% were common mutations (>5:100 in the population), 48,437 or 21.1% were low frequency mutations (1:1000-5:100 in the population), and 153,719 or 66.8% were rare mutations (<1:1000 in the population). The Bonferroni cutoff for statistical significance was determined to be $P < 2 \times 10^{-7}$ for single variants, given the ~250,000 variants on the exome array available for analysis. After performing conditional analysis, we identified 79 variants associated with LDL, 133 variant associated with HDL, and 97 variants associated with TG that met Bonferroni cutoff of exome array wide significance. The majority of these findings represented GWAS variants that had been reported in the NHGRI catalog and placed on the array. We identified low frequency and rare variants for each trait (10 SNPs for LDL, 29 SNPs for HDL, 15 SNPs for TG) that remained statistically significant after conditional analysis.

The strongest associations for each trait replicated well-established associations, which serve as positive controls (**Table 5**). For LDL-C, variants in *APOE* (R176C, frequency = 7:100, effect=-22.1 mg/dl, P < 9 x 10^{-288}) and PCSK9 (R46L, frequency = 2:100, effect=-17.6 mg/dl, P = 1 x 10^{-190}) were the top coding mutation signals. For HDL-C, the top hits were in *CETP* (A390P, frequency = 4:100, effect = -4.0 mg/dl, P = 1 x 10^{-178}) and *LPL* (S474X, frequency = 1:10, effect = +2.5 mg/dl, P = 4 x 10^{-143}). For TG, the top hits were in *APOA5* (S19W, frequency = 6:100, effect = +14.9 mg/dl, P =8 x 10^{-184}) and *LPL* (S474X, 1:10, effect = 11.6 mg/dl, P = 1 x 10^{-170}). The roles of all of these genes have been well characterized in lipid metabolism.[62,97,98,111,112] All of these variants occur at low to common frequencies resulting in strong statistical signals.

Additionally, we identified 11 new genes and variants associated with plasma lipid levels (**Table 6**). For LDL-C, we identified a variant in *ABCA6* (C1359R, frequency = 1:100, effect=+8.2 mg/dl, P=9.7 x 10^{-32}), a member of the ATP-binding cassette family, associated with increased LDL-C levels. We also identified a variant in *SERPINA* (E366K, frequency = 2:100, effect = +3.1 mg/dl, P=2.3 x 10^{-7}), which is a serine protease inhibitor, and a variant in *REST* (R645W, frequency = 6:10000, effect = +13.7 mg/dl, P=5.0 x 10^{-7}), which is a RE1-silencing transcription factor, associated with increased LDL-C levels. Furthermore we identified a variant in *FBLN1* (H695R, frequency = 2:100, effect = -2.7 mg/dl, P=5.3 x 10^{-7}), which is tumor suppressor gene, and a variant in *CCDC117* (T232I, frequency = 9:1000, effect = -4.3 mg/dl, P=7.3 x 10^{-7}), which is a coiled coil domain, associated with decreased LDL-C.

For HDL, we identified a variant in *TMED6* (F6L, frequency = 4:100, effect = -0.8 mg/dl, P=4.4 x 10^{-9}), which is involved in transmembrane transport , and a variant in *CDC25A* (Q24H, frequency = 3:100, effect = -1.0 mg/dl, P=8.4 x 10^{-8}), which is a phosphatase involved in cell cycle regulation, associated with decreased HDL-C levels. Furthermore, we identified a variant in *MAP1A*, a microtubule-associated protein, (P2349L, frequency = 3:100) associated with decreased HDL-C levels (effect = -1.4mg/dl, P=3.9 x 10^{-14}) and increased triglyceride levels (effect = +8.4mg/dl, P=3.2 x 10^{-26}).

For TG we identified a variant in *PRRC2A* (S1219Y, frequency = 2:100, effect = +6.6 mg/dl, P=4.6 x 10^{-17}), which is a proline rich coiled-coil , and a variant in *COL18A1* (V125I, frequency = 1:1000, effect = +18.0 mg/dl, P=1.3 x 10^{-7}), which encodes a type of collagen, associated with increased triglyceride levels. We also identified a variant in *EDEM3* (P746S, frequency = 1:100, effect = -5.4 mg/dl, P=2.4 x 10^{-7}), which is an ER degredation enhancer, associated with decreased triglyceride levels.

Table 5: Top Exome Array Single Variant Associations

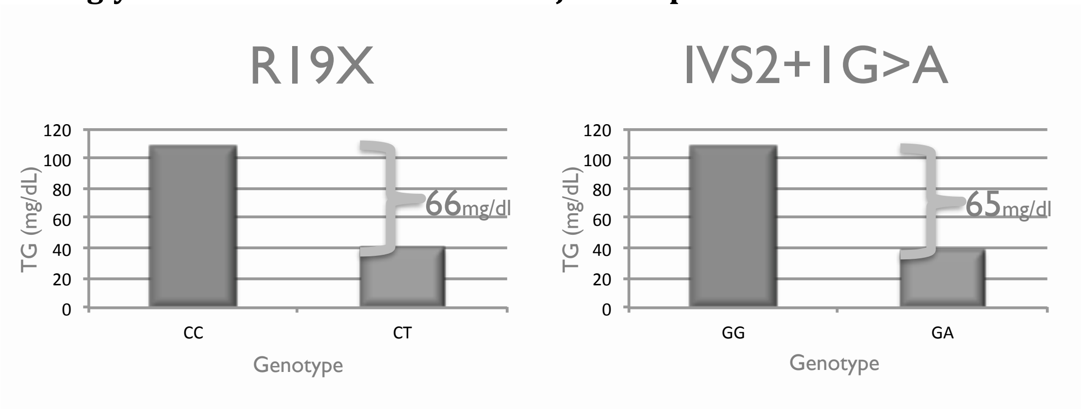
Trait	Chromsome: Position	Gene	Coding Change	P Value	Allele Frequency	Effect (mg/dl)
LDL	chr19:45412079	*APOE*	ARG176CYS	$<9.1 \times 10^{-288}$	7:100	-22.1
LDL	chr1:55505647	*PCSK9*	ARG46LEU	1.0×10^{-190}	2:100	-17.6
HDL	chr16:57015091	*CETP*	ALA390PRO	1.4×10^{-178}	4:100	-4.0
HDL	chr8:19819724	*LPL*	SER474stop	3.8×10^{-143}	1:10	+2.5
TG	chr11:116662407	*APOA5*	SER19TRP	7.7×10^{-184}	6:100	+14.9
TG	chr8:19819724	*LPL*	SER474stop	1.1×10^{-170}	1:10	-11.6

Table 6: Novel Exome Array Single Variant Associations

Trait	Chromosome: Position	Gene	Coding Change	P Value	Allele Frequency	Effect (mg/dl)
LDL	chr17:67081278	*ABCA6*	CYS1359ARG	9.7×10^{-32}	1:100	+8.2
LDL	chr14:94844947	*SERPINA1*	GLU366LYS	2.3×10^{-7}	2:100	+3.1
LDL	chr4:57796957	*REST*	ARG645TRP	5.0×10^{-7}	6:10000	+13.7
LDL	chr22:45996298	*FBLN1*	HIS695ARG	5.3×10^{-7}	2:100	-2.7
LDL	chr22:29182169	*CCDC117*	THR232ILE	7.3×10^{-7}	9:1000	-4.3
HDL	chr11:116701354	*APOC3*	IVS2+1 G>A	3.5×10^{-42}	1:1000	+10.6
HDL	chr11:116701353	*APOC3*	ARG19stop	9.9×10^{-12}	3:10000	+11.0
HDL	chr15:43820717	*MAP1A*	PRO2349LEU	3.9×10^{-14}	3:100	-1.4
HDL	chr16:69385641	*TMED6*	PHE6LEU	4.4×10^{-9}	4:100	-0.8
HDL	chr3:48229366	*CDC25A*	GLN24HIS	8.4×10^{-8}	3:100	-1.0
TG	chr11:116701353	*APOC3*	ARG19stop	5.8×10^{-23}	3:10000	-65.9
TG	chr11:116701354	*APOC3*	IVS2+1 G>A	2.0×10^{-81}	1:1000	-65.2
TG	chr15:43820717	*MAP1A*	PRO2349LEU	3.2×10^{-26}	3:100	+8.4
TG	chr6:31600106	*PRRC2A*	SER1219TYR	4.6×10^{-17}	2:100	+6.6
TG	chr21:46875817	*COL18A1*	VAL125ILE	1.3×10^{-7}	1:1000	+18.0
TG	chr1:184672098	*EDEM3*	PRO746SER	2.4×10^{-7}	1:100	-5.4

In addition to these novel associations with genes, we also discovered a new splice site variant in the gene *APOC3* (IVS2+1 G>A, frequency = 1:1000) that is significantly associated with increased HDL-C levels (effect=+10.6mg/dl, P=3.5 x 10^{-42}) and decreased TG levels (effect=-65.2mg/dl, P=2.0 x 10^{-81}). Furthermore, we report population level data for an established variant in *APOC3* (R19X, frequency = 3:10,000) that introduces a premature stop codon and truncates the protein leading to a loss of function. This variant has been reported to occur at a 5% frequency in the Lancaster Amish population and has been associated with lower serum TG, higher levels of HDL-C, lower levels of LDL-C, and lower levels of subclinical atherosclerosis as measured by coronary artery calcification.[113] In our population level data we found that the *APOC3* R19X variant is much rarer in the general population and that it is associated with increased HDL-C levels (effect=+11mg/dl, P=9.9 x 10^{-12}) and decreased triglyceride levels (effect=-65.9mg/dl, P=5.8 x 10^{-23}) (**Figure 6**). The majority of the novel associations occur with variants with low to very rare frequencies.

Figure 6: Triglyceride Levels in R19X and Adjacent Splice Mutation Carriers



Carriers of the adjacent R19X or IVS2+IG>A mutations in *APOC3* have dramatically decreased triglycerides.

Gene-level Association Analysis Using Exome Array Genotype Data

As the majority of the variants available for analysis on the exome array are rare, we pooled the missense, nonsense, and splice site mutations in each gene and performed gene burden testing. The analysis of exome array genotyping data using the variable threshold test identified 10 genes associated with LDL-C, 23 genes associated with HDL-C, and 13 genes associated with TG. The top associations were between LDL-C and *PCSK9* (P = 4.8 x 10^{-46}), HDL-C and *CETP* (2.8 x 10^{-73}), and TG and *APOC3* (P = 5.3 x 10^{-82}) (**Table 7**). The roles of all of these genes with the strongest associations have been well characterized in lipid metabolism.[62,97,114]

To account for opposite effects of variants within a gene region in calculating the burden results, we performed the SKAT test using nonsense, missense, and splice site variants with a minor allele frequency lower than 5%. The analysis of exome array genotyping data using the SKAT burden test identified 8 genes associated with LDL-C, 18 genes associated with HDL-C, and 7 genes associated with TG after accounting for the number of genes tested. The top associations were between LDL-C and *PCSK9* (P = 2.4 x 10^{-94}), HDL-C and *ANGPTL4* (6.7 x 10^{-66}), and TG and *ANGPTL4* (P = 2.1 x 10^{-73}) (**Table 8**). The roles of all of these genes with the strongest associations have also been well characterized in lipid metabolism.[62,115]

Table 7: Top Variable Threshold Gene Burden Associations

Trait	Gene	P Value	Effect Size	Variants
LDL	*PCSK9*	8.6×10^{-59}	-0.2	26
LDL	*LDLR*	3.1×10^{-12}	0.7	28
HDL	*CETP*	2.8×10^{-73}	-0.2	15
HDL	*LPL*	1.2×10^{-70}	-0.2	12
TG	*APOC3*	5.3×10^{-82}	-1.0	4
TG	*ANGPTL4*	2.2×10^{-47}	-0.2	11

Effect size reported in standard deviation units. Variants denote number of single variants in gene contributing to burden result.

Table 8: Top SKAT Gene Burden Associations

Trait	Gene	P Value	Effect Size	Variants
LDL	*PCSK9*	2.4×10^{-94}	-0.2	26
LDL	*APOE*	3.5×10^{-8}	0.1	4
HDL	*ANGPTL4*	6.7×10^{-66}	0.2	11
HDL	*LPL*	5.3×10^{-47}	-0.2	12
TG	*ANGPTL4*	2.1×10^{-73}	-0.2	11
TG	*APOC3*	8.1×10^{-68}	-1.0	4

Effect size reported in standard deviation units. Variants denote number of single variants in gene contributing to burden result.

DISCUSSION

We used an exome array to genotype over 130,000 individuals at almost 250,000 sites to assess the role of low-frequency and rare variants at the population level. After meta-analyzing the results from all 58 participating studies, we discovered several new genes and variants significantly associated with plasma lipid levels in humans. We report for the first time variants in the genes *ABCA6, SERPINA1, REST, FBLN1,* and *CCDC117* associated with LDL-C levels, variants in *TMED6* and CDC25A associated with HDL-C levels, variants in *PRRC2A, COL18A1,* and *EDEM3* associated with TG and variants in *MAP1A* and *APOC3* associated with HDL-C and TG. None of these genes have previously been associated with plasma lipid levels, and their role in lipid metabolism remains largely unknown.

These results permit several conclusions. First, this study was successful in replicating findings of coding mutations in genes with well-established associations with plasma lipids. The associations of *APOE* and *PCSK9* with LDL-C, *CETP* and *LPL* with HDL-C, and *APOA5* and *GCKR* with TG have been well established and their functions well characterized.[62,97-99] The study also successfully replicated plasma lipid associations with GWAS variants that were included on the exome array. This establishes the validity of the genotyping as well as the analytic methods employed. Furthermore, conditional analysis based on nearby loci with strong associations was successful in filtering out numerous false single variant signals that arose based on proximity and linkage to associated loci.

Second, we discovered only a few new rare coding variants using this approach. The collective sample size of over 130,000 individuals provides sufficient power to detect weak associations of low frequency and rare alleles with small effect. In the meta-analysis, we had over 80% statistical power to identify variants with an effect size of 0.2 standard deviations down to 0.35% frequency and variants with an effect size of 0.4 standard deviations down to a 0.085% frequency at an alpha level of 2 x 10^{-7}. Therefore we can reasonably conclude that within the context of the exome array, low-frequency and rare coding variants of large effect in the frequency range of (0.085-5%) do not contribute significantly to the overall variation in plasma lipids at the population level, and larger sample sizes will be needed in order to detect variants with smaller effects. Given that the frequency of recessive alleles responsible for Mendelian dyslipidemias are commonly much lower than the detection threshold of the exome array, it is not surprising that so few very rare variants were discovered in this analysis.

Third, rare and low frequency variants collectively explain only a small fraction of the missing heritability in plasma lipid levels. Despite the relatively large effect sizes on the lipid traits by some of the mutations, the paucity and overall low frequency of the variants with novel associations collectively only explain 0.002% of the variance in LDL-C, 0.003% of the variance in HDL-C, 0.006% of the variance in TG. However, in our single variant analysis we used only the additive model for simplicity. Factoring interactions and other modes of inheritance may explain a collectively larger fraction of the missing heritability.

Several limitations of the present study need to be considered. We were not able to evaluate extremely rare variants that may be unique to individuals; exome or whole genome sequencing is needed to capture this type of variation. As such, it remains possible that a burden of such very rare mutations could contribute to plasma lipid variation. If so, the sample sizes required to yield new rare variant discoveries are likely to be extraordinarily large.

The Exome Array is constrained to the coding and splice site variation observed in the ~12,000 individuals who comprised the initial exome sequencing discovery set. Furthermore, ~20% of the content contributed for design failed to be converted into genotyping assays and thus, these variants are not present on the Exome Array.

Additionally, the majority of the individuals who were genotyped are of European ancestry. The uniformity in allele frequencies and haplotype blocks associated with predominantly single ancestry analysis limits the power to detect true associations limited to non-European groups. However, our group also conducted a similar exome array association study with lipids and myocardial infarction (MI) in 56,000 individuals of African and European ancestry and identified only 4 additional low frequency variants associated with HDL-C and TG but not with LDL-C or MI risk.[116]

Using the exome array to genotype a large cohort of individuals, we were able to discover several rare variants in genes associated with plasma lipid levels. Looking forward, it will be important to further investigate if these variants are associated with coronary heart disease. Very little is known about the function of the newly discovered genes, and understanding their role in lipid metabolism may provide new insights into treating patients with dyslipidemias.

CONCLUSIONS AND FUTURE DIRECTIONS

After performing targeted sequencing of genes surrounding GWAS loci in participants with extremely high or low HDL-C levels, we did not discover any new rare coding variants with a strong effect on HDL-C. These results suggest that rare coding variants may not be significantly contributing to the original GWAS signals and that targeted exome sequencing has limited utility in discovering functional variants at these loci. Adding more samples to the study will better power the analysis, but this is unlikely to translate into new gene discoveries.

As the targeted sequencing data are unable to link a specific rare coding mutation to nearby GWAS loci in most cases, it will be important to redirect attention to the actual GWAS SNPs that were initially discovered. Investigating the role of possible functional or regulatory elements at these sites in disease-specific cell lines as an extension of available ENCODE data may be fruitful in identifying mechanistic schema and guiding future experiments.[101] Although targeted sequencing has had some success in suggesting coding variants that are likely contributing to the GWA signal for some traits, revisiting and systematically studying the initially discovered non-coding variants in the implicated loci will be necessary to better understand the biologic underpinnings of these associations.

Using the exome array to genotype rare variants in over 130,000 individuals, we have discovered 11 new genes associated with plasma lipids. We report novel associations of variants in the genes *ABCA6, SERPINA1, REST, FBLN1,* and *CCDC117* with LDL-C levels, variants in *TMED6* and *CDC25A* with HDL-C levels, variants in *PRRC2A, COL18A1,* and *EDEM3* with TG and variants in *MAP1A* and *APOC3* with HDL-

C and TG. We also reported 2 variants in *APOC3* with new, strong associations with HDL-C and TG. These results suggest that rare and low frequency variants explain a small portion of plasma lipid level variance at the population level. The generalizability of these conclusions with regards to genetic architecture is limited by the content of the exome array and the exomes from which the array SNPs were gathered; however, given the large scale of the undertaking, these results suggest that only a handful of other rare variants may be found even if the analysis were to be significantly extended in sample size or if such methodology was applied to other complex traits.

We will next assess the role of these variants with CHD. Our group is also leading an exome array meta-analysis consortium investigating the role of rare and low frequency variants in causing early onset myocardial infarction (MI). We will intersect the findings from the lipids meta-analysis with the exome array analysis for MI to inquire if the newly discovered genes contribute directly to coronary artery disease. Finally, very little is known about the genes and variants with novel associations discovered using the exome array, so we will carry out functional studies for each protein. Cell-based assays testing for the LDL trait are in progress for the top variant associations with plasma LDL-C levels. Determination of the function these discoveries in lipid metabolism may provide new insights into treating patients with dyslipidemias and coronary heart disease.

AUTHOR CONTRIBUTIONS

The research presented in this thesis is the product of an international collaboration of physicians, scientists, students, and study organizers, namely the Global Lipids Genetics Consortium (GLGC). All sequencing and genotyping was performed at genetics cores of academic facilities around the world such as the Broad Institute. APP wrote all parts of this manuscript.

Part I: Targeted Sequencing of GWAS Loci in Extremes of the High-density Lipoprotein Cholesterol Distribution

APP performed all variant calling, quality control, and analyses for the discovery phase targeted sequencing and replication phase exome chip genotyping experiments with guidance from GMP, DJR, and SK. JPP performed preliminary analysis on the dataset. DBL assisted in gathering phenotype data for targeted sequencing and exome array genotyping phases under the supervision of DJR.

Part II: Plasma Lipids Association Study Including ~130,000 Individuals Genotyped Using the Exome Array

APP prepared the analysis plan for all of the participating studies to follow under the guidance of SK and GMP. APP outlined the instructions for raw genotyping data processing, genotype quality control, phenotype modeling, and analysis and guided analysts of participating groups of the GLGC Exome Chip Working Group through data analysis. APP and HT performed all steps of analysis for six of the 58 participating studies (~20,000 individuals). DJL performed the final steps of meta-analysis of the single-variant and gene based results. DJL and APP led all consortium-wide discussions of data processing and interpretation with guidance from GA and SK.

REFERENCES

1. Finegold JA, Asaria P, Francis DP. Mortality from ischaemic heart disease by country, region, and age: Statistics from World Health Organisation and United Nations. International Journal of Cardiology 2013;168:934-45.
2. Hoyert DL, Xu J. Deaths: Preliminary data for 2011. Centers for Disease Control and Prevention; 2012:1-52.
3. FastStats: Heart Disease (U.S.). Centers for Disease Control and Prevention 2013. 2014, at http://www.cdc.gov/nchs/fastats/heart.htm.)
4. Kannel WB, Dawber TR, Kagan A, Revotskie N, Stokes 3rd J. Factors of risk in the development of coronary heart disease--six year follow-up experience. The Framingham Study. Annals of internal medicine 1961;55:33-50.
5. Wilson PWF, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. Circulation 1998;97:1837-47.
6. Lloyd-Jones DM, Larson MG, Beiser A, Levy D. Lifetime risk of developing coronary heart disease. Lancet 1999;353:89-92.
7. Berry JD, Dyer A, Cai X, et al. Lifetime risks of cardiovascular disease. New England Journal of Medicine 2012;366:321-9.
8. Vionnet N, Stoffel M, Takeda J, et al. Nonsense mutation in the glucokinase gene causes early-onset non-insulin-dependent diabetes mellitus. Nature 1992;356:721-2.
9. Froguel P, Zouali H, Vionnet N, et al. Familial hyperglycemia due to mutations in glucokinase - Definition of a subtype of diabetes mellitus. New England Journal of Medicine 1993;328:697-702.
10. Walston J, Silver K, Bogardus C, et al. Time of onset of non-insulin-dependent diabetes mellitus and genetic variation in the β3-adrenergic-receptor gene. New England Journal of Medicine 1995;333:343-7.
11. Widen E, Lehto M, Kanninen T, Walston J, Shuldiner AR, Groop LC. Association of a polymorphism in the β3-adrenergic-receptor gene with features of the insulin resistance syndrome in finns. New England Journal of Medicine 1995;333:348-51.
12. Yamagata K, Furuta H, Oda N, et al. Mutations in the hepatocyte nuclear factor-4α gene in maturity-onset diabetes of the young (MODY1). Nature 1996;384:458-60.
13. Yamagata K, Oda N, Kaisaki PJ, et al. Mutations in the hepatocyte nuclear factor-1α gene in maturity-onset diabetes of the young (MODY3). Nature 1996;384:455-8.
14. Horikawa Y, Iwasaki N, Hara M, et al. Mutation in hepatocyte nuclear factor-1 beta gene (TCF2) associated with MODY. Nature genetics 1997;17:384-5.
15. Stoffers DA, Ferrer J, Clarke WL, Habener JF. Early-onset type-II diabetes mellitus (MODY4) linked to IPF1. Nature genetics 1997;17:138-9.
16. Gloyn AL, Pearson ER, Antcliff JF, et al. Activating mutations in the gene encoding the ATP-sensitive potassium-channel subunit Kir6.2 and permanent neonatal diabetes. New England Journal of Medicine 2004;350:1838-49.

17. Saxena R, Voight BF, Lyssenko V, et al. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. Science 2007;316:1331-6.

18. Scott LJ, Mohlke KL, Bonnycastle LL, et al. A genome-wide association study of type 2 diabetes in finns detects multiple susceptibility variants. Science 2007;316:1341-5.

19. Sladek R, Rocheleau G, Rung J, et al. A genome-wide association study identifies novel risk loci for type 2 diabetes. Nature 2007;445:881-5.

20. Zeggini E, Scott LJ, Saxena R, Voight BF. Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. Nature genetics 2008;40:638-45.

21. Zeggini E, Weedon MN, Lindgren CM, et al. Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. Science 2007;316:1336-41.

22. Voight BF, Scott LJ, Steinthorsdottir V, et al. Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. Nature genetics 2010;42:579-89.

23. Levy D, Ehret GB, Rice K, et al. Genome-wide association study of blood pressure and hypertension. Nature genetics 2009;41:677-87.

24. Ehret GB, Munroe PB, Rice KM, et al. Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. Nature 2011;478:103-9.

25. Newton-Cheh C, Johnson T, Gateva V, et al. Genome-wide association study identifies eight loci associated with blood pressure. Nature genetics 2009;41:666-76.

26. Lifton RP, Dluhy RG, Powers M, et al. A chimaeric 11β-hydroxylase/aldosterone synthase gene causes glucocorticoid-remediable aldosteronism and human hypertension. Nature 1992;355:262-5.

27. Lifton RP, Dluhy RG, Powers M, et al. Hereditary hypertension caused by chimaeric gene duplications and ectopic expression of aldosterone synthase. Nature genetics 1992;2:66-74.

28. Shimkets RA, Warnock DG, Bositis CM, et al. Liddle's syndrome: Heritable human hypertension caused by mutations in the β subunit of the epithelial sodium channel. Cell 1994;79:407-14.

29. Hansson JH, Nelson-Williams C, Suzuki H, et al. Hypertension caused by a truncated epithelial sodium channel γ subunit: Genetic heterogeneity of Liddle syndrome. Nature genetics 1995;11:76-82.

30. Mune T, Rogerson FM, Nikkila H, Agarwal AK, White PC. Human hypertension caused by mutations in the kidney isozyme of 11β- hydroxysteroid dehydrogenase. Nature genetics 1995;10:394-9.

31. Geller DS, Farhi A, Pinkerton N, et al. Activating mineralocorticoid receptor mutation in hypertension exacerbated by pregnancy. Science 2000;289:119-23.

32. Lifton RP, Gharavi AG, Geller DS. Molecular mechanisms of human hypertension. Cell 2001;104:545-56.

33. Wilson FH, Disse-Nicodème S, Choate KA, et al. Human hypertension caused by mutations in WNK kinases. Science 2001;293:1107-12.

34. Boyden LM, Choi M, Choate KA, et al. Mutations in kelch-like 3 and cullin 3 cause hypertension and electrolyte abnormalities. Nature 2012;482:98-102.

35. Choi M, Scholl UI, Yue P, et al. K + channel mutations in adrenal aldosterone-producing adenomas and hereditary hypertension. Science 2011;331:768-72.

36. Friedewald WT, Levy RI, Fredrickson DS. Estimation of the concentration of low-density lipoprotein cholesterol in plasma, without use of the preparative ultracentrifuge. Clinical Chemistry 1972;18:499-502.

37. Saland JM, Ginsberg HN. Lipoprotein metabolism in chronic renal insufficiency. Pediatric nephrology (Berlin, Germany) 2007;22:1095-112.

38. Mitchell RNS, Frederick J. Robbins and Cotran Pathologic Basis of Disease. In: Kumar V, ed. 8 ed: Saunders; 2009.

39. Semenkovich CF. Goldman's Cecil Medicine. In: Goldman L, ed. 24 ed: Elsevier; 2012.

40. Badimón JJ, Ibáñez B. Increasing High-Density Lipoprotein as a Therapeutic Target in Atherothrombotic Disease. Incremento de las HDL como arma terapéutica en la aterotrombosis 2010;63:323-33.

41. deGoma EM, deGoma RL, Rader DJ. Beyond High-Density Lipoprotein Cholesterol Levels. Evaluating High-Density Lipoprotein Function as Influenced by Novel Therapeutic Approaches. Journal of the American College of Cardiology 2008;51:2199-211.

42. Garber AM, Avins AL. Triglyceride concentration and coronary heart disease [11]. British Medical Journal 1994;309:1440-1.

43. Nigon F, Lesnik P, Rouis M, Chapman MJ. Discrete subspecies of human low density lipoproteins are heterogeneous in their interaction with the cellular LDL receptor. Journal of Lipid Research 1991;32:1741-53.

44. Rosenson RS, Shott S, Tangney CC. Hypertriglyceridemia is associated with an elevated blood viscosity Rosenson: Triglycerides and blood viscosity. Atherosclerosis 2002;161:433-9.

45. MRC/BHF Heart Protection Study of cholesterol-lowering with simvastatin in 5963 people with diabetes: A randomised placebo-controlled trial. Lancet 2003;361:2005-16.

46. Cannon CP, Braunwald E, McCabe CH, et al. Intensive versus Moderate Lipid Lowering with Statins after Acute Coronary Syndromes. New England Journal of Medicine 2004;350:1495-504.

47. Ford ES, Ajani UA, Croft JB, et al. Explaining the decrease in U.S. deaths from coronary disease, 1980-2000. New England Journal of Medicine 2007;356:2388-98.

48. Ridker PM, Danielson E, Fonseca FAH, et al. Rosuvastatin to prevent vascular events in men and women with elevated C-reactive protein. New England Journal of Medicine 2008;359:2195-207.

49. Pedersen TR. Randomised trial of cholesterol lowering in 4444 patients with coronary heart disease: The Scandinavian Simvastatin Survival Study (4S). Lancet 1994;344:1383-9.

50. Bosch J, Gerstein HC, Dagenais GR, et al. n-3 fatty acids and cardiovascular outcomes in patients with dysglycemia. New England Journal of Medicine 2012;367:309-18.

51. Rubins HB, Robins SJ, Collins D, et al. Gemfibrozil for the secondary prevention of coronary heart disease in men with low levels of high-density lipoprotein cholesterol. New England Journal of Medicine 1999;341:410-8.

52. Smith GD, Ebrahim S. 'Mendelian randomization': Can genetic epidemiology contribute to understanding environmental determinants of disease? International Journal of Epidemiology 2003;32:1-22.
53. Do R, Willer CJ, Schmidt EM, et al. Common variants associated with plasma triglycerides and risk for coronary artery disease. Nature genetics 2013;45:1345-53.
54. Voight BF, Peloso GM, Orho-Melander M, et al. Plasma HDL cholesterol and risk of myocardial infarction: A mendelian randomisation study. The Lancet 2012;380:572-80.
55. Namboodiri KK, Kaplan EB, Heuch I, et al. The collaborative lipid research clinics family study: Biological and cultural determinants of familial resemblance for plasma lipids and lipoproteins. Genetic Epidemiology 1985;2:227-54.
56. Kathiresan S, Manning AK, Demissie S, et al. A genome-wide association study for blood lipid phenotypes in the Framingham Heart Study. BMC Medical Genetics 2007;8.
57. Metzker ML. Sequencing technologies the next generation. Nature Reviews Genetics 2010;11:31-46.
58. Belmont JW, Boudreau A, Leal SM, et al. A haplotype map of the human genome. Nature 2005;437:1299-320.
59. Lander ES, Linton LM, Birren B, et al. Initial sequencing and analysis of the human genome. Nature 2001;409:860-921.
60. Kathiresan S, Srivastava D. Genetics of human cardiovascular disease. Cell 2012;148:1242-57.
61. Lehrman MA, Schneider WJ, Sudhof TC, Brown MS, Goldstein JL, Russell DW. Mutation in LDL receptor: Alu-Alu recombination deletes exons encoding transmembrane and cytoplasmic domains. Science 1985;227:140-6.
62. Abifadel M, Varret M, Rabès JP, et al. Mutations in PCSK9 cause autosomal dominant hypercholesterolemia. Nature genetics 2003;34:154-6.
63. Soria LF, Ludwig EH, Clarke HRG, Vega GL, Grundy SM, McCarthy BJ. Association between a specific apolipoprotein B mutation and familial defective apolipoprotein B-100. Proceedings of the National Academy of Sciences of the United States of America 1989;86:587-91.
64. Berge KE, Tian H, Graf GA, et al. Accumulation of dietary cholesterol in sitosterolemia caused by mutations in adjacent ABC transporters. Science 2000;290:1771-5.
65. Garcia CK, Wilund K, Arca M, et al. Autosomal recessive hypercholesterolemia caused by mutations in a putative LDL receptor adaptor protein. Science 2001;292:1394-8.
66. Musunuru K, Pirruccello JP, Do R, et al. Exome sequencing, ANGPTL3 mutations, and familial combined hypolipidemia. New England Journal of Medicine 2010;363:2220-7.
67. Kathiresan S, Melander O, Guiducci C, et al. Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. Nature genetics 2008;40:189-97.
68. Willer CJ, Sanna S, Jackson AU, et al. Newly identified loci that influence lipid concentrations and risk of coronary artery disease. Nature genetics 2008;40:161-9.

69. Kathiresan S, Willer CJ, Peloso GM, et al. Common variants at 30 loci contribute to polygenic dyslipidemia. Nature genetics 2009;41:56-65.

70. Teslovich TM, Musunuru K, Smith AV, et al. Biological, clinical and population relevance of 95 loci for blood lipids. Nature 2010;466:707-13.

71. Global Lipids Genetics C, Willer CJ, Schmidt EM, et al. Discovery and refinement of loci associated with lipid levels. Nature genetics 2013;45:1274-83.

72. Musunuru K, Strong A, Frank-Kamenetsky M, et al. From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. Nature 2010;466:714-9.

73. Chasman DI, Paré G, Mora S, et al. Forty-three loci associated with plasma lipoprotein size, concentration, and cholesterol content in genome-wide analysis. PLoS Genetics 2009;5.

74. Aulchenko YS, Ripatti S, Lindqvist I, et al. Loci influencing lipid levels and coronary heart disease risk in 16 European population cohorts. Nature genetics 2009;41:47-55.

75. Sabatti C, Service SK, Hartikainen AL, et al. Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. Nature genetics 2009;41:35-46.

76. Nejentsev S, Walker N, Riches D, Egholm M, Todd JA. Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. Science 2009;324:387-9.

77. Galarneau G, Palmer CD, Sankaran VG, Orkin SH, Hirschhorn JN, Lettre G. Fine-mapping at three loci known to affect fetal hemoglobin levels explains additional genetic variation. Nature genetics 2010;42:1049-51.

78. Raychaudhuri S, Iartchouk O, Chin K, et al. A rare penetrant mutation in CFH confers high risk of age-related macular degeneration. Nature genetics 2011;43:1232-6.

79. Helgason H, Sulem P, Duvvari MR, et al. A rare nonsynonymous sequence variant in C3 is associated with high risk of age-related macular degeneration. Nature genetics 2013;45:1371-6.

80. Seddon JM, Yu Y, Miller EC, et al. Rare variants in CFI, C3 and C9 are associated with high risk of advanced age-related macular degeneration. Nature genetics 2013;45:1366-73.

81. Van De Ven JPH, Nilsson SC, Tan PL, et al. A functional variant in the CFI gene confers a high risk of age-related macular degeneration. Nature genetics 2013;45:813-7.

82. Zhan X, Larson DE, Wang C, et al. Identification of a rare coding variant in complement 3 associated with age-related macular degeneration. Nature genetics 2013;45:1375-81.

83. Rivas MA, Beaudoin M, Gardet A, et al. Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. Nature genetics 2011;43:1066-73.

84. Gnirke A, Melnikov A, Maguire J, et al. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nature Biotechnology 2009;27:182-9.

85. Depristo MA, Banks E, Poplin R, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature genetics 2011;43:491-501.

86. Cingolani P, Platts A, Wang le L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly 2012;6:80-92.

87. Goldstein JI, Crenshaw A, Carey J, et al. zCall: a rare variant caller for array-based genotyping: genetics and population analysis. Bioinformatics 2012;28:2543-5.

88. Neale BM, Rivas MA, Voight BF, et al. Testing for an unusual distribution of rare variants. PLoS Genetics 2011;7.

89. Price AL, Kryukov GV, de Bakker PIW, et al. Pooled Association Tests for Rare Variants in Exon-Resequencing Studies. American Journal of Human Genetics 2010;86:832-8.

90. Lee S, Wu MC, Lin X. Optimal tests for rare variant effects in sequencing association studies. Biostatistics 2012;13:762-75.

91. Price AL, Kryukov GV, de Bakker PI, et al. Pooled association tests for rare variants in exon-resequencing studies. American journal of human genetics 2010;86:832-8.

92. Team RDC. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2010.

93. McKenna A, Hanna M, Banks E, et al. The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research 2010;20:1297-303.

94. Purcell S, Neale B, Todd-Brown K, et al. PLINK: A tool set for whole-genome association and population-based linkage analyses. American Journal of Human Genetics 2007;81:559-75.

95. PLINK/SEQ: A Library for the Analysis of Genetic Variation Data. 2012. at https://atgu.mgh.harvard.edu/plinkseq/.)

96. Acton S, Rigotti A, Landschulz KT, Xu S, Hobbs HH, Krieger M. Identification of scavenger receptor SR-BI as high density lipoprotein receptor. Science 1996;271:518-20.

97. Kondo I, Berg K, Drayna D, Lawn R. DNA polymorphism at the locus for human cholesteryl ester transfer protein (CETP) is associated with high density lipoprotein cholesterol and apolipoprotein levels. Clinical Genetics 1989;35:49-56.

98. Patsch JR, Prasad S, Gotto AM, Patsch W. High density lipoprotein2. Relationship of the plasma levels of this lipoprotein species to its composition, to the magnitude of postprandial lipemia, and to the activities of lipoprotein lipase and hepatic lipase. Journal of Clinical Investigation 1987;80:341-7.

99. Aouizerat BE, Kulkarni M, Heilbron D, et al. Genetic analysis of a polymorphism in the human apoA-V gene: Effect on plasma lipids. Journal of Lipid Research 2003;44:1167-73.

100. Jaye M, Lynch KJ, Krawiec J, et al. A novel endothelial-derived lipase that modulates HDL metabolism. Nature genetics 1999;21:424-8.

101. Consortium EP, Dunham I, Kundaje A, et al. An integrated encyclopedia of DNA elements in the human genome. Nature 2012;489:57-74.

102. Hunt KA, Mistry V, Bockett NA, et al. Negligible impact of rare autoimmune-locus coding-region variants on missing heritability. Nature 2013;498:232-5.
103. Purcell S, Cherny SS, Sham PC. Genetic power calculator: Design of linkage and association genetic mapping studies of complex traits. Bioinformatics 2003;19:149-50.
104. Keating BJ, Tischfield S, Murray SS, et al. Concept, design and implementation of a cardiovascular gene-centric 50 K SNP array for large-scale genomic association studies. PLoS ONE 2008;3.
105. Trynka G, Hunt KA, Bockett NA, et al. Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. Nature genetics 2011;43:1193-201.
106. Voight BF, Kang HM, Ding J, et al. The Metabochip, a Custom Genotyping Array for Genetic Studies of Metabolic, Cardiovascular, and Anthropometric Traits. PLoS Genetics 2012;8.
107. Exome Chip Design. University of Michigan, 2013. 2014, at http://genome.sph.umich.edu/wiki/Exome_Chip_Design.)
108. Ionita-Laza I, Lee S, Makarov V, Buxbaum JD, Lin X. Sequence kernel association tests for the combined effect of rare and common variants. American Journal of Human Genetics 2013;92:841-53.
109. Mantel N, Haenszel W. Statistical Aspects of the Analysis of Data From Retrospective Studies of Disease. Journal of the National Cancer Institute 1959;22:719-48.
110. Liu DJ, Peloso GM, Zhan X, et al. Meta-analysis of gene-level tests for rare variant association. Nature genetics 2013.
111. Pennacchio LA, Olivier M, Hubacek JA, et al. An apolipoprotein influencing triglycerides in humans and mice revealed by comparative sequencing. Science 2001;294:169-73.
112. Zhang SH, Reddick RL, Piedrahita JA, Maeda N. Spontaneous hypercholesterolemia and arterial lesions in mice lacking apolipoprotein E. Science 1992;258:468-71.
113. Pollin TI, Damcott CM, Shen H, et al. A null mutation in human APOC3 confers a favorable plasma lipid profile and apparent cardioprotection. Science 2008;322:1702-5.
114. Jong MC, Hofker MH, Havekes LM. Role of apoCs in lipoprotein metabolism: Functional differences between ApoC1, ApoC2, and ApoC3. Arteriosclerosis, Thrombosis, and Vascular Biology 1999;19:472-84.
115. Sukonina V, Lookene A, Olivecrona T, Olivecrona G. Angiopoietin-like protein 4 converts lipoprotein lipase to inactive monomers and modulates lipase activity in adipose tissue. Proceedings of the National Academy of Sciences of the United States of America 2006;103:17450-5.
116. Peloso GM, Auer PL, Bis JC, et al. Association of low-frequency and rare coding-sequence variants with blood lipids and coronary heart disease in 56,000 whites and blacks. American Journal of Human Genetics 2014;94:223-32.