Insights into the genetic architecture underlying plasma lipids and related phenotypes from genome-wide human genetic variation

by

Ellen Marie Schmidt

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Bioinformatics)
in The University of Michigan
2016

Doctoral Committee:

Assistant Professor Cristen J. Willer, Chair
Professor Michael L. Boehnke
Professor Charles F. Burant
Professor Margit Burmeister
Assistant Professor Hyun Min Kang

To my parents, for fostering lifelong curiosity and creativity, and providing endless opportunity, encouragement, and love.

ACKNOWLEDGEMENTS

I am honored to have conducted research as part of the University of Michigan Bioinformatics Graduate Program. My success as a graduate student would not have been possible without the dedicated mentorship of my thesis advisor, Cristen Willer. She has been a tremendously positive influence on me through her insight, guidance, and enthusiasm for research, and I will practice the countless lessons I have learned from her throughout my career. I am especially grateful for the extra opportunities that have enriched my graduate school experience including conferences, short courses, and an internship.

I am greatly appreciative for the support and direction I received from my thesis committee members, Michael Boehnke, Charles Burant, Margit Burmeister, and Hyun Min Kang. Their guidance has been instrumental in my research success, and I am honored to have learned from such brilliant scientists. I would particularly like to thank Margit Burmeister for allowing me my first research opportunity at UM and encouraging me to pursue my degree early on. I am also appreciative to Michael Boehnke for giving me the opportunity to complete my first research rotation as a bioinformatics student in his lab. Thank you especially to Hyun Min Kang, from whom I learned an incredible amount about sequencing data analysis. I am grateful to many others who have gone out of their way to share research and career guidance including Ryan Mills, Stephen Parker, Jeffrey de Wet, Daniel Burns, Brian Athey, and members of the Center for Statistical Genetics. I am also very

lucky to have worked with many wonderful collaborators in both The Global Lipids Genetics Consortium and The HUNT Study.

It has been a pleasure working with and learning from such remarkably talented lab members and peers. Thank you to Jin Chen for sharing his programming knowledge and his dedicated efforts in developing GREGOR with me. I am grateful for the wonderful collaboration with Sebanti Sengupta on the GLGC project. Matthew Flickinger provided many helpful scripts on the wiki for making beautiful figures. I am especially grateful for the friendships of my fellow graduate students Wei Zhou, Mallory Freeberg, Artur Veloso, Kraig Stevenson, and my entire bioinformatics and biostatistics student cohorts. You have inspired and challenged me to work my hardest, and your camaraderie and encouragement throughout graduate school has been tremendously supportive.

My graduate school journey would not have been the same without the many lifelong friendships formed out of the St. Mary Student Parish community. Thank you also to Carol Radovic and members of Ann Arbor Ballet Theatre who have re-inspired my passion for dance and shown me a fulfilling creative outlet during graduate school. I am also grateful to Brunch Club, with whom I spent many Saturday mornings sharing delightful food and fellowship.

I am so blessed to have completed this journey with my incredibly supportive partner, Aaron Skiba. Your love, encouragement, patience, and steadfast confidence in me have given me immense strength to take on challenges. Finally, thank you to my wonderful family, who have supported and loved me unconditionally throughout my life. I have been especially influenced by my grandparents, my parents Henry and Carol Schmidt, and my sister Christina Schmidt, who have all led by example and taught me that hard work, perseverance, and faith are key to facing any obstacles.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF ABBREVIATIONS

aCGH	array comparative genomic hybridization
BMI	body mass index
BP	blood pressure
CAD	coronary artery disease
ChIP-Seq	chromatin immunoprecipitation followed by high-throughput DNA sequencing
CNV	copy number variant
DEL	deletion
DHS	DNase hypersensitive site
DNA	deoxyribonucleic acid
DUP	duplication
ENCODE	ENCyclopedia Of DNA Elements
eQTL	expression quantitative trait locus
GREGOR	Genomic Regulatory Elements and Gwas Overlap AlgoRithm
GWAS	genome wide association study
HDL-C	high density lipoprotein cholesterol
INV	inversion
LD	linkage disequilibrium
LDL-C	low density lipoprotein cholesterol
MAF	minor allele frequency
MI	myocardial infarction
NGS	next generation sequencing
SNP	single nucleotide polymorphism
SV	structural variant
T2D	type 2 diabetes
TC	total cholesterol
TF	transcription factor
TFBS	transcription factor binding site
TG	triglycerides

CHAPTER I

Introduction

When the Human Genome Project was launched 25 years ago, the stage was set for an exciting era in biology and medicine. Pioneering discoveries of the most fundamental information concerning gene structure, function, and regulation have furthered our understanding of the role of genetics in health and disease. Technological advances in high-throughput sequencing and methods for interpreting large-scale genetic data are rapidly shaping the breadth of human genetics research. The complexity and diversity of the human genome still presents considerable challenges, however. We can harness the information from human genetic variation to address enduring challenges in understanding complex disorders, the motivation for which I present the following dissertation.

1.1 Population-based association studies

Complex traits are multifactorial, presenting complexities beyond single-gene disorders with classic Mendelian inheritance patterns. The considerable challenges in understanding complex phenotypes have driven the development of study designs that rely on comparisons of unrelated affected and unaffected individuals. A genome wide association study (GWAS) tests the relationship between genetic marker predic-

Modified from: Schmidt and Willer (2015)

tors across the genome and a single case-control or quantitative phenotype response by employing logistic or linear regression, respectively. Typically, each study participant is genotyped for a set of genome-wide independent markers on a commercial genotyping array. In humans there are stretches of DNA that segregate together more often than is expected by chance, resulting in non-independent markers in linkage disequilibrium (LD). To address multiple testing and LD properties, a typical GWAS considers one million independent single nucleotide polymorphisms (SNPs), resulting in a Bonferroni significance cut-off of association P-value $\leq 5\text{x}10^{-8}$. Increasing sample sizes provide greater power to detect associations at the genome-wide level. Currently, there are catalogued over 10,000 single markers that reach genome-wide significance for hundreds of phenotypes (Welter et al., 2014).

A major consideration in the design of case-control GWA studies is the choice of a control group. For a trait with high prevalence for example, it is important to carefully choose healthy matched controls rather than picking a random sample of control individuals from the population. GWAS designs involving quantitative traits should ensure that the trait is normally distributed, which often requires a logarithmic or inverse normal transformation of the trait before association testing. Appropriately adjusting for confounders such as age and sex can also be critical to prevent spurious or false positive associations and to maximize power. Population structure presents another challenge that is typically accounted for using principle components (Price et al., 2006) or mixed-model approaches (Kang et al., 2010). We expect association P-values that deviate from the null uniform distribution to represent true positive associations. Inflation from the null distribution indicates additional batch effect or population structure that was not accounted for in the association analysis, and suggests that false positive results may be present. Test

statistics can be inflated by a factor lambda (λ), which is defined as the ratio of the median of the observed distribution of the test statistic to the median of the expected distribution (0.455, ${\chi^2}_{\text{df=1}}$) (Devlin and Roeder, 1999). Genomic control adjustment by this lambda reduces inflation and the risk of false positive associations. With careful study design including consideration of confounding factors, GWAS is a powerful tool to discover true causal relationships in which genetic marker alleles or nearby linked alleles influence susceptibility.

Genome-wide screening for single marker associations is generally used for identifying common variation (minor allele frequency (MAF) $> 5\%$), but loses power in efforts to identify associations with low frequency ($0.5\% < \text{MAF} \leq 5\%$) or rare (MAF $\leq 0.5\%$) variants. Protein-coding variants with deleterious function are likely to be rarer in the human population due to natural selection acting against them, and have mostly arisen recently in evolutionary history (Fu et al., 2013). Because of the rare nature of most variants with functional consequence, studies carefully designed to uncover rare variant associations are crucial (Lee et al., 2014; Zuk et al., 2014). Recent advances in exome sequencing and exome array technologies have facilitated larger and more accurate studies for interrogating the protein-coding 1-2% of the genome. However, single variant association tests commonly used by GWAS carry a heavy multiple testing burden and still lack power when applied to rare variants of high impact. Additional challenges for finding rare variation using traditional GWAS single-variant approaches include poor coverage on arrays and difficulties with imputing. Thus, aggregation-based tests that group multiple variants by a single gene or functional unit have become standard for rare variant association testing.

Several regression-based approaches have been developed in recent years to optimize rare variant discovery. In a simple burden test, multiple rare variants are

collapsed into a genetic score representing the cumulative effect of those variants in a single unit. Then, the score is tested for association with a trait or disease. This idea, which has been implemented by numerous investigators (Morgenthaler and Thilly, 2007; Li and Leal, 2008; Madsen and Browning, 2009; Morris and Zeggini, 2010; Asimit et al., 2012), assumes that all variants in a single unit are causal and that all alleles affect the phenotype with the same magnitude and direction of effect. More robust modifications of a simple burden test introduce adaptive weights or thresholds (Han and Pan, 2010; Hoffmann et al., 2010; Liu and Leal, 2010; Price et al., 2010; Ionita-Laza et al., 2011; Lin and Tang, 2011). We can account for the protective or deleterious impact of alleles on phenotype by considering the magnitude and direction of effect using variance-component tests (Pan, 2009; Neale et al., 2011; Wu et al., 2011). Finally, we can combine burden and component tests (Lee et al., 2012; Derkach et al., 2013; Sun et al., 2013) or score statistics (Chen et al., 2012) to achieve more robust power. Choosing the optimal strategy for grouping rare variants is flexible and may depend on the genetic architecture of a particular trait (Ladouceur et al., 2012).

Meta-analysis is a powerful tool to jointly analyze GWA datasets from multiple studies, especially when individual-level data are not available (Chapman et al., 2011). In fact, the statistical power achieved by meta-analysis of summary statistics is quite comparable to that achieved from the cumbersome pooling of individual-level data (Lin and Zeng, 2010). Fisher's method (Fisher et al., 1970) for combining P-values is one simple method, but it neither weights by sample size nor considers magnitude or direction of effect. This approach is impractical when individual studies are unequal in size and/or the number of studies becomes large. We can combine evidence for association by converting P-values into a signed Z-score weighted by

sample size, or by weighting effect size estimates by their estimated standard errors (Stouffer et al., 1949; Willer et al., 2010). When studies provide score statistics for each variant and a variance-covariance matrix, a fixed-effects meta-analysis will achieve improved power (Hu et al., 2013; Lee et al., 2013; Liu et al., 2014).

The association findings from GWAS provide an initial guide for the development of medical treatments by pointing to a genomic region of interest. Within a single locus however, there may be several or more genes and hundreds of linked genetic variants. Identifying the putative causal variant and unraveling the underlying functional mechanism at a single locus often requires finer interrogation and experimental follow-up. Given the non-independent nature of markers across the genome, it is common to only genotype a subset of independent markers in a GWA study. However, commercial SNP genotyping panels only assay a small fraction of variants that contribute to complex disease. Meta-analysis of multiple GWA studies by assigning SNPs genotyped in one study as proxies for SNPs genotyped in another study is complicated. Genome sequencing is more comprehensive, but can be expensive for studies involving thousands of participants. Imputation provides a cost-effective *in silico* strategy for accurately guessing un-typed markers without directly genotyping every variant across the genome (Marchini et al., 2007; Li et al., 2010). Under the assumption that unrelated individuals share short stretches of haplotypes inherited from distant common ancestors, we can use a subset of typed markers measured in only one individual to impute into another. Targeted dense genotyping of carefully selected loci is an effective follow-up strategy, as demonstrated by arrays such as Metabochip (Voight et al., 2012a) and Immunochip (Cortes and Brown, 2011) that are tailored for trait-specific follow-up.

1.2 Insights into plasma lipids from genetic discovery

Cardiovascular disease is the leading cause of death in the United States and throughout the world, representing a significant human health burden (Mozaffarian et al., 2015). Genetic studies of lipid levels, known risk factors for heart disease with heritability ranging from 40% to 60% in humans (Weiss et al., 2006), are logical targets in efforts to prevent and treat heart disease. Modulation of these quantitative lipid traits, which include low-density lipoprotein cholesterol (LDL-C), high-density lipoprotein cholesterol (HDL-C), triglycerides (TG), and total cholesterol (TC), can be effective therapeutically. Lipid traits are not independent, making it challenging to untangle the effects of specific lipids on disease risk. LDL-C and TC generally act in the same direction of effect on heart disease risk since the majority of TC is composed of LDL-C. In addition, high TG is associated with high LDL-C and low HDL-C, while LDL-C and HDL-C are positively and inversely associated with heart disease risk, respectively (Emerging Risk Factors Collaboration et al., 2009).

Much of our current understanding of the genetics of blood lipids and implications in human health has originated from genome-wide association discoveries. Early GWA studies with modest sample sizes (<10,000) uncovered common variants (MAF>5%) with large effect sizes (Kathiresan et al., 2008; Willer et al., 2008). The power to discover new lipid-associated variants increased with sample size (Kathiresan et al., 2009; Teslovich et al., 2010). In Chapter II, I describe a joint meta-analysis of nearly 180,000 samples which uncovered 62 novel independent genetic loci containing lipid-associated variants (Global Lipids Genetics Consortium et al., 2013). Most of the associated variants are non-protein-coding, suggesting a regulatory role (Welter et al., 2014). An illustration of the regulatory role of a noncoding LDL-

C-associated variant is evident at the *SORT1* locus. Researchers demonstrated in human-derived hepatocytes that variant rs12740374 at an LDL-C-associated GWAS locus creates a C/EBP transcription factor binding site, causing altered expression of the nearby *SORT1* gene (Musunuru et al., 2010). Still, common variants generally have limited functional consequence.

Common variants identified by GWAS explain only a fraction (∼20-25%) of the heritable trait variance for lipids (Teslovich et al., 2010). In effort to explain some of the missing heritability, we turn to low frequency and rare variation. The Common Disease-Rare Variant hypothesis proposes that the combined effect of a number of low frequency variants with large effect sizes accounts for some of the missing heritability (Pritchard and Cox, 2002). Indeed, early sequencing studies of candidate genes supported the contribution of multiple rare alleles to plasma HDL-C levels (Cohen et al., 2004).

Mendelian family studies involving large pedigrees are valuable for rare variant genetic studies. In this design, the co-segregation of variants among affected family members can be traced. Investigators recently used whole-exome sequencing in a multi-generation family to uncover a rare variant in a highly conserved codon of *SLC25A40* that is associated with TG, giving insight into a previously unknown biological mechanism of hypertriglyceridemia (Rosenthal et al., 2013). My subsequent focus will include findings from large-scale array and sequencing studies for complex lipid traits.

Interrogating the protein-coding genome through re-sequencing coding regions (Kryukov et al., 2009) and whole-exome sequencing (Do et al., 2012) can reveal rare mutations with a large effect on phenotype. A splice variant in *APOC3* associated with TG was identified using whole-genome sequencing, representing one of the first

rare variants of large effect to be found using this sequencing approach at the population scale (Timpson et al., 2014). Sequencing the exome is more cost-effective than whole-genome sequencing, however, and fewer statistical tests are performed, reducing the multiple testing burden. The potential of exome sequencing has resulted in studies powered for the discovery of novel rare variation implicated in blood lipids. For example, investigators of the National Heart, Lung, and Blood Institute Exome Sequencing Project used exome sequencing to identify the burden of rare variants in four genes (*PNPLA5*, *PCSK9*, *LDLR* and *APOB*) significantly associated with LDL-C (Lange et al., 2014). By contrast to modest effect sizes observed from individual SNPs identified by GWAS, the burdens of rare variants in these genes have substantially higher effect sizes (Figure 1.1). Association testing using the more cost-effective study design of genotyping and successfully imputing SNPs has also led to novel insights into the impact of rare variants on lipids (Surakka et al., 2015). Tables 1.1 and 1.2 summarize the contribution of low frequency and rare variation to lipids from single variant and burden tests, respectively.

The exome chip custom genotyping array allows for large-scale efficient genotyping of low frequency coding variants with large effect sizes. Exome wide association studies for lipids and related diseases revealed several significant variants at both established and previously unknown lipid loci. Rare variants at *ANGPTL4*, *LIPC* and *LIPG*, for example, were found to be associated with TG and HDL-C (Holmen et al., 2014) (Table 1.1). In addition, a more common variant in the protein-coding gene, *TM6SF2*, was found to be associated with total cholesterol and myocardial infarction risk (Holmen et al., 2014; Kozlitina et al., 2014). Functional follow-up revealed that modulation of Tm6sf2 in mice alters lipid levels, providing the causal gene at a GWAS locus that was previously intractable for follow-up due to a large

number of genes in the associated region. These findings illustrate the value of interrogating changes in the exome in guiding our exploration of the functional gene at lipid loci.

Because we expect that rare variants of large effects are more likely to occur in coding regions, detection of rare noncoding variants with comparable effect sizes will generally require much larger sample sizes. Whole-genome sequencing of nearly 1,000 individuals revealed more about the genetic architecture of HDL-C (Morrison et al., 2013). Morrison et al. (2013) found that common variation explains a whole 61.8% of the trait heritability for HDL-C, and individuals with extreme HDL-C harbor rare variants with large effect sizes. Dense genotyping on custom arrays can guide discovery of rare functional variants that were not previously interrogated and help narrow the association signal (Sanna et al., 2011; Wu et al., 2013). Fine mapping on Metabochip in Europeans highlighted the LDL-C associated rare R46L (p.Arg46Leu) variant (allele frequency 0.03) at *PCSK9* to refine the GWAS signal (Global Lipids Genetics Consortium et al., 2013). In addition, methods to evaluate enrichment of noncoding variants in regulatory regions of the genome will give insight into the biological mechanisms involved and help prioritize rare functional variants (Lo et al., 2014; Schmidt et al., 2015).

1.3 Translation of plasma lipid levels to disease risk

Understanding the relationship between plasma lipid concentrations and heart disease risk is paramount in addressing human health. Several lipid loci contain variants associated with risk for diseases such as coronary artery disease (CAD) (CARDIoGRAMplusC4D Consortium et al., 2013), highlighting the causal role of blood lipids on CAD. Although the correlation between increased triglyceride levels

and increased risk of CAD is well established (Sarwar et al., 2007), the causality of this relationship is a separate question that has claimed recent attention. After adjusting for the effects of both LDL-C and HDL-C, Do et al. (2013) found correlation between the effect size of TG-associated SNPs and the magnitude of the effect on CAD risk, suggesting causality. Exome sequencing revealed a set of rare variants, including one missense and 3 loss-of-function driver mutations in *APOC3* associated with low plasma TG (TG and HDL Working Group of the Exome Sequencing Project, National Heart, Lung, and Blood Institute et al., 2014). Carriers of such mutations showed significantly reduced risk for CAD.

Although the causal relationship between low LDL-C and reduced risk for myocardial infarction (MI) is well established and used in treatment (Baigent et al., 2005), the causality of high HDL-C with a similar outcome has been challenged (Voight et al., 2012b). A Mendelian randomization study (n >25,000 participants) involving a low frequency variant p.Asn396Ser (allele frequency 0.026) in *LIPG* that is associated with high HDL-C showed that carriers did not have a significantly reduced risk for MI (Voight et al., 2012b). This questions the utility of high plasma HDL-C to either predict or treat heart disease.

Researchers sequenced the exons of the *NPC1L1* gene and found naturally occurring mutations that can mimic the activity of an LDL-C lowering drug and thus reduce coronary heart disease (CHD) risk (Myocardial Infarction Genetics Consortium Investigators et al., 2014). Low frequency variants in *PCSK9* are also known to be associated with low LDL-C and reduced CHD risk (Cohen et al., 2005, 2006; Cohen and Hobbs, 2013). This prompted the interrogation of rare coding variation to find similar occurrences elsewhere in the genome. Using an exome array, researchers identified four low-frequency variants associated with HDL-C (in *ANGPTL8*, *PAFAH1B2*, and

PCSK7) and TG levels (in *COL18A1*), but not CHD (Peloso et al., 2014) (Table 1.1).

Lipid loci demonstrate a significant degree of pleiotropy, in which a single locus can result in multiple phenotypes. Figure 1.2 illustrates the pleiotropic nature of lipid-associated loci, many of which were discovered in studies described in this thesis. This complexity introduces challenges for translation to human health, presenting motivation for human genetics research of complex traits. Studies of heritable variation will be instrumental in guiding physicians toward appropriate risk prevention, diagnosis, and treatment of cardiovascular disease.

1.4 Interpreting noncoding genetic variation

Understanding the underlying biology of noncoding human genetic variation and translating association findings into clinical practice remains a universal challenge. In contrast to protein-coding mutations that cause an amino acid change to potentially alter protein function, noncoding variation typically acts by altering the DNA sequence to which transcription factors (TF) and other proteins bind. Changing TF binding affinity can affect gene expression levels, thus contributing to phenotypic variation. The noncoding nature of most trait-associated variation identified by GWAS suggests that these polymorphisms play an important role in transcriptional regulation (Hindorff et al., 2009). To understand this, we can explore epigenomic changes across the genome such as histone modifications and DNA methylation that impact gene expression levels by changing chromatin structure. High throughput technologies have allowed us to investigate protein interactions with DNA through chromatin immunoprecipitation followed by massively parallel DNA sequencing (ChIP-Seq). The information about protein binding sites together with noncoding trait-associated variation from GWAS provides further insight into the mechanisms acting to alter

phenotype (see Chapter III).

Several large-scale efforts such as the National Institutes of Health Roadmap Epigenomics Project (Roadmap Epigenomics Consortium et al., 2015) and the National Human Genome Research Institute ENCODE Project (ENCODE Project Consortium, 2012) have been launched with the goal of developing a comprehensive catalogue of all human genomic functional elements. The data generated by these public repositories are key to addressing the challenges of interpreting noncoding variation.

We can harness the heritable nature of drug response to explore the genetic mechanisms of noncoding mutations contributing to drug response variability. For example, genetic variation within regulatory elements can disrupt TF binding and drug response by altering a drug-targeted affinity. This concept has been demonstrated by Soccio et al. (2015), showing that variants within binding sites of PPARγ, a nuclear receptor target for anti-diabetic therapy, alter PPARγ and cofactor occupancy. This in turn alters response to the drug rosiglitazone, demonstrating a new mechanism by which noncoding variation leads to drug response variability. Still, the heterogeneity in drug response remains a major challenge facing physicians when prescribing treatment.

1.5 Discovery of structural variation

The complexity of the human genome extends beyond single nucleotide polymorphisms and indels to larger structural variation that can span thousands of bases. This class of variation presents unique challenges in discovery and functional interpretation. Structural variation (SV) historically refers to chromosomal rearrangements of >1 kb in size, but due to sequencing technologies broadening the spectrum of discovery, can now be expanded to events of >50 base pairs (bp) in length (Alkan

et al., 2011). Structural variants can be balanced or unbalanced, changing the number of base pairs in the genome and thus referred to as copy number variants (CNVs). Pang et al. (2010) estimates that CNVs within a single genome result in 1.2% difference from the consensus reference sequence. In contrast to ∼83% of the total detected genetic variation in gene expression that results from SNPs, CNVs capture ∼17% (Stranger et al., 2007). Mills et al. (2011) found 22,025 deletions and 6,000 additional SVs including insertions and tandem duplications in 185 sequenced human genomes. The latest 1000 Genomes Project effort discovered and genotyped ∼14,000 large deletions (48bp-995kb) in a diverse set of over 1,000 sequenced individuals (1000 Genomes Project Consortium et al., 2012). The Database of Genomic Variants archive (DGVa) catalogues structural variant data in a public repository for dissemination to the wider research community (`http://www.ebi.ac.uk/dgva`).

Methods of detecting structural variation have evolved with changing technology and study designs. Early studies used hybridization-based technologies such as array CGH (comparative genomic hybridization) to capture CNVs (Iafrate et al., 2004; McCarroll et al., 2008; Conrad et al., 2010). Although microarrays are cost-effective, they are low throughput and low resolution technologies that cannot identify balanced structural variants. Advances in next generation sequencing (NGS) technologies and the routine use of genome- and exome-sequencing data sets have allowed us to call complex events with unprecedented resolution. For NGS-based discovery, mapped sequence reads are compared to a reference genome to find patterns that can be classified into various SV types such as deletions, duplications, inversions, and translocations.

There are a number of NGS-based computational approaches for SV discovery, all of which present bioinformatics challenges and each with its own set of strengths

and weaknesses (Mills et al., 2011; Alkan et al., 2011). The read-pair mapping approach compares the consistency of orientation of read pairs to a reference genome, and can detect most classes of variation. Discordantly matched paired-ends with an alignment distance, or insert size, that deviates significantly from the expected distance on the genome are used to estimate SV coordinates (Korbel et al., 2007; Chen et al., 2009). Another method observes significantly higher or lower read depth compared to a random distribution of mapping depth to call duplications or deletions, respectively (Yoon et al., 2009). A split-read method identifies the exact breakpoint of a structural variant by observing regions where the read alignment to the genome is broken (Ye et al., 2009). GenomeSTRiP integrates read-depth, read-pair, and split-read approaches for discovering and genotyping deletions in a population (Handsaker et al., 2011, 2015). Another SV caller, DELLY, combines read-pair and split-read analysis for calling both balanced and unbalanced events (Rausch et al., 2012). Lastly, *de novo* assembly of contigs that are compared to a reference genome can be used to discover various classes of SVs (Zerbino and Birney, 2008; Li, 2015). When various NGS-based SV discovery approaches are directly compared on a distinct set of samples, the number of events called uniquely by a single method is as high as 80% of all SVs discovered by that method (Mills et al., 2011; Alkan et al., 2011). This comparison illustrates the complementary nature of these various discovery techniques. In addition, there are sensitivity vs. specificity trade-offs among popular tools, and a comparison of their performance on calling deletions over 100 bp reveals false negative rates ranging from 0.31 to 0.79 and false positive rates ranging from 0.09 to 0.37 (Li, 2015).

1.6 Functional impact of structural variation in complex disorders

Our current understanding of the role of structural variation in health-related phenotypes is limited relative to less complex forms of genetic variation. Indeed, SVs are known to contribute to many disease types ranging from Mendelian disorders (*e.g.* Charcot-Marie-Tooth Disease (Lupski et al., 1991)) to sporadic developmental syndromes (*e.g.* autism (Sebat et al., 2007)) to common complex disease (*e.g.* psoriasis (Hollox et al., 2008), systemic lupus erythematosus (Yang et al., 2007)).

Genome-wide association studies, used primarily for identifying associations between SNPs and quantitative traits or disease phenotypes, are less commonly implemented for association with structural variation. Our ability to sequence large numbers of individuals and accurately call and genotype SVs has made genome-wide SV associations more informative. Discovery and association analysis of the functional impact of structural variation in myocardial infarction (MI) from whole genome sequencing is the subject of Chapter IV. Previous discovery of SV's using array CGH found common copy number variants that are well tagged by trait-associated SNPs, presenting plausible functional candidates (Conrad et al., 2010). For example, Conrad et al. (2010) identified a CNV in LD (r^2=1) with an MI-associated single variant (rs6725887) at the *WDR12* locus first reported by Myocardial Infarction Genetics Consortium et al. (2009). With the improved resolution of SV detection from sequencing technologies, we can develop a more comprehensive map of structural variation to better understand the genetic landscape of complex disease.

1.7 Dissertation outline

My research objective is to understand how human genetic variation causes phenotypic differences and individual disease risk, even when only a small fraction of

this variation is protein-coding. The subsequent chapters explore multiple facets of human genetic variation to further our understanding of the genetic landscape of quantitative lipid traits and cardiovascular disorders. From single genetic variants with a focus on understanding noncoding transcriptional regulation to structural variation classified from whole genome sequencing, this work contributes important insights into complex disease and will help tailor strategies for translation to human health.

Novel lipid-associated variants with small effect sizes can be identified at genome-wide significance through targeted genotyping of additional individuals on Metabochip followed by meta-analysis with the original lipids GWAS results. In Chapter II, I describe the largest genome-wide meta-analysis for lipids to date, involving nearly 100,000 additional participants phenotyped for lipids and genotyped on Metabochip. I report 62 novel genetic loci associated with lipids and through various downstream bioinformatics analyses, provide evidence for the biological relevance of these loci to help inform potential functional follow-up.

Because of the noncoding nature of most trait-associated variants identified by GWAS, including those associated with lipids examined in Chapter II, I hypothesize that a majority are involved in transcriptional regulation rather than altering protein function to induce phenotypic change. I develop the open source tool GREGOR (Genomic Regulatory Elements and Gwas Overlap AlgoRithm) to quantify enrichment of trait-specific GWAS variants in regulatory features. I find evidence of enrichment of lipid-associated variants in regulatory features in liver, and see analogous enrichment of other trait-associated variants in features of biologically relevant tissues. This method, described in Chapter III, gives further insight into the mechanisms of transcriptional regulation by which trait-

associated variants are acting. In addition, I evaluate regulatory feature overlap of linked variants at a set of individual lipid-associated loci to hypothesize the functionality of particular variants, and present experimental results to support my computational predictions.

Lastly, I hypothesize that there are different frequencies of genomic structural variants in myocardial infarction cases compared to controls and apply established and complementary SV detection algorithms to identify and genotype deletions, duplications, and inversions. Chapter IV examines the functional impact of structural variation on MI through whole genome sequencing in a Norwegian sample, and provides the results of genome-wide association testing of SVs for MI status and quantitative lipid traits.

Table 1.1: Contribution of low frequency and rare genetic variation to lipid levels from single variant tests

Locus	Variant annotation	Variant ID	Variant type	MAF (%)	Effect size	Trait[a]	Ethnicity[b]	Reference
ABCA6	p.Cys1359Arg	rs77542162	Missense	2.0	0.220, 0.179	LDL-C, TC	European	Surakka et al. (2015)
ANGPTL4	p.Glu40Lys	rs116843064	Missense	2.9	26.9	TG	Norwegian	Holmen et al. (2014)
ANGPTL8	p.Gln121X	rs145464906	Nonsense	0.01, 0.1	10 mg/dL	HDL-C	AA, EA	Peloso et al. (2014)
APOC3	IVS2+1G→A	rs138326449	Splicing	0.25	-1.43	TG	British	Timpson et al. (2014)
APOC3	p.Arg19X	rs76353203	Nonsense	1.9	1.471, 0.513	HDL-C, TG	Greek	Tachmazidou et al. (2013)
C3orf14	p.Leu33Met	chr3:62307648	Missense	0.06, 0.2	0.648	TG	AA, EA	TG and HDL Working Group of the Exome Sequencing Project, National Heart, Lung, and Blood Institute et al. (2014)
CD300LG	p.Arg82Cys	rs72836561	Missense	2.7	0.23	TG	European	Surakka et al. (2015)
COL14A1	p.Ala1197Thr	chr8:121292281	Missense	0.06, 0.1	0.702	TG	AA, EA	TG and HDL Working Group of the Exome Sequencing Project, National Heart, Lung, and Blood Institute et al. (2014)
COL18A1	p.Gly111Arg	rs114139997	Missense	1.9, 0.003	16%	TG	AA, EA	Peloso et al. (2014)
DSEL	p.Ala124Thr	chr18:65181506	Missense	0.03, 0.2	-0.698	TG	AA, EA	TG and HDL Working Group of the Exome Sequencing Project, National Heart, Lung, and Blood Institute et al. (2014)
FAM175A	p.Arg252Gln	chr4:84384688	Missense	0, 0.2	-0.724	TG	AA, EA	TG and HDL Working Group of the Exome Sequencing Project, National Heart, Lung, and Blood Institute et al. (2014)
FAM179A	p.Val852Ala	chr2:29259543	Missense	0.5, 2	-0.183	TG	AA, EA	TG and HDL Working Group of the Exome Sequencing Project, National Heart, Lung, and Blood Institute et al. (2014)
GIN1	p.Asn515Asp	chr5:102423628	Missense	1, 0.02	-0.371	TG	AA, EA	TG and HDL Working Group of the Exome Sequencing Project, National Heart, Lung, and Blood Institute et al. (2014)
GLI1	p.Arg382Trp	chr12:57863433	Missense	0.03, 0.4	-0.476	TG	AA, EA	TG and HDL Working Group of the Exome Sequencing Project, National Heart, Lung, and Blood Institute et al. (2014)
HIVEP3	p.Arg2001Gln	chr1:41978890	Missense	1.0, 0	0.288	TG	AA, EA	TG and HDL Working Group of the Exome Sequencing Project, National Heart, Lung, and Blood Institute et al. (2014)
LIPC	p.Thr405Met	rs113298164	Missense	0.75	9.6	HDL-C	Norwegian	Holmen et al. (2014)
LIPG	p.Asn396Ser	rs77960347	Missense	1.4	5.8	HDL-C	Norwegian	Holmen et al. (2014)
MYCT1	p.Thr54Ala	chr6:153019197	Missense	0.5, 0	0.586	TG	AA, EA	TG and HDL Working Group of the Exome Sequencing Project, National Heart, Lung, and Blood Institute et al. (2014)
NAV2	p.Thr447Met	chr11:19955322	Missense	0.3, 1.0	-0.254	TG	AA, EA	TG and HDL Working Group of the Exome Sequencing Project, National Heart, Lung, and Blood Institute et al. (2014)
PAFAH1B2	p.Ser161Leu	rs186808413	Missense	0.2, 1.1	3 mg/dL	HDL-C	AA, EA	Peloso et al. (2014)
PCSK7	p.Arg504His	rs142953140	Missense	0.2, 0	17 mg/dL	HDL-C	AA, EA	Peloso et al. (2014)
PCSK9	p.Arg46Leu	rs11591147	Missense	3.2, 0.6	21% decrease	LDL-C	White, Black	Cohen et al. (2006)
PCSK9	p.Cys679X	rs28362286	Nonsense	1.4, <0.1, <0.2	40% decrease	LDL-C	AA, EA, Hisp	Cohen et al. (2005)
PCSK9	p.Tyr142X	rs67608943	Nonsense	0.4	40% decrease	LDL-C	AA	Cohen et al. (2005)
RAE1	p.Pro129Ser	chr20:55941872	Missense	0.1, 0.3	0.563	TG	AA, EA	TG and HDL Working Group of the Exome Sequencing Project, National Heart, Lung, and Blood Institute et al. (2014)
TM6SF2	p.Leu156Pro	rs187429064	Missense	3.6	0.25	TC	European	Surakka et al. (2015)
ZPBP2	p.Lys262Glu	chr17:38031648	Missense	0.03, 0.5	-0.457	TG	AA, EA	TG and HDL Working Group of the Exome Sequencing Project, National Heart, Lung, and Blood Institute et al. (2014)

[a] TG, triglycerides; HDL-C, high-density lipoprotein cholesterol; LDL-C, low-density lipoprotein cholesterol; TC, total cholesterol
[b] AA, African American; EA, European American; Hisp, Hispanic

Table 1.2: Contribution of low frequency and rare genetic variation to lipid levels from burden tests

Locus	**Burden test**[a]	**Variant type**[b]	**Burden frequency (%)**	**Burden effect size**	**Trait**[c]	**Ethnicity**[d]	**Reference**
APOB	CMC	LoF<5%	0.4, 0.1	-1.9, -1.5	LDL-C	EA, AA	Lange et al. (2014)
APOC3	Gene-based	missense, nonsense, or splice-site<1%	0.326, 0.341	-0.55, -0.38	TG	EA, AA	TG and HDL Working Group of the Exome Sequencing Project, National Heart, Lung, and Blood Institute et al. (2014)
LDLR	CMC	NS<0.1%	2.4, 2.8	0.9, 0.6	LDL-C	EA, AA	Lange et al. (2014)
LIPC	Gene-based	NS≤0.5%	0.26		HDL-C	European	Surakka et al. (2015)
NPC1L1	Gene-based	nonsense, splice-site, or frameshift		-13, -12	TC, LDL-C	European	Myocardial Infarction Genetics Consortium Investigators et al. (2014)
PCSK9	CMC	LoF<5%	0, 2.1	NA, -1.2	LDL-C	EA, AA	Lange et al. (2014)
PCSK9	CMC	NS<5%	3.2	-0.6	LDL-C	EA	Lange et al. (2014)
PNPLA5	CMC	NS<0.1%	1, 1.5	0.5, 1.1	LDL-C	EA, AA	Lange et al. (2014)
SLC25A40	Gene-based	missense		0.42	TG	EA + AA	Rosenthal et al. (2013)

[a] CMC, combined multivariate and collapsing method
[b] LoF, loss-of-function; NS, nonsynonymous variants
[c] TG, triglycerides; HDL-C, high-density lipoprotein cholesterol; LDL-C, low-density lipoprotein cholesterol; TC, total cholesterol
[d] AA, African American; EA, European American

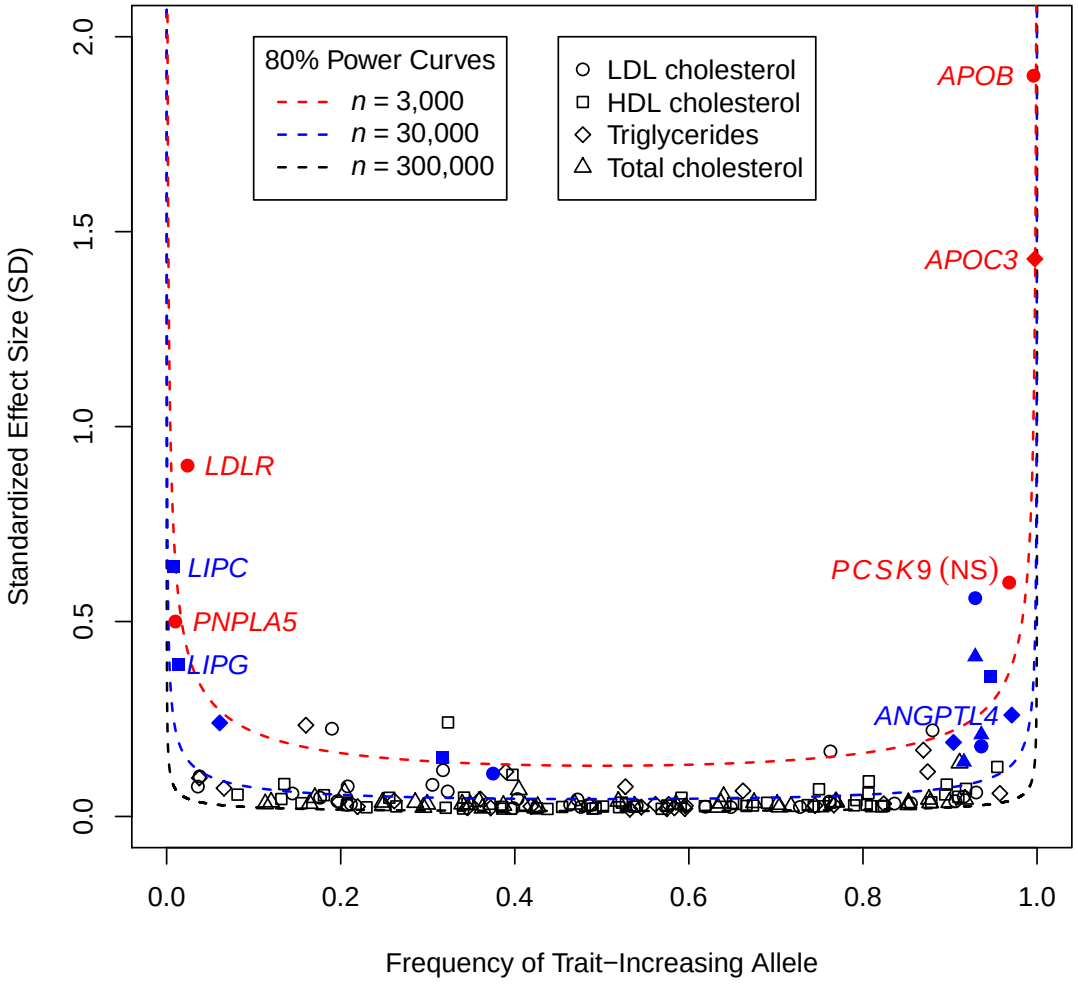


Figure 1.1: Power to detect lipid-associated loci for different study designs. Dotted lines represent 80% power curves and points represent lipid-associated loci from sequencing (Lange et al., 2014; Timpson et al., 2014) (red), exome chip (Holmen et al., 2014) (blue), and GWAS (Global Lipids Genetics Consortium et al., 2013) (black) study designs. Colored loci in red represent standardized effect sizes and burden frequencies estimated from European American or British samples (Lange et al., 2014; Timpson et al., 2014) (red) and Norwegian samples (Holmen et al., 2014) (blue). SD, standard deviation units.

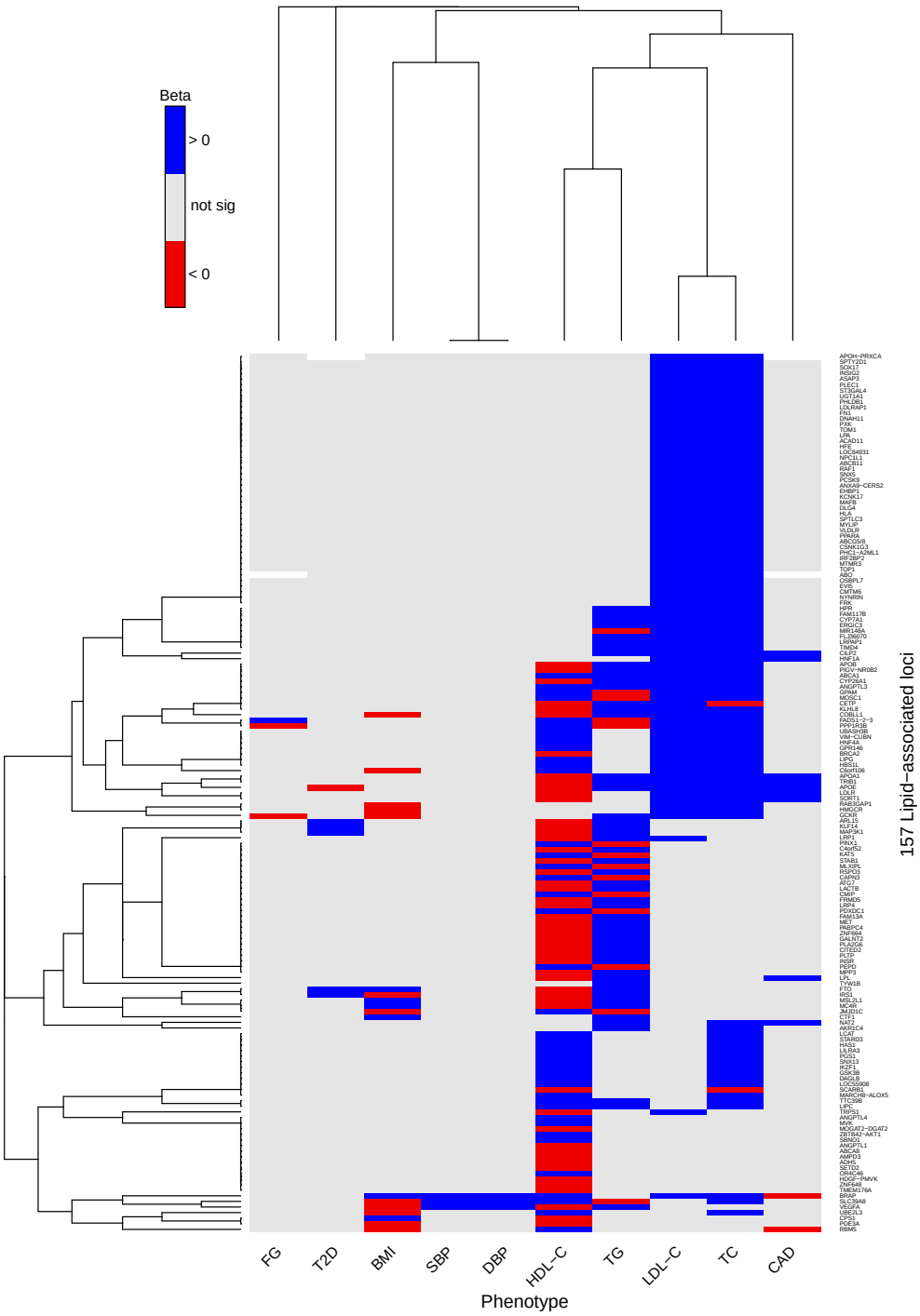


Figure 1.2: Effects of lipid-associated loci on related phenotypes. Effect sizes were obtained from Global Lipids Genetics Consortium et al. (2013) (HDL-C, LDL-C, TC, TG), Locke et al. (2015) (BMI, body mass index), International Consortium for Blood Pressure Genome-Wide Association Studies et al. (2011) (DBP and SBP, diastolic and systolic blood pressure), CARDIoGRAMplusC4D Consortium et al. (2013) (CAD, coronary artery disease), Morris et al. (2012) (T2D, type 2 diabetes), and Scott et al. (2012) (FG, fasting glucose). Blue and red colors represent positive and negative direction of effect, respectively; gray represents not significant after Bonferroni correction for 157 independent lipid loci (P>0.0003); white represents missing data.

CHAPTER II

Metabochip meta-analysis for discovery and refinement of genetic loci associated with plasma lipid levels

2.1 Abstract

Levels of low-density lipoprotein cholesterol (LDL-C), high-density lipoprotein cholesterol (HDL-C), triglycerides (TG) and total cholesterol (TC) are heritable, modifiable risk factors for coronary artery disease. To identify new loci and refine known loci influencing these lipids, we examined 188,577 individuals using genome-wide and custom genotyping arrays. We identify and annotate 157 loci associated with lipid levels at $P<5\text{x}10^{-8}$, including 62 loci not previously associated with lipid levels in humans. Using dense genotyping in individuals of European, East Asian, South Asian and African ancestry, we narrow association signals in 12 loci. We find that loci associated with blood lipid levels are often associated with cardiovascular and metabolic traits, including coronary artery disease, type 2 diabetes, blood pressure, waist-hip ratio and body mass index. Our results demonstrate the value of using genetic data from individuals of diverse ancestry and provide insights into the biological mechanisms regulating blood lipids to guide future genetic, biological and therapeutic research.

Official citation: Global Lipids Genetics Consortium et al. (2013)

2.2 Introduction

Blood lipids are heritable, modifiable risk factors for coronary artery disease (CAD) (Kannel et al., 1961; Castelli, 1988), a leading cause of death (Lloyd-Jones et al., 2010). Human genetic studies of lipid levels can identify targets for new therapies for cholesterol management and the prevention of heart disease and can complement studies in model organisms (Teslovich et al., 2010; Barter and Rye, 2012). Studies of naturally occurring genetic variation can proceed through large-scale association analyses focused on unrelated individuals or through the investigation of mendelian forms of dyslipidemia in families (Rahalkar and Hegele, 2008). We previously identified 95 loci associated with blood lipids, accounting for ∼10-12% of total trait variance (Teslovich et al., 2010), and showed that variants with small effects can indicate pathways and therapeutic targets that enable clinically important changes in blood lipid levels (Teslovich et al., 2010; Musunuru et al., 2010).

Here we report on studies of naturally occurring variation in 188,577 European-ancestry individuals and 7,898 non-European-ancestry individuals. Our analyses identify 157 loci associated with lipid levels at $P<5\text{x}10^{-8}$, including 62 new loci. Thirty of the 62 loci do not include genes implicated in lipid biology by previous literature. We tested lipid-associated SNPs for association with mRNA expression levels, carried out pathway analyses to uncover relationships between loci and compared the locations of lipid-associated SNPs with those of genes and other functional elements in the genome. These results provide direction for biological and therapeutic research into risk factors for CAD.

2.3 Results

2.3.1 New loci associated with blood lipid levels

We examined subjects of European ancestry, including 94,595 individuals from 23 studies genotyped with genome-wide association study (GWAS) arrays (Teslovich et al., 2010) and 93,982 individuals from 37 studies genotyped with the Metabochip array (Voight et al., 2012a) (Figure 2.1). The Metabochip includes variants representing promising loci from our previous GWAS (14,886 SNPs) and from GWAS of other CAD risk factors and related traits (50,459 SNPs), variants from the 1000 Genomes Project (1000 Genomes Project Consortium et al., 2010) and focused resequencing (Sanna et al., 2011) efforts in 64 previously associated loci (28,923 SNPs) and fine-mapping variants in 181 loci associated with other traits (93,308 SNPs). In cases where Metabochip and GWAS array data were available for the same individuals, we used Metabochip data to ensure that key variants were directly genotyped rather than imputed.

We excluded individuals known to be on lipid-lowering medications and evaluated the additive effect of each SNP on blood lipid levels after adjusting for age and sex. Genomic control values (Devlin and Roeder, 1999) for the initial meta-analyses were 1.10-1.15, low for a sample of this size, indicating that population stratification should have had only a minor impact on our results (Figure 2.2). After genomic control correction, 157 loci associated with blood lipid levels were identified ($P<5\text{x}10^{8}$), including 62 newly associated loci (Figure 2.3, Tables 2.1, 2.2, 2.3, and 2.4). Loci were >1 Mb apart and nearly independent ($r^2<0.10$). Of the 62 newly associated loci, 24 demonstrated the strongest evidence of association with HDL cholesterol levels, 15 demonstrated the strongest evidence of association with LDL cholesterol levels, 8 demonstrated the strongest evidence of association with triglyceride levels, and 15

demonstrated the strongest evidence of association with total cholesterol (Figure 2.4). Several of these loci were validated by a similar extension based on published Global Lipids Genetics Consortium GWAS results (Asselbergs et al., 2012).

The effects of newly identified loci were generally smaller than in earlier GWAS (Figure 2.5). For the 62 newly identified variants, trait variance explained in the Framingham offspring was 1.6% for HDL cholesterol levels, 2.1% for triglyceride levels, 2.4% for LDL cholesterol levels and 2.6% for total cholesterol levels.

2.3.2 Overlap of genetic discoveries and previous knowledge

To investigate connections between our new loci and known lipid biology, we first catalogued genes within 100 Kb of the peak associated SNPs and searched PubMed and Online Mendelian Inheritance in Man (OMIM) for occurrences of these gene names and their aliases in the context of relevant keywords. After manual curation, we identified at least 1 strong candidate in 32 of the 62 loci (52%) (Tables S2.1 and S2.2). For the remaining 30 loci, we found no literature support for the role of a nearby gene in regulating blood lipid levels. This search highlighted genes whose connections to lipid metabolism have been extensively documented in mouse models (such as *VLDLR* and *LRPAP1* (Welch et al., 1996)) and human cell lines (such as *VIM* (Sarria et al., 1992)), as well as candidates whose connection to lipid levels is more recent, such as *VEGFA*. With respect to the latter, recent studies of *VEGFB* have suggested that vascular endothelial growth factors have an unexpected role in the targeting of lipids to peripheral tissues (Hagberg et al., 2010), which we corroborate by associating variants near *VEGFA* with blood triglyceride and HDL cholesterol levels.

Multiple types of evidence supported several literature-identified candidates (Tables S2.3 and S2.4. For example, VLDLR is categorized by Gene Ontology (GO)

(Ashburner et al., 2000) in the retinoid X nuclear receptor (RXR) activation pathway, which also includes genes (*APOB*, *APOE*, *CYP7A1*, *APOA1*, *HNF1A* and *HNF4A*) in previously implicated loci (Teslovich et al., 2010). However, because these additional sources of evidence build on overlapping knowledge, they are not truly independent.

To estimate the probability of finding ≥32 literature-supported candidates after automated search and manual review of results, we repeated our text-mining literature search using 100 permutations of SNPs matched for allele frequency, distance to the nearest gene and number of proxies in linkage disequilibrium (LD). To approximate manual curation of the text-mining results, we focused on genes implicated by 3 or more publications (25 in observed data, 8.7 on average in control SNP sets, P=8x10^{8}).

2.3.3 Pathway Analyses

We performed a gene set enrichment analysis, using MAGENTA (Segrè et al., 2010) to evaluate the over-representation of biological pathways among associated loci. Across the 157 loci, MAGENTA identified 71 enriched pathways. These pathways included at least 1 gene in 20 of our newly identified loci. Examples included *DAGLB* (connected to previously associated loci by genes in the triglyceride lipase activity pathway), *INSIG2* (connected to previously associated loci by the cholesterol and steroid metabolic process pathways), *AKR1C4* (connected to previously associated loci by the steroid metabolic process and bile acid biosynthesis pathways), *VLDLR* (connected to previously associated loci by the retinoic X receptor activation and lipid transport pathways, among others) and *PPARA*, *ABCB11* and *UGT1A1* (three genes assigned to pathways implicated in the activation of nuclear hormone receptors, which have an important role in lipid metabolism through the transcrip-

tional regulation of genes in sterol metabolic pathways (Fitzgerald et al., 2002)). Of the 16 loci where literature review and pathway analysis both suggested a candidate, the predictions overlapped 14 times (Tables S2.3 and S2.4; by chance, we expected 6.6 overlapping predictions; P=1x10^{5}).

2.3.4 Protein-protein interactions

We assessed evidence for physical interactions between proteins encoded near our associated SNPs using DAPPLE (Rossin et al., 2011). We found an excess of direct protein-protein interactions for genes in loci associated with LDL cholesterol levels (ten interactions; P=0.0002), HDL cholesterol levels (eight interactions; P=0.002) and total cholesterol levels (six interactions; P=0.017) but not for triglyceride levels (two interactions; P=0.27). Most of the interactions involved genes at known loci (such as the interaction network connecting *PLTP*, *APOE*, *APOB* and *LIPC*) or highlighted the same genes as the literature and pathway analyses (such as those connecting *VLDLR*, *APOE*, *APOB*, *CETP* and *LPL*). Among the new loci, we identified a link between *AKT1* and *GSK3B*. *GSK3B* has been shown to have a role in energy metabolism (Plyte et al., 1992), and its activity is regulated by *AKT1* through phosphorylation (Toker and Cantley, 1997). Literature review also supported a role in the regulation of blood lipid levels for these two genes.

2.3.5 Regulation of gene expression by associated variants

Many variants associated with complex traits act through the regulation of gene expression. We examined whether our 62 newly identified variants were associated with the expression levels of nearby genes in liver, omental fat or subcutaneous fat. Fifteen variants were associated with the transcript levels of a nearby gene at a significance of P<5x10^{-8} (Table S2.5), and seven lipid-associated variants were in

strong LD (r^2>0.8) with the strongest expression quantitative trait locus (eQTL) for the region (r^2>0.8). In three of these loci, literature searches also prioritized candidate genes. In all three, eQTL analysis and literature review identified the same candidate (*DAGLB*, *SPTLC3* and *PXK*; P=0.05). For the remaining four loci (near *RBM5*, *ADH5*, *TMEM176A* and *GPR146*), analysis of expression levels identified candidates that were not supported by literature or pathway analyses.

2.3.6 Coding variation

In some loci where previous association studies of coding variants were inconclusive, we now found convincing evidence of association, demonstrating the benefits of the large sample sizes achievable through collaboration. For example, in the *APOH* locus (Kaprio et al., 1991), our most strongly associated variant was rs1801689 (APOH p.Cys325Gly; P=1x10^{-11} for LDL cholesterol levels). Overall, at 15 of the 62 new loci, there was at least 1 nonsynonymous variant within 100 kb of and in strong LD (r^2>0.8) with the index SNP (Table S2.6) (18 loci when there was no restriction on distance). This ~30% overlap between associated loci and coding variation is similar to that for other complex traits (1000 Genomes Project Consortium et al., 2010). Unexpectedly, in the 11 loci where a candidate was suggested by literature review and by examination of coding variation, the candidates from these methods coincided 7 times (P=0.03 compared to the expected overlap by chance of 3.8 times); thus, agreement between literature review and examination of coding variation was less significant than for eQTL studies and analyses of pathways or protein-protein interactions.

2.3.7 Overlap between association signals and regulators of transcription in liver

Despite our efforts, 18 of the 62 newly identified loci remain without prioritized candidate genes. The liver is an important hub of lipid biosynthesis, and there is evidence that lipid-associated variants might be associated with changes in gene regulation in liver cells (Ernst et al., 2011). Using Encyclopedia of DNA Elements (ENCODE) data (ENCODE Project Consortium, 2011), we evaluated whether associated SNPs overlapped experimentally annotated functional elements identified in HepG2 cells, a commonly used model of human hepatocytes. To determine significance, we generated 100,000 lists of permuted SNPs matched for minor allele frequency (MAF), distance to the nearest gene and number of SNPs in LD (r^2>0.8) (Section 2.5.12). In HepG2 cells, lipid-associated SNPs were enriched in 8 of the 15 functional chromatin states defined by Ernst et al. (2011) (P<1x10^{-5}; Table S2.7). The strongest enrichment was in regions with 'strong enhancer activity' (3.7-fold enrichment; P=2x10^{-25}; Table S2.8). In the other eight cell types examined by Ernst et al. (2011), no more than three functional chromatin states showed evidence for enrichment (and, when present, enrichment was weaker).

We proceeded to investigate the overlap between lipid-associated loci and functional marks in HepG2 cells in more detail Table S2.8). Notable regulatory elements showing significant overlap with lipid-associated loci included histone marks associated with active regulatory regions (acetylation of histone H3 at lysine 27 (H3K27ac), P=3x10^{-20}; acetylation of histone H3 at lysine 9 (H3K9ac), P=3x10^{-22}), promoters (trimethylation of histone H3 at lysine 4 (H3K4me3), P=2x10^{-15}; dimethylation of histone H3 at lysine 4 (H3K4me2), P=8x10^{-12}), transcribed regions (trimethylation of histone H3 at lysine 36 (H3K36me3), P=4x10^{-14}), indicators of open chromatin (FAIRE (formaldehyde-assisted isolation of regulatory elements), P=5x10^{-9};

DNase I sensitivity, P=2x10^{-4}) and regions that interact with the transcription factors HNF4A (P=6x10^{-10}) and CEBP/B (P=1x10^{-5}). Overall, 56 of our 62 new loci contained at least 1 SNP that overlapped a functional mark (ENCODE Project Consortium, 2011) and/or chromatin state (Ernst et al., 2011) highlighted in Table S2.8, including all but 3 of the loci where no candidates were suggested by literature review or analyses of pathways, coding variation or gene expression.

2.3.8 Initial fine mapping of 65 lipid-associated loci

Previous fine mapping of five LDL cholesterol-associated loci found that variants with the strongest association were often substantially different in frequency and effect size from those identified by GWAS (Sanna et al., 2011). Metabochip genotypes enabled us to carry out an initial fine-mapping analysis for 65 loci: 60 selected for fine mapping on the basis of our previous study (Teslovich et al., 2010) and 5 nominated for fine mapping because of association with other traits.

For each of these loci, we identified the most strongly associated Metabochip variant and evaluated whether it (i) reached genome-wide significant evidence for association (to avoid chance fluctuations in regions where the signal was relatively weak) and (ii) was different from the GWAS index SNP in terms of frequency and effect size (operationalized to r^2<0.8 with the GWAS index SNP). In the European samples, fine mapping identified eight loci where the fine-mapping signal was clearly different from the GWAS signal (Table S2.9). The two largest differences were at the loci near *PCSK9* (top GWAS variant with MAF (f)=0.24, P=9x10^{-24}; fine-mapping variant with f=0.03, P=2x10^{-136}) and *APOE* (GWAS variant f=0.20, P=3x10^{-44}; fine-mapping variant f=0.07, P=3x10^{-651}), consistent with results from Sanna et al. (2011). Large differences were also observed near *LRP4* (GWAS f=0.17, P=8x10^{-14}; fine-mapping f=0.35, P=1x10^{-26}), *IGF2R* (GWAS f=0.16, P=7x10^{-9}; fine-mapping

f=0.37, P=2x10^{-13}), *NPC1L1* (GWAS f=0.27, P=2x10^{-5}; fine-mapping f=0.24, P=1x10^{-12}), *ST3GAL4* (GWAS f=0.26, P=2x10^{-6}; fine-mapping f=0.07, P=6x10^{-11}), *MED1* (GWAS f=0.37, P=3x10^{-5}; fine-mapping f=0.24, P=2x10^{-10}) and *COBLL1* (GWAS f=0.12, P=2x10^{-6}; fine-mapping f=0.11, P=6x10^{-9}). Thus, although the large changes observed by Sanna et al. (2011) after fine mapping are by no means unique, they are not typical. Except for the p.Arg46Leu variant encoded in *PCSK9*, the variants showing the strongest association in fine-mapped loci all had MAF>0.05.

We also attempted fine mapping in samples with African (n=3,263), East Asian (n=1,771) and South Asian (n=4,901) ancestry. Despite comparatively small sample sizes, ancestry-specific analyses identified associated SNPs clearly distinct from the original GWAS variant in five loci (Table S2.9). These loci included *APOE*, consistent with the analyses in individuals of European ancestry, three loci where differences in LD between populations enabled fine mapping in samples of African (*SORT1* and *LDLR*) or East Asian (*APOA5*) ancestry and *CETP*, where an African ancestry-specific variant was present. For *CETP*, *SORT1* and *APOA5*, results are consistent with those of other fine-mapping and functional studies (Musunuru et al., 2010; Buyske et al., 2012; Palmen et al., 2008).

2.3.9 Association of lipid-related loci with metabolic and cardiovascular traits

To evaluate the role of the 157 loci identified here in related traits, we evaluated the most strongly associated SNPs for each locus in genetic studies of CAD (n=114,590 including 37,653 cases) (Schunkert et al., 2011; Coronary Artery Disease (C4D) Genetics Consortium, 2011), type 2 diabetes (T2D; n=47,117 including 8,130 cases), (Voight et al., 2010) body mass index (BMI; n=123,865 individuals) (Speliotes et al., 2010) and waist-hip ratio (WHR; n=77,167 individuals) (Heid et al., 2010), systolic and diastolic blood pressure (SBP and DBP; n=69,395 individuals) (Inter-

national Consortium for Blood Pressure Genome-Wide Association Studies et al., 2011) and fasting glucose levels (n=46,186 non-diabetic individuals) (Dupuis et al., 2010). We observed an excess of SNPs nominally associated (P<0.05) with all these traits, including a 5.1-fold excess for CAD (40 nominally significant loci; P=2x10^{-19}), a 4.1-fold excess for BMI (32 loci; P=1x10^{-11}), a 3.7-fold excesses for DBP (29 loci; P=1x10^{-9}), a 3.4-fold excess for WHR (27 loci; P=1x10^{-9}), a 2.5-fold excess for SBP (20 loci; P=1x10^{-4}), a 2.3-fold excess for T2D (18 loci; P=0.001) and a 2.2-fold excess for fasting glucose levels (17 loci; P=3x10^{-3}). Interestingly, for the new loci, we observed greater overlap with BMI, SBP and DBP (nine overlapping loci each) than with CAD (eight overlapping loci). Of the new loci, the two SNPs showing the strongest association with CAD mapped near *RBM5* (rs2013208: P_{HDL}=9x10^{-12}, P_{CAD}=7x10^{-5}) and *CMTM6* (rs7640978: P_{LDL}=1x10^{-8}, P_{CAD}=4x10^{-4}).

We tested whether the LDL cholesterol-, total cholesterol- or triglyceride- increasing allele or the HDL cholesterol- decreasing allele was associated with increased risk of cardiovascular disease or related metabolic outcomes; the direction of effect of each locus was categorized according to the primary association signal at the locus, as in Tables 2.1, 2.2, 2.3, and 2.4. We observed association with increased CAD risk (104/149; P=1x10^{-6}), SBP (96/155; P=2.7x10^{-3}) and WHR adjusted for BMI (92/154; P=0.019). There were many instances where a single locus was associated with many traits. These included variants near *FTO*, consistent with previous reports (Freathy et al., 2008); near *VEGFA* (associated with triglyceride levels, CAD, T2D, SBP and DBP); near *SLC39A8* (associated with HDL cholesterol levels, BMI, SBP and DBP); and near *MIR581* (associated with HDL cholesterol levels, BMI, T2D and DBP). In some cases, such as *FTO*, a strong association with BMI or another phenotype generated weaker association signals for other metabolic traits

(Freathy et al., 2008). In other cases, such as *SORT1*, a primary effect on lipid levels might mediate secondary association with other traits, such as CAD (Musunuru et al., 2010).

2.3.10 Association of lipid traits with CAD

Epidemiological studies consistently show that high total cholesterol and LDL cholesterol levels are associated with increased risk of CAD, whereas high HDL cholesterol levels are associated with reduced risk of CAD (Clarke et al., 2007). In genetic studies, the connection between LDL cholesterol levels and CAD is clear, whereas the results for HDL cholesterol levels are more equivocal (Willer et al., 2008; Voight et al., 2012b; Frikke-Schmidt et al., 2008). In our data, trait-increasing alleles at the loci showing the strongest association with LDL cholesterol levels (31 loci), triglyceride levels (30 loci) or total cholesterol levels (38 loci) were associated with increased risk of CAD (P=2x10^{-12}, 2x10^{-16} and 0.006, respectively). Conversely, trait-decreasing alleles at loci showing the strongest association with HDL cholesterol levels (64 loci) were associated with increased CAD risk at P=0.02. When we focused on loci uniquely associated with LDL cholesterol levels (12 loci where P>0.05 for other lipids), triglyceride levels (6 loci) or HDL cholesterol levels (14 loci), only the association with LDL cholesterol remained significant (P=0.03).

To better explore how associations with individual lipid levels were related to CAD risk, we used linear regression to test whether association with lipid levels could predict impact on CAD risk. In this analysis, the effect on CAD of 149 lipid-associated loci (CAD results were not available for 8 SNPs) was correlated with LDL cholesterol (Pearson's r=0.74; P=7x10^{-6}) and triglyceride (Pearson's r=0.46; P=0.02) effect sizes but not with HDL cholesterol effect sizes (Pearson's r=-9x10^{-4}; P=0.99; (Figure 2.6). Because most variants affect multiple lipid fractions (Figure 2.3), dissecting

the relationship between lipid level and CAD effects requires multivariate analysis. In a companion manuscript in this issue, we use multivariate analysis and detailed examination of triglyceride-associated loci to show that increased LDL cholesterol and triglyceride levels but not HDL cholesterol levels appear to be causally related to CAD risk (Do et al., 2013).

2.3.11 Evidence for additional loci not yet reaching genome-wide significance

To evaluate evidence for loci not yet reaching genome-wide significance, we compared the directions of effect in GWAS and Metabochip analyses of non-overlapping samples outside the 157 genome-wide significant loci. For independent variants (r^2<0.1) with association P<0.1 in the GWAS-only analysis, a significant excess was concordant in the direction of effect for HDL cholesterol levels (62.9% of 1,847 SNPs; P<1x10^{-16}), LDL cholesterol levels (58.6% of 1,730 SNPs; P<1x10^{-16}), triglyceride levels (59.1% of 1,783 SNPs; P<1x10^{-16}) and total cholesterol levels (61.0% of 1,904 SNPs; P<1x10^{-16}), suggesting that there are many additional loci to be discovered in future studies.

2.4 Discussion

Molecular understanding of the genes and pathways that modify blood lipid levels in humans will facilitate the design of new therapies for cardiovascular and metabolic disease. This understanding can be gained from studies of model organisms, *in vitro* experiments, bioinformatic analyses and human genetic studies. Here we demonstrate association between blood lipid levels and 62 new loci, bringing the total number of lipid-associated loci to 157 (Tables 2.1, 2.2, 2.3, 2.4, and Figure 2.3). All but one of the loci identified here include protein-coding genes within 100 kb of the SNP showing the strongest association. Whereas 38 of the 62 new loci include genes

whose role in the regulation of blood lipid levels is supported by literature review or analysis of curated pathway databases, the remainder include only genes whose role in such regulation has not been documented.

In total, there are 240 genes within 100 kb of 1 of our 62 new lipid-associated loci-providing a daunting challenge for future functional studies. Prioritizing on the basis of literature review, pathway analysis, regulation of mRNA expression levels and protein-altering variants suggests that 70 genes in 44 of the 62 new loci might be the focus of the first round of functional studies (summarized in Tables S2.3 and S2.4). Although we found significant overlap, different sources of prioritization sometimes disagreed. This result suggests that truly understanding causality will be very challenging. We include an interpreted digest of genes highlighted by our study in Table S2.10. Clearly, a range of approaches will be needed to follow up these findings. To illustrate possibilities, consider US Patent Application 20090036394 disclosing that, in the mouse, knockout of *Gpr146* modifies blood lipid levels. Here we show that variants near the human homolog of this gene, *GPR146*, are associated with the levels of total cholesterol-providing an added incentive for studies of GPR146 inhibitors in humans. *GPR146* encodes a G protein-coupled receptor, an attractive pharmaceutical target, so it is tempting to speculate that, one day, pharmaceutical inhibition of GPR146 may modify cholesterol levels and reduce risk of heart disease.

Each associated locus typically includes many strongly associated (and potentially causal) variants. Our fine-mapping results illustrate how genetic analysis of large samples and individuals of diverse ancestry can help focus the search for causal variants. In our fine-mapping analysis of 65 lipid-associated loci, we were able to separate the strongest signal in a region from the previous GWAS-identified signal in 12 instances. In 3 of these 12 instances, fine-mapping was enabled by the analysis

of a few thousand individuals of African or East Asian ancestry, whereas, in the remaining instances, fine mapping was possible through the examination of nearly 100,000 individuals of European ancestry. A more detailed fine-mapping exercise, including imputation of variants from emerging, very large reference panels, may help refine the locations of additional signals.

Lipid-associated loci were strongly associated with CAD, T2D, BMI, SBP and DBP. In univariate analyses, we found that effects on LDL cholesterol and triglyceride levels all predicted association with CAD, but HDL cholesterol levels did not. In a companion paper, more detailed multivariate investigation shows that our data are consistent with the hypothesis that both LDL cholesterol and triglyceride levels but not HDL cholesterol levels are causally related to CAD risk. HDL cholesterol, LDL cholesterol and triglyceride levels summarize aggregate levels of different lipid particles, each with potentially distinct consequences for CAD risk. We evaluated the association of our loci with lipid subfractions in 2,900 individuals from the Framingham Heart Study (Figure S2.1 and Table S2.11) and with sphingolipids, which are components of lipid membranes in cells, in 4,034 individuals from 5 samples of European ancestry (Table S2.12). The results suggest that HDL cholesterol-associated variants can have a markedly different impact on these subphenotypes. For example, among HDL cholesterol-associated loci, variants near *LIPC* were strongly associated with plasmalogen levels ($P<1\text{x}10^{-40}$), variants near *ABCA1* were associated with sphingomyelin levels ($P<1\text{x}10^{-5}$), and variants near *CETP*, which show the strongest association with HDL cholesterol levels overall, were associated with neither of these. Detailed genetic dissection of these subphenotypes in larger samples could lead to functional groupings of HDL cholesterol-associated variants that reconcile the results of genetic studies (which show no clear connection between HDL

cholesterol-associated variants and CAD risk) and epidemiological studies (which show clear association between plasma HDL cholesterol levels and CAD risk).

In summary, we report the largest genetic association study of blood lipid levels yet conducted. The large number of loci identified, the many candidate genes they contain and the diverse proteins they encode generate new leads and insights into lipid biology. It is our hope that the next round of genetic studies will build on these results, using new sequencing, genotyping and imputation technologies to examine rare loss-of-function alleles and other variants of clear functional impact to accelerate the translation of these leads into mechanistic insights and improved treatments for CAD.

2.5 Methods

2.5.1 Samples studied

We collected summary statistics for Metabochip SNPs from 45 studies. Of these, 37 studies consisted primarily of individuals of European ancestry, including both population-based studies and case-control studies of CAD and T2D. Another 8 studies consisted primarily of individuals with non-European ancestry, including 2 studies of individuals of South Asian descent, AIDHS/SDS (n=1,516) and PROMIS (n=3,385); 2 studies of individuals of East Asian descent, CLHNS (n=1,771) and TAI-CHI (n=7,044); and 5 studies of individuals of recent African ancestry, MRC/UVRI GPC (n=1,687) from Uganda, SEY (n=426) from the Caribbean, and FBPP (n=1,614; triglyceride results unavailable), GXE (n=397) and SPT (n=838) from the United States. Each contributing study individually obtained ethics approval for their data generation and analyses.

2.5.2 Genotyping

We genotyped 196,710 genetic variants prioritized on the basis of previous GWAS for cardiovascular and metabolic phenotypes using the Illumina iSelect Metabochip (Voight et al., 2012a) genotyping array. To design the Metabochip, we used our previous GWAS of 100,000 individuals (Teslovich et al., 2010) to prioritize 5,023 SNPs for HDL cholesterol, 5,055 SNPs for LDL cholesterol, 5,056 SNPs for triglycerides and 938 SNPs for total cholesterol. These independent SNPs represent most loci with $P<0.005$ in our original GWAS for HDL cholesterol, LDL cholesterol and triglycerides and with $P<0.0005$ for total cholesterol. An additional 28,923 SNPs were selected for fine mapping of 65 previously identified lipid loci. The Metabochip also included 50,459 SNPs prioritized on the basis of GWAS of non-lipid traits and 93,308 SNPs selected for fine mapping of loci associated with non-lipid traits (5 of these loci were associated with blood lipids by the analyses described here).

2.5.3 Phenotypes

Blood lipid levels were typically measured after >8 hours of fasting. Individuals known to be on lipid-lowering medication were excluded when possible. LDL cholesterol levels were directly measured in ten studies (24% of total study individuals) and were estimated using the Friedewald formula (Friedewald et al., 1972) in the remaining studies. Trait residuals within each study cohort were adjusted for age, age^2 and sex and were then quantile normalized. Explicit adjustments for population structure using principal-component (Price et al., 2006) or mixed-model approaches (Kang et al., 2010) were carried out in 24 studies (35% of study individuals); all studies were adjusted using genomic control before meta-analysis (Devlin and Roeder, 1999). In studies ascertained on diabetes or cardiovascular disease status, cases and

controls were analyzed separately. All meta-analyses were limited to a single ancestry group (for example, European only).

2.5.4 Primary statistical analysis

Individual SNP association tests were performed using linear regression with the inverse normal transformed trait values as the dependent variable and the expected allele count for each individual as the independent variable. These analyses were performed using PLINK (26 samples; 53% of the total number of individuals), SNPTEST (4 samples; 20% of the total number of individuals), EMMAX (9 samples; 14% of the total number of individuals), Merlin (4 samples; 9% of the total number of individuals), GENABEL (1 sample; 3% of the total number of individuals) and MMAP (1 sample; 1% of the total number of individuals).

2.5.5 Meta-analysis

Meta-analysis was performed using the Stouffer method (Stouffer et al., 1949; Willer et al., 2010) with weights proportional to the square root of the sample size for each sample. To correct for inflated test statistics due to potential population stratification, we first applied genomic control to each sample and then repeated the procedure with initial meta-analysis results. For GWAS samples, we used all available SNPs when estimating the median test statistic and inflation factor λ. For Metabochip samples, we used a subset of SNPs (n=7,168) that had P-values of >0.50 for all lipid traits in the original GWAS, expecting that the majority of these would not be associated with lipids and would behave as null variants in the Metabochip samples. Signals were considered to be novel if they reached a P-value of $<5\text{x}10^{-8}$ in the combined GWAS and Metabochip meta-analysis and were >1 Mb away from the nearest previously described lipid-associated locus and other new loci. We used

only European samples for the discovery of new genome-wide significant loci. Non-European samples were used only for meta-analysis and examination of fine-mapping analyses.

2.5.6 Quality control

To flag potentially erroneous analyses, we carried out a series of quality control steps. Average standard errors for association statistics from each study were plotted against study sample size to identify outlier studies. We inspected allele frequencies to ensure all analyses used the same strand assignment of alleles. We evaluated whether reported statistics and allelic effects were consistent with published findings for known loci. Genomic control values for study-specific analyses were inspected, and all were <1.20. Finally, within each study, we excluded variants for which the minor allele was observed <7 times.

2.5.7 Proportion of trait variance explained

We estimated the increase in trait variance explained by new loci in the Framingham cohort (n=7,132) using 3 models for each trait residual: (i) lead and secondary SNPs from the previously published loci (Teslovich et al., 2010); (ii) previously published lipid loci plus newly reported loci; and (iii) newly reported loci. We regressed lipid residuals on these sets of SNPs using the lme kinship package in R.

2.5.8 Initial automated review of the published literature

An initial list of candidates within each locus was generated with Snipper and then subjected to manual review. For each locus, Snipper first generates a list of nearby genes and then checks for the co-occurrence of the corresponding gene names and selected search terms ("cholesterol", "lipids", "HDL", "LDL" or "triglycerides") in published literature and OMIM. We supplemented this approach with traditional

literature searches using PubMed and Google.

2.5.9 Generating permuted sets of non-associated SNPs

To estimate the expected chance overlap between literature searches and our loci, we generated lists of permuted SNPs. To generate these lists, we first identified all non-associated lipid-related SNPs (P>0.10 for any of the four lipid traits) and created bins on the basis of three statistics: MAF, distance to the nearest gene and number of SNPs with r^2>0.8. For each index SNP, we identified 500 non-lipid-associated SNPs that fell within the same 3 bins and randomly selected 1 SNP for each permuted list.

2.5.10 Pathway analyses

To investigate whether lipid-associated variants overlapped previously annotated pathways, we used gene set enrichment analysis (GSEA), as implemented in MAGENTA (Segrè et al., 2010) using the meta-analysis of all studies, including GWAS and Metabochip SNPs. Briefly, MAGENTA first assigns SNPs to a given gene when within 110 kb upstream or 40 kb downstream of transcript boundaries. The most significant SNP P-value within this interval is then adjusted for confounders (gene size, marker density and LD) to create a gene association score. When the same SNP is assigned to multiple genes, only the gene with the lowest score is kept for downstream analyses. Subsequently, MAGENTA attaches pathway terms to each gene using several annotation resources, including GO, PANTHER, Ingenuity and KEGG. Finally, the genes are ranked on the basis of their gene association scores, and a modified GSEA test is used to test the null hypothesis that all gene score ranks above a given rank cutoff are randomly distributed with regard to a given pathway term (and compared to multiple randomly sampled gene sets of identical size).

We evaluated enrichment using a rank cutoff of 5% of the total number of genes. A minimum of 10,000 gene set permutations were performed, and up to 1,000,000 permutations were performed for GSEA *P*-values below $1x10^{-4}$.

We used the Disease Association Protein-Protein Link Evaluator package (DAPPLE) to examine evidence for protein-protein interaction networks connecting genes across different lipid-related loci. This analysis included the 62 new loci as well as the 95 previously known loci; we focus our discussion on pathways that included 1 or more genes from new loci.

2.5.11 *Cis*-expression quantitative trait locus analysis

To determine whether lipid-associated SNPs might act as *cis* regulators of nearby genes, we examined association with the expression levels of 39,280 transcripts in 960 human liver samples, 741 human omental fat samples and 609 human subcutaneous fat samples. Tissue samples were collected postmortem or during surgical resection from donors; tissue collection, DNA and RNA isolation, expression profiling and genotyping were performed as described (Keating et al., 2008). MACH was used to obtain imputed genotypes for ~2.6 million SNPs in HapMap release 22 for each of the samples. We examined the correlation between each of the 62 new index SNPs and all transcripts within 500 kb of the SNP position, performing association analyses as previously described (Schadt et al., 2008).

2.5.12 Functional annotation of associated variants

We attempted to identify lipid-associated SNPs that fell in important regulatory domains. We initially created a list of all potentially causal variants by selecting index SNPs at loci identified in this study or in Teslovich et al. (2010). We then selected any variant in strong LD (r^2>0.8 from the 1000 Genomes Project or HapMap

data) with each index SNP. We compared the positions of the index SNPs and their proxies to previously described functional marks (Ernst et al., 2011; ENCODE Project Consortium, 2011). To assess the expected overlap with functional marks, we created 100,000 permuted sets of non-associated SNPs (see Section 2.5.9) and evaluated permuted SNP lists for overlap with functional domains. We estimated a P-value for each functional domain as the proportion of permuted sets with an equal or greater number of loci overlapping functional domains (for large P-values). For small P-values, we used a normal approximation to the empirical overlap distribution to estimate P-values.

2.5.13 Association with lipid subfractions

Lipoprotein fractions in samples from the Women's Genome Health Study (WGHS) (n=23,170) were measured using the LipoProtein-II assay (Liposcience), and Framingham Heart Study Offspring samples (n=2,900) were measured with the LipoProtein-I assay (Liposcience) (Chasman et al., 2009). Additional information on subfraction measurements can be found in Figures S2.1 and S2.2. Log transformations were used for non-normalized traits. All models were adjusted for age, sex and principal components. The genetic association analysis of WGHS used SNP genotypes imputed from the HapMap release 22 CEU (Utah residents of Northern and Western European ancestry) reference panel using MACH. Of the 23,170 WGHS participants, 16,730 were fasting for 8 hours before blood draw (72.2%).

2.5.14 URLs

Summary results for our studies are available. We hope that they will facilitate continued research into the genetics of blood lipid levels and, eventually, help identify improved treatments for CAD. To browse the full result set, go to `http://www.`

sph.umich.edu/csg/abecasis/public/lipids2013/. Snipper, http://csg.sph.umich.edu/boehnke/snipper/; DAPPLE, http://www.broadinstitute.org/mpg/dapple/dapple.php.

2.6 Acknowledgements

We especially thank the more than 196,000 volunteers who participated in our study. Detailed acknowledgement of funding sources is provided in the Global Lipids Genetics Consortium et al. (2013) Supplementary Note.

Table 2.1: New loci primarily associated with HDL cholesterol discovered from joint GWAS and Metabochip meta-analysis

Locus	Markername	Chr.	hg19 position (Mb)	Associated trait(s)	MAF	Minor/ major allele	Effect of A1	Joint n (x 1,000)	Joint *P*-value
PIGV-NR0B2	rs12748152	1	27.14	HDL-C, LDL-C, TG	0.09	T/C	0.051, 0.050, 0.037	187, 173, 178	$1x10^{-15}$, $3x10^{-12}$, $1x10^{-9}$
HDGF-PMVK	rs12145743	1	156.70	HDL-C	0.34	G/T	0.020	181	$2x10^{-8}$
ANGPTL1	rs4650994	1	178.52	HDL-C	0.49	G/A	0.021	187	$7x10^{-9}$
CPS1	rs1047891	2	211.54	HDL-C	0.33	A/C	-0.027	182	$9x10^{-10}$
ATG7	rs2606736	3	11.40	HDL-C	0.39	C/T	0.025	129	$5x10^{-8}$
SETD2	rs2290547	3	47.06	HDL-C	0.20	A/G	-0.030	187	$4x10^{-9}$
RBM5	rs2013208	3	50.13	HDL-C	0.50	T/C	0.025	170	$9x10^{-12}$
STAB1	rs13326165	3	52.53	HDL-C	0.21	A/G	0.029	187	$9x10^{-11}$
GSK3B	rs6805251	3	119.56	HDL-C	0.39	T/C	0.020	186	$1x10^{-8}$
C4orf52	rs10019888	4	26.06	HDL-C	0.18	G/A	-0.027	187	$5x10^{-8}$
FAM13A	rs3822072	4	89.74	HDL-C	0.46	A/G	-0.025	187	$4x10^{-12}$
ADH5	rs2602836	4	100.01	HDL-C	0.44	A/G	0.019	187	$5x10^{-8}$
RSPO3	rs1936800	6	127.44	HDL-C, TG[a]	0.49	C/T	0.020, -0.020	187, 168	$3x10^{-10}$, $3x10^{-8}$
DAGLB	rs702485	7	6.45	HDL-C	0.45	G/A	0.024	187	$6x10^{-12}$
SNX13	rs4142995	7	17.92	HDL-C	0.38	T/G	-0.026	165	$9x10^{-12}$
IKZF1	rs4917014	7	50.31	HDL-C	0.32	G/T	0.022	187	$1x10^{-8}$
TMEM176A	rs17173637	7	150.53	HDL-C	0.12	C/T	-0.036	184	$2x10^{-8}$
MARCH8-ALOX5	rs970548	10	46.01	HDL-C, TC	0.26	C/A	0.026, 0.025	187, 187	$2x10^{-10}$, $8x10^{-9}$
OR4C46	rs11246602	11	51.51	HDL-C	0.15	C/T	0.034	176	$2x10^{-10}$
KAT5	rs12801636	11	65.39	HDL-C	0.23	A/G	0.024	187	$3x10^{-8}$
MOGAT2-DGAT2	rs499974	11	75.46	HDL-C	0.19	A/C	-0.026	187	$1x10^{-8}$
ZBTB42-AKT1	rs4983559	14	105.28	HDL-C	0.40	G/A	0.020	184	$1x10^{-8}$
FTO	rs1121980	16	53.81	HDL-C, TG[b]	0.43	A/G	-0.020, 0.021	186, 155	$7x10^{-9}$, $3x10^{-8}$
HAS1	rs17695224	19	52.32	HDL-C	0.26	A/G	-0.029	185	$2x10^{-13}$

[a] The secondary trait TG was most strongly associated with a different SNP, rs719726 (within 1 Mb of rs1936800, r^2=0.74)

[b] The secondary trait TG was most strongly associated with a different SNP, rs9930333 (within 1 Mb of rs1121980, r^2=0.99)

[*] Effect sizes are given with respect to the minor allele (A1) in SD units. For loci associated with two or more traits at genome-wide significance, the trait corresponding to the strongest *P*-value is listed first.

Table 2.2: New loci primarily associated with LDL cholesterol discovered from joint GWAS and Metabochip meta-analysis

Locus	Markername	Chr.	hg19 position (Mb)	Associated trait(s)	MAF	Minor/ major allele	Effect of A1	Joint n (x 1,000)	Joint *P*-value
ANXA9-CERS2	rs267733	1	150.96	LDL-C	0.16	G/A	-0.033	165	$5x10^{-9}$
EHBP1	rs2710642	2	63.15	LDL-C	0.35	G/A	-0.024	173	$6x10^{-9}$
INSIG2	rs10490626	2	118.84	LDL-C, TC[a]	0.08	A/G	-0.051, -0.042	173, 184	$2x10^{-12}$, $6x10^{-9}$
LOC84931	rs2030746	2	121.31	LDL-C, TC	0.40	T/C	0.021, 0.020	173, 187	$9x10^{-9}$, $4x10^{-8}$
FN1	rs1250229	2	216.30	LDL-C	0.27	T/C	-0.024	173	$3x10^{-8}$
CMTM6	rs7640978	3	32.53	LDL-C, TC	0.09	T/C	-0.039, -0.038	172, 186	$1x10^{-8}$, $2x10^{-8}$
ACAD11	rs17404153	3	132.16	LDL-C, HDL-C[b]	0.14	T/G	-0.034, -0.028	172, 187	$2x10^{-9}$, $5x10^{-9}$
CSNK1G3	rs4530754	5	122.86	LDL-C, TC	0.46	G/A	-0.028, -0.023	173, 187	$4x10^{-12}$, $2x10^{-9}$
MIR148A	rs4722551	7	25.99	LDL-C, TG[c], TC	0.20	C/T	0.039, 0.023, 0.029	173, 178, 187	$4x10^{-14}$, $9x10^{-11}$, $7.0x10^{-9}$
SOX17	rs10102164	8	55.42	LDL-C, TC	0.21	A/G	0.032, 0.030	173, 187	$4x10^{-11}$, $5x10^{-11}$
BRCA2	rs4942486	13	32.95	LDL-C	0.48	T/C	0.024	172	$2x10^{-11}$
APOH-PRXCA	rs1801689	17	64.21	LDL-C	0.04	C/A	0.103	111	$1x10^{-11}$
SPTLC3	rs364585	20	12.96	LDL-C	0.38	A/G	-0.025	172	$4x10^{-10}$
SNX5	rs2328223	20	17.85	LDL-C	0.21	C/A	0.03	171	$6x10^{-9}$
MTMR3	rs5763662	22	30.38	LDL-C	0.04	T/C	0.077	163	$1x10^{-8}$

[a] The secondary trait TC was most strongly associated with a different SNP, rs17526895 (within 1 Mb of rs10490626, r^2=0.98)

[b] The secondary trait HDL-C was most strongly associated with a different SNP, rs13076253 (within 1 Mb of rs17404153, r^2=0.00)

[c] The secondary trait TG was most strongly associated with a different SNP rs4719841 (within 1 Mb of rs4722551, r^2=0.10)

[*] Effect sizes are given with respect to the minor allele (A1) in SD units. For loci associated with two or more traits at genome-wide significance, the trait corresponding to the strongest *P*-value is listed first.

Table 2.3: New loci primarily associated with total cholesterol discovered from joint GWAS and Metabochip meta-analysis

Locus	Markername	Chr.	hg19 position (Mb)	Associated trait(s)	MAF	Minor/ major allele	Effect of A1	Joint n (x 1,000)	Joint *P*-value
ASAP3	rs1077514	1	23.77	TC	0.15	C/T	-0.03	184	$6x10^{-9}$
ABCB11	rs2287623	2	169.83	TC	0.41	G/A	0.027	184	$4x10^{-12}$
FAM117B	rs11694172	2	203.53	TC	0.25	G/A	0.028	187	$2x10^{-9}$
UGT1A1	rs11563251	2	234.68	TC, LDL-C	0.12	T/C	0.037, 0.034	187, 173	$1x10^{-9}$, $5x10^{-8}$
PXK	rs13315871	3	58.38	TC	0.10	A/G	-0.036	187	$4x10^{-8}$
KCNK17	rs2758886	6	39.25	TC	0.30	A/G	0.023	187	$3x10^{-8}$
HBS1L	rs9376090	6	135.41	TC	0.28	C/T	-0.025	187	$3x10^{-9}$
GPR146	rs1997243	7	1.08	TC	0.16	G/A	0.033	183	$3x10^{-10}$
VLDLR	rs3780181	9	2.64	TC, LDL-C	0.08	G/A	-0.044, -0.044	186, 172	$7x10^{-10}$, $2x10^{-9}$
VIM-CUBN	rs10904908	10	17.26	TC	0.43	G/A	0.025	187	$3x10^{-11}$
PHLDB1	rs11603023	11	118.49	TC	0.42	T/C	0.022	187	$1x10^{-8}$
PHC1-A2ML1	rs4883201	12	9.08	TC	0.12	G/A	-0.035	187	$2x10^{-9}$
DLG4	rs314253	17	7.09	TC, LDL-C	0.37	C/T	-0.023, -0.024	184, 170	$3x10^{-10}$, $3x10^{-10}$
TOM1	rs138777	22	35.71	TC	0.36	A/G	0.021	185	$5x10^{-8}$
PPARA	rs4253772	22	46.63	TC, LDL-C[a]	0.11	T/C	0.032, 0.031	185, 171	$1x10^{-8}$, $3x10^{-8}$

[a] The secondary trait LDL-C was most strongly associated with a different SNP, rs4253776 (within 1 Mb of rs4253772, r^2=0.95)

[*] Effect sizes are given with respect to the minor allele (A1) in SD units. For loci associated with two or more traits at genome-wide significance, the trait corresponding to the strongest *P*-value is listed first.

Table 2.4: New loci primarily associated with triglycerides discovered from joint GWAS and Metabochip meta-analysis

Locus	Markername	Chr.	hg19 position (Mb)	Associated trait(s)	MAF	Minor/ major allele	Effect of A1	Joint n (x 1,000)	Joint *P*-value
LRPAP1	rs6831256	4	3.47	TG, TC[a], LDL-C[a]	0.42	G/A	0.026, 0.025, 0.022	177, 187, 173	$2x10^{-12}$, $1x10^{-10}$, $2x10^{-8}$
VEGFA	rs998584	6	43.76	TG, HDL-C	0.49	A/C	0.029, -0.026	175, 184	$3x10^{-15}$, $2x10^{-11}$
MET	rs38855	7	116.36	TG	0.47	G/A	-0.019	178	$2x10^{-8}$
AKR1C4	rs1832007	10	5.25	TG	0.18	G/A	-0.033	178	$2x10^{-12}$
PDXDC1	rs3198697	16	15.13	TG	0.43	T/C	-0.020	176	$2x10^{-8}$
MPP3	rs8077889	17	41.88	TG	0.22	C/A	0.025	176	$1x10^{-8}$
INSR	rs7248104	19	7.22	TG	0.42	A/G	-0.022	176	$5x10^{-10}$
PEPD	rs731839	19	33.90	TG, HDL-C	0.35	G/A	0.022, -0.022	176, 185	$3x10^{-9}$, $3x10^{-9}$

[a] The secondary traits TC and LDL-C were most strongly associated with a different SNP, rs6818397 (within 1 Mb of rs6831256, r^2=0.18)

[*] Effect sizes are given with respect to the minor allele (A1) in SD units. For loci associated with two or more traits at genome-wide significance, the trait corresponding to the strongest *P*-value is listed first.

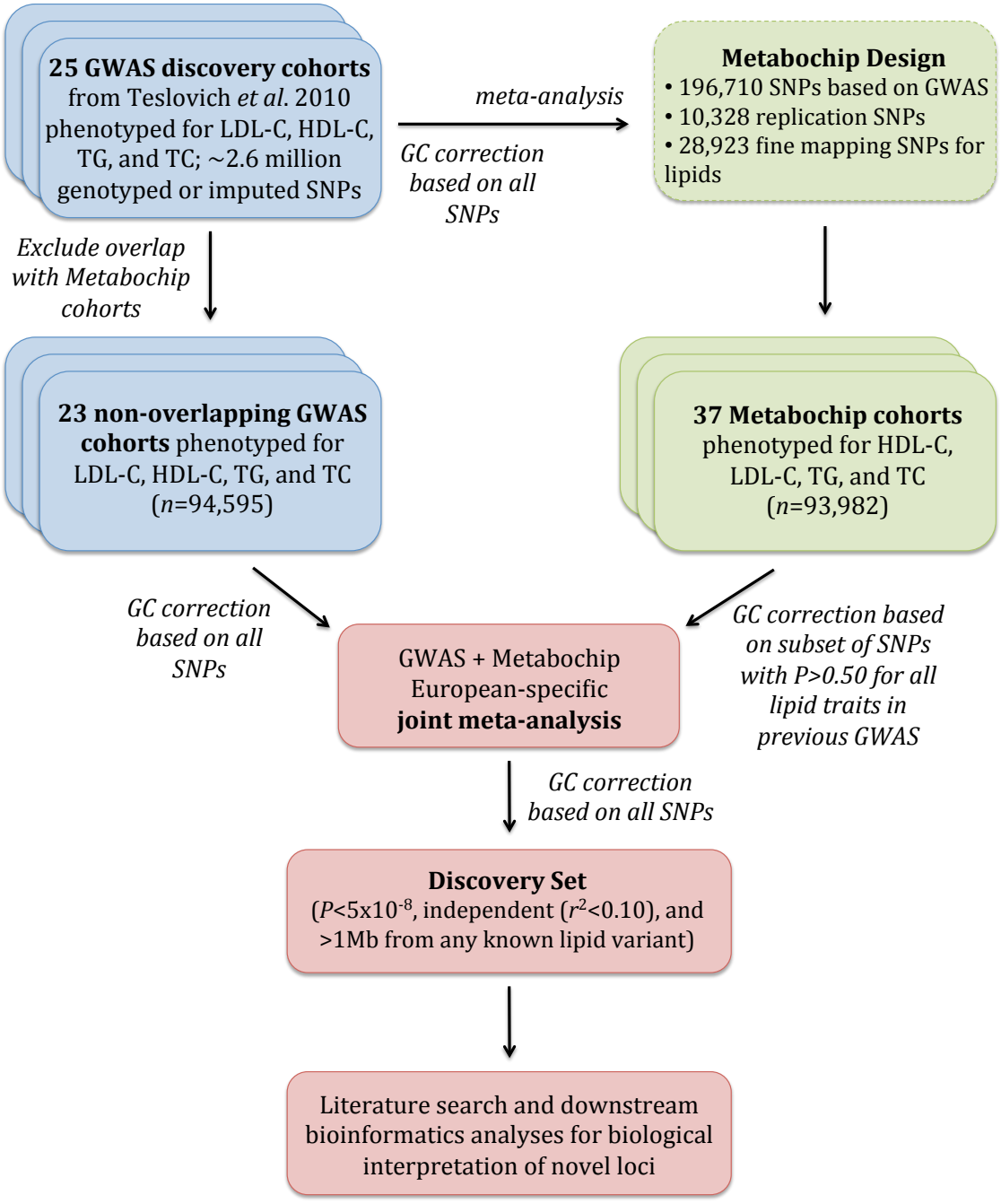
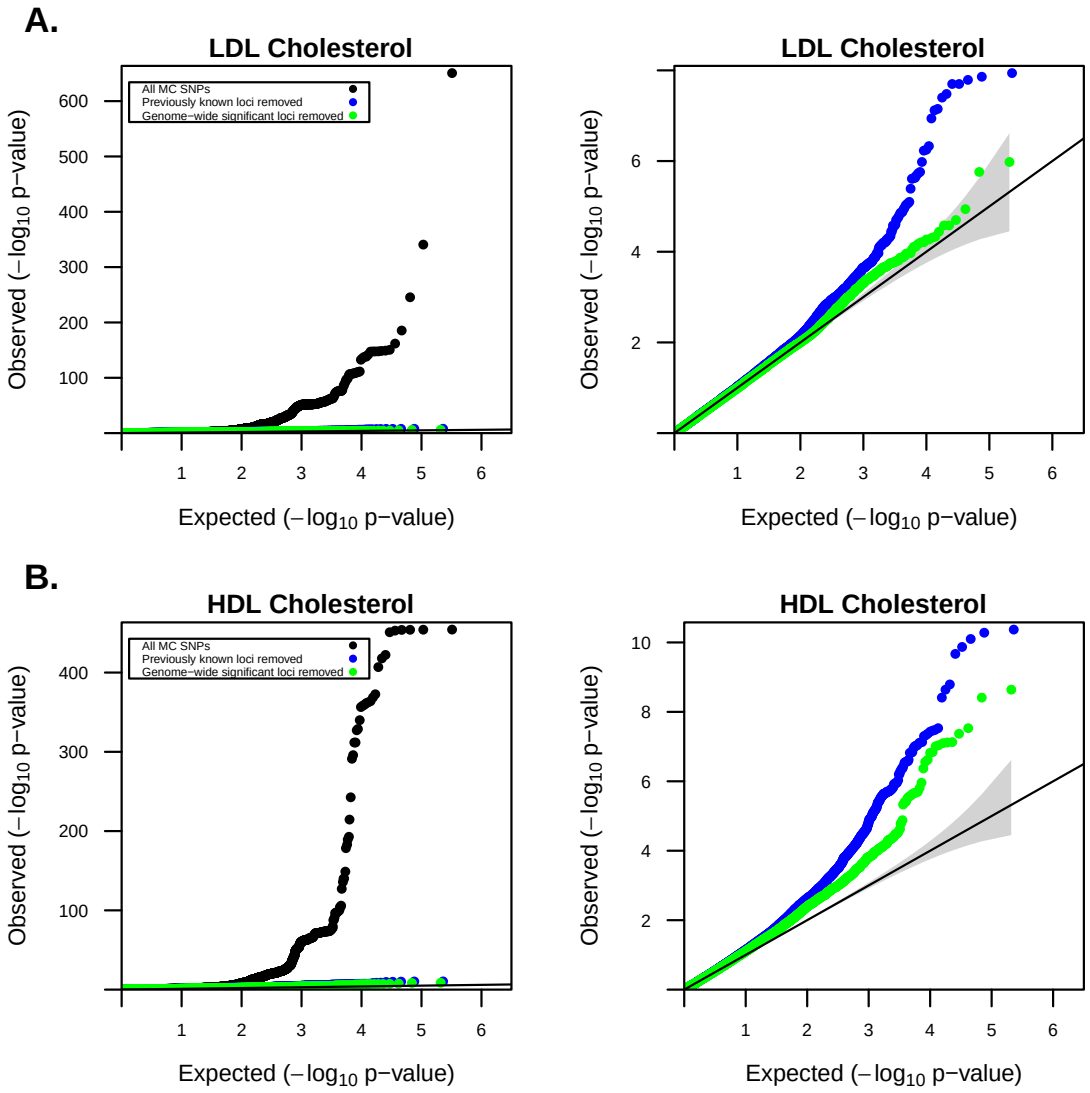
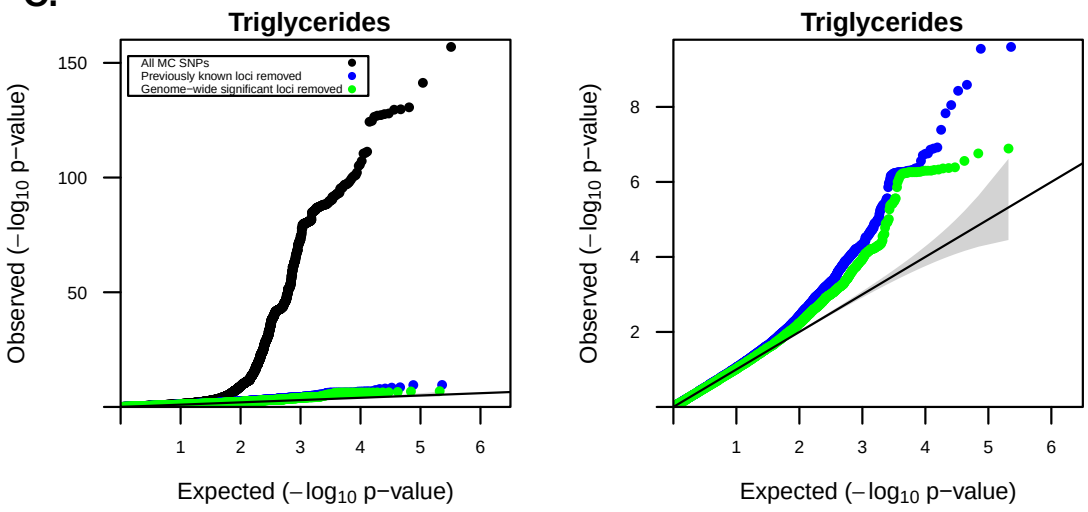


Figure 2.1: GLGC metabochip meta-analysis study design.

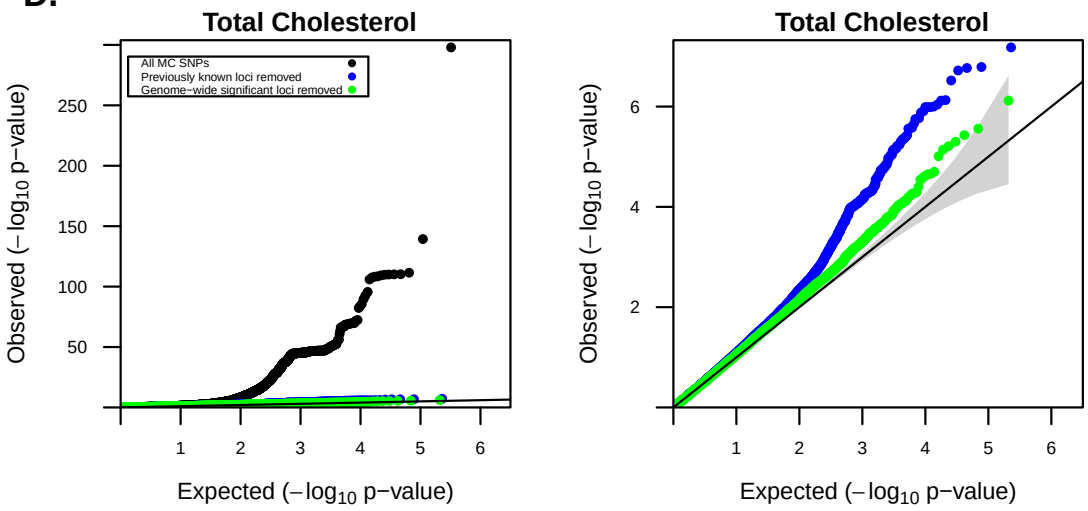
Figure 2.2: Quantile-quantile plots of trait-specific meta-analysis P-value distributions for (A) LDL cholesterol, (B) HDL cholesterol, (C) Triglycerides, and (D) Total cholesterol. Points in blue represent the P-value distribution after removing ±1 Mb of previously known lipid loci. There is reduced inflation of P-values after removing ±1 Mb of all genome-wide significant loci (shown in green). Genomic control lambda (λ_{GC}) values for all Metabochip SNPs were between 1.19 (triglyceride levels) and 1.28 (HDL cholesterol) and reflect the enrichment of associated SNPs in the genotyping array. After removing SNPs within 1 Mb of previously reported associated variants, the lambda values ranged from 1.00 (LDL cholesterol) to 1.10 (HDL cholesterol). After removing SNPs in newly genome-wide significant loci, lambda values reached 1.00 for LDL cholesterol and triglycerides, 1.05 for total cholesterol, and 1.07 for HDL cholesterol.



C.


Triglycerides
All MC SNPs
Previously known loci removed
Genome−wide significant loci removed
Observed (−log10 p−value)
Expected (−log10 p−value)
Triglycerides
Observed (−log10 p−value)
Expected (−log10 p−value)

D.


Total Cholesterol
All MC SNPs
Previously known loci removed
Genome−wide significant loci removed
Observed (−log10 p−value)
Expected (−log10 p−value)
Total Cholesterol
Observed (−log10 p−value)
Expected (−log10 p−value)

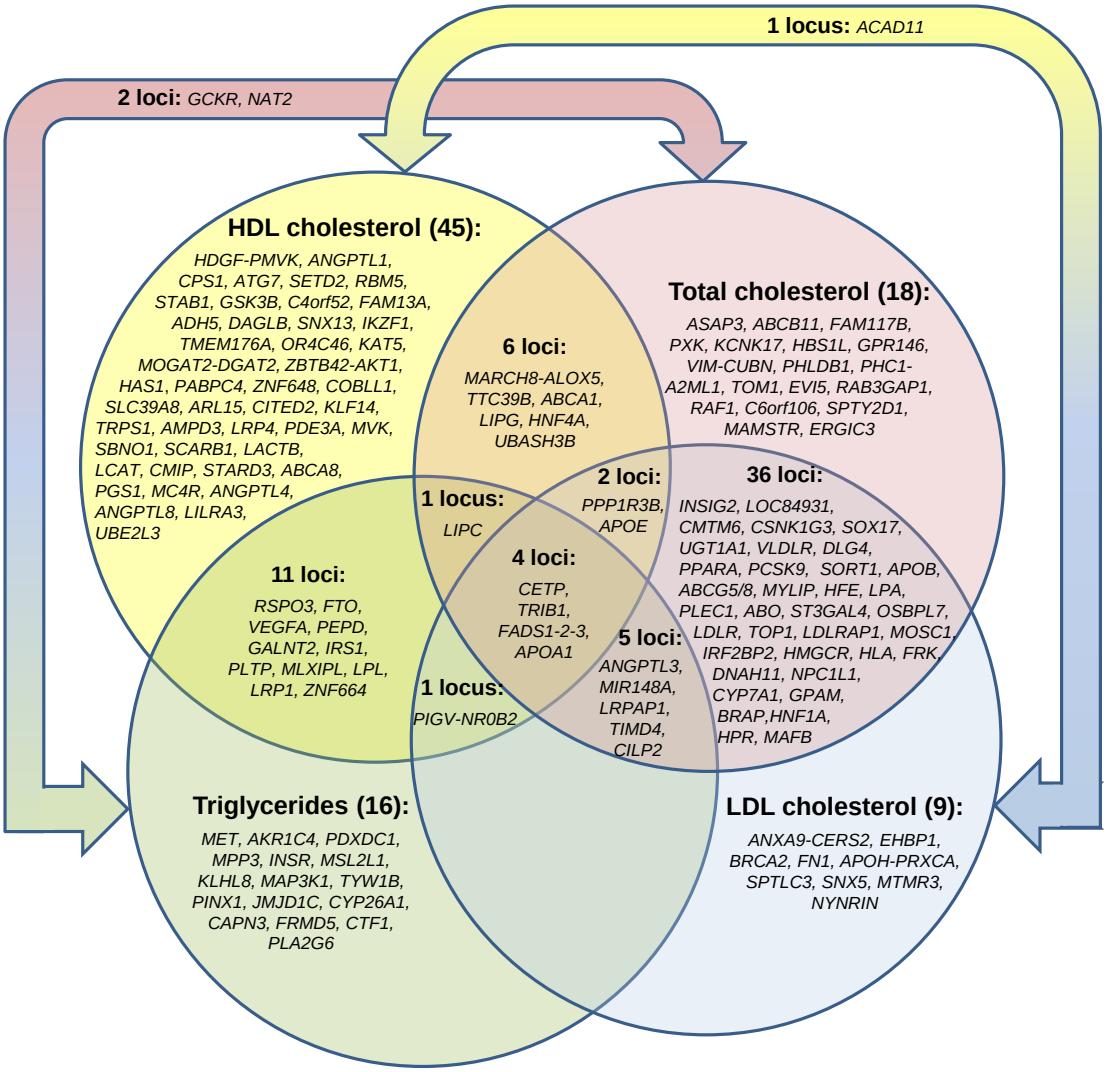
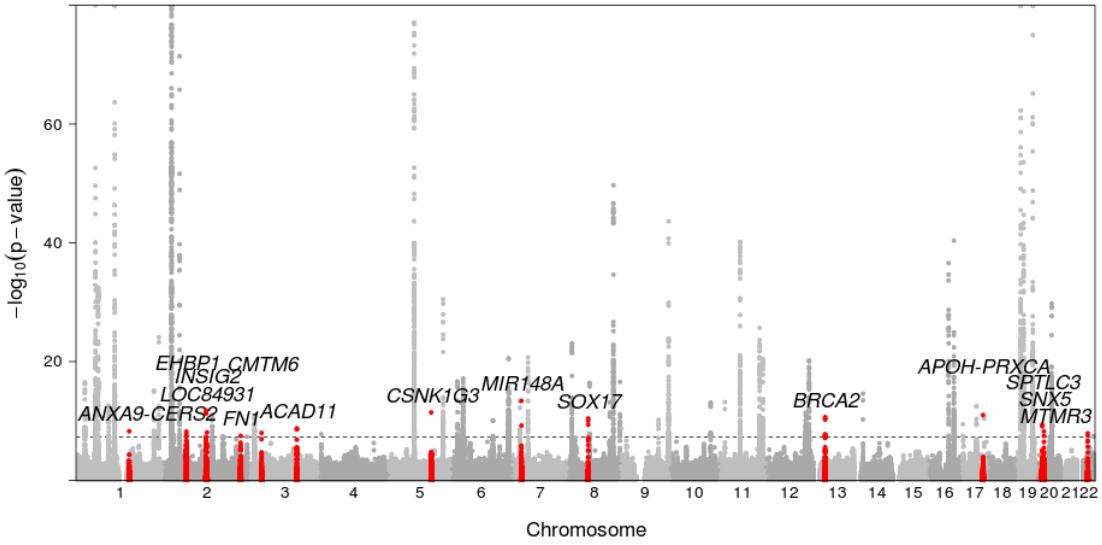


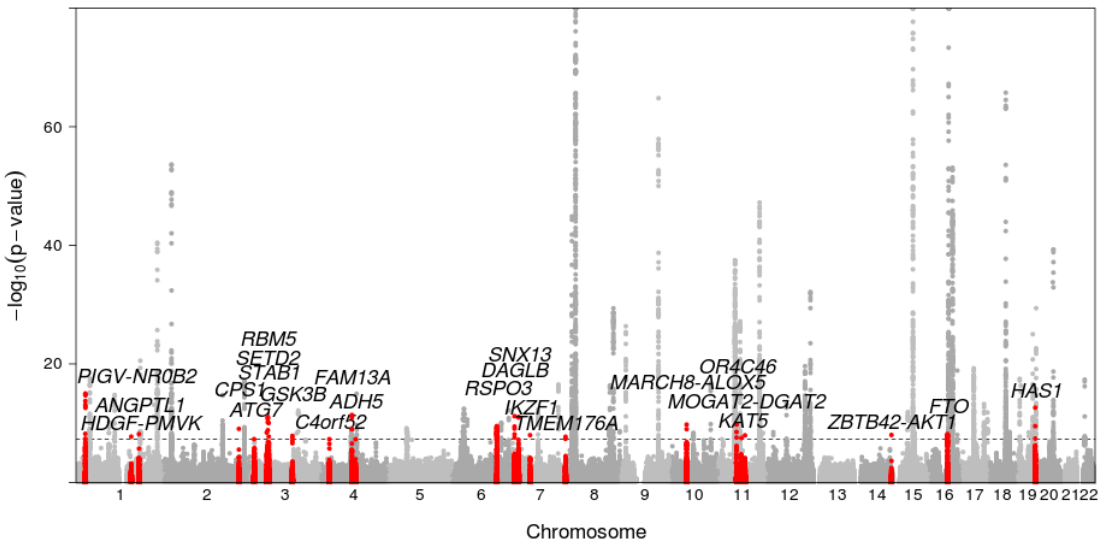
Figure 2.3: Schematic summary of known lipid-associated loci reported from GWAS. The Venn diagram illustrates overlap of genetic loci associated with different lipid traits. The number of loci primarily associated with only one trait is reported in parentheses after the trait name and locus names are listed below in italics. Loci that show association with two or more traits are shown in the appropriate overlapping segments.

Figure 2.4: Manhattan plots highlighting novel genome-wide significant lipid loci. Trait-specific loci that reach genome-wide significance (P<5x10^{-8}) from the European joint meta-analysis are shown in red for (A) LDL cholesterol, (B) HDL cholesterol, (C) Triglycerides, and (D) Total cholesterol. P-values are truncated at 1x10^{-80}.

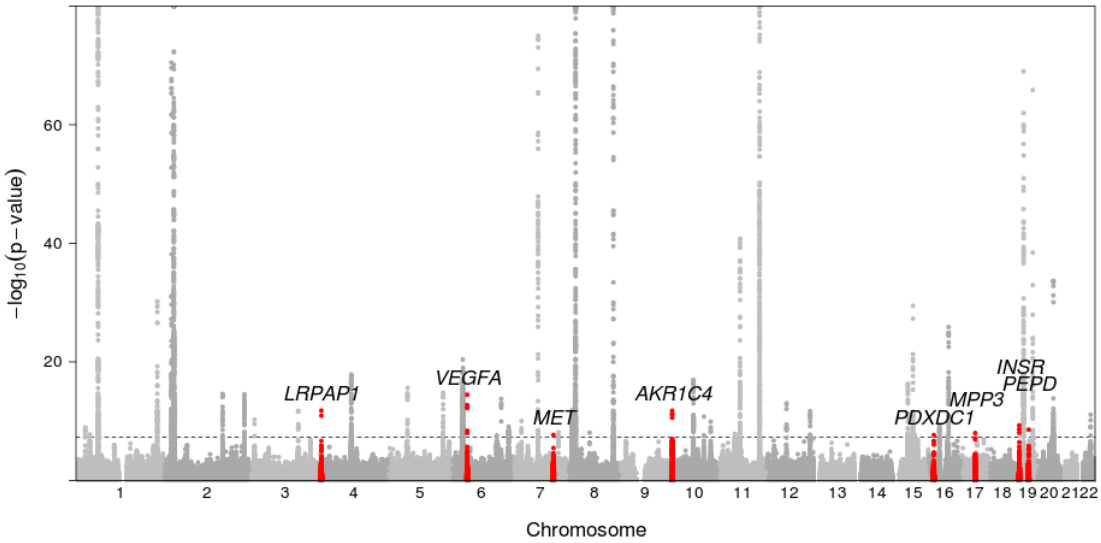
A. LDL Cholesterol



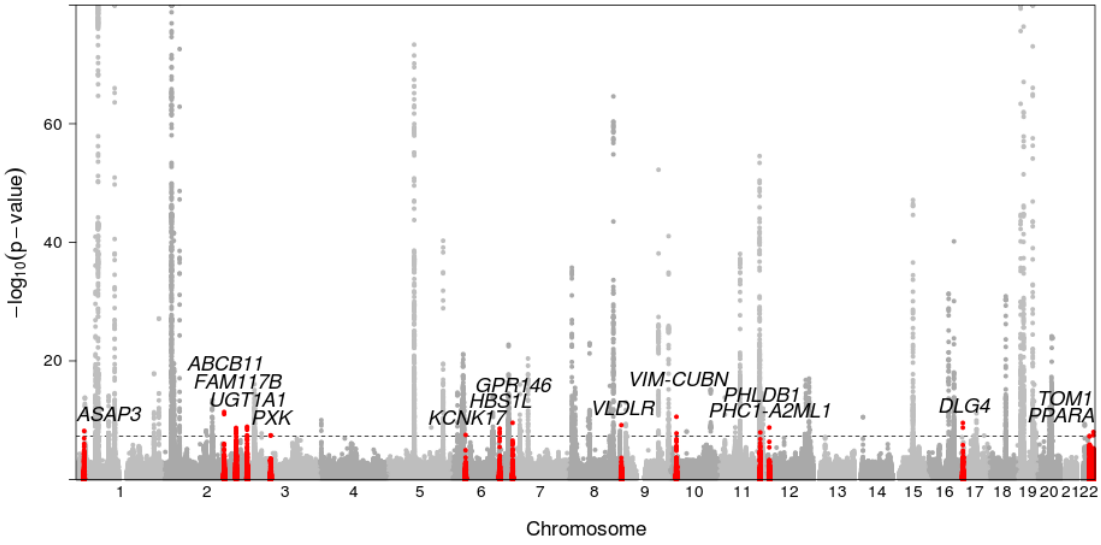
B. HDL Cholesterol



C. Triglycerides



D. Total cholesterol



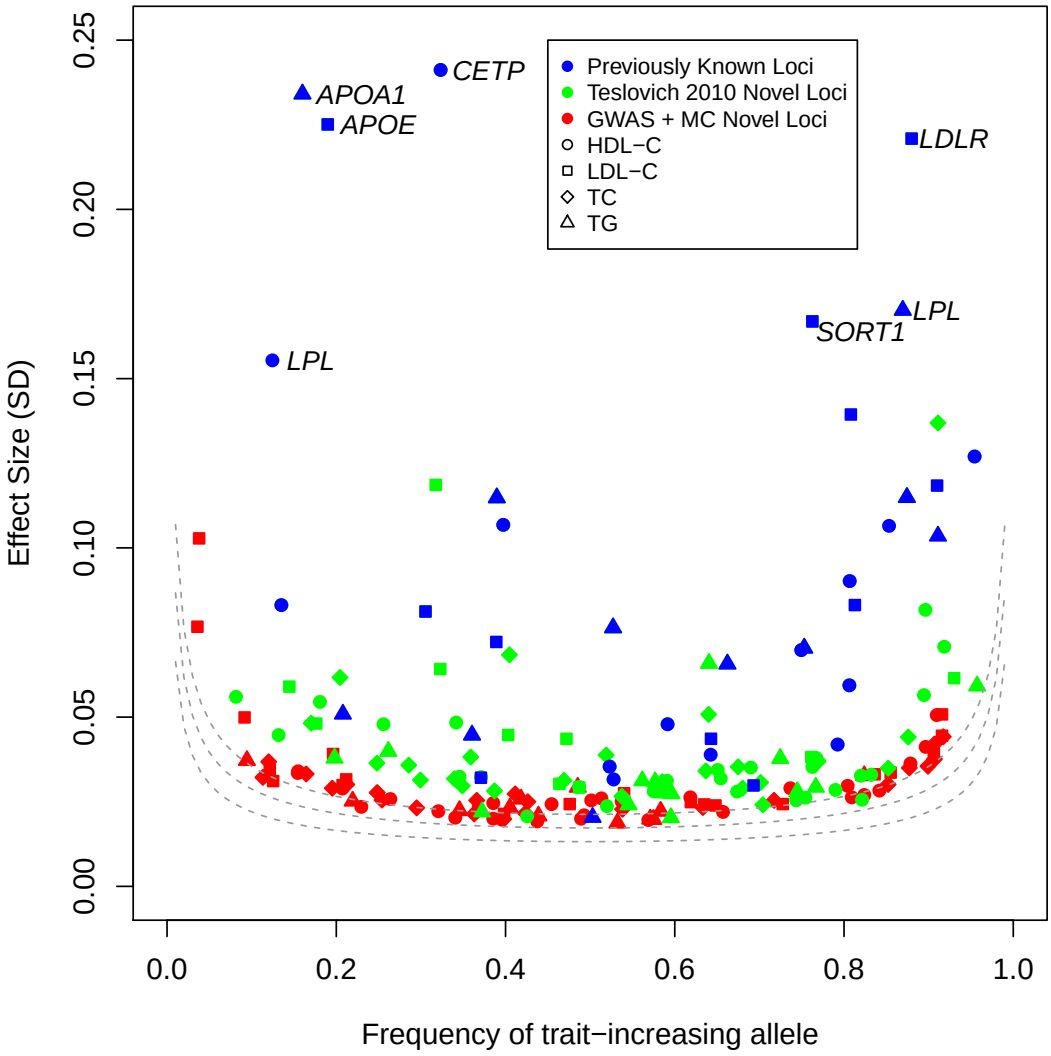


Figure 2.5: Power to detect variants of different allele frequencies and effect sizes. Lipid effect sizes of SNPs in the GWAS + Metabochip meta-analysis are shown in red (novel lipid loci) in comparison to SNPs discovered by previous GWAS efforts (shown in blue and green). Dotted lines represent power curves for the minimum effect sizes that could be identified for a given effect-allele frequency with 10%, 50%, and 90% power, assuming sample size 200,000 and alpha level 5×10^{-8}.

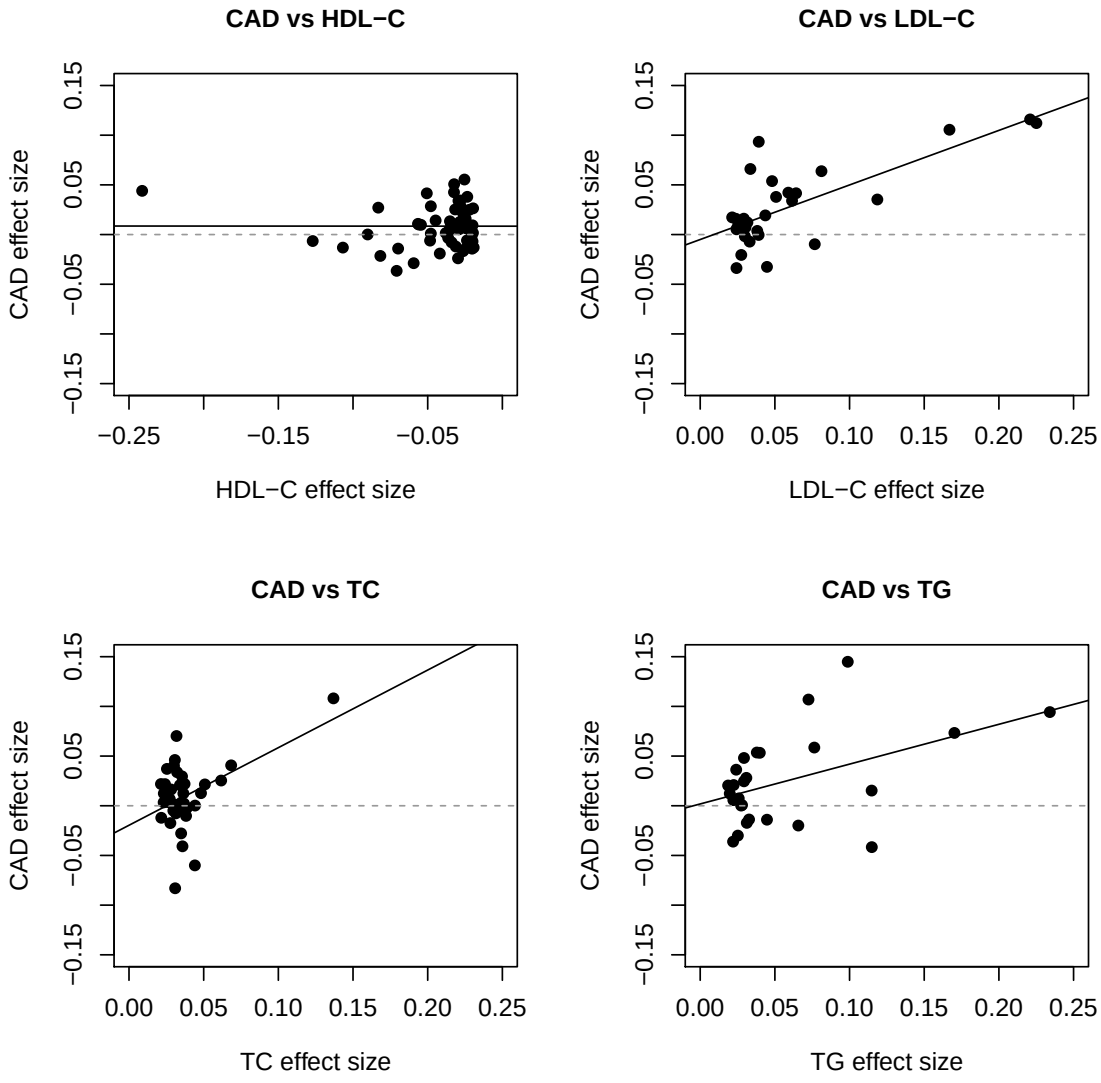
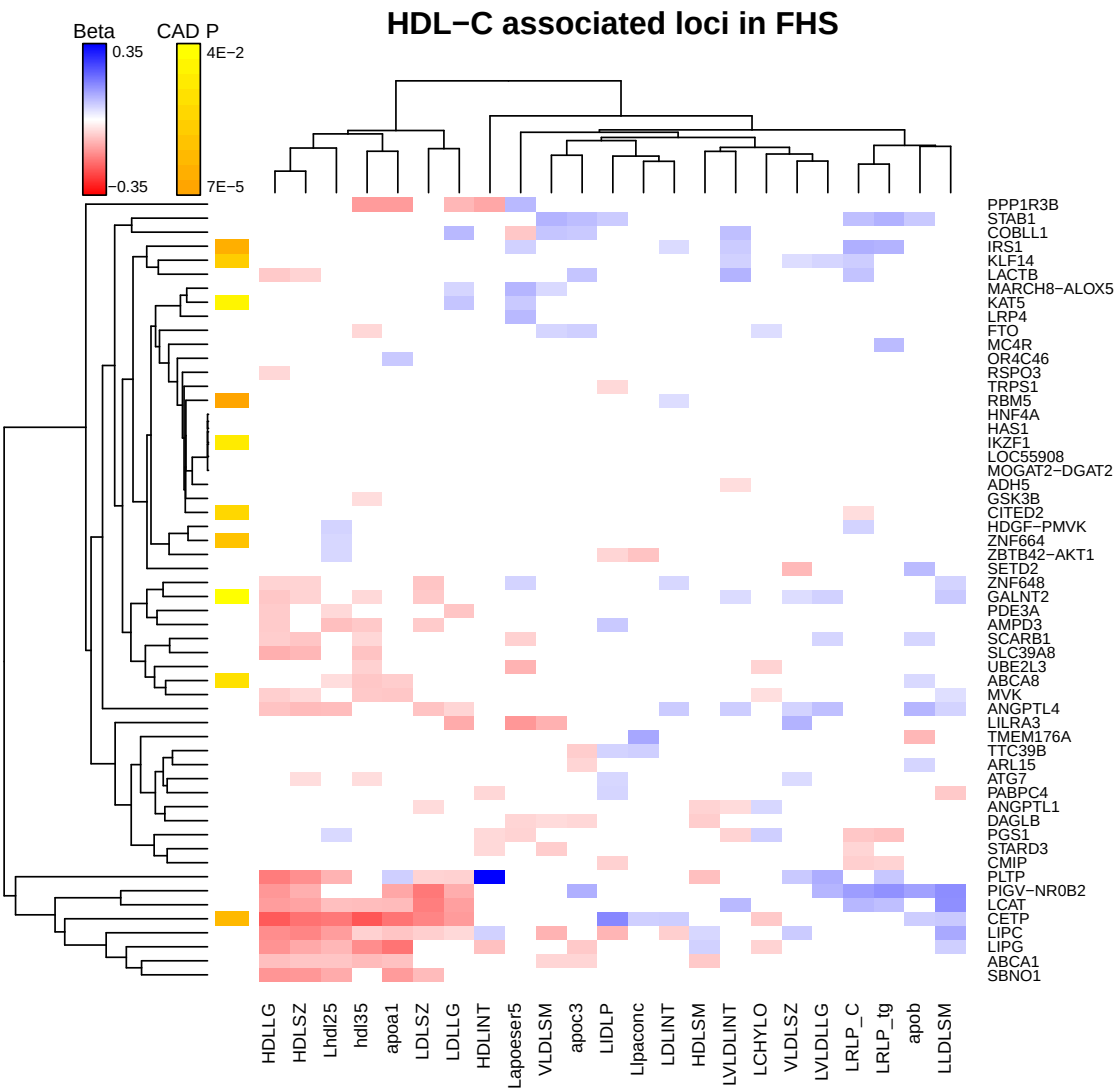


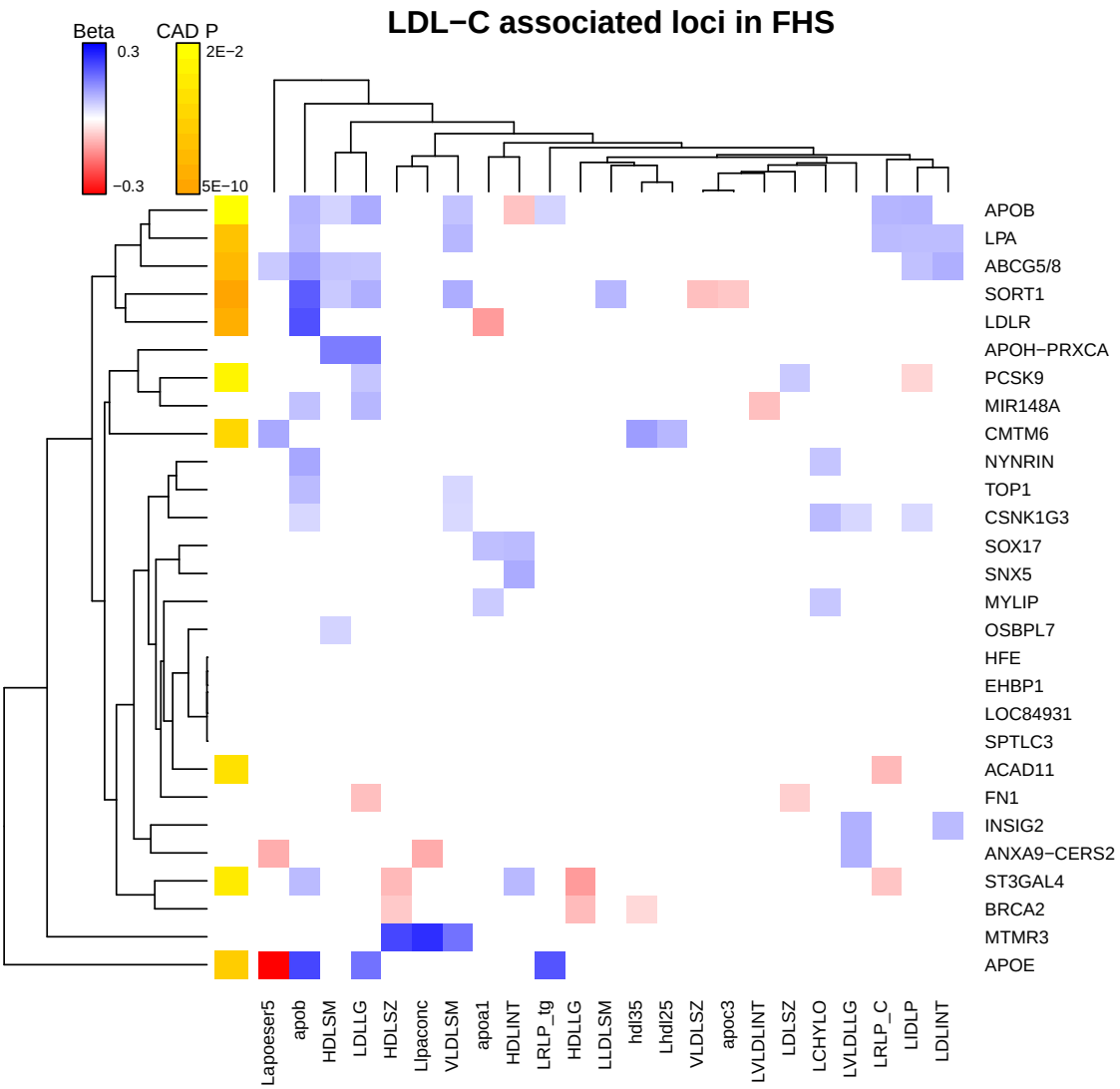
Figure 2.6: Effect size correlations of lipid- and CAD- associated variants. Plots show coronary artery disease (CAD) effect sizes against lipid effect sizes for SNPs showing primary association with each lipid trait. All effect sizes were oriented to the lipid trait-increasing (LDL-C, TG, TC) or trait-decreasing (HDL-C) allele. Diagonal lines represent regressions of predictor lipid effect sizes by outcome CAD effect sizes for SNPs that show primary association with each trait including both previously known and newly reported index SNPs. LDL-C effect sizes were strongly associated with CAD effect sizes (Pearson r=0.74, P=7x10^{-6}). The correlation between CAD effect size and triglyceride effect size (Pearson r=0.46, P=0.02) was higher than that observed for HDL-C (Pearson r=-9x10^{-4}, P=0.99). Lipid effect sizes were transformed into SD units.

Figure S2.1: Association with lipid subfractions in Framingham Heart Study. Heatmaps show effect sizes for association (P<0.10) with 23 lipid subfractions (Chasman et al., 2009) in Framingham Heart Study (FHS) offspring with respect to the trait-decreasing allele of **(A)** HDL-C and trait-increasing allele of **(B)** LDL-C, **(C)** TC, and **(D)** TG. Significant associations (P<0.05) of lipid-associated SNPs with coronary artery disease (CAD) are annotated on the y-axis at both known and novel genetic loci primarily associated with each trait. Dendrogram clustering of loci (y-axis) and lipid subfraction phenotypes (x-axis) based on the effect sizes (beta) are also shown. Figure **(E)** is a heatmap of correlations for the 23 lipid subfractions in Framingham.

A.

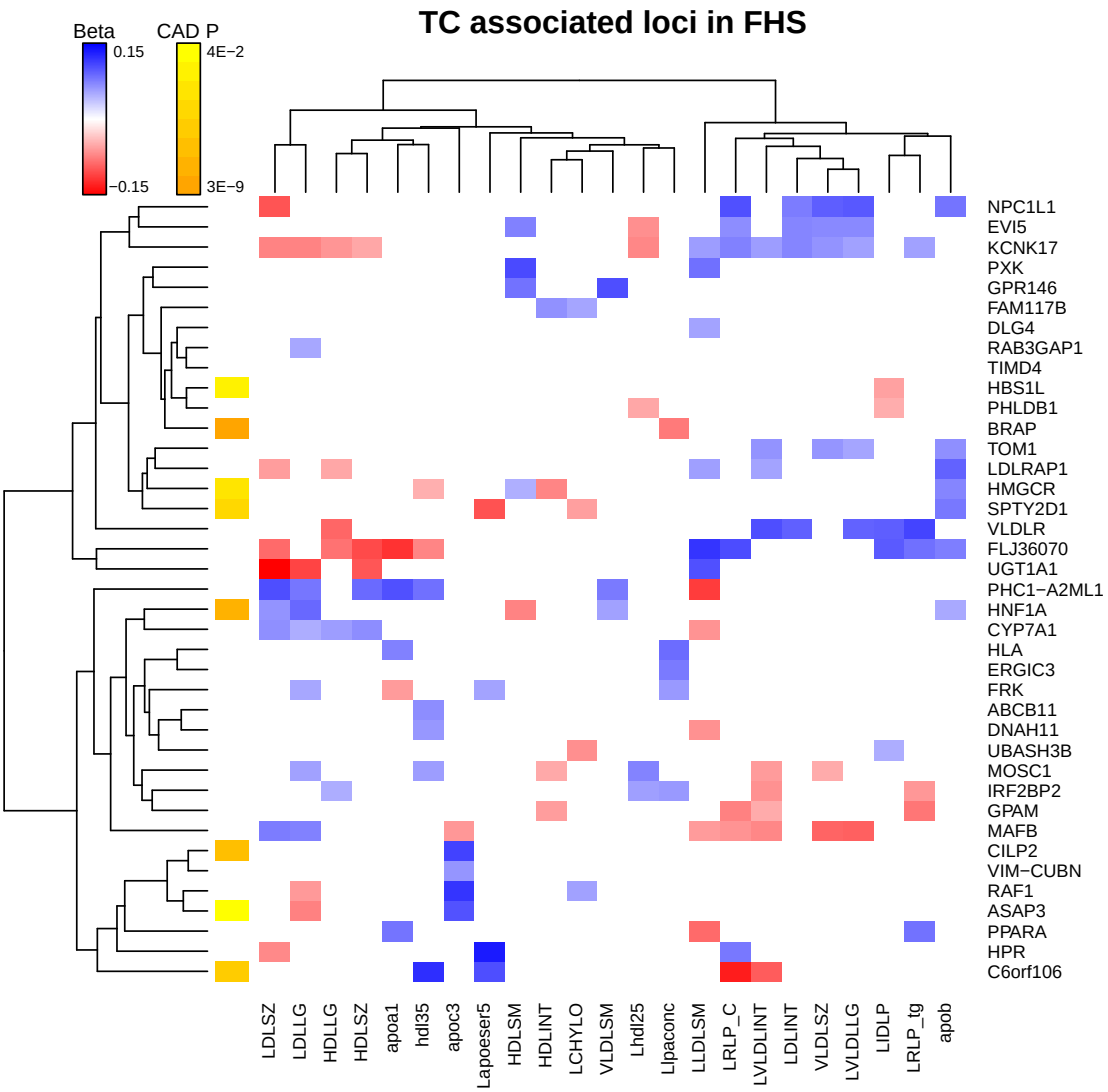


B.

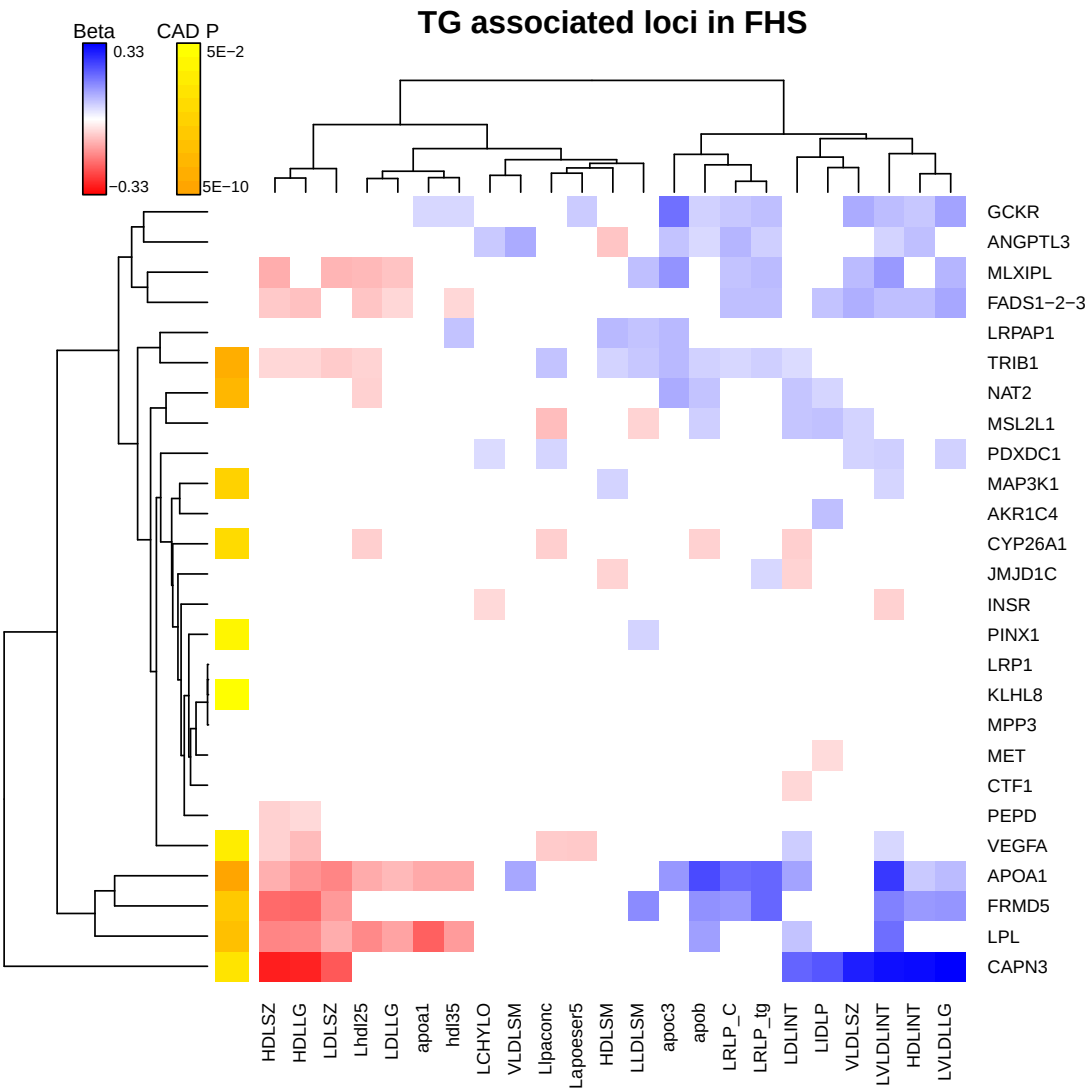


*The beta for the strongest association observed, rs4420638 at the *APOE* locus and Lapoeser5apc (beta=-0.62), is displayed as the minimum (-0.3) so that the color scale for the heatmap is more comparable to the heatmaps from the other 3 lipid traits.

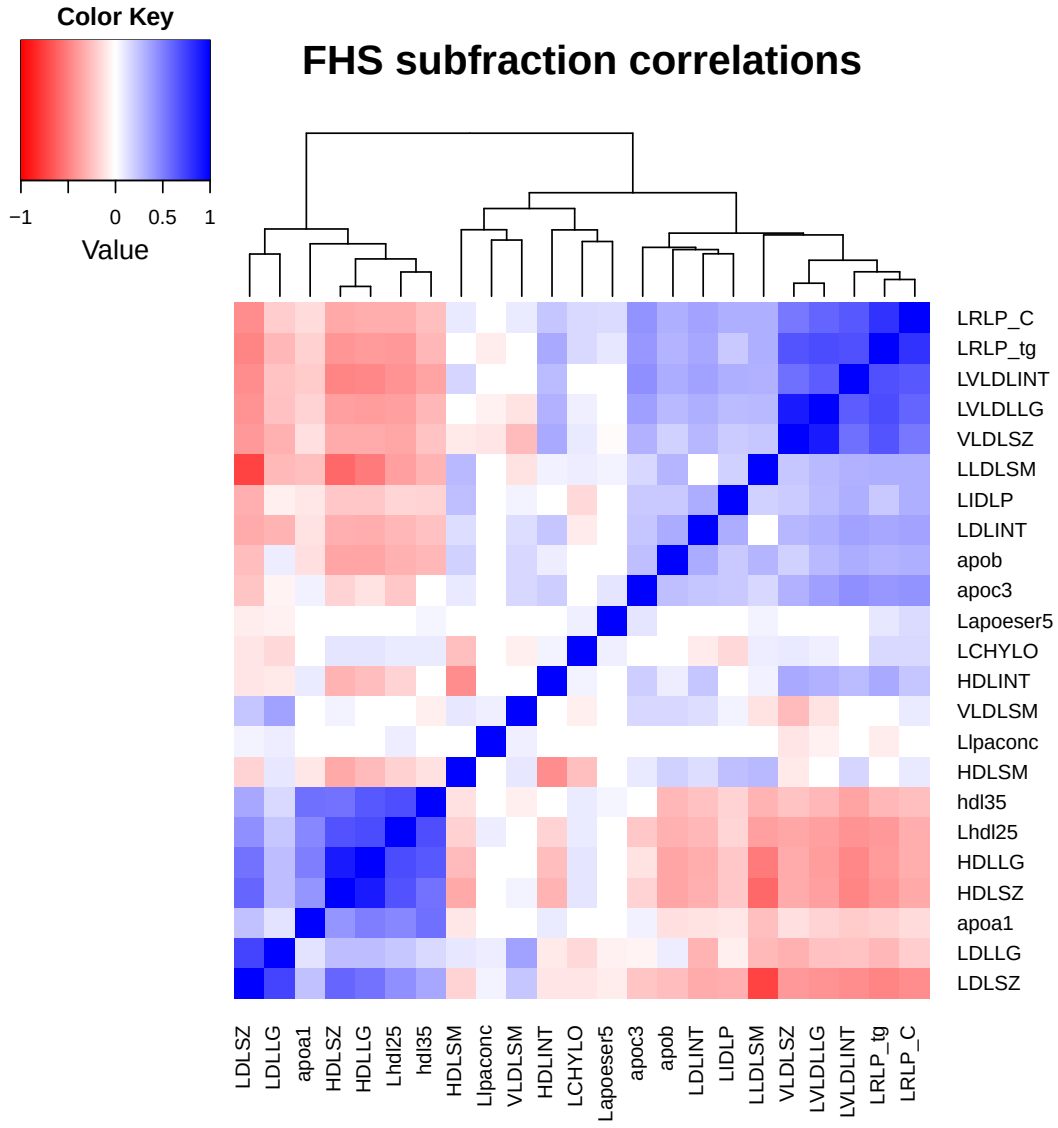
C.


TC associated loci in FHS
Beta
0.15
−0.15
CAD P
4E−2
3E−9
NPC1L1
EVI5
KCNK17
PXK
GPR146
FAM117B
DLG4
RAB3GAP1
TIMD4
HBS1L
PHLDB1
BRAP
TOM1
LDLRAP1
HMGCR
SPTY2D1
VLDLR
FLJ36070
UGT1A1
PHC1−A2ML1
HNF1A
CYP7A1
HLA
ERGIC3
FRK
ABCB11
DNAH11
UBASH3B
MOSC1
IRF2BP2
GPAM
MAFB
CILP2
VIM−CUBN
RAF1
ASAP3
PPARA
HPR
C6orf106
LDLSZ
LDLLG
HDLLG
HDLSZ
apoa1
hdl35
apoc3
Lapoeser5
HDLSM
HDLINT
LCHYLO
VLDLSM
Lhdl25
Llpaconc
LLDLSM
LRLP_C
LVLDLINT
LDLINT
VLDLSZ
LVLDLLG
LIDLP
LRLP_tg
apob

D.


TG associated loci in FHS
Beta
0.33
−0.33
CAD P
5E−2
5E−10
GCKR
ANGPTL3
MLXIPL
FADS1−2−3
LRPAP1
TRIB1
NAT2
MSL2L1
PDXDC1
MAP3K1
AKR1C4
CYP26A1
JMJD1C
INSR
PINX1
LRP1
KLHL8
MPP3
MET
CTF1
PEPD
VEGFA
APOA1
FRMD5
LPL
CAPN3
HDLSZ
HDLLG
LDLSZ
Lhdl25
LDLLG
apoa1
hdl35
LCHYLO
VLDLSM
Lipaconc
Lapoeser5
HDLSM
LLDLSM
apoc3
apob
LRLP_C
LRLP_tg
LDLINT
LIDLP
VLDLSZ
LVLDLINT
HDLINT
LVLDLLG

E.



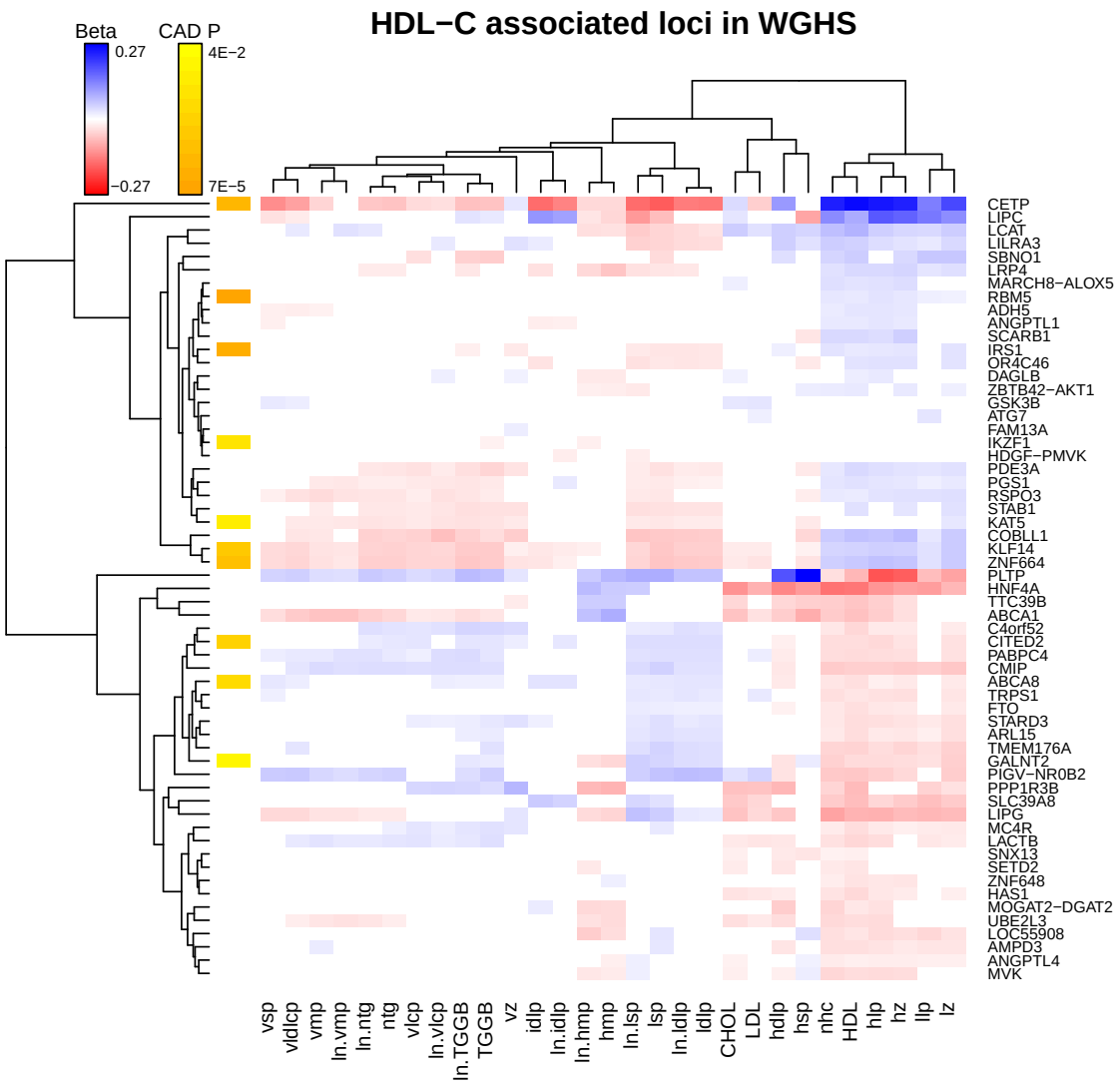
Lipid Subfraction Abbreviations

HDLLG	Large particles of high density lipoproteinconcentrations determined by NMR, Exam 4
HDLSM	Small particles of high density lipoproteinconcentrations determined by NMR, Exam 4
HDLSZ	Weighted average for HDL size based on measurements of HDLP1 through HDLP6, Exam 4
Lapoeser5*	ApoE concentrations in mg/dL using immunochemical technique by Servia, Exam 5
LCHYLO*	Chylomicron particles size >220 nm (expressed as TG concentrations in mg/dl) and determined using NMR, Exam 4
LDLINT	Medium particles of low density lipoprotein determined by NMR, Exam 4
LDLLG	Large particles of low density lipoprotein determined by NMR, Exam 4
LDLSZ	Weighted average for LDL size based on measurements of LDLP1 through LDLP6 determined by NMR, Exam 4
Lhdl25*	HDL2 cholesterol subfractions after chemical precipitation
LIDLP*	Intermediate density lipoprotein determined by NMR, Exam 4
LLDLSM	Small particles of low density lipoprotein determined by NMR, Exam 4
Llpaconc	Lipoprotein(a) concentration, Exam 3
LRLP-C*	Remnant like particles measured using selective immunoseparation of lipoproteins using the Otsuka kit. Expressed as cholesterol in mg/dL, Exam 4
LRLP-tg*	Remnant like particles measured using selective immunoseparation of lipoproteins using the Otsuka kit. Expressed as triglycerides in mg/dL, Exam 4
LVLDLINT*	Medium particles of very low density lipoprotein determined by NMR, Exam 4
LVLDLLG*	Large particles of very low density lipoprotein determined by NMR, Exam 4
VLDLSM	Small particles of very low density lipoprotein determined by NMR, Exam 4
VLDLSZ	Weighted average for VLDL size based on measurements of VLDLP1 through VLDLP6 determined by NMR, Exam 4

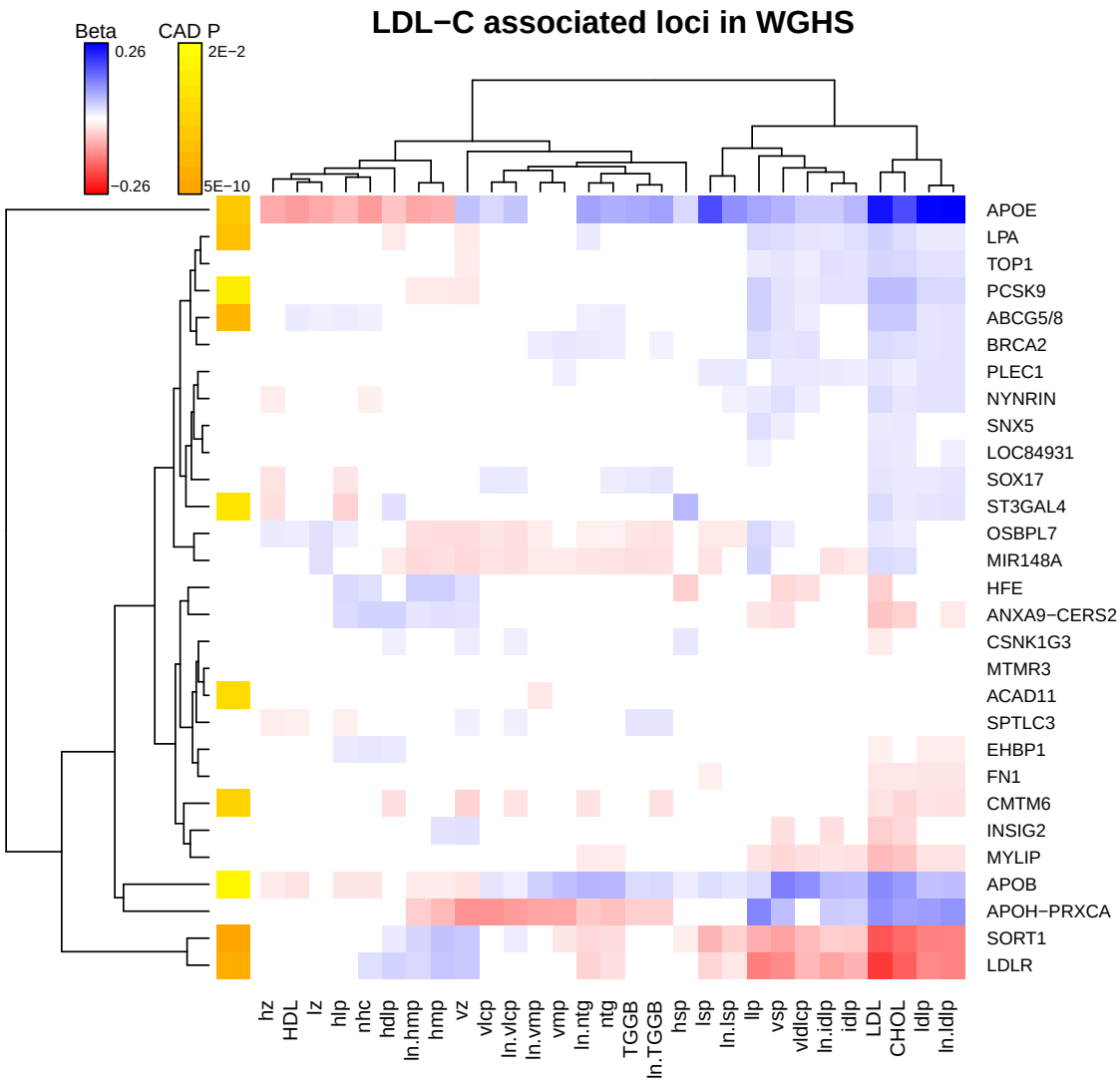
*Log transformed. All models were adjusted for age, sex and PCs. Low-, high-, intermediate- and very low-density lipoprotein particle concentrations were measured by nuclear magnetic resonance (NMR).

Figure S2.2: Association with lipid subfractions in Women's Genome Health Study. Heatmaps show effect sizes for association (P<0.10) with 23 lipid subfractions (Chasman et al., 2009) in the Women's Genome Health Study (WGHS) with respect to the trait-decreasing allele of **(A)** HDL-C and trait-increasing allele of **(B)** LDL-C, **(C)** TC, and **(D)** TG. Significant associations (P<0.05) of lipid-associated SNPs with coronary artery disease (CAD) are annotated on the y-axis at both known and novel genetic loci primarily associated with each trait. Dendrogram clustering of loci (y-axis) and lipid subfraction phenotypes (x-axis) based on the effect sizes (beta) are also shown. Figure **(E)** is a heatmap of correlations for the 23 lipid subfractions in WGHS.

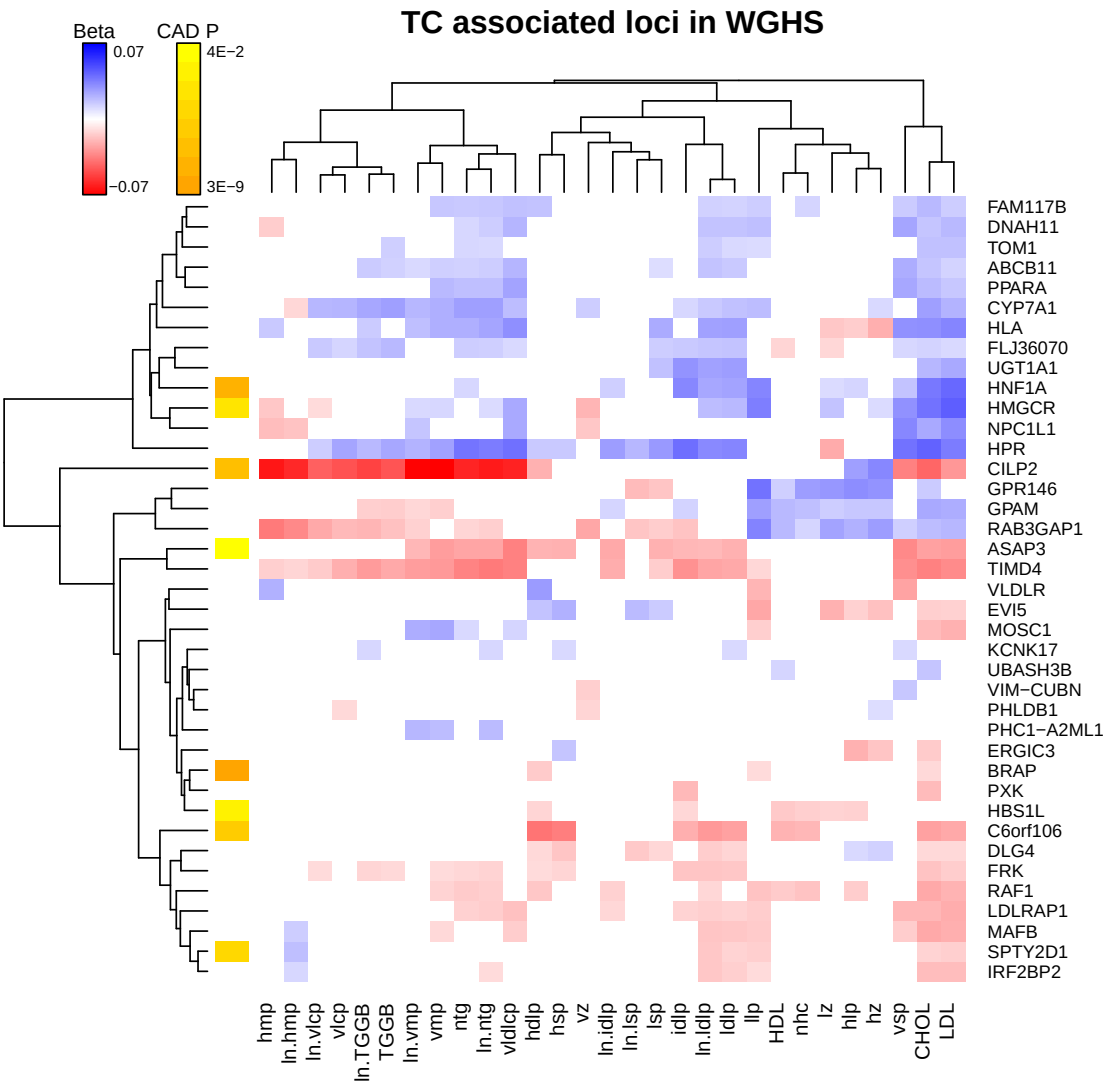
A.



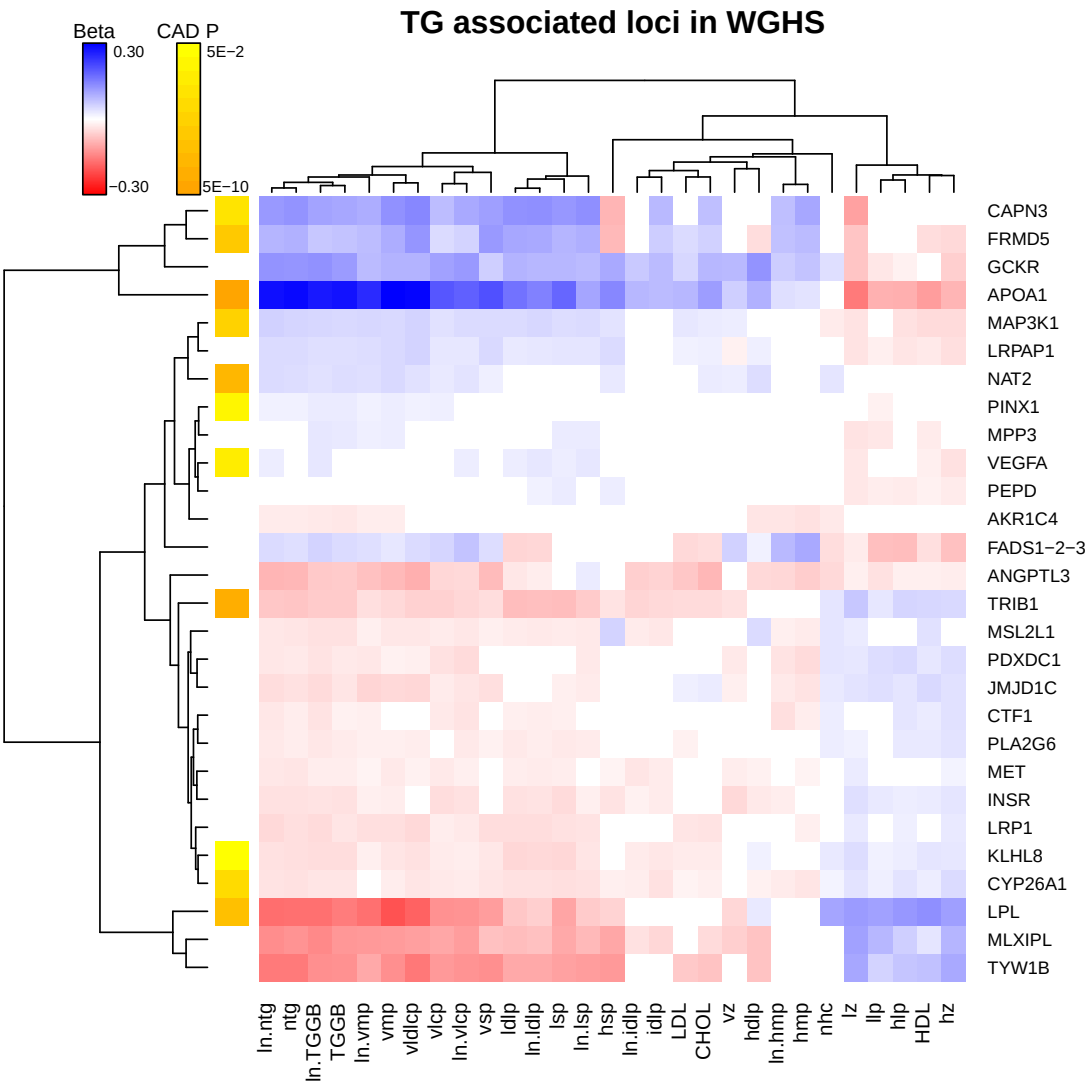
B.


LDL−C associated loci in WGHS
Beta
0.26
−0.26
CAD P
2E−2
5E−10
APOE
LPA
TOP1
PCSK9
ABCG5/8
BRCA2
PLEC1
NYNRIN
SNX5
LOC84931
SOX17
ST3GAL4
OSBPL7
MIR148A
HFE
ANXA9−CERS2
CSNK1G3
MTMR3
ACAD11
SPTLC3
EHBP1
FN1
CMTM6
INSIG2
MYLIP
APOB
APOH−PRXCA
SORT1
LDLR
hz
HDL
lz
hlp
nhc
hdlp
ln.hmp
hmp
vz
vlcp
ln.vlcp
ln.vmp
vmp
ln.ntg
ntg
TGGB
ln.TGGB
hsp
lsp
ln.lsp
llp
vsp
vldlcp
ln.idlp
idlp
LDL
CHOL
ldlp
ln.ldlp

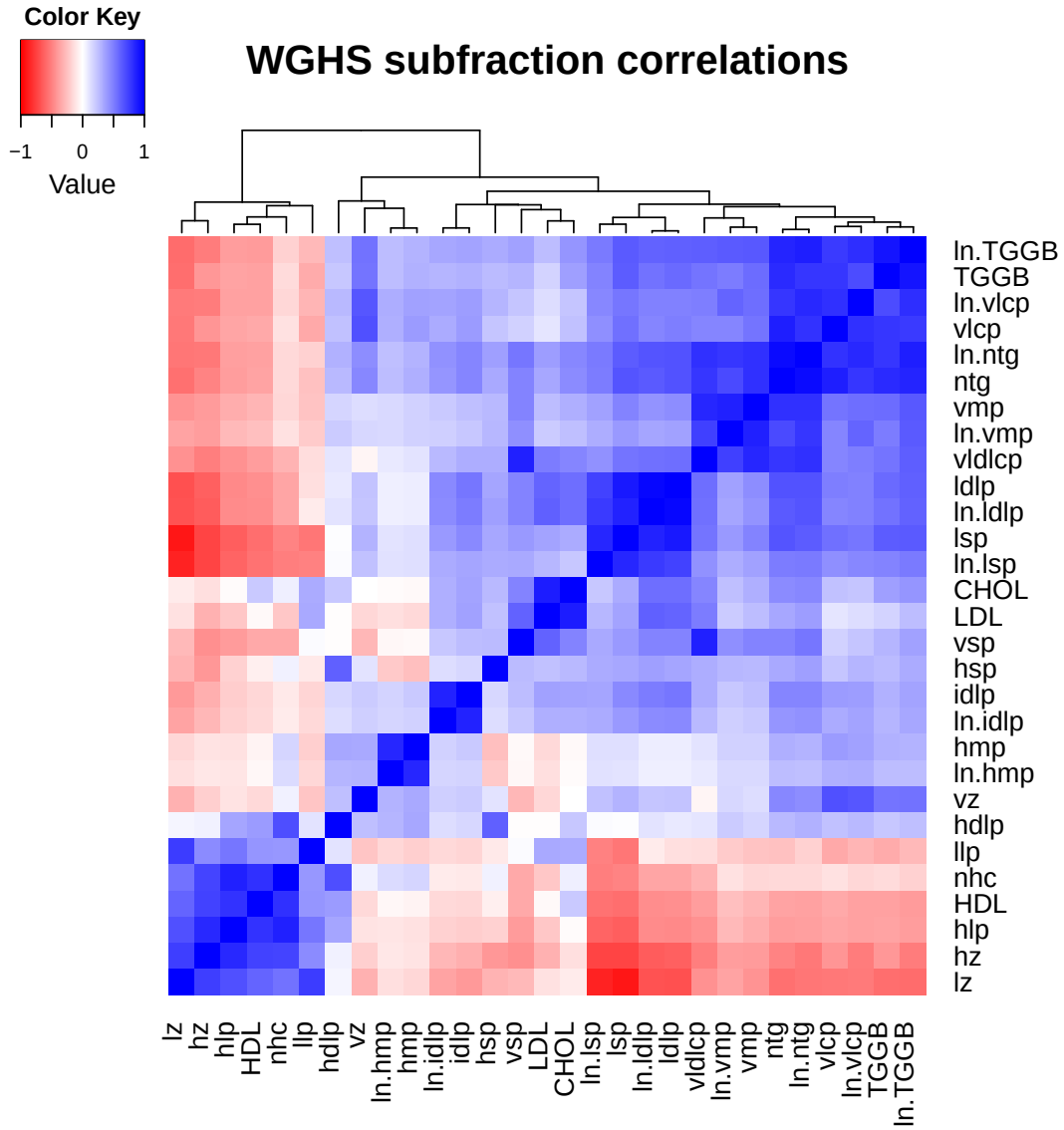
C.


TC associated loci in WGHS
Beta
0.07
−0.07
CAD P
4E−2
3E−9
FAM117B
DNAH11
TOM1
ABCB11
PPARA
CYP7A1
HLA
FLJ36070
UGT1A1
HNF1A
HMGCR
NPC1L1
HPR
CILP2
GPR146
GPAM
RAB3GAP1
ASAP3
TIMD4
VLDLR
EVI5
MOSC1
KCNK17
UBASH3B
VIM−CUBN
PHLDB1
PHC1−A2ML1
ERGIC3
BRAP
PXK
HBS1L
C6orf106
DLG4
FRK
RAF1
LDLRAP1
MAFB
SPTY2D1
IRF2BP2
hmp
ln.hmp
ln.vlcp
vlcp
ln.TGGB
TGGB
ln.vmp
vmp
ntg
ln.ntg
vldlcp
hdlp
hsp
vz
ln.idlp
ln.lsp
lsp
idlp
ln.idlp
ldlp
llp
HDL
nhc
lz
hlp
hz
vsp
CHOL
LDL

D.


TG associated loci in WGHS
Beta
0.30
−0.30
CAD P
5E−2
5E−10
CAPN3
FRMD5
GCKR
APOA1
MAP3K1
LRPAP1
NAT2
PINX1
MPP3
VEGFA
PEPD
AKR1C4
FADS1−2−3
ANGPTL3
TRIB1
MSL2L1
PDXDC1
JMJD1C
CTF1
PLA2G6
MET
INSR
LRP1
KLHL8
CYP26A1
LPL
MLXIPL
TYW1B
ln.ntg
ntg
ln.TGGB
TGGB
ln.vmp
vmp
vldlcp
vlcp
ln.vlcp
vsp
ldlp
ln.ldlp
lsp
ln.lsp
hsp
ln.idlp
idlp
LDL
CHOL
vz
hdlp
ln.hmp
hmp
nhc
lz
llp
hlp
HDL
hz

E.



Lipid Subfraction Abbreviations

llp: LDL large	**hmp:** HDL medium	**vsp:** VLDL small
lsp: LDL small	**ln.hmp:** ln[HDL medium]	**vz:** VLDL mean size
ln.lsp: ln[LDL small]	**hsp:** HDL small	**ntg:** TG by NMR
lz: LDL mean size	**hz:** HDL mean size	**ln.ntg:** ln[TG by NMR]
idlp: IDL total	**nhc:** HDL-C by NMR	**TGGB:** TG assay
ln.idlp: ln[IDL total]	**HDL:** HDL-C assay	**ln.TGGB:** ln[TG assay]
ldlp: LDL total	**vldlcp:** VLDL total	**CHOL:** Total Cholesterol
ln.ldlp: ln[LDL total]	**vlcp:** VLDL large	**ln.vlcp:** ln[VLDL large]
LDL: LDL-C assay	**hdlp:** HDL total	**vmp:** VLDL medium
hlp: HDL large	**ln.vmp:** ln[VLDL medium]	

*NMR, nuclear magnetic resonance

Table S2.1: Literature investigation of novel LDL-C and TC associated loci

Locus	Markername	Associated trait(s)	Literature Candidate	Complete Gene Name	Reference
			Loci Primarily Associated with LDL Cholesterol		
ANXA9-CERS2	rs267733	LDL-C	*CERS2*	ceramide synthase 2	PMID20940143, PMID20110363, PMID19801672
EHBP1	rs2710642	LDL-C	*EHBP1*	EH domain binding protein 1	PMID21332221
INSIG2	rs10490626	LDL-C, TC	*INSIG2*	insulin induced gene 2	PMID22143767, PMID20817058, PMID20090767
LOC84931	rs2030746	LDL-C, TC			
FN1	rs1250229	LDL-C	*FN1*	fibronectin 1	PMID16150826
CMTM6	rs7640978	LDL-C, TC			
ACAD11	rs17404153	LDL-C, HDL-C			
CSNK1G3	rs4530754	LDL-C, TC			
MIR148A	rs4722551	LDL-C, TG, TC			
SOX17	rs10102164	LDL-C, TC			
BRCA2	rs4942486	LDL-C			
APOH-PRXCA	rs1801689	LDL-C	*APOH, PRKCA*	apolipoprotein H, protein kinase C, alpha	PMID12740481, PMID20692055, PMID12952980
SPTLC3	rs364585	LDL-C	*SPTLC3*	serine palmitoyltransferase, long chain base subunit 3	PMID19648650
SNX5	rs2328223	LDL-C	*SNX5*	sorting nexin 5	PMID15561769
MTMR3	rs5763662	LDL-C			
			Loci Primarily Associated with Total Cholesterol		
ASAP3	rs1077514	TC			
ABCB11	rs2287623	TC	*ABCB11*	ATP-binding cassette, sub-family B (MDR/TAP), member 11	PMID21726512, PMID19228692
FAM117B	rs11694172	TC			
UGT1A1	rs11563251	TC, LDL-C	*UGT1A1/3/4/5/6/7/8/9/20*	UDP glucuronosyltransferase 1 family, polypeptide A1	PMID17908920
PXK	rs13315871	TC	*PXK*	PX domain containing serine/threonine kinase	PMID20086096, PMID17178602
KCNK17	rs2758886	TC			
HBS1L	rs9376090	TC			
GPR146	rs1997243	TC			
VLDLR	rs3780181	TC,LDL-C	*VLDLR*	very low density lipoprotein receptor	PMID8827514
VIM-CUBN	rs10904908	TC	*VIM, CUBN*	vimentin, cubilin	PMID22535769, PMID7706405, PMID1527066, PMID10371504
PHLDB1	rs11603023	TC			
PHC1-A2ML1	rs4883201	TC	*A2ML1*	alpha-2-macroglobulin-like 1	PMID18648652
DLG4	rs314253	TC, LDL-C	*ACADVL*	acyl-CoA dehydrogenase, very long chain	PMID19889959
			CTDNEP1	CTD nuclear envelope phosphatase 1	PMID22134922
			SLC2A4	solute carrier family 2, member 4	PMID16096283
TOM1	rs138777	TC	*HMOX1*	hemeoxygenase (decycling) 1	PMID22004613
PPARA	rs4253772	TC,LDL-C	*PPARA*	peroxisome proliferator-activated receptor alpha	PMID21540177, PMID21487230

Table S2.2: Literature investigation of novel HDL-C and TG associated loci

Locus	Markername	Associated trait(s)	Literature Candidate	Complete Gene Name	Reference
Loci Primarily Associated with HDL Cholesterol					
PIGV-NR0B2	rs12748152	HDL-C, LDL-C, TG	*PIGV*	phosphatidylinositol glycan anchor biosynthesis, class V	PMID20802478, PMID15623507
			NR0B2	nuclear receptor subfamily 0, group B, member 2	PMID22577560, PMID20375098
HDGF-PMVK	rs12145743	HDL-C	*HDGF*	hepatoma-derived growth factor	PMID14635185
			CRABP2	cellular retinoic acid binding protein 2	PMID17484622
ANGPTL1	rs4650994	HDL-C			
CPS1	rs1047891	HDL-C			
ATG7	rs2606736	HDL-C			
SETD2	rs2290547	HDL-C			
RBM5	rs2013208	HDL-C			
STAB1	rs13326165	HDL-C	*STAB1*	stabilin 1	PMID21480214, PMID19726632, PMID21030611
			NISCH	nischarin	PMID21484668
GSK3B	rs6805251	HDL-C	*GSK3B*	glycogen synthase kinase 3 beta	PMID21334395
			NR1I2	nuclear receptor subfamily 1, group I, member 2	PMID21295138
C4orf52	rs10019888	HDL-C			
FAM13A	rs3822072	HDL-C			
ADH5	rs2602836	HDL-C			
RSPO3	rs1936800	HDL-C, TG			
DAGLB	rs702485	HDL-C	*DAGLB*	diacylglycerol lipase, beta	PMID21949825
SNX13	rs4142995	HDL-C	*SNX13*	sorting nexin 13	PMID12461558
IKZF1	rs4917014	HDL-C	*IKZF1*	IKAROS family zinc finger 1 (Ikaros)	PMID18483254
TMEM176A	rs17173637	HDL-C			
MARCH8-ALOX5	rs970548	HDL-C, TC	*ALOX5*	arachidonate 5-lipoxygenase	PMID22293202
OR4C46	rs11246602	HDL-C			
KAT5	rs12801636	HDL-C	*KAT5*	K(lysine) acetyltransferase 5	PMID18096664, PMID17996965
MOGAT2-DGAT2	rs499974	HDL-C	*MOGAT2*	monoacylglycerol O-acyltransferase 2	PMID21734185, PMID14966132
			DGAT2	diacylglycerol O-acyltransferase 2	PMID22493088, PMID21317108, PMID22155452
ZBTB42-AKT1	rs4983559	HDL-C	*AKT1*	v-akt murine thymoma viral oncogene homolog 1	PMID18054314, PMID20054340
FTO	rs1121980	HDL-C, TG			
HAS1	rs17695224	HDL-C	*HAS1*	hyaluronan synthase 1	PMID9933623
Loci Primarily Associated with Triglycerides					
LRPAP1	rs6831256	TG, TC, LDL-C	*LRPAP1*	low density lipoprotein receptor-related protein associated protein 1	PMID16973241
VEGFA	rs998584	TG, HDL-C	*VEGFA*	vascular endothelial growth factor A	PMID21348596, PMID18789802
MET	rs38855	TG			
AKR1C4	rs1832007	TG	*AKR1C4*	aldo-ketoreductase family 1, member C4	PMID18024509
PDXDC1	rs3198697	TG			
MPP3	rs8077889	TG			
INSR	rs7248104	TG			
PEPD	rs731839	TG, HDL-C	*CEBPG*	CCAAT/enhancer binding protein (C/EBP), gamma	PMID12177065

Table S2.3: Biological candidate genes at novel LDL-C and TC associated loci based on nonsynonymous substitutions, gene expression levels (eQTLs) and pathway analyses

Locus	Markername	Nearest Gene (Kb away)	No. of Genes within 100 Kb	Nonsynonymous SNP (r^2)	Gene with nonsynonymous SNP	Amino Acid Change	PolyPhen Score[a]	eQTL Gene ($P<5x10^{-8}$)	Pathway Analysis
Loci Primarily Associated with LDL Cholesterol									
ANXA9-CERS2	rs267733	*ANXA9* (0)	10	rs267733	*ANXA9*	Asp166Gly	0.99		*ANXA9*
EHBP1	rs2710642	*EHBP1* (0)	1						
INSIG2	rs10490626	*INSIG2* (10.2)	2	rs17512204 (1.00)	*CCDC93*	Pro228Leu	0.01		*INSIG2*
LOC84931	rs2030746	*LOC84931* (85.6)	1						
FN1	rs1250229	*FN1* (3.6)	2	rs1250259 (1.00)	*FN1*	Gln15Leu	0.00		
CMTM6	rs7640978	*CMTM6* (0)	3	rs2303857 (.91)	*DYNC1LI1*	Gln277Arg	0.02		
ACAD11	rs17404153	*DNAJC13* (0)	2	rs41272321 (0.85)	*ACAD11**	Lys414Thr	NA		
CSNK1G3	rs4530754	*CSNK1G3* (0)	2						
MIR148A	rs4722551	*MIR148A* (2.2)	1						
SOX17	rs10102164	*SOX17* (48.2)	1						
BRCA2	rs4942486	*BRCA2* (0)	5						*BRCA2*
APOH-PRXCA	rs1801689	*APOH* (0)	3	rs1801689	*APOH*	Cys325Gly	1.00		*APOH*
SPTLC3	rs364585	*SPTLC3* (26.9)	1					*SPTLC3*	
SNX5	rs2328223	*SNX5* (76.3)	2						
MTMR3	rs5763662	*MTMR3* (0)	2						
Loci Primarily Associated with Total Cholesterol									
ASAP3	rs1077514	*ASAP3* (0)	6						
ABCB11	rs2287623	*ABCB11* (0)	4	rs2287622 (1.00)	*ABCB11*	Val444Ala	0.00		*ABCB11*
FAM117B	rs11694172	*FAM117B* (0)	2						
UGT1A1	rs11563251	*UGT1A1* (0)	12						*UGT1A1*
PXK	rs13315871	*PXK* (0)	4					*PXK*	
KCNK17	rs2758886	*KCNK17* (15.9)	4						
HBS1L	rs9376090	*HBS1L* (35.2)	2						
GPR146	rs1997243	*C7orf50* (0)	7	rs11761941 (1.00)	*GPR146*	Gly11Glu	NA	*GPR146*	
VLDLR	rs3780181	*VLDLR* (0)	3						*VLDLR*
VIM-CUBN	rs10904908	*VIM* (10.0)	3						*CUBN*
PHLDB1	rs11603023	*PHLDB1* (0)	7						
PHC1-A2ML1	rs4883201	*PHC1* (0)	4						
DLG4	rs314253	*DLG4* (1.6)	13						*DLG4*
TOM1	rs138777	*TOM1* (0)	4	rs1053593 (.92)	*HMGXB4*	Gly165Val	0.01		
PPARA	rs4253772	*PPARA* (0)	6						*PPARA*

* Genes with a non-synonymous SNP in linkage disequilibrium with the lead SNP for the locus, but more than 100Kb away

[a] PolyPhen-2 classifier estimates the probability an amino-acid change is damaging to the encoded protein. For markers labeled NA, PolyPhen scores were not available.

Table S2.4: Biological candidate genes at novel HDL-C and TG associated loci based on nonsynonymous substitutions, gene expression levels (eQTLs) and pathway analyses

Locus	Markername	Nearest Gene (Kb away)	No. of Genes within 100 Kb	Nonsynonymous SNP (r^2)	Gene with nonsynonymous SNP	Amino Acid Change	PolyPhen Score[a]	eQTL Gene ($P<5x10^{-8}$)	Pathway Analysis
Loci Primarily Associated with HDL Cholesterol									
PIGV-NR0B2	rs12748152	*PIGV* (13.5)	7	rs17360994 (1.00),	*C1orf172**	Gln100Arg	0.20		*NR0B2*
				rs7545442 (.90)	*NUDC**	Thr68Met	NA		
				rs6659176 (1.00)	*NR0B2*	Gly171Ala	0.99		
*HDGF-PMVK**	rs12145743	*RRNAD1* (0)	10	rs4399146 (1.00)	*HDGF*	Pro201Leu	0.00		
*ANGPTL1**	rs4650994	*C1orf220* (0)	3						
CPS1	rs1047891	*CPS1* (0)	2	rs1047891	*CPS1*	Thr1412Asn	0.01		*CPS1*
ATG7	rs2606736	*ATG7* (0)	2						
SETD2	rs2290547	*SETD2* (0)	4	rs2305637 (.94)	*NBEAL2*	Ser2054Phe	0.99		
RBM5	rs2013208	*RBM5* (0)	4	rs2230590 (.89)	*MST1R**	Gln523Arg	0.00	*RBM5*	
				rs1062633 (.93)		Arg1335Gly	0.00		
STAB1	rs13326165	*STAB1* (0)	10	rs887515 (.85)	*NISCH*	Ala1056Val	0.00		
GSK3B	rs6805251	*GSK3B* (0)	3						*GSK3B*
C4orf52	rs10019888	*C4orf52** (131.5)	0						
FAM13A	rs3822072	*FAM13A* (0)	2						
ADH5	rs2602836	*ADH5* (4.9)	4					*ADH5*	
RSPO3	rs1936800	*RSPO3* (4)	1						
DAGLB	rs702485	*DAGLB* (0)	5					*DAGLB*	*DAGLB*
SNX13	rs4142995	*SNX13* (0)	1						
IKZF1	rs4917014	*IKZF1* (0)	1						
TMEM176A	rs17173637	*ABP1* (20.1)	5					*TMEM176A*	
MARCH8-ALOX5	rs970548	*MARCH8* (0)	3	rs2291429 (.95)	*MARCH8*	Leu269Trp	NA		
				rs2291428 (.95)	*MARCH8*	Phe277Leu	NA		
OR4C46	rs11246602	*OR4C46* (3.2)	2	rs12419022 (.97)	*OR5W2**	His65Arg	0.01		
				rs11230983 (.97)	*OR5D13**	Arg124His	0.02		
				rs12224086 (.94)	*OR5AS1**	Arg122Leu	0.90		
KAT5	rs12801636	*PCNXL3* (0)	12						
MOGAT2-DGAT2	rs499974	*MOGAT2* (12.7)	4						
ZBTB42-AKT1	rs4983559	*ZBTB42* (6.2)	7						*AKT1*
FTO	rs1121980	*FTO* (0)	2						
HAS1	rs17695224	*FPR3* (0)	6						
Loci Primarily Associated with Triglycerides									
LRPAP1	rs6831256	*DOK7* (0)	4						*LRPAP1*
VEGFA	rs998584	*VEGFA* (3.7)	1						*VEGFA*
MET	rs38855	*MET* (0)	1						
AKR1C4	rs1832007	*AKR1C4* (0)	2	rs3829125 (1.00)	*AKR1C4*	Ser145Cys	0.00		
				rs17134592 (1.00)	*AKR1C4*	Leu311Val	0.00		*AKR1C4*
PDXDC1	rs3198697	*PDXDC1* (0)	4						
MPP3	rs8077889	*MPP3* (0)	6						*MPP3*
INSR	rs7248104	*INSR* (0)	1						*INSR*
PEPD	rs731839	*PEPD* (0)	2						

* Gene selected as locus label was judged to be an especially worthy candidate > 100Kb away, or no genes within 100Kb of the lead SNP were available

* Genes with a non-synonymous SNP in linkage disequilibrium with the lead SNP for the locus, but more than 100Kb away

[a] PolyPhen-2 classifier estimates the probability an amino-acid change is damaging to the encoded protein. For markers labeled NA, PolyPhen scores were not available.

Table S2.5: Overlap between eQTL loci and new lipid-associated loci

Index SNP	**Position**	**Transcript**	**Index SNP *P*-value**	**Expression Increasing Allele**	**Top eQTL SNP**	**Top eQTL SNP *P*-value**	**r^2**	**Conditional *P*-value (Index SNP)**	**Conditional *P*-value (Top eQTL SNP)**
eQTLs in Loci Primarily Associated with HDL-C									
rs2013208	chr3 at 50.1Mb	*RBM5* in Omental Fat	$3x10^{-30}$	T	rs2353579	$7x10^{-33}$	0.93	1.00	0.60
rs2013208	chr3 at 50.1Mb	*RBM5* in Subcutaneous Fat	$5x10^{-22}$	T	rs4688758	$2x10^{-23}$	0.93	0.93	0.63
rs2602836	chr4 at 100.2Mb	*ADH5* in Omental Fat	$7x10^{-27}$	G	rs1800759	$4x10^{-47}$	0.82	0.09	$7x10^{-9}$
rs2602836	chr4 at 100.2Mb	*ADH5* in Subcutaneous Fat	$5x10^{-17}$	G	rs1800759	$7x10^{-31}$	0.80	0.20	$6x10^{-4}$
rs702485	chr7 at 6.4Mb	*DAGLB* in Omental Fat	$6x10^{-26}$	G	rs13238780	$3x10^{-27}$	0.94	0.99	0.79
rs702485	chr7 at 6.4Mb	*DAGLB* in Subcutaneous Fat	$2x10^{-13}$	G	rs836556	$1x10^{-15}$	0.92	0.93	0.61
rs17173637	chr7 at 150.2Mb	*TMEM176A* in Subcutaneous Fat	$2x10^{-13}$	C	Index SNP				
eQTLs in Loci Primarily Associated with LDL-C									
rs364585	chr20 at 12.9Mb	*SPTLC3* in Liver	$8x10^{-37}$	A	rs168622	$1x10^{-38}$	0.97	0.95	0.88
eQTLs in Loci Primarily Associated with Total Cholesterol									
rs13315871	chr3 at 58.4Mb	*PXK* in Liver	$7x10^{-17}$	A	rs13066269	$7x10^{-17}$	0.99	1.00	1.00
rs1997243	chr7 at 1.1Mb	*GPR146* in Omental Fat	$7x10^{-33}$	A	Index SNP				
rs1997243	chr7 at 1.1Mb	*GPR146* in Subcutaneous Fat	$9x10^{-18}$	A	rs2363286	$9x10^{-18}$	1.00	1.00	1.00

The table lists index SNPs for new lipid-associated loci that are also eQTLs (with $P<5x10^{-8}$) for a nearby transcript in liver, omental fat, or subcutaneous fat. The top eQTL associated SNP in the region is also listed, together with its eQTL association *P*-value and linkage disequilibrium with the lipid-associated SNP. Conditional *P*-values for the index SNP are from an analysis that includes the top eQTL SNP as a covariate (and vice-versa). Only loci for which the r^2 linkage disequilibrium coefficient between the index GWAS SNP and top eQTL SNP was >0.50 are listed.

Table S2.6: Nonsynonymous variants in linkage disequilibrium with index SNPs at novel loci

Lead SNP	**Chr**	**hg19 Position (Mb)**	**Lead Trait**	**Non-synonymous SNP**	**r^2**	**Gene with Nonsynonymous SNP**	**Amino Acid Change**	**PolyPhen-2 Classifier***
rs12748152	1	27.14	HDL-C	rs17360994	1.00	*C1orf172*	Gln100Arg	0.20
				rs7545442	.90	*NUDC*	Thr68Met	NA
				rs6659176	1.00	*NR0B2*	Gly171Ala	0.99
rs12145743	1	156.70	HDL-C	rs4399146	1.00	*HDGF*	Pro201Leu	0.00
rs1047891	2	211.54	HDL-C	rs1047891	–	*CPS1*	Thr1412Asn	0.01
rs2290547	3	47.06	HDL-C	rs2305637	.94	*NBEAL2*	Ser2054Phe	0.99
rs2013208	3	50.13	HDL-C	rs2230590	.89	*MST1R*	Gln523Arg	0.00
				rs1062633	.93	*MST1R*	Arg1335Gly	0.00
rs13326165	3	52.53	HDL-C	rs887515	.85	*NISCH*	Ala1056Val	0.00
rs970548	10	46.01	HDL-C	rs2291429	.95	*MARCH8*	Leu269Trp	NA
				rs2291428	.95	*MARCH8*	Phe277Leu	NA
rs11246602	11	55.20	HDL-C	rs12419022	.97	*OR5W2*	His65Arg	0.01
				rs11230983	.97	*OR5D13*	Arg124His	0.02
				rs12224086	.94	*OR5AS1*	Arg122Leu	0.90
rs267733	1	150.96	LDL-C	rs267733	–	*ANXA9*	Asp166Gly	0.99
rs10490626	2	118.84	LDL-C	rs17512204	1.00	*CCDC93*	Pro228Leu	0.01
rs1250229	2	216.30	LDL-C	rs1250259	1.00	*FN1*	Gln15Leu	0.00
rs7640978	3	32.53	LDL-C	rs2303857	.91	*DYNC1LI1*	Gln277Arg	0.02
rs17404153	3	132.16	LDL-C	rs41272321	.85	*ACAD11*	Lys414Thr	NA
rs1801689	17	64.21	LDL-C	rs1801689	–	*APOH*	Cys325Gly	1.00
rs2287623	2	169.83	TC	rs2287622	1.00	*ABCB11*	Val444Ala	0.00
rs1997243	7	1.08	TC	rs11761941	1.00	*GPR146*	Gly11Glu	NA
rs138777	22	35.71	TC	rs1053593	.92	*HMGXB4*	Gly165Val	0.01
rs1832007	10	5.25	TG	rs3829125	1.00	*AKR1C4*	Ser145Cys	0.00
				rs17134592	1.00	AKR1C4	Leu311Val	0.00

*The PolyPhen-2 classifier estimates the probability that the amino-acid change is damaging to the encoded protein. For markers labeled NA, PolyPhen scores were not available from the PolyPhen web service at: `http://genetics.bwh.harvard.edu/pph2/bgi.shtml`

Table S2.7: Overlap of SNPs at known and novel lipid loci with chromatin states in 9 different cell types

Cell Type	**Observed Number of Chromatin States* Showing Excess Overlap with Lipid Loci**	**Chromatin States* Showing Excess Overlap with Lipid Loci**
H1 embryonic stem cells (H1 ES)	2	Transcription Transition (HMM9) P=4x10^{-10} Transcription Elongation (HMM10) P=5x10^{-10}
B-lymphoblastoid cells (GM12878)	0	
Umbilical vein endothelial cells (HUVEC)	2	Transcription Transition (HMM9) P=2x10^{-7} Transcription Elongation (HMM10) P=6x10^{-7}
Skeletal muscle myoblasts (HSMM)	1	Transcription Elongation (HMM10) P=6x10^{-8}
Mammary epithelial cells (HMEC)	2	Transcription Transition (HMM9) P=6x10^{-11} Transcription Elongation (HMM10) P=2x10^{-9}
Normal epidermal keratinocytes (NHEK)	2	Transcription Elongation (HMM10) P=2x10^{-8} Weak Transcription (HMM11) P=3x10^{-6}
Normal lung fibroblasts (NHLF)	2	Transcription Elongation (HMM10) P=2x10^{-10} Transcription Transition (HMM9) P=8x10^{-8}
Erythrocyticleukaemia cells (K562)	3	Weak Transcription (HMM11) P=1x10^{-11} Weak Enhancer (HMM7) P=2x10^{-10} Strong Enhancer (HMM5) P=4x10^{-8}
Hepatocellular carcinoma cells (HepG2)	8	Strong Enhancer (HMM4) P=2x10^{-25} Weak Enhancer (HMM7) P=4x10^{-14} Weak Transcription (HMM11) P=2x10^{-11} Strong Enhancer (HMM5) P=5x10^{-11} Transcription Elongation (HMM10) P=3x10^{-10} Weak Enhancer (HMM6) P=1x10^{-7} Active Promoter (HMM1) P=4x10^{-7} Weak Promoter (HMM2) P=7x10^{-7}

*Chromatin states were described previously (Ernst et al., 2011) based on hidden Markov models (HMM) of histone methylation and acetylation marks from 9 cell types. SNPs in high linkage disequilibrium (r^2>0.8 in 1000 Genomes Project European ancestry samples) with known or novel lipid loci were compared to matched sets of HapMap SNPs (see Section 2.5.12).

Table S2.8: Overlap with chromatin states, histone marks and transcription factor ChIP-Seq in HepG2 Cells

	Known and Novel Lipid loci (n=157)			Only Novel Lipid Loci (n=62)		
	Observed Number of Loci with ≥ 1 SNP in a Regulatory Region	**Expected Number of Loci**	***P*-value**	**Observed Number of Loci with ≥ 1 SNP in a Regulatory Region**	**Expected Number of Loci**	***P*-value**
Overlap with Chromatin States from Ernst et al. (2011) (13 tested)*						
Strong Enhancer (HMM4)	49	13.7	$2x10^{-25}$	20	6.2	$9x10^{-10}$
Weak Enhancer (HMM7)	60	26.9	$4x10^{-14}$	25	11.9	$3x10^{-5}$
Weak Transcription (HMM11)	99	62.1	$2x10^{-11}$	41	26.4	$9x10^{-5}$
Strong Enhancer (HMM5)	34	12.8	$5x10^{-11}$	10	5.6	$5x10^{-2}$
Transcription Elongation (HMM10)	65	35.4	$3x10^{-10}$	26	15.4	$1x10^{-3}$
Weak Enhancer (HMM6)	57	33.5	$1x10^{-7}$	21	14.5	.013
Active Promoter (HMM1)	39	20.3	$4x10^{-7}$	14	8.8	.039
Weak Promoter (HMM2)	45	24.8	$7x10^{-7}$	15	10.6	.088
Transcription Transition (HMM9)	37	18.7	$3x10^{-5}$	18	8.0	$4x10^{-4}$
Overlap with Histone Marks (5 tested)						
H3K9ac	97	47.3	$3x10^{-22}$	37	20.1	$6x10^{-8}$
H3K27ac	84	39.2	$3x10^{-20}$	34	16.7	$4x10^{-8}$
H3K4me3	88	47.9	$2x10^{-15}$	34	20.1	$7x10^{-5}$
H3K36me3	104	62.3	$4x10^{-14}$	41	26.1	$2x10^{-5}$
H3K4me2	111	74.3	$8x10^{-12}$	44	31.1	$7x10^{-5}$
Overlap with Open Chromatin (2 tested)						
FAIRE	51	26.5	$5x10^{-9}$	19	11.3	$8x10^{-3}$
DNase hypersensitivity	33	18.3	$2x10^{-4}$	12	8.1	.09
Overlap with Transcription Factor ChIP-Seq (11 tested)						
HNF4	38	16.2	$6x10^{-10}$	14	7.1	$6x10^{-3}$
CEBP/β	40	20.4	$1x10^{-5}$	16	9.1	.010
CTCF	55	37.6	$4x10^{-4}$	21	16.2	.055
HSF1	9	2.6	$1x10^{-3}$	4	1.1	.024

*Chromatin states were described previously (Ernst et al., 2011) based on hidden Markov models (HMM) of histone methylation and acetylation marks from 9 cell types. Data for histone marks, open chromatin, and transcription factor ChIP-seq were obtained from the ENCODE Project (ENCODE Project Consortium, 2011). SNPs in high linkage disequilibrium (r^2>0.8 in 1000 Genomes Project European ancestry samples) with known or novel lipid loci were compared to matched sets of HapMap SNPs (see Section 2.5.12). This table lists only regulatory elements that exhibited a significant excess overlap ($P<1x10^{-3}$ to account for 31 HepG2 regulatory elements tested). FAIRE, Formaldehyde-Assisted Isolation of Regulatory Elements.

Table S2.9: Fine mapping results in different ancestries

Chr	Fine Mapping Interval (hg19 Mb)	Locus Name	Top GWAS SNP	# LD Proxies in European	*P*value	N	% Var	Freq	*P*value	N	% Var	Freq	Top Metabochip SNP	# LD Proxies	EUR r^2 with GWA SNP	Other r^2 with GWA SNP	*P*value	N	% Var	Freq
HDL Cholesterol																				
African																				
16	56.98-57.02	*CETP*	rs173539	12	$9x10^{-370}$	92,820	2.48	0.34	$3x10^{-3}$	2,738	0.37	0.38	rs17231520	3	NA	0.11	$2x10^{-16}$	4,420	3.03	0.08
European																				
2	165.5-165.73	*COBLL1*	rs12328675	9	$1x10^{-10}$	94,311	0.06	0.86	$2x10^{-6}$	92,781	0.03	0.88	rs355863	13	0.43	0.43	$6x10^{-9}$	90,652	0.04	0.11
11	46.33-47.35	*LRP4*	rs3136441	80	$7x10^{-18}$	94,311	0.10	0.81	$8x10^{-14}$	92,664	0.08	0.83	rs10838692	55	0.28	0.28	$1x10^{-26}$	92,742	0.16	0.65
17	37.39-38.07	*MED1* (*PPP1R1B*)	rs881844	55	$3x10^{-14}$	92,820	0.06	0.34	$3x10^{-5}$	92,574	0.02	0.37	rs10445306	270	0.44	0.44	$2x10^{-10}$	92,699	0.05	0.24
LDL Cholesterol																				
African																				
1	109.66-110.31	*SORT1*	rs629301	11	$2x10^{-168}$	89,888	1.19	0.75	$4x10^{-5}$	3,940	0.93	0.65	rs12740374	2	1	0.63	$3x10^{-10}$	2,555	1.84	0.24
19	11.18-11.26	*LDLR*	rs6511720	43	$3x10^{-115}$	87,565	1.05	0.13	$8x10^{-6}$	2,652	0.89	0.13	rs115594766	17	0.97	0.6	$9x10^{-10}$	2,636	1.73	0.81
19	45.40-45.44	*APOE-C1-C2-C4*	rs4420638	6	$1x10^{-140}$	77,643	1.52	0.81	0.697	2,628	0.01	0.81	rs7412 (e2)	1	0.02	0.02	$1x10^{-50}$	2,594	9.64	0.11
European																				
1	55.50-55.51	*PCSK9*	rs17111503	1	$2x10^{-27}$	89,888	0.22	0.75	$9x10^{-24}$	83,102	0.14	0.76	rs11591147 (R46L)	1	0	0	$2x10^{-136}$	77,417	1.38	0.03
6	160.47-160.58	*IGF2R*	rs1564348	4	$2x10^{-16}$	89,873	0.11	0.81	$7x10^{-9}$	83,116	0.05	0.84	rs2297374	15	0.11	0.11	$2x10^{-13}$	83,090	0.07	0.37
7	44.37-44.68	*NPC1L1*	rs217406	6	$6x10^{-11}$	86,806	0.12	0.79	$2x10^{-5}$	82,799	0.03	0.73	rs2073547	5	0.39	0.39	$1x10^{-12}$	83,083	0.08	0.76
11	126.22-126.27	*ST3GAL4*	rs11220463	24	$4x10^{-15}$	89,888	0.12	0.85	$2x10^{-6}$	83,068	0.04	0.74	rs59379014	11	0.35	0.35	$6x10^{-11}$	83,083	0.06	0.07
19	45.40-45.44	*APOE-C1-C2-C4*	rs4420638	6	$1x10^{-140}$	77,643	1.52	0.81	$3x10^{-44}$	15,460	1.71	0.8	rs7412 (e2)	2	0.02	0.02	$2x10^{-651}$	82,533	4.63	0.07
Triglycerides																				
East Asian																				
11	116.53-116.67	*APOA5-A4-C3-A1*	rs2160669	20	$3x10^{-128}$	91,013	0.96	0.90	$3x10^{-27}$	8,743	1.37	0.79	rs651821	16	0.85	0.76	$2x10^{-55}$	8,743	2.83	0.73

Table S2.10: Candidate genes at novel loci

ABCB11 **(ATP-binding cassette, sub-family B, member 11)** is involved in the ATP-dependent secretion of bile salts (MIM 603201). Hepatic overexpression of *Abcb11* in mice increased absorption of cholesterol and promoted diet-induced obesity and hypercholesterolemia. *G6PC2* encodes a glucose-6-phosphatase catalytic subunit (MIM 608058). Variants at this locus have been implicated in liver enzyme and fasting glucose levels.
ACAD11 **(acyl-CoA dehydrogenase family, member 11)** is involved in the β-oxidation of long-chain fatty acids in muscle and heart (MIM 614288).
ADH5 **(alcohol dehydrogenase 5 (class III), chi polypeptide)** encodes a protein involved in oxidation of long-chain primary alcohols and which catalyzes a step in the elimination of formaldehyde (MIM 103710).
AKR1C4 **(aldo-keto reductase family 1, member C4)** encodes a protein that produces intermediates in bile acid biosynthesis and inactivates circulating steroid hormones (MIM 600451). *AKR1C4* is expressed exclusively in the liver and is transcriptionally regulated by LXRA.
ANGPTL1 **(angiopoietin-like 1 gene)** is a member of the angiopoietin family involved in angiogenesis, and widely expressed in highly vascularized tissues (MIM 603874).
ANXA9 **(annexin A9) and *CERS2* (ceramide synthase 2)**. *ANXA9* is a calcium-dependent phospholipid-binding protein (MIM 603319). *CERS2* is involved in regulation of long acyl chain and sphingolipid metabolism (MIM 606920).
APOH **(Apolipoprotein H, also known as beta-2 glycoprotein I) and *PRKCA* (protein kinase C, alpha)**. *APOH* is a glycoprotein that is involved in the activation of lipoprotein lipase and which neutralizes negatively charged phospholipids (MIM 138700). *PRKCA* is activated by *APOA1* and diacylglycerol during cholesterol mobilization (MIM 176960).
ASAP3 **(ArfGAP with SH3 domain, ankyrin repeat and PH domain 3)** is a GTPase-activating protein that promotes cell differentiation and migration and has been implicated in cancer cell invasion.
ATG7 (autophagy related 7) encodes a protein that is part of the autophagy machinery (MIM 608760). Dysfunction in autophagy canimpact systems related to intracellular energy utilization and promote apoptotic cell death.
BRCA2 **(breast cancer 2, early onset)** is involved in maintenance of genome stability, specifically the homologous recombination pathway for repair of double stranded DNA. Variants in the region can increase risk of breast and other types of cancer (MIM 600185).
C4orf52 **(chromosome 4 open reading frame 52)**. The nearest gene to the lead signal is an uncharacterized gene with unknown function, and there are no other obvious candidate genes in the locus.
CMTM6 **(CKLF-like MARVEL)**. This gene belongs to the chemokine-like factor gene superfamily, but the exact function of the encoded protein is unknown (MIM 607889).
CPS1 **(carbamoyl-phosphate synthase 1, mitochondrial)** encodes a mitochondrial enzyme that catalyzes the first committed step of the urea cycle (MIM 608307). The lead variant encodes a threonine to asparagine substitution previously associated with levels of homocysteine and fibrinogen.
CSNK1G3 **(casein kinase 1, gamma 3)** encodes a serine/threonine-protein kinase that is involved in a number of cellular processes including DNA repair, cell division, nuclear localization and membrane transport (MIM 604253).
DAGLB **(diacylglycerol lipase, beta)** catalyzes the hydrolysis of diacylglycerol (DAG) to 2-arachidonoyl-glycerol, an abundant endocannabinoid (MIM 614016). Endocannabinoids function signaling molecules, regulate axonal growth, and drive adult neurogenesis.
DLG4 **(discs, large homolog 4)** encodes a membrane-associated guanylate kinase and may function at postsynaptic sites (MIM 602887). Nearby, *DVL2* may also play a role in signal transduction (MIM 602151) and *CTDNEP1* is involved in a phosphatase cascade regulating nuclear membrane biogenesis (MIM 610684). *SLC2A4* is an insulin-regulated glucose transporter (MIM 138190). The variant identified here was previously associated with alkaline phosphatase levels in plasma.
EHBP1 **(EH domain binding protein 1)**. The mouse homologue of *EHBP1* was down-regulated in a transgenic *Pcsk9* mouse model and up-regulated in a *Pcsk9* knockout mouse.
FAM13A **(family with sequence similarity 13, member A)**. *FAM13A* has a putative role in signal transduction, and gene expression has been shown to be increased in response to hypoxia in cell lines from several tissues (MIM 613299).
FAM117B **(family with sequence similarity 117, member B)** is an uncharacterized protein. Nearby, *BMPR2* encodes a bone morphogenetic protein receptor (MIM 600799). Defects in *BMPR2* cause primary pulmonary hypertension.

***FN1* (fibronectin 1)** is a glycoprotein involved in cell adhesion and migration processes including embryogenesis, wound healing, blood coagulation, host defense, and metastasis (MIM 135600). Fibronectin is one of the first extracellular matrix proteins deposited at atherosclerosis-prone sites, and is central in the formation of atherosclerotic lesions.

***FTO* (fat mass and obesity associated)** contributes to the regulation of the global metabolic rate, energy expenditure and energy homeostasis (MIM 610966). Variants in this gene have been repeatedly associated with obesity-related phenotypes, and it may act through hypothalamic regulation of food intake.

***GPR146* (G protein-coupled receptor 146)** is an orphan G protein-coupled receptor. While no ligand has yet been identified, knockout mice exhibit reduced cholesterol levels (U. S. Patent Filing 20090036394). The adjacent gene, *GPER* encodes the intracellular G protein-coupled estrogen receptor 1 (MIM 601805).

***GSK3B* (glycogen synthase kinase 3 beta)** encodes a kinase involved in energy metabolism, neuronal cell development, and body pattern formation (MIM 605004). In mice, Gsk3b activity regulates pancreatic islet beta cell growth64. Nearby, *NR1I2* encodes a nuclear receptor that can form a heterodimer with retinoic acid receptor RXR and involved with homeostasis of numerous metabolites, including lipids (MIM 603065).

***HAS1* (hyaluronan synthase 1)** is one of three isozymes that synthesize hyaluronic acid, produced during wound healing and tissue repair to provide a framework for growth of blood vessels and fibroblasts (MIM 601463). The nearest gene, *FPR3* (formyl peptide receptor 3) is involved in host defense and inflammation (MIM 136539).

***HBS1L* (HBS1-like, S. cerevisiae)** encodes a member of the GTP-binding elongation factor family (MIM 612450). Variants at this locus regulate persistence of fetal hemoglobinin adults and other haematological traits.

***HDGF* (hepatoma derived growth factor) and *PMVK* (phosphomevalonate kinase)**. HDGF is a growth factor that may be involved in cell proliferation and differentiation (MIM 600339). *PMVK* catalyzes the fifth reaction of the cholesterol biosynthetic pathway (MIM 607622). Nearby, *CRABP2* (cellular retinoic acid binding protein 2) encodes a cytosol-to-nuclear shuttling protein involved in the retinoid signaling pathway (MIM 180231).

***IKZF1* (IKAROS family zinc finger 1)** is a transcription factor that regulates the low-density lipoprotein receptor in certain cell types.

***INSIG2* (insulin induced gene 2)**. *INSIG2* influences cholesterol metabolism, lipogenesis, and glucose homeostasis in diverse tissues (MIM 608660).

***INSR* (insulin receptor)** is a transmembrane tyrosine kinase receptor that binds insulin and stimulates glucose uptake (MIM 147670). The receptor activates several downstream pathways.

***LOC84931* (uncharacterized gene)**. The nearest gene to the lead signal is an uncharacterized gene with unknown function, and there are no obvious candidate genes in the region.

***LRPAP1* (low density lipoprotein receptor-related protein associated protein 1)** encodes a chaperone for the lipoprotein receptorrelated proteins (MIM 104225). *Lrpap1* knockout mice exhibit impaired export of LRP2 and VLDL receptors from the endoplasmic reticulum.

***KAT5* (K(lysine) acetyltransferase 5)**. *KAT5* is a positive regulator of PPARG transcription involved in adipogenesis.

***KCNK17* (potassium channel, subfamily K, member 17)** passes outward current under physiological potassium concentrations (MIM 607370). Variants 50 kb away at *KCNK16* have been implicated in type 2 diabetes.

***MARCH8* (membrane-associated ring finger (C3HC4) 8, E3 ubiquitin protein ligase) and *ALOX5* (arachidonate 5-lipoxygenase)**. *MARCH8* induces the internalization of several membrane glycoproteins (MIM 613335). *ALOX5* is a lipid metabolism enzyme that catalyzes the conversion of arachidonic acid to leukotrienes, inflammatory mediators implicated in atherosclerosis and several cancers (MIM 152390).

***MET* (met proto-oncogene (hepatocyte growth factor receptor))** encodes a receptor tyrosine kinase that regulates hepatocyte cell proliferation, migration and survival (MIM 164860).

***MIR148A* (microRNA 148a)**. MicroRNAs are short non-coding RNAs involved in post-transcriptional regulation of gene expression. miR-148a has been implicated in several cancers (MIM 613786).

***MOGAT2* (monoacylglycerol O-acyltransferase 2) and *DGAT2* (diacylglycerol O-acyltransferase 2)**. *MOGAT2* plays a central role in absorption of dietary fat in the small intestine76. *DGAT2* encodes one of two enzymes that catalyze the final reaction in the synthesis of triglycerides, in which diacylglycerol is covalently bound to long chain fatty acyl-CoA (MIM 606983).

***MPP3* (membrane protein, palmitoylated 3)** is a membrane-associated guanylate kinase that regulates trafficking and processing of cell-cell adhesion molecule nectin-1/*alpha* (MIM 601114).

***MTMR3* (myotubularin related protein 3)** encodes a phosphatase that binds to phosphoinositide lipids (MIM 603558).

***OR4C46* (olfactory receptor, family 4, subfamily C, member 46)**. This signal is located in a cluster of G-protein-coupled olfactory receptors, including OR5W2, OR5D13, and OR5AS1 (MIM 614273).

***PDXDC1* (pyridoxal-dependent decarboxylase domain containing 1)**. Little is known about this decarboxylase (MIM 614244). Variants at this locus have been shown previously to be associated with circulating sphingolipid levels. About 300 kb away, *PLA2G10* encodes a protein that releases arachidonic acid from cell membrane phospholipids (MIM 603603).

***PEPD* (peptidase D)** encodes an enzyme that hydrolyzes peptides with C-terminal proline or hydroxyproline residues and helps recycle proline (MIM 613230). Also at this locus are the genes encoding transcription factors CCAAT/enhancer binding protein alpha and gamma (*CEBPA* (MIM 116897), *CEBPG* (MIM 138972)), involved in adipogenesis. Variants in this locus are associated with adiponectin levels and type 2 diabetes in East Asians.

***PHC1* (polyhomeotic homolog 1) and *A2ML1* (alpha-2-macroglobulin-like 1)** is required to maintain the transcriptionally repressed state of many genes (MIM 602978). *A2ML1* is an inhibitor for several proteases and binds to low density lipoprotein receptor-related protein 1 (MIM 610627).

***PHLDB1* (pleckstrin homology-like domain, family B, member 1)**. *PHLDB1* is an insulin-responsive protein that enhances Akt activation, and *PHLDB1* expression is increased during adipocyte differentiation (MIM 612834).

***PIGV* (phosphatidylinositol glycan anchor biosynthesis, class V) and *NR0B2* (nuclear receptor subfamily 0, group B, member 2)**. *PIGV* is a mannosyltransferase that plays a role in multiple cellular processes, including protein sorting and signal transduction (MIM 610274). *NR0B2* is a transcriptional regulator involved in cholesterol, bile acid, and fatty acid metabolism and glucose-energy homeostasis.

***PPARA* (peroxisome proliferator activated receptor alpha)** encodes a nuclear transcription factor that regulates fatty acid synthesis, and oxidation and gluconeogenesis (MIM 170998). *PPARA* regulates the expression of lipoprotein receptors and cholesterol transporters involved in the reverse cholesterol transport pathway.

***PXK* (PX domain containing serine/threonine kinase)** plays a critical role in epidermal growth factor receptor trafficking by modulating ubiquitination of the receptor (MIM 611450).

***RBM5* (RNA binding motif protein 5)** is an hypothetical tumour suppressor gene encoding a nuclear RNA binding protein involved in the induction of cell cycle arrest and apoptosis (MIM 606884). Nearby, *MST1R* encodes macrophage stimulating 1 receptor and is involved in host defense (MIM 600168).

***RSPO3* (R-spondin 3)**. *RSPO3* encodes a protein that regulates beta-catenin signaling, promotes angiogenesis and vascular development (MIM 610574). In mouse, *Rspo3* is required for *Vegf* expression and endothelial cell proliferation. Variants in this locus are associated with waist-hip ratio, bone mineral density and renal traits.

***SETD2* (SET domain containing 2)** encodes a histone methyltransferase specific for lysine-36 of histone H3, a mark associated with active chromatin (MIM 612778). Nearby, *NBEAL2* encodes neurobeachin-like 2, which may play a role in megakaryocyte alpha-granule biogenesis (MIM 614169).

***SNX5* (sorting nexin 5)** encodes a protein that binds to phosphatidylinositol 4,5-bisphosphate and is involved in intracellular transport of cargo receptors from endosomes to the trans-Golgi network (MIM 605937).

***SNX13* (sorting nexin 13)**. This gene belongs to the sorting nexin (SNX) family and the regulator of G protein signaling (RGS) family (MIM 606589). It may be involved in several stages of intracellular trafficking.

***SOX17* (SRY (sex determining region Y)-box 17)** encodes a transcription regulator that plays a key role in the regulation of embryonic development and is required for normal looping of the embryonic heart tube (MIM 610928).

***SPTLC3* (serine palmitoyltransferase, long chain base subunit 3)**. SPTLC3 catalyzes the rate-limiting step of the de novo synthesis of sphingolipids (MIM 611120). Variants at this locus are associated with circulating sphingolipid levels.

***STAB1* (stabilin 1)** encodes a large, transmembrane receptor involved in angiogenesis, lymphocyte homing, cell adhesion, and receptor scavenging (MIM 608560). STAB1 mediates endocytosis of various ligands, including low-density lipoprotein. Variants at this locus have been associated with waist-hip ratio.

***TMEM176A* (transmembrane protein 176A)** is a transmembrane protein (MIM 610334).

***TOM1* (target of myb1)**. *TOM1* shares its N-terminal domain in common with proteins associated with vesicular trafficking at the endosomes (MIM 604700). Nearby, *HMOX1* encodes an essential enzyme in heme catabolism (MIM 141250). *Hmox1* knockout mice have low plasma triglycerides and altered composition of HDL.

***UGT1A1* (UDP glucuronosyltransferase 1 family, polypeptide A1)**. This complex locus encodes several glycosyltransferases that transform small lipophilic molecules, such as steroids, bilirubin, hormones, and drugs, into water-soluble excretable metabolites (MIM 191740). Variants at this locus are associated with serum bilirubin levels.

***VEGFA* (vascular endothelial growth factor A)** encodes a growth factor active in angiogenesis and endothelial cell growth, promoting cell migration, and inhibiting apoptosis (MIM 192240). Variants in this locus are associated with waist-hip ratio.

***VIM* (vimentin) and *CUBN* (cubilin, intrinsic factor-cobalamin receptor)**. VIM is an intermediate filament that controls the transport of LDL-derived cholesterol from a lysosome to the site of esterification (MIM 193060). CUBN is a receptor for high-density lipoproteins/apolipoprotein A-I, intrinsic factor-vitamin B_{12}, and albumin (MIM 602997).

***VLDLR* (very low density lipoprotein receptor)** binds VLDL and other lipoproteins and transports them into cells (MIM 192977). *VLDLR* is expressed on the capillary endothelium of skeletal muscle, heart, and adipose tissue.

***ZBTB42* (zinc finger and BTB domain containing 42) and *AKT1* (v-akt murine thymoma viral oncogene homolog 1)**. *ZBTB42* is a DNA-binding transcriptional repressor (MIM 613915). *AKT1* is a serine-threonine protein kinase that is activated by platelet-derived growth factor (MIM 164730). The Akt signaling pathway controls multiple cellular functions in the cardiovascular system, and murine Akt1 has an atheroprotective role.

Table S2.11: Overlap of lipid subfractions in Framingham with lipid associated loci

Locus	SNP	Lipid subfraction trait	A1/A2	N	MAF	Beta	*P*-value	Novel lipid locus	Lipid *P*-value
Overlap of Lipid Subfractions with HDL-C Associated Loci									
LIPC	rs1532085	HDL2 cholesterol subfraction	A/G	2,900	0.38	0.13	$2x10^{-6}$	No	$1x10^{-188}$
LIPC	rs1532085	HDL size	A/G	2,742	0.38	0.17	$4x10^{-9}$	No	$1x10^{-188}$
LIPC	rs1532085	Large particles of HDL	A/G	2,742	0.38	0.16	$6x10^{-8}$	No	$1x10^{-188}$
CETP	rs3764261	Intermediate density lipoprotein	A/C	2,742	0.31	-0.16	$9x10^{-8}$	No	$1x10^{-769}$
CETP	rs3764261	HDL2 cholesterol subfraction	A/C	2,900	0.31	0.18	$1x10^{-9}$	No	$1x10^{-769}$
CETP	rs3764261	LDL size	A/C	2,742	0.31	0.17	$7x10^{-8}$	No	$1x10^{-769}$
CETP	rs3764261	Large particles of LDL	A/C	2,742	0.31	0.14	$9x10^{-6}$	No	$1x10^{-769}$
CETP	rs3764261	HDL size	A/C	2,742	0.31	0.19	$6x10^{-10}$	No	$1x10^{-769}$
CETP	rs3764261	Large particles of HDL	A/C	2,742	0.31	0.22	$4x10^{-13}$	No	$1x10^{-769}$
CETP	rs3764261	HDL3 cholesterol subfraction	A/C	2,900	0.31	0.23	$1x10^{-14}$	No	$1x10^{-769}$
CETP	rs3764261	Apoliprotein AI concentration	A/C	2,885	0.31	0.19	$4x10^{-10}$	No	$1x10^{-769}$
LIPG	rs7241918	Apoliprotein AI concentration	G/T	2,885	0.17	-0.19	$2x10^{-7}$	No	$1x10^{-44}$
PLTP	rs6065906	Large particles of HDL	C/T	2,742	0.18	-0.18	$1x10^{-6}$	No	$5x10^{-40}$
PLTP	rs6065906	Medium particles of HDL	C/T	2,742	0.18	0.35	$1x10^{-21}$	No	$5x10^{-40}$
Overlap of Lipid Subfractions with LDL-C Associated Loci									
SORT1	rs629301	Apolipoprotein B concentration	G/T	2,821	0.21	-0.19	$2x10^{-8}$	No	$5x10^{-241}$
ApoE	rs4420638	ApoE concentration	G/A	2,260	0.16	-0.62	$9x10^{-10}$	No	$2x10^{-178}$
Overlap of Lipid Subfractions with Triglyceride Associated Loci									
GCKR	rs1260326	Apolipoprotein CIII concentration	T/C	2,484	0.45	0.18	$2x10^{-10}$	No	$2x10^{-239}$
LPL	rs12678919	Apoliprotein AI concentration	G/A	2,885	0.1	0.2	$1x10^{-5}$	No	$2x10^{-199}$
APOA1	rs964184	Medium particles of VLDL	G/C	2,742	0.14	0.26	$2x10^{-10}$	No	$7x10^{-224}$
APOA1	rs964184	Remnant like particles expressed as triglycerides	G/C	2,385	0.14	0.2	$5x10^{-6}$	No	$7x10^{-224}$
APOA1	rs964184	Remnant like particles expressed as cholesterol	G/C	2,468	0.14	0.19	$7x10^{-6}$	No	$7x10^{-224}$
APOA1	rs964184	Apolipoprotein B concentration	G/C	2,821	0.14	0.23	$4x10^{-9}$	No	$7x10^{-224}$

*The threshold used for significance is $P \leq 1.4x10^{-5}$. This corresponds to a Bonferroni correction for 23 subfractions and 151 SNPs found in the lipid subfraction dataset (0.05/(23*151)). LDL, low density lipoprotein; HDL, high density lipoprotein; VLDL, very low density lipoprotein

Table S2.12: Overlap of sphingolipids with lipid loci

					sphingomyelin (SPM)		**phosphatidyl-choline1 (PC)**		**phosphatidyl-choline2 (PC)**		**lysophosphatidyl-choline (LPC)**		**phosphatidyl-ethanolamine (PLPE)**		**PE based plasmologens (PE)**	
SNP	**Locus**	**Lipid trait**	**Novel/ known**	**Allele**	***P*-value**	**lipid side chain**	***P*-value**	**lipid side chain**	***P*-value**	**lipid side chain**	***P*-value**	**lipid side chain**	***P*-value**	**lipid side chain**	***P*-value**	**lipid side chain**
rs3905000	*ABCA1*	HDL	Known	A	$6.6x10^{-6}$	saturated SPM										
rs1532085	*LIPC*	HDL	Known	A			$2.5x10^{-7}$	PC O 34:1	$4.0x10^{-7}$	Total PC					$1.4x10^{-41}$	PE 36:4
rs1800775	*CETP*	HDL	Known	A					$1.8x10^{-6}$	Ether PC						
rs261332	*LIPC*	HDL	Known	A											$3.9x10^{-28}$	PE 38:5
rs174556	*FADS1-FADS2*	LDL	Known	T	$1.2x10^{-12}$	SPM 16:1	$5.6x10^{-101}$	PC 38:4	$6.6x10^{-30}$	PC 40:5	$3.5x10^{-75}$	LPC 20:4	$1.6x10^{-20}$	PLPE 18:0/20:4	$6.1x10^{-20}$	PE 34:2
rs364585	*SPTLC3*	LDL	Novel	A	$6.2x10^{-10}$	SPM 17:0										
rs1367117	*APOB*	LDL	Known	A	$3.5x10^{-7}$	SPM 16:0										
rs1864163	*CETP*	LDL	Known	A			$1.2x10^{-6}$	PC O 38:5	$4.0x10^{-6}$	Ether PC						
rs4970834	*CELSR2*	TC	Known	T	$5.8x10^{-6}$	SPM 16:0										
rs12916	*HMGCR*	TC	Known	T			$4.2x10^{-6}$	PC 34:2								
rs2304130	*NCAN*	TC	Known	A					$3.6x10^{-7}$	PC 40:7						
rs1260326	*GCKR*	TG	Known	T			$1.4x10^{-7}$	PC 34:4	$1.2x10^{-6}$	PC 32:2	$3.0x10^{-6}$	LPC 16:1				
rs3198697	*PDXDC1*	TG	Novel	T	$9.5x10^{-13}$	PC 38:3			$4.8x10^{-9}$	LPC 20:3						

*The threshold used for significance is $P \leq 1.5x10^{-5}$. This corresponds to a Bonferroni correction for 23 principal components and 145 SNPs found in the sphingolipids dataset (0.05/(23*145)).

CHAPTER III

GREGOR: evaluating global enrichment of trait-associated variants in epigenomic features using a systematic, data-driven approach

3.1 Abstract

The majority of variation identified by genome wide association studies falls in non-coding genomic regions and is hypothesized to impact regulatory elements that modulate gene expression. Here we present a statistically rigorous software tool GREGOR (**G**enomic **R**egulatory **E**lements and **G**was **O**verlap algo**R**ithm) for evaluating enrichment of any set of genetic variants with any set of regulatory features. Using variants from five phenotypes, we describe a data-driven approach to determine the tissue and cell types most relevant to a trait of interest and to identify the subset of regulatory features likely impacted by these variants. Last, we experimentally evaluate six predicted functional variants at six lipid-associated loci and demonstrate significant evidence for allele-specific impact on expression levels. GREGOR systematically evaluates enrichment of genetic variation with the vast collection of regulatory data available to explore novel biological mechanisms of disease and guide us toward the functional variant at trait-associated loci.

GREGOR, including source code, documentation, examples, and executables, is

Official citation: Schmidt et al. (2015)

available at `http://genome.sph.umich.edu/wiki/GREGOR`.

3.2 Introduction

The list of common genetic variants associated with complex disease continues to grow as a result of increasingly powered genome wide association studies (GWAS) (Welter et al., 2014). A large proportion of the associated variants are non-coding and it has proven difficult to identify the functional variant at loci with many variants in tight linkage disequilibrium (LD). In addition, these loci often account for only a small percentage of the trait heritability which makes any minor alteration of transcript levels difficult to detect. Although eQTLs (expression quantitative trait loci) in relevant tissues can highlight loci where variants likely impact transcription of nearby genes, fine-mapping of the causal variant is plagued by the same LD patterns that impact disease association studies. Common variation located outside of protein-coding regions modulates regulatory elements in a cell-type specific manner (Claussnitzer et al., 2014; Ernst et al., 2011; Kichaev et al., 2014; Lo et al., 2014; Maurano et al., 2012; Parker et al., 2013; Pickrell, 2014; Thurman et al., 2012; Trynka et al., 2013). Examining disease-associated variants in relation to genomic regions of functional importance can give insight into the molecular mechanisms leading to disease phenotypes, particularly when all associated variants are considered in aggregate.

Our understanding of the location of regulatory elements in the genome has expanded with the advent of chromatin immunoprecipitation followed by high-throughput DNA sequencing (ChIP-Seq) technology and the Encyclopedia of DNA Elements (ENCODE) Project (ENCODE Project Consortium, 2012). However, it is challenging to untangle meaningful biological understanding in a systematic manner, given

the diverse set of data available from hundreds of cell types and tissues. With the notion that non-coding genetic variation plays a role in transcriptional regulation via regulatory epigenomic features, we can harness these data to gain knowledge of important biological mechanisms. For example, genetic variation that impacts local chromatin or methylation states and DNA accessibility can impact transcription in a given cell. In a majority of associated genomic regions, the SNP (single nucleotide polymorphism) supported by ENCODE data is a SNP in strong LD with the top reported GWAS SNP (Schaub et al., 2012). Systematic chromatin profiling has revealed that variants linked with the GWAS index variant, defined here as the most strongly associated variant, are often positioned within enhancer elements active in relevant cell types (Ernst et al., 2011). Furthermore, the overlap of particular histone methylation marks with trait associated variants is cell type-specific, suggesting that gene regulation is influenced by trait alleles in a cell type-specific manner (Trynka et al., 2013). Previous work has used chromatin profiles and other ChIP-seq experimental data to investigate GWAS variation and predict the impact of candidate variants in particular genomic regions (Boyle et al., 2012; Claussnitzer et al., 2014; Kichaev et al., 2014; Lo et al., 2014; Maurano et al., 2012; Pickrell, 2014; Thurman et al., 2012; Ward and Kellis, 2012). However, these methods often do not consider an appropriate control set for evaluating enrichment and do not always carefully evaluate the most relevant tissues or cell types for enrichment of trait-specific variation. Identifying causal variants and mechanisms at GWAS loci remains a universal scientific challenge.

With this motivation, we developed a statistically rigorous approach to quantify enrichment of trait-associated variants in experimentally annotated functional elements such as open chromatin states, histone marks and protein-binding sites in

relevant cell types to develop a clearer understanding of the underlying regulatory mechanisms. We apply an algorithm and systematic scientific method for prioritizing functional candidate variants at genome-wide significant trait-associated loci. Our aims are threefold:

(i) elucidate the important tissue/cell types in which genetic variation impacts transcription for a particular trait,

(ii) narrow our focus of the regulatory features underlying transcription disrupted by trait-associated variants, and

(iii) use positional overlap with selected regulatory domains to identify potential functional candidates at trait-associated loci.

To address these aims, we evaluate genetic variation identified by GWAS for five metabolic phenotypes: blood pressure (International Consortium for Blood Pressure Genome-Wide Association Studies et al., 2011), (C. Newton-Cheh and P. Munroe, unpublished data), body mass index (Locke et al., 2015), coronary artery disease (Coronary Artery Disease (C4D) Genetics Consortium, 2011; Schunkert et al., 2011), lipids (Global Lipids Genetics Consortium et al., 2013) and type 2 diabetes (Morris et al., 2012). We present GREGOR (**G**enomic **R**egulatory **E**lements and **G**was **O**verlap algo**R**ithm), an open source tool for evaluating enrichment as a method to query the vast array of ENCODE data for the design of functional experiments, enabling scientists with non-computational backgrounds to prioritize variants and loci for functional follow-up (Figure 3.1).

3.3 Methods

We hypothesize that the index variant reported by GWAS is not necessarily the causal variant, owing to LD at associated regions. To account for this, we first create a list of all potential causal variants by selecting variants in strong LD (r^2>0.7) with trait-associated index SNPs in whole genome sequenced samples: the 1000 Genomes Phase 1 version 2 European Panel (1000 Genomes Project Consortium et al., 2010). Reference data from non-European populations from the 1000 Genomes Project are also available with GREGOR for selection of LD proxies. Although many indicators of regulatory potential exist for non-coding regions, we select DNase hypersensitive sites (DHSs) as a general marker of functional importance to address our first scientific question: which cell type shows strongest enrichment of trait-associated loci? We gather data from the ENCODE Project and when experimental replicates are available, we calculate the union of DHSs derived from the same tissue (Table S3.1). We then examine overlap of these potential causal SNPs with DHSs from various different tissue categories. By the same approach, we later evaluate the position of the index SNPs and their LD proxies relative to histone methylation marks and ChIP-seq transcription factor binding sites (TFBS) , as well as previously defined functional chromatin states.

We calculate the total number of trait-associated loci at which either the index SNP or at least one of its LD proxies overlaps with a regulatory region across the genome. In order to evaluate the significance of this observed overlap at each individual regulatory feature, we estimate the probability of the observed overlap of GWAS SNPs relative to expectation using a set of matched control variants. For each GWAS index SNP, we identify a set of 500 control SNPs randomly selected from across the

genome that match the index SNP for: (i) number of variants in LD, (ii) minor allele frequency (±1%) and (iii) distance to the nearest gene. When two or more GWAS index SNPs match each other following the three criteria above, they share a set of control SNPs. We consider that the number of index SNPs within its matched control set of SNPs that overlaps a given feature follows a binomial distribution with two parameters: (i) the number of GWAS index SNPs present in the control set (1 or greater), and (ii) the proportion of SNPs within the control set or their LD proxies that physically overlaps a feature. Considering the number of index SNPs that overlaps with a feature, we compute the sum of independent binomial random variables. Then for each regulatory feature, we calculate the fold-enrichment over expectation and an enrichment P-value that represents the probability that the overlap of control SNPs represented as a cumulative probability distribution is greater than or equal to the observed overlap that we see from GWAS index SNPs (Figure S3.1, Table 3.1).

We evaluated the performance of our method using a range of parameters including different numbers of variants in LD in the matched control sets, and matched control set size (Figure S3.2). The magnitude of enrichment is generally consistent across ranges of these parameters, and the subsequent results use r^2=0.7 with matched control set size of > 500. P-values generated based on randomly permuted sets of non-associated matched control SNPs are highly concordant with estimated P-values (Section 3.7.4, Figure S3.3).

We attempted to evaluate the type I error rate of our enrichment method. We tested enrichment of 50 sets of randomly selected SNPs in DHSs of different tissues. SNP sets were matched with lipid-associated SNPs on 3 properties: number of LD proxies, minor allele frequency and distance to the nearest gene. A QQ plot reveals P-values that closely follow the null uniform distribution, whereas the P-value dis-

tribution for lipid-associated variants sharply deviates from the null (Figure 3.2a). Additionally, we investigated type I error by first partitioning DHSs of each tissue into genic landmark categories (Parker et al., 2013) and then randomly shuffling within each category. After re-combining the DHS categories for each tissue, we evaluated enrichment of the lipid-associated variants and again compared the results to the original P-value distribution (Figure 3.2b).

3.4 Results

3.4.1 Prioritizing tissue types for five phenotypes using DNase hypersensitivity sites

Our first objective is to use available epigenomic data to identify which tissues are the most biologically relevant to the trait-specific genetic variation identified by GWAS. We evaluated enrichment of independent GWAS loci for five related phenotypes: 99 blood pressure loci (BP; 2.2% trait variance explained), 97 body mass index loci (BMI; 2.7% trait variance explained) (Locke et al., 2015), 36 coronary artery disease loci (CAD; 10% trait variance explained) (Schunkert et al., 2011), 157 lipid loci (high- and low- density lipoprotein cholesterol, total cholesterol and triglycerides; 10-12% trait variance explained) (Global Lipids Genetics Consortium et al., 2013) and 65 type 2 diabetes loci (T2D; 10.7% trait variance explained) (Morris et al., 2012). DHSs are open regions of DNA accessible to protein binding, and are important in the transcriptional activity within a given cell. ENCODE has experimentally identified DHSs using DNase-seq in hundreds of cell types. We evaluate enrichment of GWAS loci in the union DHSs of cell types derived from the same tissue (Table S3.1). By testing five sets of trait-associated SNPs in DHSs of 41 tissue types, we set a Bonferroni corrected threshold for significance at $P<2.4\text{x}10^{-4}$.

GWAS loci were significantly enriched in DHSs of tissues that are remarkably consistent with our biological understanding of the trait (Figure 3.3). For example,

BP-associated variants are highly enriched in DHSs in cell types derived from blood vessel (P=1.2x10^{-9}; fold enrichment 1.5) and heart (P=5.3x10^{-8}; fold enrichment 1.6); CAD-associated variants in DHSs from heart (P=2.3x10^{-5}; fold enrichment 1.7) and blood (P=5.6x10^{-5}; fold enrichment 1.4); lipid-associated variants in DHSs from liver (P=2.0x10^{-14}; fold enrichment 1.6), monocytes (P=7.1x10^{-13}; fold enrichment 1.9) and blood (P=4.7x10^{-11}; fold enrichment 1.4); and BMI-associated variants in DHSs in frontal cortex (P=8.8x10^{-5}; fold enrichment 1.7). We also find enrichment of BMI-associated variants in DHSs of human olfactory neurosphere-derived cells from mucosal biopsies (P=4.2x10^{-5}; fold enrichment 1.7), suggesting a plausible link between olfaction and food intake. However, there are other cases in which we find enrichment of trait-associated variants in unexpected tissue types. For example, although we observe significant enrichment of T2D-associated variants in pancreatic tissue as expected (P=1.0x10^{-4}; fold enrichment 1.6), we see stronger evidence for enrichment in heart tissue (P=1.4x10^{-6}; fold enrichment 1.7) and embryonic stem cells (P=2.5x10^{-6}; fold enrichment 1.5). We used this knowledge to guide subsequent enrichment analysis of other epigenomic features by focusing on the most significant cell types to reduce the multiple testing burden in subsequent assessments of additional regulatory features. This data-driven approach to reduction of a large set of potentially relevant regulatory elements in a myriad of cell lines and tissues can be used for phenotypes where little is known about the biology, and may also identify novel tissues where these GWAS loci are actively transcribed. Alternatively, investigators might bypass this step and instead use *a priori* biological knowledge to focus on a specific tissue or cell type. One could also integrate the two approaches to choose some empirically-selected cell types but up-weight biologically relevant cell types.

We additionally investigate whether enrichment is tissue type-specific. Given the

wealth of DHS data available, often in replicates and for multiple cell types from the same tissue type, we hypothesize that each cell type has some level of missing data and artifacts. To address this, we define consensus regions of open chromatin that are commonly shared among at least 50% of all cell types within a single tissue group, and re-evaluate enrichment of lipid-associated GWAS variants. We additionally compare results for consensus thresholds (proportion of cell types required to show a DHS at that genomic position) of 100%, 75%, 25% and the union of cell types derived from the same tissue. We found that when we used stricter definitions to select functional regions (*e.g.* 100% of cell types were required to share the DHS), we typically observed higher fold enrichment, but less significant enrichment P-values (Figure S3.4). Conversely, when we relaxed the criterion to allow DHSs observed in only 25% of cell types, we typically observed stronger P-values but lower fold enrichment. This is likely due to inclusion of more artefactual DHSs using the relaxed definition, but exclusion of true DHSs under the strict definition. In subsequent analyses, we used the most relaxed definition of regulatory elements by including any element observed in at least one replicate or cell type within each tissue category. We opted to be more inclusive to allow for the most complete identification of DHSs.

3.4.2 Prioritizing regulatory elements in selected tissues

Following prioritization of important tissue types for GWAS of a specific phenotype, we next selected specific regulatory elements that were enriched for GWAS variants in cell types derived from relevant tissues, focusing solely on the tissues selected in Section 3.4.1. We evaluated enrichment of trait-associated variants in chromatin states predicted from histone methylation marks and a learned multivariate hidden Markov model (Ernst et al., 2011) (Figure 3.4). Confirming previous reports (Maurano et al., 2012), we found significant enrichment of genetic variation in weak

and strong enhancer states for nearly all phenotypes tested. Trait-associated variants are most highly enriched in active promoters commonly marked by H3K4me2, H3K4me3, acetylation, or H2A.Z. There is less striking enrichment in domains that contain repressed genes such as H3K9me2, H3K9me3 or H3K27me3.

We further evaluated enrichment in TFBS and histone modifications identified by ChIP-Seq. We investigated any Tier 1 or 2 ENCODE cell types available for relevant tissues identified in Section 3.4.1, taking the union of experimental replicates when available. For cell types HepG2, Monocytes CD14+ (RO01746), GM12878, K562 and CD20+ (RO01778), we find significant enrichment of lipid-associated variation for key transcriptional machinery including RNA Polymerase II (P=6.2x10^{-24}; fold enrichment 2.0) and the ubiquitous transcription factor SP1 (P=1.510^{-15}; fold enrichment 3.0). In addition, lipid-associated variants are highly enriched in binding sites of RCOR1 (P=1.8x10^{-16}; fold enrichment 2.1), EP300 (P=1.2x10^{-14}; fold enrichment 2.0), JUND (P=2.2x10^{-14}; fold enrichment 2.0) and H3K4me3 (P=1.4x10^{-13}; fold enrichment 2.1). We tested a total of 158 regulatory features, 75 of which reach Bonferroni significance with P<3.2x10^{-4} (Table S3.2). We are particularly interested in 15 known lipid gene regulators as well as 16 transcription factors and 4 histone markers associated with lipid change in the literature. Of the 75 Bonferroni significant regulatory features, 18 of these are among this *a priori*-defined lipid-related list of 35 elements.

3.4.3 Prioritizing candidate functional variants using selected regulatory elements in relevant tissues

As we gain knowledge about the transcriptional machinery that acts in concert with trait-associated genetic variation, we can make more informed predictions about potential functional variants at a single locus. We hypothesize that variants present

within multiple regulatory domains are more likely to play a role in transcriptional regulation within a cell. Subsequently, we can use this information in combination with functional protein-coding information, transcript level annotation, and deleteriousness scoring to prioritize loci and individual variants for functional follow-up.

We proceeded to prioritize potential functional variants in the 157 known lipid-associated loci (Figure 3.5). With the assumption that a protein-coding variant is likely the functional driver of transcription at a given locus, we excluded any lipid-associated loci from follow-up consideration that contains at least one non-synonymous variant in LD (r^2>0.7) with the GWAS index SNP. This resulted in 103 remaining loci for further evaluation. We next examined our results from Step 2 to focus on the selected transcription factors and histone marks, and prioritized loci at which multiple transcription factors bind in blood, monocytes, or liver. In a data-driven approach, we flagged variants that overlap with a subset of significantly enriched regulatory domains as plausible functional candidate SNPs (n=23). We evaluated overlap of GWAS variants at candidate loci in lipid gene regulators as well as transcription factors and histone marks involved with lipid change in the literature. Variants at a set of five of these loci that overlap with at least eight (25%) lipid-related regulatory features were commonly found using both the data-driven and biological-driven selection of regulatory features, including the known functional variant rs12740374 at *SORT1* (Musunuru et al., 2010). Many of these candidate variants are also eQTLs in liver, omental fat or subcutaneous fat or had at least one surrogate SNP in LD (r^2>0.7) with the eQTL SNP at that locus (eQTL P<1x10^{-3}) (Schadt et al., 2008). In addition to considering these various data, we counted the number of variants at each locus and focused on loci with relatively few numbers of variants to increase the likelihood of identifying the functional variant.

Thus, we narrowed down the list of lipid loci that likely have a strong impact on regulating transcription to guide us to promising candidates for functional follow-up.

After analyzing the overlap of non-coding variants with biological TFBS from ChIP-seq and using our criteria of non-coding variants, eQTLs, and number of variants at a locus, we chose five loci and picked one SNP from each region that had some evidence of being the functional variant due to overlap with the most regulatory regions for further study (*FAM117B*: rs11692610; *ANGPTL8*: rs737337; *SPTLC3*: rs1321940; *IRF2BP2*: rs526936; *ADH5*: rs1800759) (Figure 3.6). At each locus, we selected an additional variant with no predicted C/EBP binding site overlapping as an internal control (*FAM117B*: rs11694172; *ANGPTL8*: rs3810308; *SPTLC3*: rs364585; *IRF2BP2*: rs514230; *ADH5*: rs2602836). Variant rs12740374 from the *SORT1* locus has previously been demonstrated to alter a C/EBP TFBS (Musunuru et al., 2010), and thus was used as a positive control here (rs629301 as the *SORT1* locus internal control).

We next attempted to directly determine the allele-specific effects of the non-coding SNP polymorphism on transcription factor binding at each of the six lipid loci. We generated luciferase constructs containing ± 300-400 bp around each genetic variant (generating the alternate allele with site-directed mutagenesis) and transfected them into HepG2 cells over-expressing C/EBP-β. We normalized luciferase activities to the pcDNA3.1-co-transfected groups (control construct with no C/EBP-β DNA inserted), and found robust luciferase activity increase in the rs12740374-T construct compared with rs12740374-G from the *SORT1* locus (fold increase=1.8, P=4x10^{-4}), which was consistent with the previous report. Similarly, for the other five loci examined, the single nucleotide changes in other predicted functional SNP sites caused significant luciferase activity differences in response to C/EBP-β over-

expression (P<0.05), indicating those non-coding variants may change transcription factor binding activity in GWAS loci and possibly affect downstream gene expression (Figure 3.7). After correction for 12 tests (2 SNPs at six loci), we still find significant differences between the two alleles (P<0.05) at the candidate functional SNP for five out of the six loci (*SORT1*: rs12740374-T; *ANGPTL8*: rs737337-C; *FAM117B*: rs11692610-T; *IRF2BP2*: rs526936-C; *SPTLC3*: rs1321940-G). In contrast, the luciferase signal changes of the internal control SNP constructs were significant at Bonferroni levels for only two of the six loci.

Our results are generally supported by *post hoc* annotations of predicted regulatory elements defined by RegulomeDB (Boyle et al., 2012). For example, the RegulomeDB score of the GWAS HDL cholesterol-associated index SNP at the *SPTLC3* locus (rs364585) is 5, indicating that there is TF binding or DNase peak epigenomic data to support its functionality. In contrast, the RegulomeDB score of our predicted functional SNP at this locus (rs1321940) is 2a, indicating that it is likely to affect binding based on evidence of TF binding, and the presence of a matched TF motif, DNase footprint, and DNase peak. Among all 18 variants within r^2>0.7 of the index SNP at this locus, only 2 SNPs have a score of 2b or better. We observe similar trends for other loci at which we performed experimental follow-up (Figure 3.6). Counterintuitively, the known functional variant rs12740374 at the *SORT1* locus has a RegulomeDB score of 2b, whereas the variant reported by GWAS in that region, rs629301, has a higher score of 1f. This result emphasizes the need to consider multiple sources of data when prioritizing functional candidates for experimental follow-up. Although annotation of individual variants is useful in predicting the potential impact of a single variant, GREGOR considers all trait-associated variants in aggregate to prioritize which functional elements and in which tissues they

are most relevant for the trait being examined. An alternative is considering the entirety of ENCODE data, much of which will represent irrelevant tissue types or highly correlated data sets.

3.5 Discussion

We have developed a systematic approach for evaluating enrichment of trait-associated variants in epigenomic features, allowing us to prioritize tissues, regulatory elements, and potential functional variants that affect transcriptional regulation (Figure 3.1). Our method takes into account all potential causal variants at a locus due to LD and estimates enrichment with particular regulatory features using matched control variants. It is an unbiased approach that can be used to narrow the focus of cell types and regulatory features that does not rely on *a priori* knowledge of biological mechanisms. The resultant findings will guide us to a more global understanding of the underlying epigenomic architecture leading to trait-specific variation.

We present here one reasonable approach for prioritizing the potential functional variant at a locus. We attempted to select loci with the best chance of demonstrating a functional variant for experimental follow-up. However, our approach is limited to only one potential functional variant per locus and does not claim to definitively identify the true or only functional variant at any locus. More comprehensive interrogation of variation within a locus will be required to fully understand the underlying molecular mechanisms involved.

Different cell types and tissues are more easily accessible than others for sequencing. This approach will become even more impactful as we develop an increasingly comprehensive and diverse interrogation of the epigenome to answer important biological questions about the regulatory role of non-coding variation. In all, this ap-

proach will help guide our knowledge of the important mechanisms occurring outside of protein-coding regions that underlie cell-type-specific transcriptional regulation.

3.6 Acknowledgements

This work was directly supported by HL094535 (CJW) and by the National Science Foundation Open Data Integrative Graduate Education and Research Traineeship (IGERT) Grant 0903629 (EMS). CJW was additionally supported by HL109946 and HL127564.

The authors thank Praveen Sethupathy and Michael Stitzel for helpful discussion at the initiation of the project. We also thank Martin Buchkovich, Charles Burant, Ruth Loos, and Stephen Parker for helpful input.

3.7 Supplementary Methods

3.7.1 Data acquisition and pre-processing

DNase-seq ENCODE data for all available cell types were downloaded in the processed narrowPeak format. The local maxima of the tag density in broad, variable-sized hotspot regions of chromatin accessibility were thresholded at FDR 1% with peaks set to a fixed width of 150 bp. Individual cell types were further grouped into 41 broad tissue categories (`http://genome.ucsc.edu/ENCODE/cellTypes.html`) by taking the union of DHSs for all related cell types and replicates (Table S3.1). We also obtained a set of BED files in hg19 assembly from the Integrative Analysis and original ENCODE analysis. These data include uniformly processed datasets in 125 cell types generated by the "Open Chromatin" (Duke University) and University of Washington (UW) ENCODE groups. Data processed during the ENCODE Integrative Analysis were downloaded for available tissues. Otherwise, data from the original ENCODE analysis were obtained. We examined the overlap of

DHSs across different cell types, and found that as expected, cell types derived from related tissues generally clustered together. In addition, we examined chromatin state segmentation by HMM generated from ENCODE/Broad in nine human cell types, as well as transcription factor binding sites by ChIP-seq from the ENCODE Analysis Working Group (AWG) including ENCODE/HudsonAlpha (HAIB), ENCODE/Stanford/Yale/Davis/Harvard (SYDH), ENCODE/Univ of Chicago, ENCODE/Open Chrom (UT Austin), and ENCODE/Univ of Washington (UW). No dataset analyzed was under embargo.

3.7.2 Selecting matched control SNPs for GWAS index SNPs

For each GWAS locus, we selected a set of matched control SNPs based on 3 criteria: **i)** number of variants in LD (r^2>0.7; ± 8 variants), **ii)** minor allele frequency (± 1%), and **iii)** distance to nearest gene (± 11,655 bp). To calculate the distance to the nearest gene, we calculated the distance to the 5' flanking gene (start and end position) and to the 3' flanking gene and then used the minimum of these 4 values. If the SNP fell within the transcribed region of a gene, the distance was 0.

3.7.3 Estimating probability of observed and expected overlap between a regulatory feature and GWAS locus

We estimated the probability that a set of GWAS loci overlap with a regulatory feature more often than we expect by chance using the following method. We considered a GWAS locus as the GWAS index SNP or a SNP in LD with the index SNP (r^2>0.7). For each regulatory feature, we counted the number of GWAS loci in which we observed physical overlap with at least one experimentally defined genomic region of the feature. The number of GWAS index SNPs in the ith matched control set that demonstrates positional overlap with a given epigenomic feature, written as s_i, follows a binomial distribution with parameters n_i and p_i. The parameter n_i

is equal to the number of index SNPs present in the ith control set. The second parameter p_i is calculated as the number of variants in the ith control set or their LD proxies that overlaps with the feature, divided by the total number of variants in the ith control set. If we assume there are r control sets in total, the number of index SNPs from all control sets that falls in a single feature is the sum of independent non-identical binomial random variables:

$$S = \sum_{i=1}^{r} s_i \tag{3.1}$$

In most cases only one index variant is assigned to a matched control set, but there are some exceptions where more than one index SNP could match on the same 3 properties. We estimate an enrichment P-value for any given s as $P(S \geq s)$. P is the cumulative right tail probability based on the distribution of S and is calculated using a saddlepoint approximation method (Eisinga et al., 2013).

3.7.4 Permutation testing to evaluate estimated *P*-values

We performed up to 100,000 permutations to evaluate our enrichment P-value estimation method and found the results to be highly concordant for permutation P-values less than $1x10^{-5}$ that could be estimated (Figure S3.3). To assess the expected overlap with a regulatory domain, we generated 100,000 random permuted sets of non-associated matched control SNPs based on the criteria described above. We selected a control variant from the control pool for each locus and identified the variants in LD, resulting in 100,000 control sets. We evaluated the random SNP lists for overlap with each functional domain by averaging the number of SNPs that fell within the experimentally annotated regions from each control set that had at least one variant overlapping a regulatory element. This approach assumes that only one variant located in a regulatory region at each locus is responsible for the

association signal. We calculated an empirical P-value for each regulatory dataset as the proportion of random sets with an equal or greater number of loci overlapping the regulatory domain than the observed set of trait-associated variants. For small P-values that could not be estimated (e.g. $P<1\text{x}10^{-5}$ for 100,000 permutations), we used a normal approximation of the empirical overlap distribution to estimate P-values.

3.7.5 Luciferase expression constructs

To characterize the intergenic region around the candidate SNPs, 600-800 bp fragments containing the SNPs from human chromosomes were cloned into the pGL4-Promoter vector (Promega), in the 5'-to-3' orientation (toward the GWAS candidate gene), upstream of the firefly luciferase gene (Table S3.3). The QuikChange Site-Directed Mutagenesis Kit (Stratagene) was used to alter single nucleotides at the targeted SNP sites. All constructs were verified by DNA sequencing.

3.7.6 Luciferase expression assays

HepG2 cultured human hepatoma cells were transfected at roughly 50% confluence and maintained in DMEM with 10% FBS. The firefly luciferase constructs were co-transfected with either the C/EBP-β expression plasmid (pcDNA3.1-C/EBP-β) or empty pcDNA3.1 vector, together with the Renilla luciferase pRL-null Vector (Promega) as internal control, using the Lipofectamine 2000 transfection reagent (Invitrogen) in the ratio 0.25μg:0.25ug:25ng:2.5μl mixed with Opti-MEM I Reduced Serum Medium (Invitrogen) for a 50μl mix used for each well of 24-well plates. Forty-eight hours after transfection, firefly and Renilla luciferase activities were measured using the Dual-Luciferase Reporter Assay System (Promega) according to the manufacturers protocol, using untransfected cells to adjust for background activity.

3.7.7 Data Access

GREGOR documentation and software download, http://genome.sph.umich.edu/wiki/GREGOR; ENCODE Consortium, http://genome.ucsc.edu/ENCODE/dataMatrix/encodeDataMatrixHuman.html; Chromatin state segmentation by HMM from ENCODE/Broad in 9 human cell types, http://genome.ucsc.edu/cgibin/hgFileUi?g=wgEncodeBroadHmm&db=hg19; GWAS results for all traits and diseases including those studied here, http://www.genome.gov/gwastudies/. Data used from the latest blood pressure GWAS are not yet published.

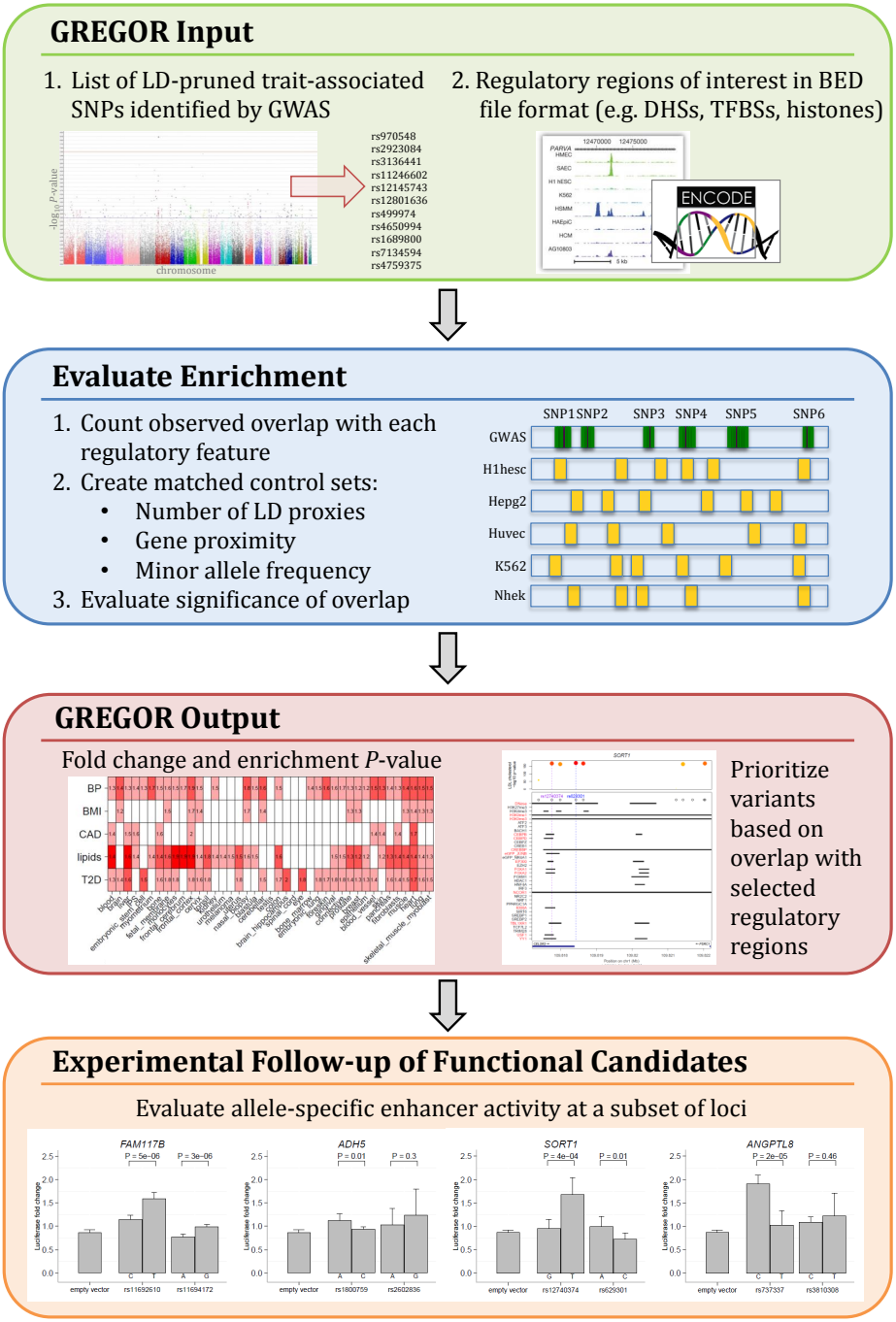


Figure 3.1: GREGOR study design.

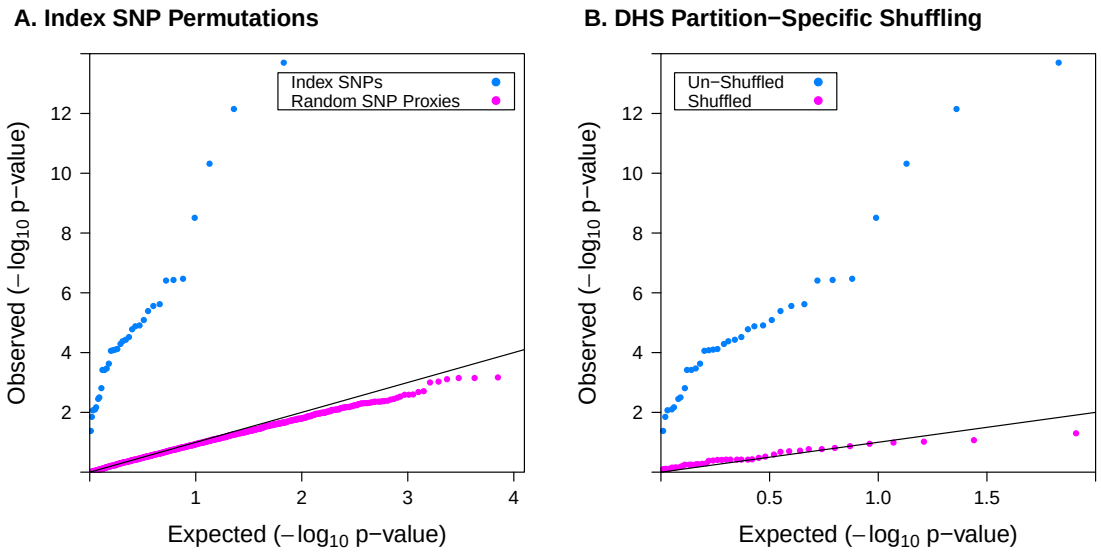


Figure 3.2: Type I error assessment of GREGOR algorithm performance. **(A)** The *P*-value distribution of enrichment in DHSs is shown for 50 SNP proxy lists (pink) together with the enrichment *P*-value distribution of the true lipid-associated variants (blue). Proxy lists were generated by choosing SNPs that matched on i) number of LD proxies, ii) minor allele frequency, and iii) gene proximity, but were otherwise randomly selected from across the genome. **(B)** DHSs were partitioned into mutually exclusive genic landmark categories based on GENCODE annotation (e.g. 3'UTR, 5'UTR, intron, coding exon, intergenic TSS distal and proximal) and randomly shuffled. After re-combining the categories for each tissue, we evaluated enrichment of lipid-associated variants (pink) and compared with the *P*-value distribution in the original DHSs (blue). *Abbreviations*: UTR, untranslated region; TSS, transcription start site.

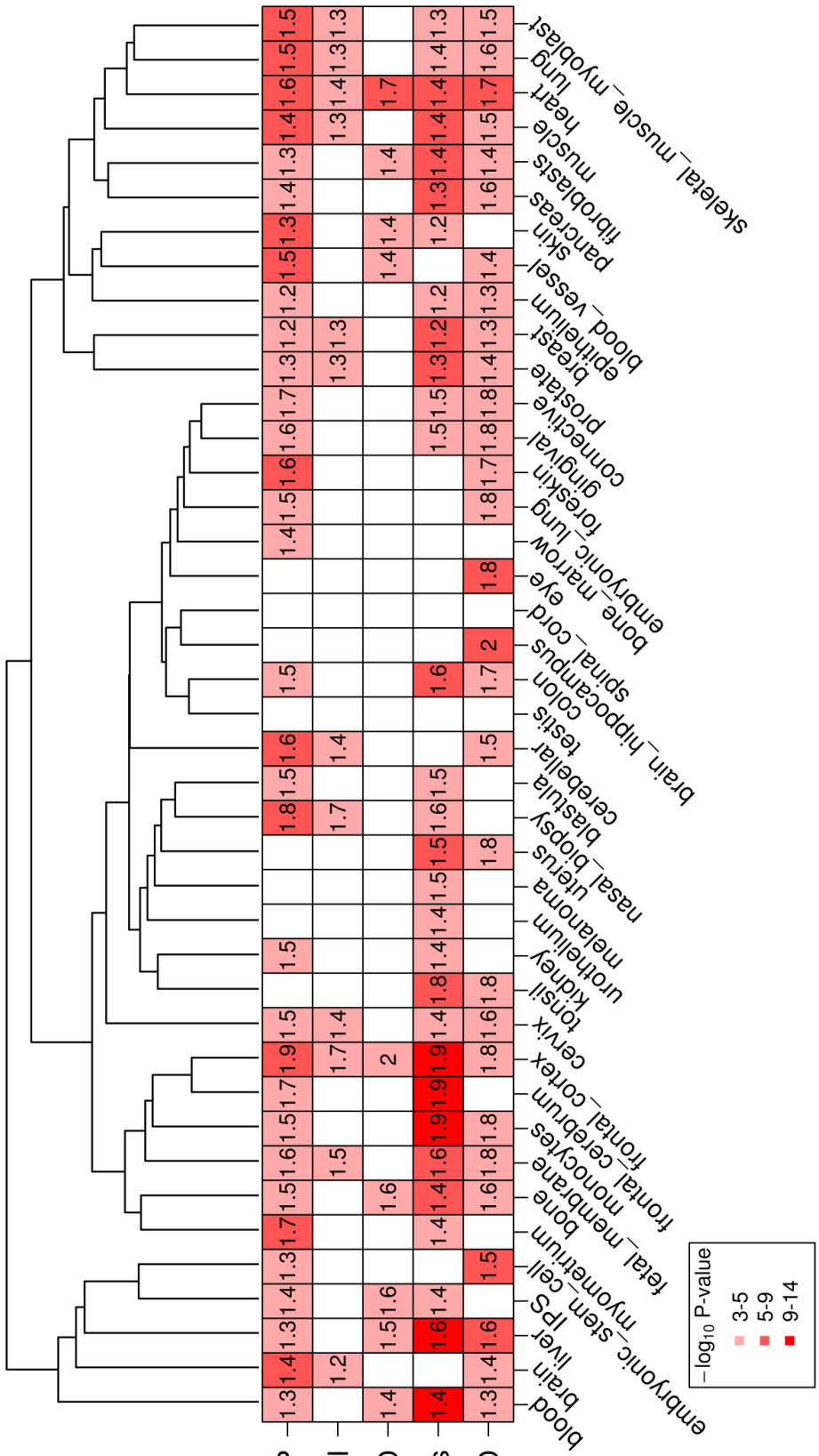
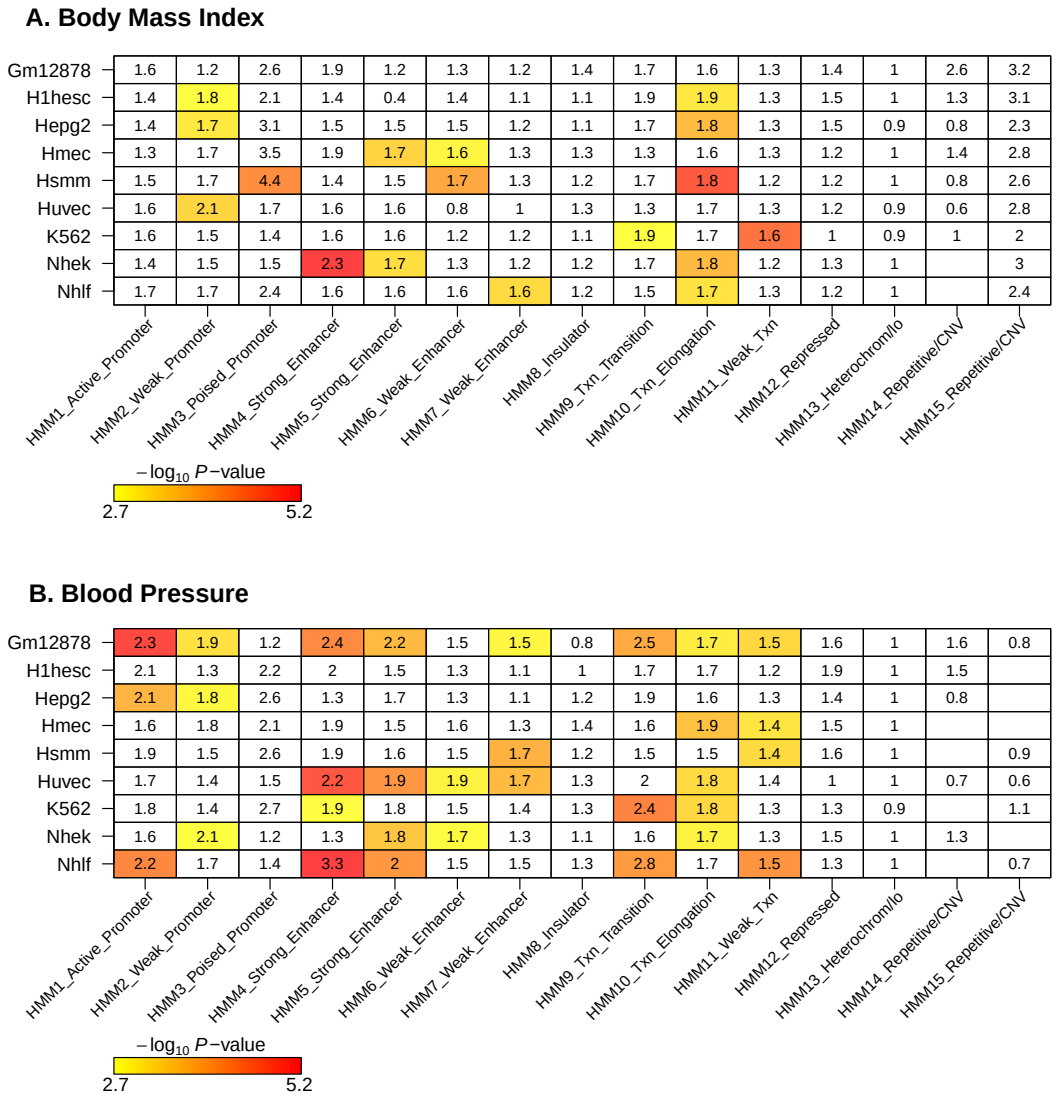
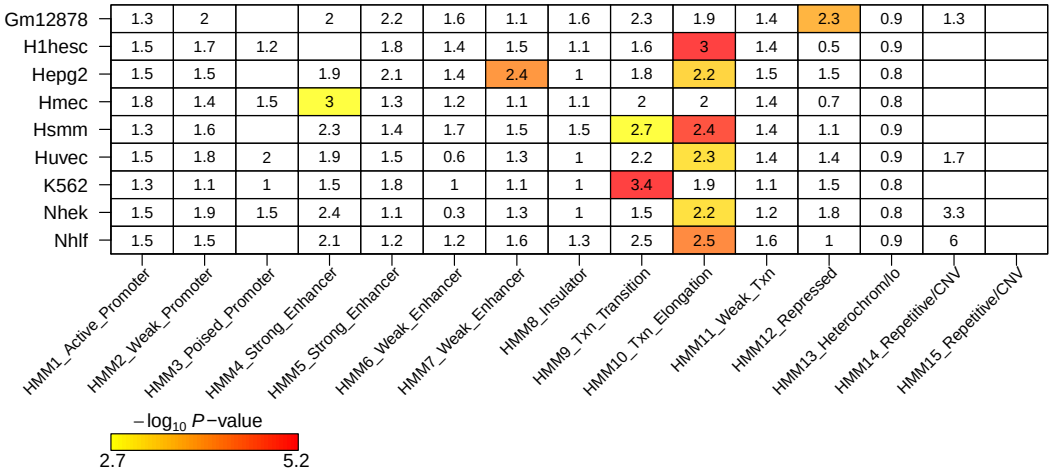


Figure 3.3: Enrichment of GWAS variants in DNase hypersensitive sites. Matrix of fold enrichment for five sets of trait-associated variants (BP, blood pressure; BMI, body mass index; CAD, coronary artery disease; T2D, type 2 diabetes) in DNase hypersensitive sites of 41 tissue groups. Bonferroni significant tissues are colored based on -$\log_{10}$ enrichment P-value. White indicates not significant after Bonferroni correction. The dendrogram shows the relationship between different cell types based on overlap of randomly selected regions of DNase hypersensitivity across the genome.

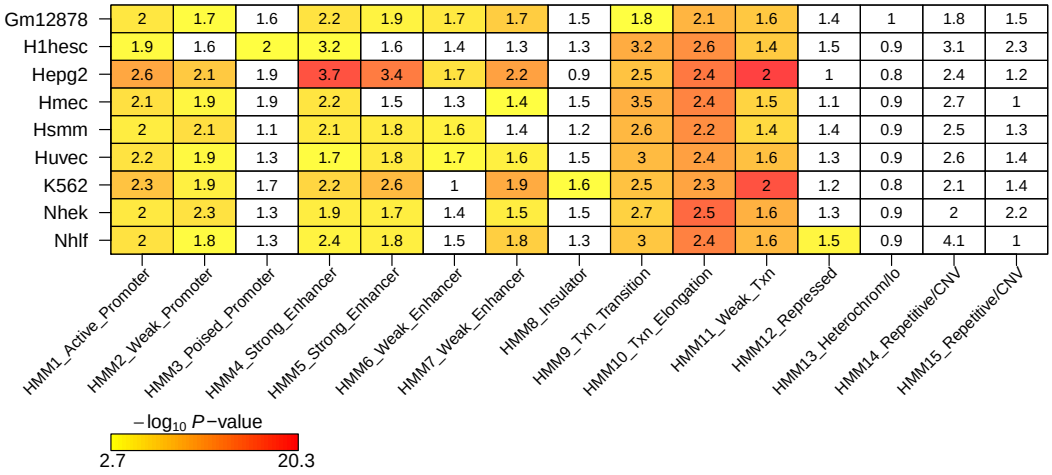
Figure 3.4: Enrichment of trait-associated variants in predicted chromatin states. Matrix of fold enrichment for five sets of GWAS variants (A. body mass index, B. blood pressure, C. coronary artery disease, D. lipids, E. type 2 diabetes) with boxes colored by -$\log_{10}$ enrichment P-value. White boxes indicate not significant after Bonferroni correction for 15 chromatin states and nine human tissues (Ernst et al., 2011). *Abbreviations*: HMM, hidden Markov model; txn, transcription; lo, low signal; CNV, copy number variation.



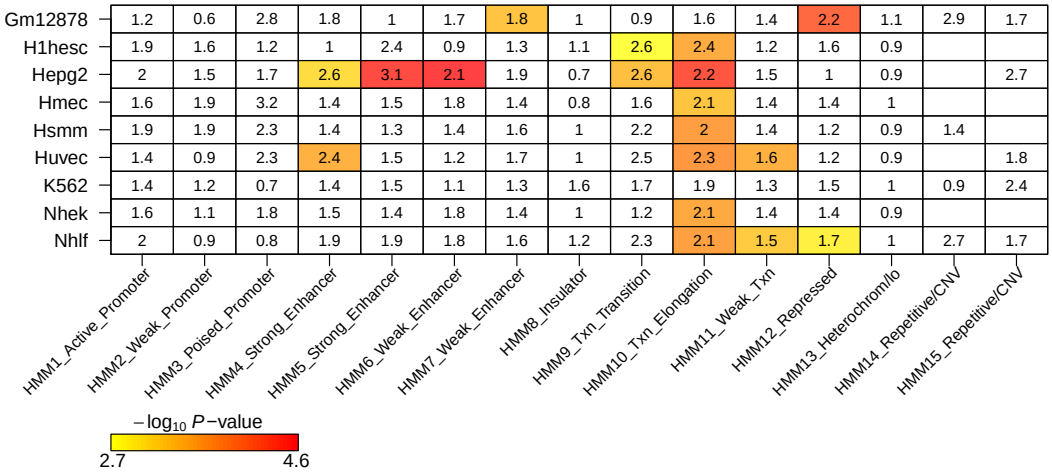
C. Coronary Artery Disease



D. Lipids



E. Type 2 Diabetes



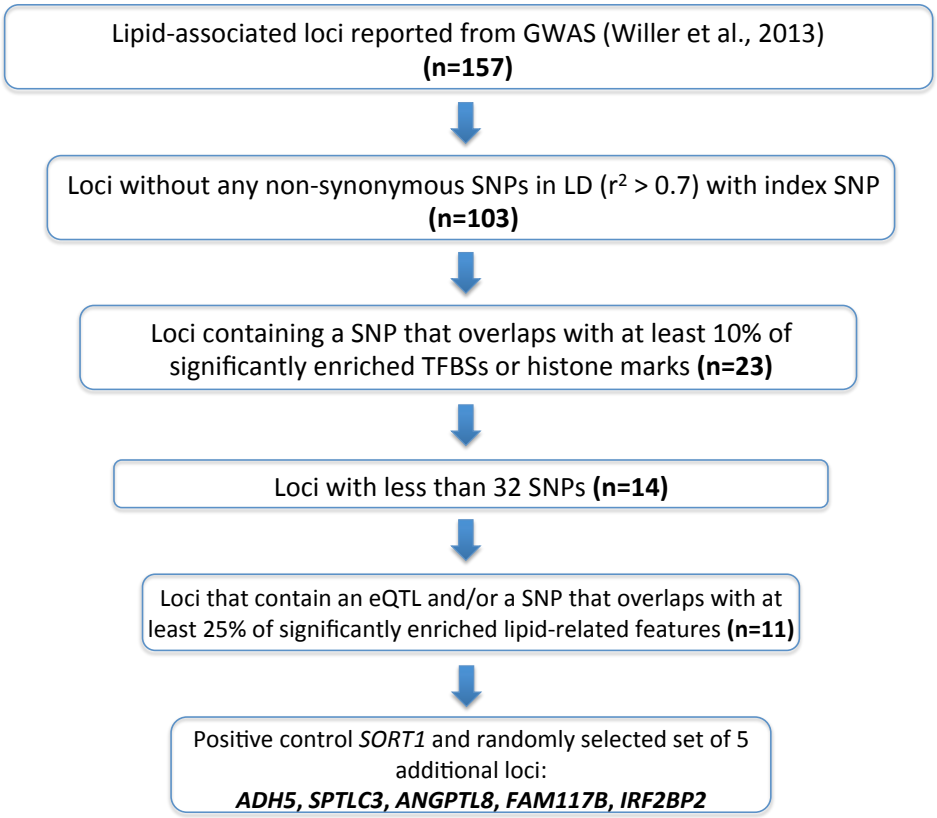
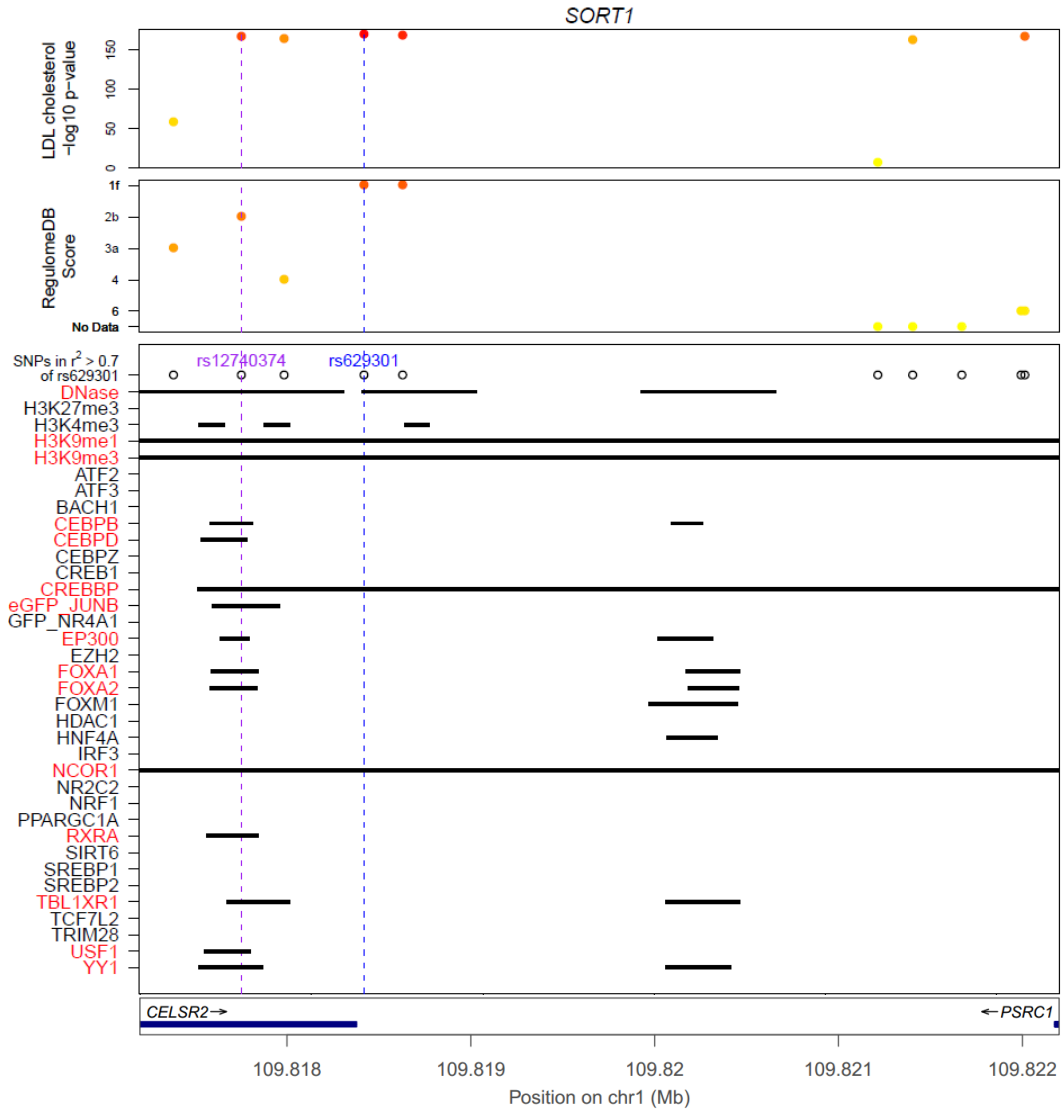


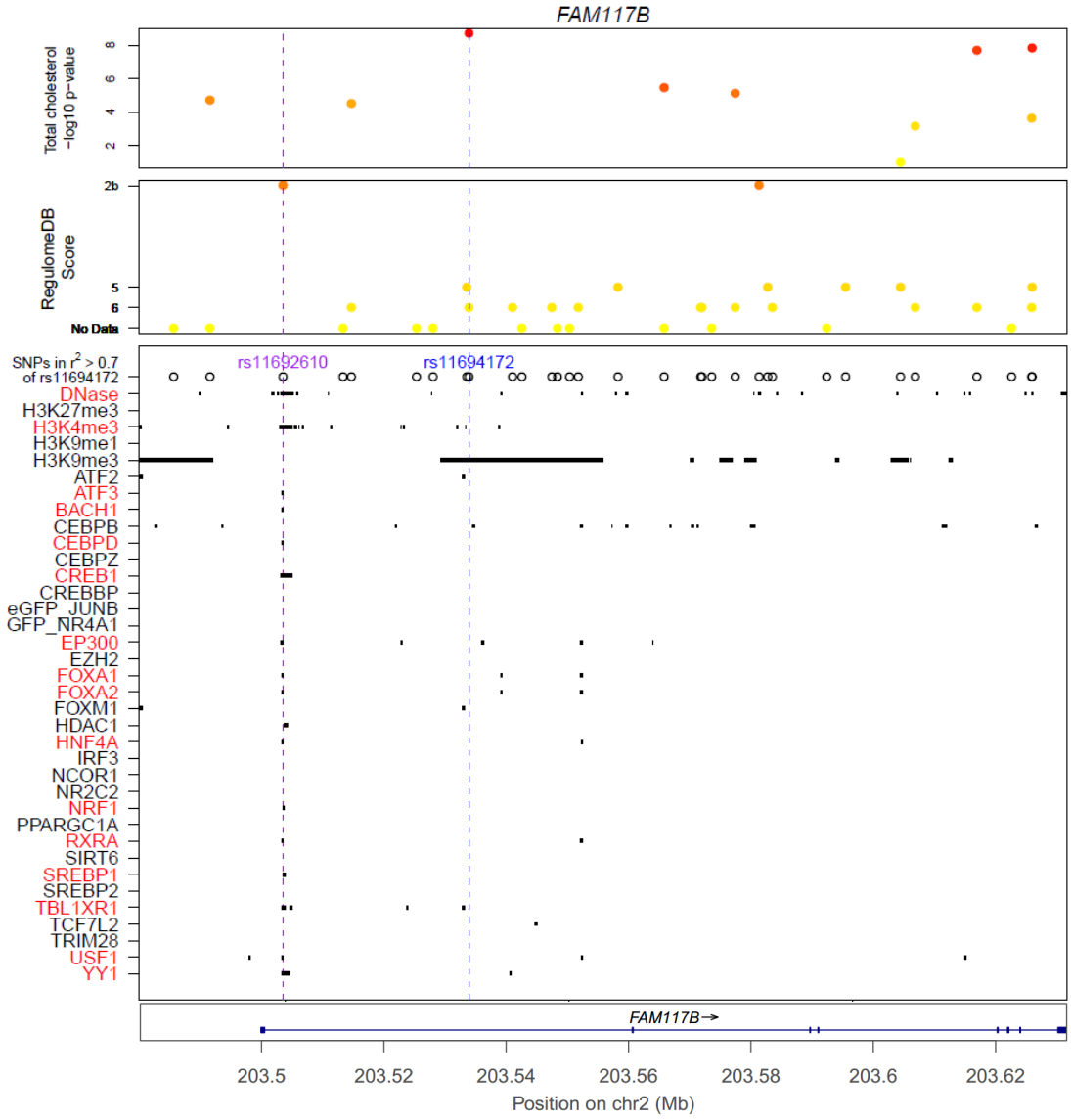
Figure 3.5: Prioritization of lipid-associated loci for functional follow-up.

Figure 3.6: Physical overlap of variants at six lipid loci with regulatory features. SNPs within $r^2 > 0.7$ of the GWAS index SNP at each locus are shown with ChIP-seq or DNase-seq binding sites of lipid-related regulatory features. GWAS -$\log_{10}$ P-values (Teslovich et al., 2010) are plotted in the top panel. RegulomeDB SNP annotation scores of predicted regulatory elements are shown in the second panel (Boyle et al., 2012). Purple dotted lines annotate the hypothesized functional variant based on physical overlap prediction. Blue dotted lines annotate the control SNP, which is usually the top most significant GWAS SNP. Regulatory elements highlighted in red annotate overlap with the candidate functional variant.

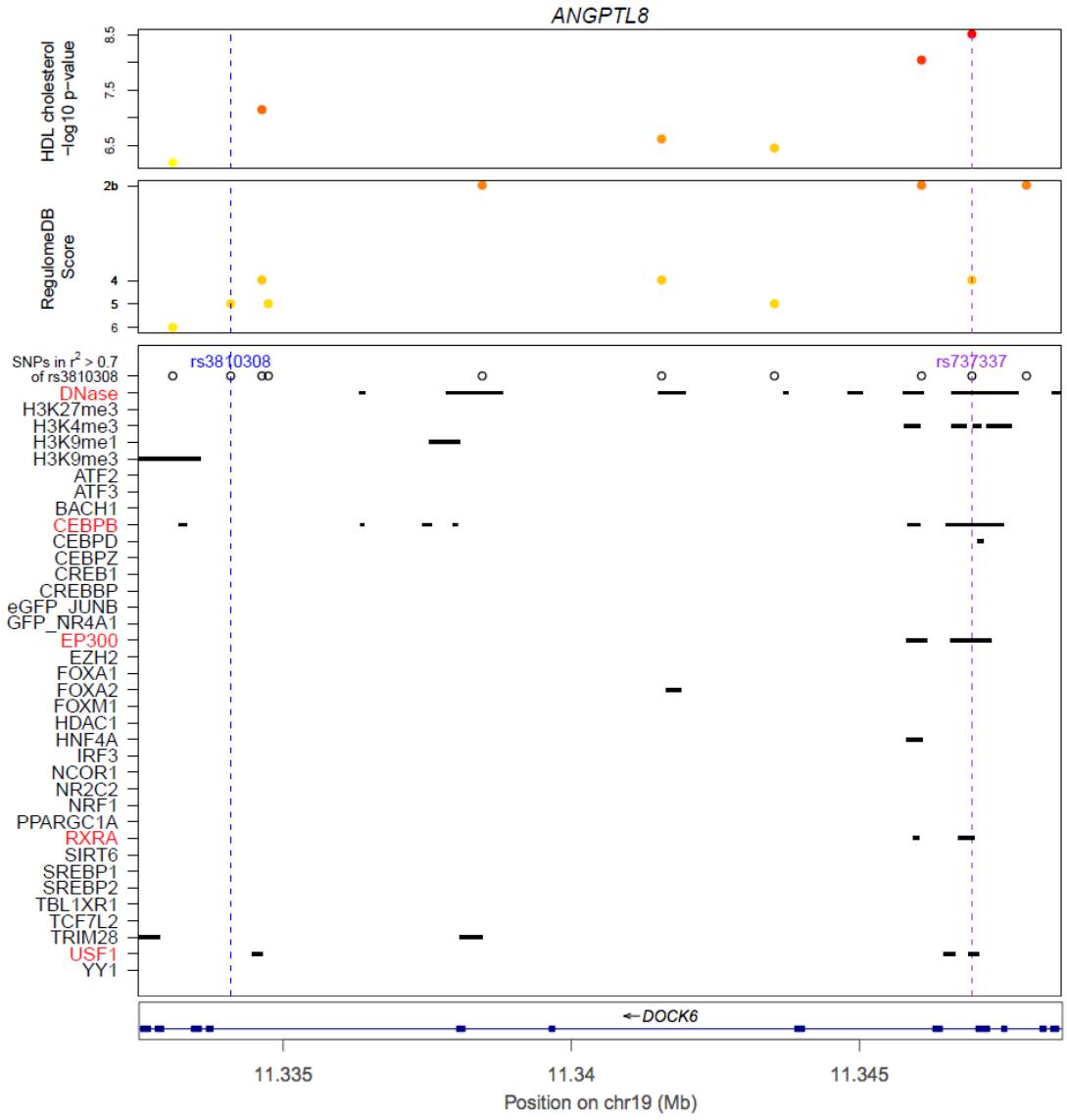
A. ***SORT1*** **(sortilin 1)**. GWAS index SNP rs629301 is associated with LDL cholesterol and total cholesterol.



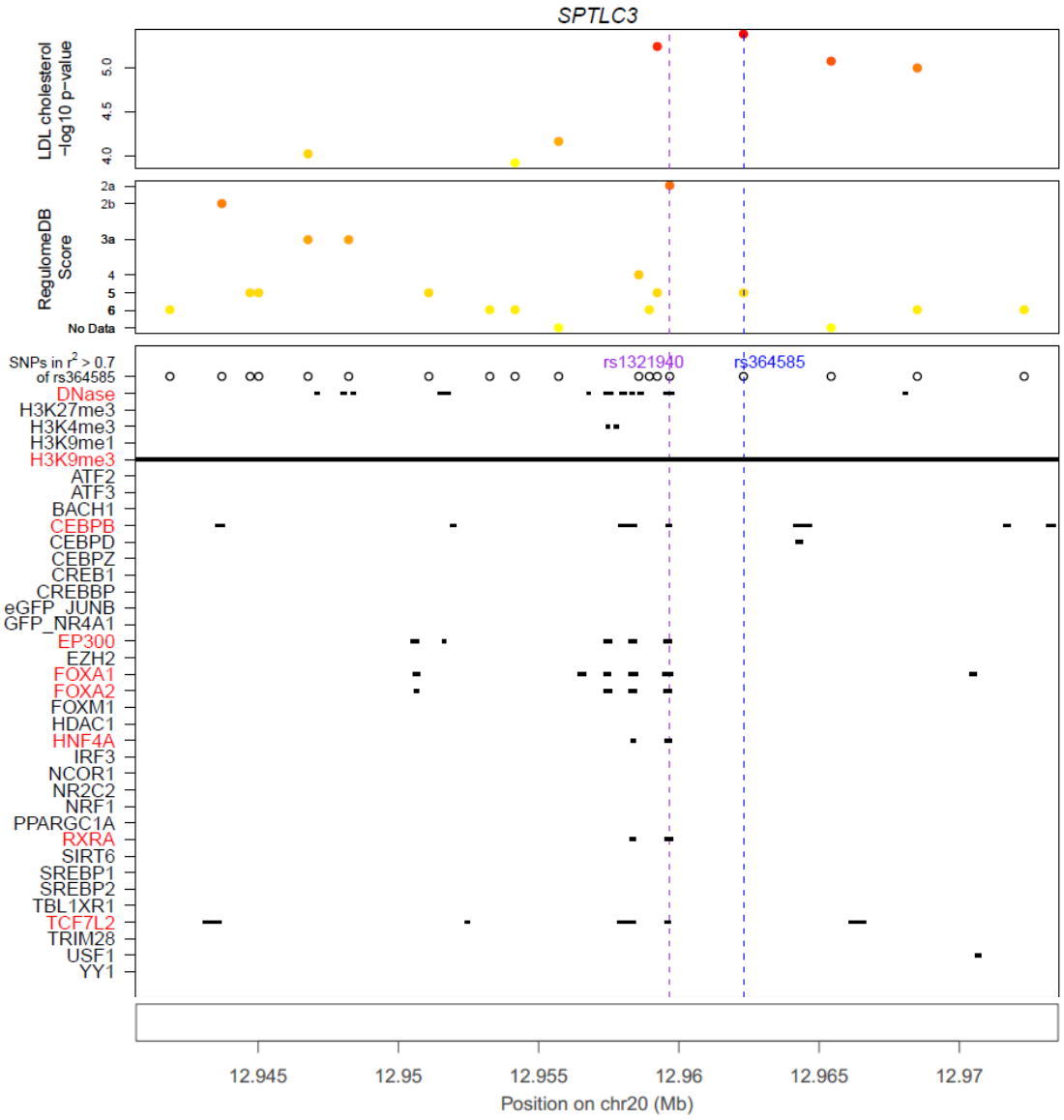
B. ***FAM117B* (family with sequence similarity 117, member B)**. GWAS index SNP rs11694172 is associated with total cholesterol.



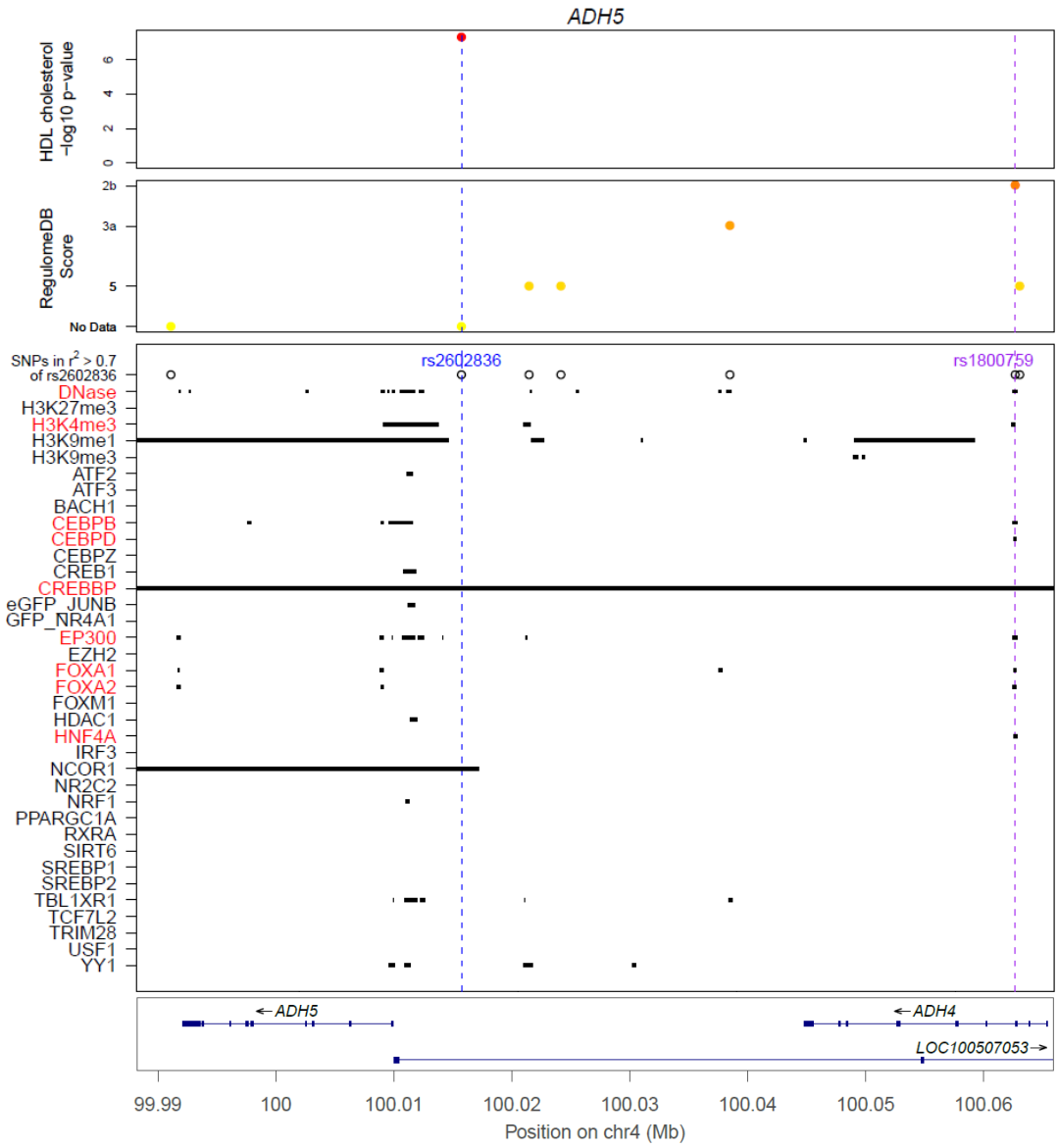
C. ***ANGPTL8*** **(Angiopoietin-like protein 8; C19orf80: chromosome 19 open reading frame 80)**. GWAS index SNP rs737337 is associated with HDL cholesterol, and the candidate functional SNP.



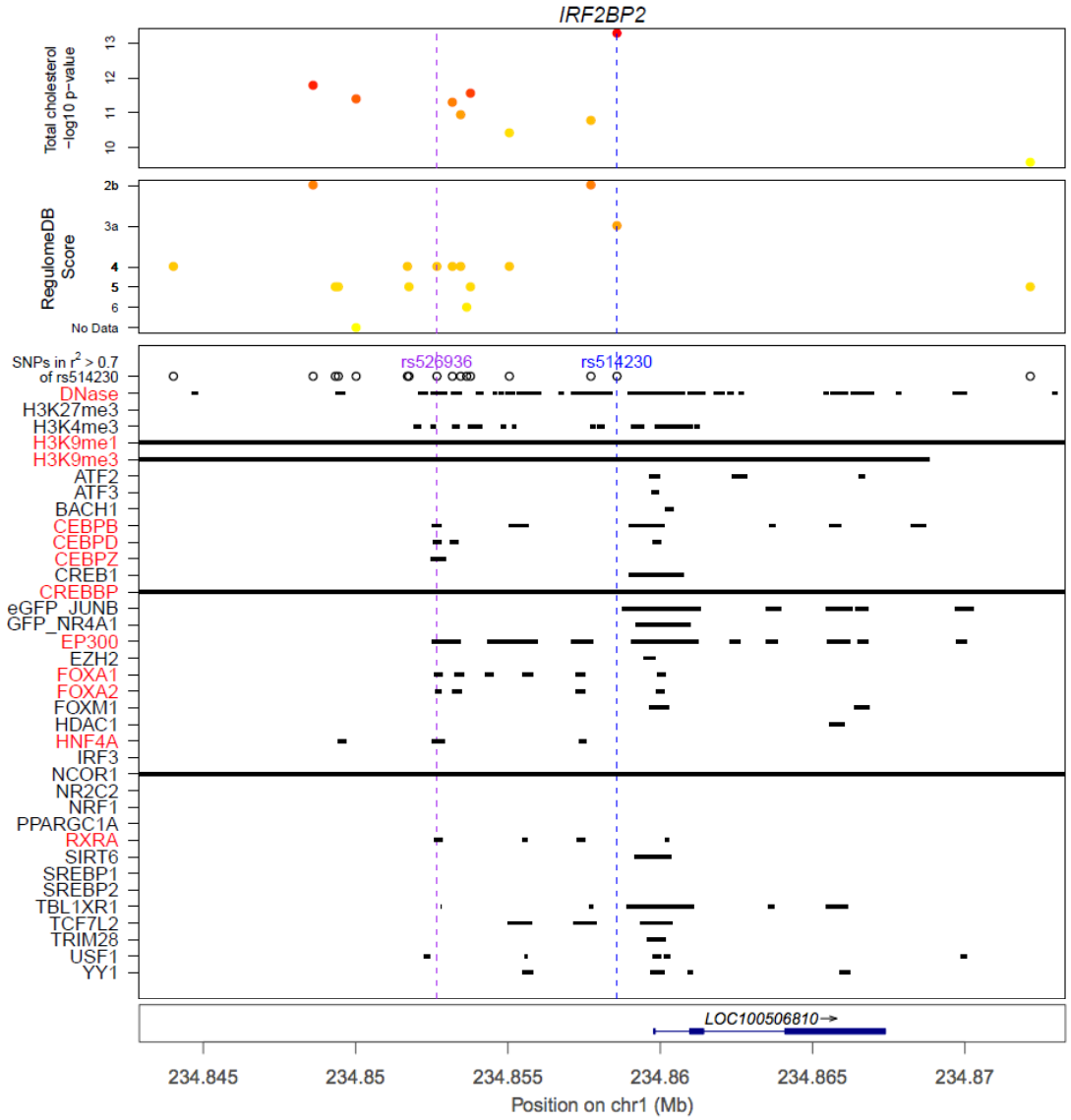
D. ***SPTLC3*** **(serine palmitoyltransferase, long chain base subunit 3)**. GWAS index SNP rs364585 is associated with LDL cholesterol.



E. ***ADH5* (alcohol dehydrogenase 5 (class III), chi polypeptide)**. GWAS index SNP rs2602836 is associated with HDL cholesterol. GWAS *P*-values here are reported from Global Lipids Genetics Consortium et al. (2013).



F. ***IRF2BP2* (interferon regulatory factor 2 binding protein 2)**. GWAS index SNP rs514230 is associated with total cholesterol and LDL cholesterol.



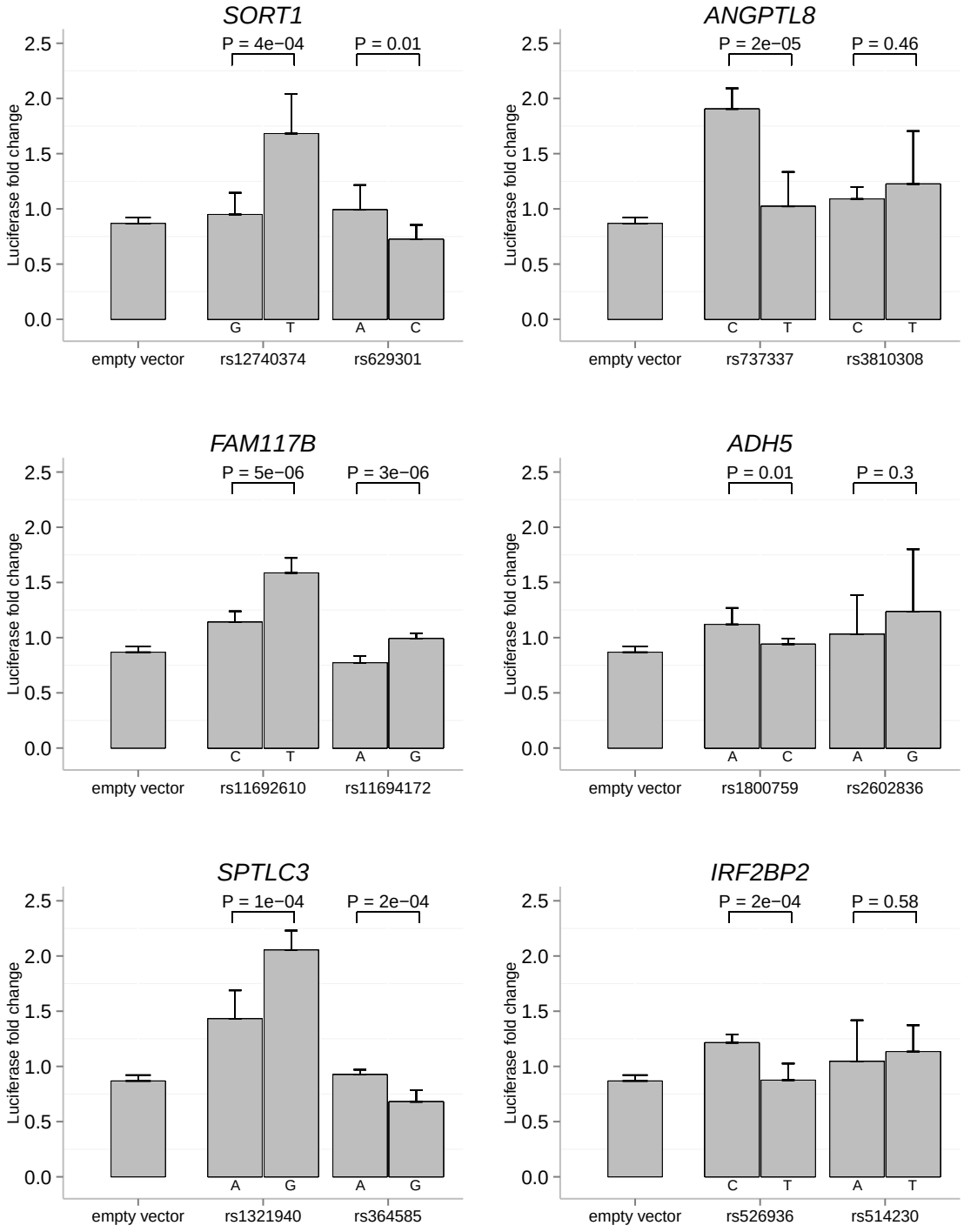


Figure 3.7: Luciferase assays with constructs containing non-coding SNP regions. Relative firefly luciferase expression from constructs with haplotypes of 600-800 bp regions was transfected into HepG2 cells. Single nucleotide alterations in each variant were introduced into constructs as indicated and all luciferase activities were normalized to their pcDNA3.1 co-transfected control groups. Nominal P-values and SD (n=8) for each SNP are shown. The PGL4 empty vector control is on the far left, while the predicted functional variant and control variant follow next in each individual locus figure.

Table 3.1: Formulae for P-value calculation

Input	A SNP set of LD-pruned r GWAS index SNPs Regulatory regions of interest formatted as BED files m = number of control SNPs selected for each index SNP
Intermediate Statistics	SNP set i ($1 \leq i \leq r$) = index SNP i and its m control SNPs $p_i = \frac{\text{number of SNPs in SNP set } i \text{ that falls in regulatory regions of interest}}{m+1}$ $S_i = \begin{cases} 1, & \text{randomly drawn SNP from SNP set } i \text{ falls in regulatory regions of interest} \\ 0, & \text{otherwise} \end{cases}$ $S_i \sim \text{Bernoulli}(p_i)$ $\sum_{i=1}^{r} S_i \sim$ sum of r independent non-identical Bernoulli distribution
Output	s = number of SNPs that falls in regulatory regions of interest in the input GWAS index SNPs Enrichment P-value = $\text{P}\left(\sum_{i=1}^{r} S_i \leq s\right)$ Expected value = $\sum_{i=1}^{r} S_i$

A SNP is considered to fall in regulatory regions of interest if itself or any of its LD proxies has positional overlap with the regions.

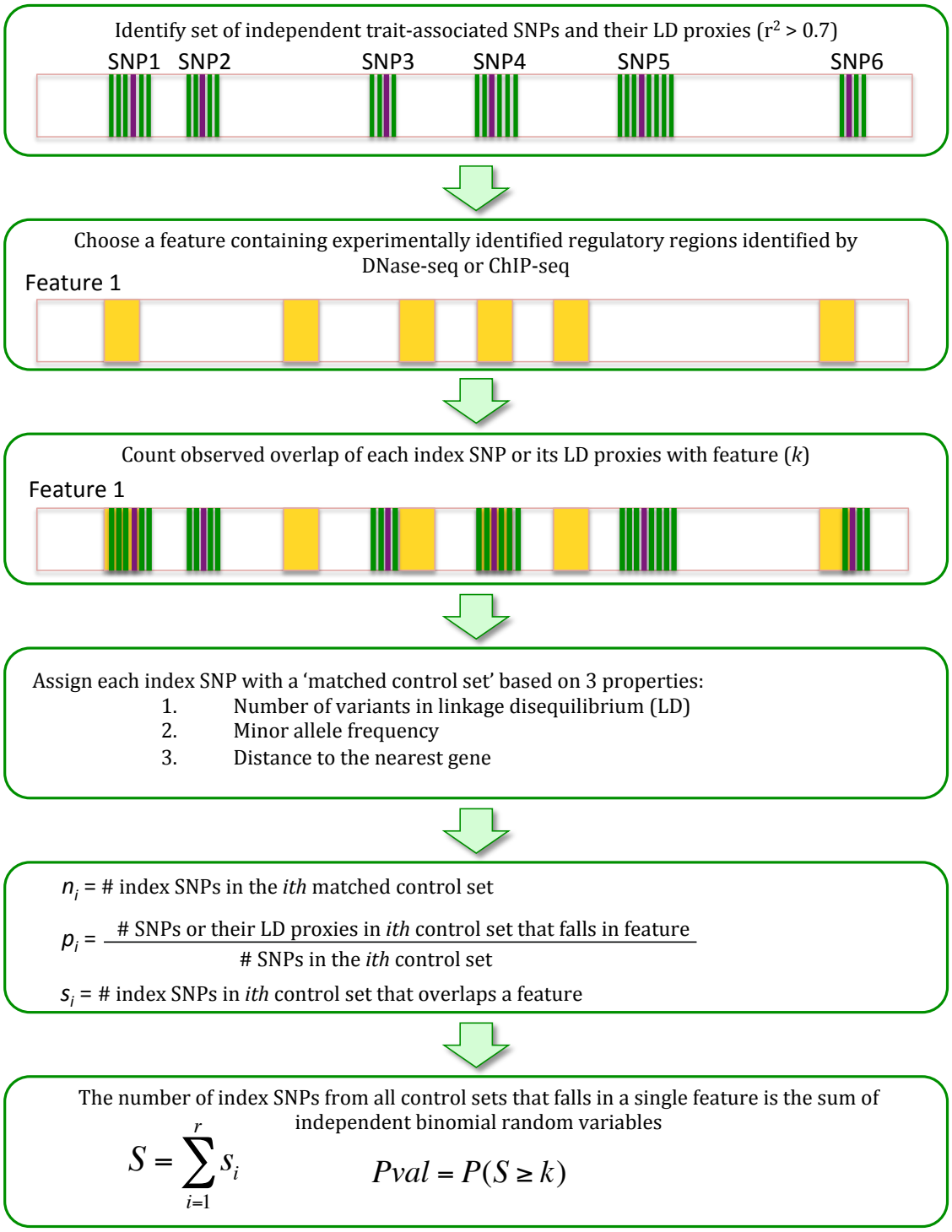
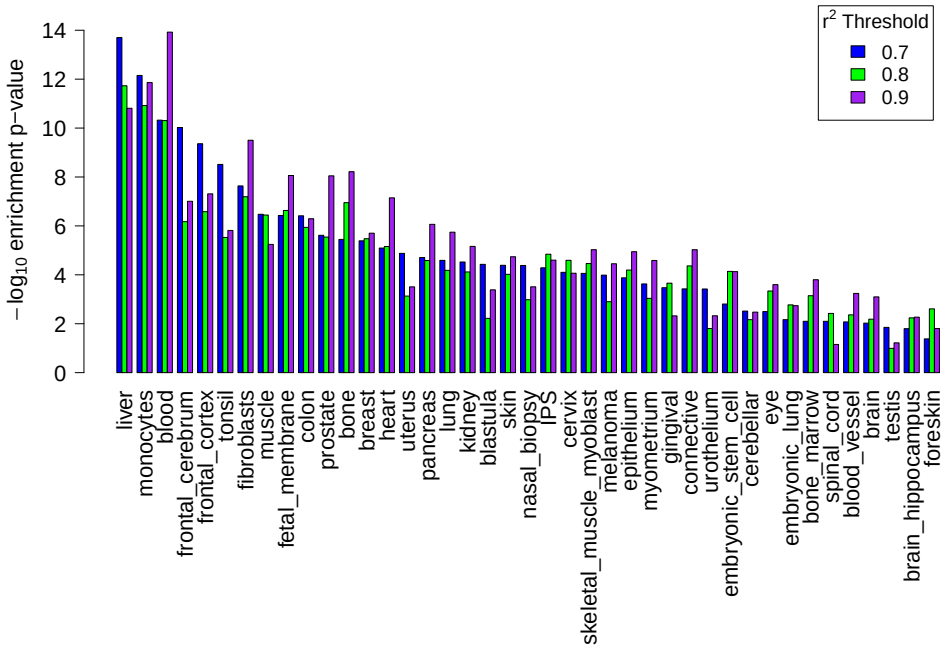


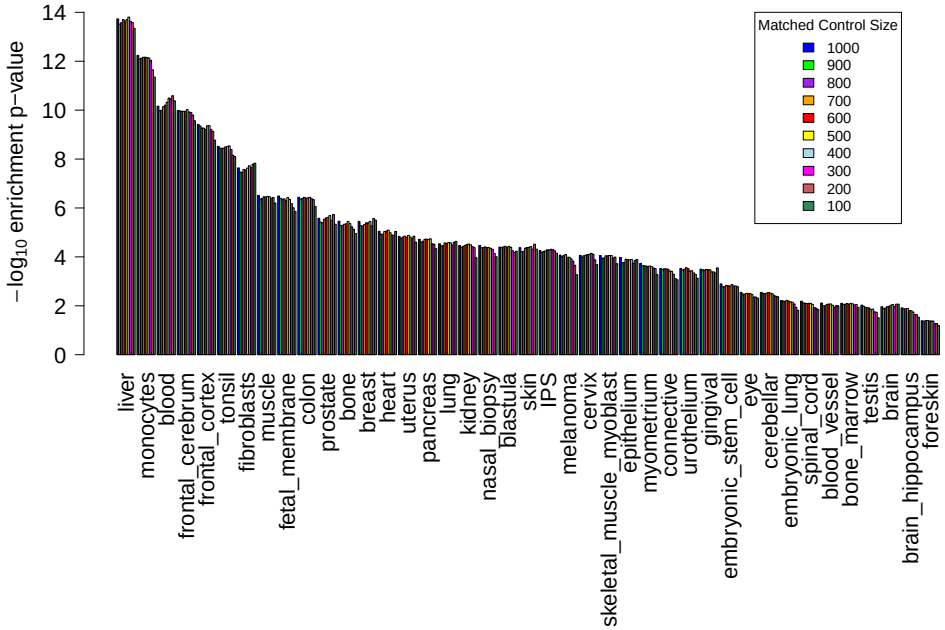
Figure S3.1: Summary of GREGOR variant enrichment method.

Figure S3.2: Enrichment of lipid-associated variation in DNase hypersensitive sites using different parameter values. Tissues are ordered by decreasing P-value significance when using the parameters r^2=0.7 and matched control set size of 500.

(A) Magnitude of enrichment for a range of r^2 thresholds. The r^2 thresholds were used to select 1) the potential functional variants in LD with index variants using 1000 Genomes CEU and 2) the control SNPs with approximately the same number of variants in LD as index variants (using the same threshold as in 1). The higher the r^2 value, the fewer variants in LD would be selected.



(B) Magnitude of enrichment for matched control sets of various sizes. Matched control sets contain variants that share the properties of 1) number of LD proxies, 2) minor allele frequency, and 3) gene proximity. The more variants selected as controls, the less close the matching.



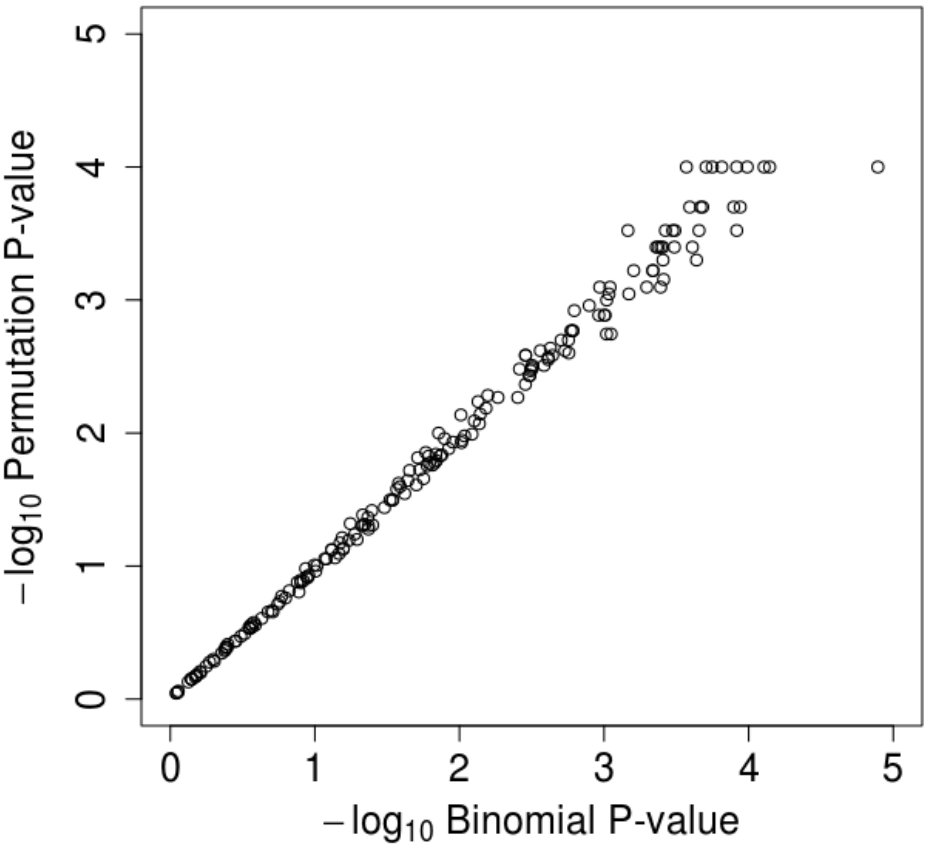
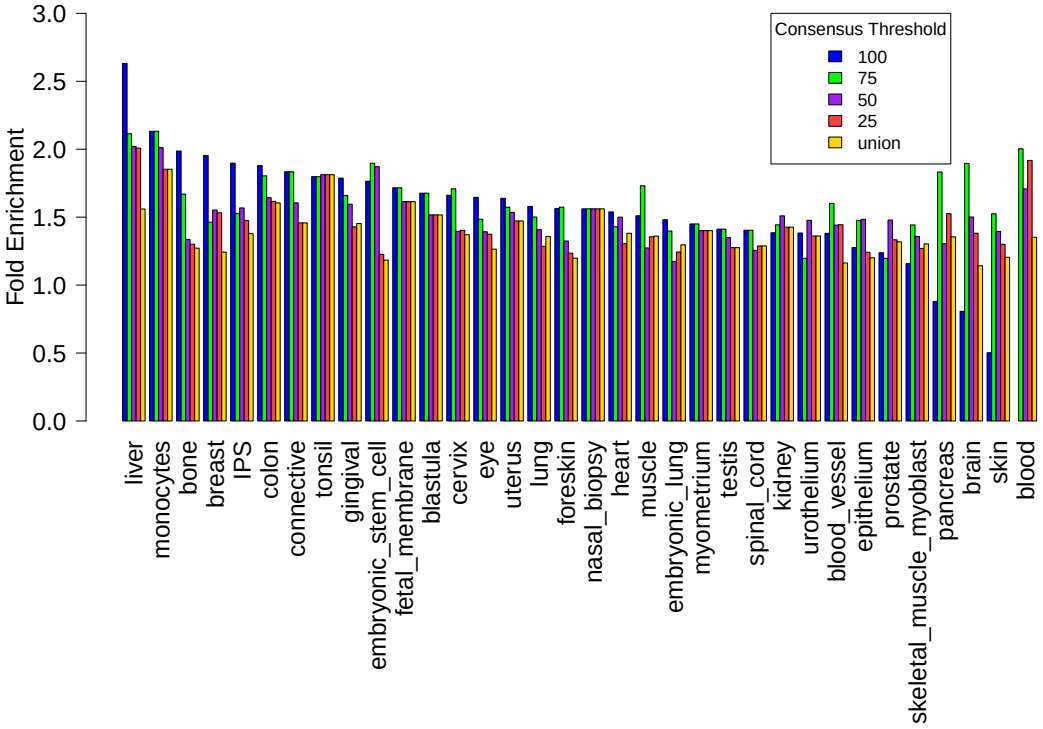


Figure S3.3: Comparison of enrichment *P*-values estimated using 10,000 permutations and the sum of binomial trials as implemented in GREGOR. *P*-values less than $1x10^{-5}$ cannot be precisely estimated by permutation testing, and so are excluded from the figure.

Figure S3.4: Fold enrichment and enrichment P-values for lipid-associated variation in DNase hypersensitive sites (DHSs) of different tissues and at different consensus thresholds. A consensus threshold is defined as the percentage of shared DHS regions among cell types derived from a given tissue.

(A) Fold Enrichment



(B) Enrichment P-values

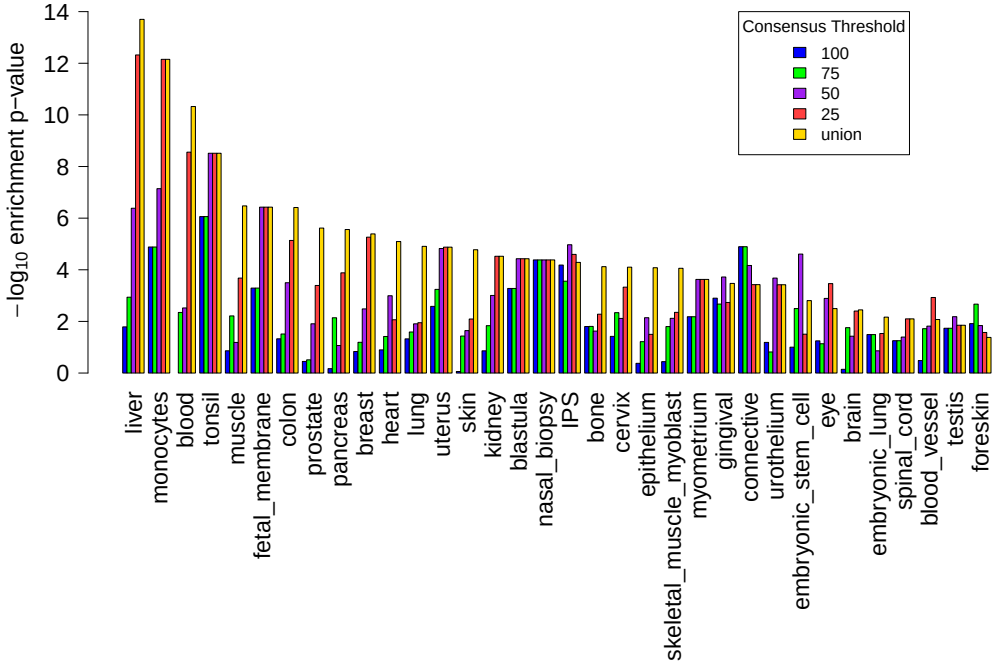


Table S3.1: Experimentally identified DNase hypersensitivity sites of various tissues from ENCODE categorized into broader tissue groups

Broad Tissue Category	**ENCODE Tissue Category**	**BED file**
Blastula	Blastula	wgEncodeAwgDnaseDukeHtr8svnUniPk.narrowPeak
	Blastula	wgEncodeOpenChromDnaseHtr8Pk.narrowPeak
Blood	Blood	wgEncodeAwgDnaseDukeCllUniPk.narrowPeak
	Blood	wgEncodeAwgDnaseDukeGm12891UniPk.narrowPeak
	Blood	wgEncodeAwgDnaseDukeGm12892UniPk.narrowPeak
	Blood	wgEncodeAwgDnaseDukeGm18507UniPk.narrowPeak
	Blood	wgEncodeAwgDnaseDukeGm19238UniPk.narrowPeak
	Blood	wgEncodeAwgDnaseDukeGm19239UniPk.narrowPeak
	Blood	wgEncodeAwgDnaseDukeGm19240UniPk.narrowPeak
	Blood	wgEncodeAwgDnaseDukeTh0UniPk.narrowPeak
	Blood	wgEncodeAwgDnaseUwCd20UniPk.narrowPeak
	Blood	wgEncodeAwgDnaseUwCd34mobilizedUniPk.narrowPeak
	Blood	wgEncodeAwgDnaseUwCmkUniPk.narrowPeak
	Blood	wgEncodeAwgDnaseUwdukeGm12878UniPk.narrowPeak
	Blood	wgEncodeAwgDnaseUwdukeK562UniPk.narrowPeak
	Blood	wgEncodeAwgDnaseUwdukeTh1UniPk.narrowPeak
	Blood	wgEncodeAwgDnaseUwGm06990UniPk.narrowPeak
	Blood	wgEncodeAwgDnaseUwGm12864UniPk.narrowPeak
	Blood	wgEncodeAwgDnaseUwGm12865UniPk.narrowPeak
	Blood	wgEncodeAwgDnaseUwHl60UniPk.narrowPeak
	Blood	wgEncodeAwgDnaseUwJurkatUniPk.narrowPeak
	Blood	wgEncodeAwgDnaseUwNb4UniPk.narrowPeak
	Blood	wgEncodeAwgDnaseUwTh2UniPk.narrowPeak
	Blood	wgEncodeOpenChromDnaseAdultcd4th0Pk.narrowPeak
	Blood	wgEncodeOpenChromDnaseAdultcd4th1Pk.narrowPeak
	Blood	wgEncodeOpenChromDnaseCd20ro01794Pk.narrowPeak
	Blood	wgEncodeOpenChromDnaseCllPk.narrowPeak
	Blood	wgEncodeOpenChromDnaseGm10248Pk.narrowPeak
	Blood	wgEncodeOpenChromDnaseGm10266Pk.narrowPeak
	Blood	wgEncodeOpenChromDnaseGm12878Pk.narrowPeak
	Blood	wgEncodeOpenChromDnaseGm12891Pk.narrowPeak
	Blood	wgEncodeOpenChromDnaseGm12892Pk.narrowPeak
	Blood	wgEncodeOpenChromDnaseGm13976Pk.narrowPeak
	Blood	wgEncodeOpenChromDnaseGm13977Pk.narrowPeak
	Blood	wgEncodeOpenChromDnaseGm18507Pk.narrowPeak
	Blood	wgEncodeOpenChromDnaseGm19238Pk.narrowPeak
	Blood	wgEncodeOpenChromDnaseGm19239Pk.narrowPeak
	Blood	wgEncodeOpenChromDnaseGm19240Pk.narrowPeak
	Blood	wgEncodeOpenChromDnaseGm20000Pk.narrowPeak
	Blood	wgEncodeOpenChromDnaseK562G1phasePk.narrowPeak
	Blood	wgEncodeOpenChromDnaseK562G2mphasePk.narrowPeak
	Blood	wgEncodeOpenChromDnaseK562NabutPk.narrowPeak
	Blood	wgEncodeOpenChromDnaseK562PkV2.narrowPeak
	Blood	wgEncodeOpenChromDnaseK562Saha1u72hrPk.narrowPeak
	Blood	wgEncodeOpenChromDnaseK562SahactrlPk.narrowPeak
	Blood	wgEncodeUwDnaseCd20ro01778PkRep1.narrowPeak
	Blood	wgEncodeUwDnaseCd20ro01778PkRep2.narrowPeak
	Blood	wgEncodeUwDnaseCd34mobilizedPkRep1.narrowPeak
	Blood	wgEncodeUwDnaseCd4naivewb11970640PkRep1.narrowPeak
	Blood	wgEncodeUwDnaseCd4naivewb78495824PkRep1.narrowPeak
	Blood	wgEncodeUwDnaseCmkPkRep1.narrowPeak
	Blood	wgEncodeUwDnaseGm06990PkRep1.narrowPeak

Blood	Blood	wgEncodeUwDnaseGm06990PkRep2.narrowPeak
	Blood	wgEncodeUwDnaseGm12864PkRep1.narrowPeak
	Blood	wgEncodeUwDnaseGm12865PkRep1.narrowPeak
	Blood	wgEncodeUwDnaseGm12865PkRep2.narrowPeak
	Blood	wgEncodeUwDnaseGm12878PkRep1.narrowPeak
	Blood	wgEncodeUwDnaseGm12878PkRep2.narrowPeak
	Blood	wgEncodeUwDnaseHl60PkRep1.narrowPeak
	Blood	wgEncodeUwDnaseHl60PkRep2.narrowPeak
	Blood	wgEncodeUwDnaseJurkatPkRep1.narrowPeak
	Blood	wgEncodeUwDnaseJurkatPkRep2.narrowPeak
	Blood	wgEncodeUwDnaseK562PkRep1.narrowPeak
	Blood	wgEncodeUwDnaseK562PkRep2.narrowPeak
	Blood	wgEncodeUwDnaseNb4PkRep1.narrowPeak
	Blood	wgEncodeUwDnaseNb4PkRep2.narrowPeak
	Blood	wgEncodeUwDnaseTh17PkRep1.narrowPeak
	Blood	wgEncodeUwDnaseTh1PkRep1.narrowPeak
	Blood	wgEncodeUwDnaseTh1PkRep2.narrowPeak
	Blood	wgEncodeUwDnaseTh1wb33676984PkRep1.narrowPeak
	Blood	wgEncodeUwDnaseTh1wb54553204PkRep1.narrowPeak
	Blood	wgEncodeUwDnaseTh1wb54553204PkRep2.narrowPeak
	Blood	wgEncodeUwDnaseTh2PkRep1.narrowPeak
	Blood	wgEncodeUwDnaseTh2PkRep2.narrowPeak
	Blood	wgEncodeUwDnaseTh2wb33676984PkRep1.narrowPeak
	Blood	wgEncodeUwDnaseTh2wb54553204PkRep1.narrowPeak
	Blood	wgEncodeUwDnaseTregwb78495824PkRep1.narrowPeak
	Blood	wgEncodeUwDnaseTregwb83319432PkRep1.narrowPeak
Blood Vessel	Blood Vessel	wgEncodeAwgDnaseDukeAosmcUniPk.narrowPeak
	Blood Vessel	wgEncodeAwgDnaseUwAoafUniPk.narrowPeak
	Blood Vessel	wgEncodeAwgDnaseUwdukeHuvecUniPk.narrowPeak
	Blood Vessel	wgEncodeAwgDnaseUwHbmecUniPk.narrowPeak
	Blood Vessel	wgEncodeAwgDnaseUwHmvecdadUniPk.narrowPeak
	Blood Vessel	wgEncodeAwgDnaseUwHmvecdbladUniPk.narrowPeak
	Blood Vessel	wgEncodeAwgDnaseUwHmvecdblneoUniPk.narrowPeak
	Blood Vessel	wgEncodeAwgDnaseUwHmvecdlyadUniPk.narrowPeak
	Blood Vessel	wgEncodeAwgDnaseUwHmvecdlyneoUniPk.narrowPeak
	Blood Vessel	wgEncodeAwgDnaseUwHmvecdneoUniPk.narrowPeak
	Blood Vessel	wgEncodeAwgDnaseUwHmveclblUniPk.narrowPeak
	Blood Vessel	wgEncodeAwgDnaseUwHmveclyUniPk.narrowPeak
	Blood Vessel	wgEncodeAwgDnaseUwHpaecUniPk.narrowPeak
	Blood Vessel	wgEncodeAwgDnaseUwHpafUniPk.narrowPeak
	Blood Vessel	wgEncodeOpenChromDnaseAosmcSerumfreePk.narrowPeak
	Blood Vessel	wgEncodeOpenChromDnaseHuvecPk.narrowPeak
	Blood Vessel	wgEncodeUwDnaseAoafPkRep1.narrowPeak
	Blood Vessel	wgEncodeUwDnaseAoafPkRep2.narrowPeak
	Blood Vessel	wgEncodeUwDnaseHbmecPkRep1.narrowPeak
	Blood Vessel	wgEncodeUwDnaseHbmecPkRep2.narrowPeak
	Blood Vessel	wgEncodeUwDnaseHbvpPkRep1.narrowPeak
	Blood Vessel	wgEncodeUwDnaseHbvsmcPkRep1.narrowPeak
	Blood Vessel	wgEncodeUwDnaseHbvsmcPkRep2.narrowPeak
	Blood Vessel	wgEncodeUwDnaseHmvecdadPkRep1.narrowPeak
	Blood Vessel	wgEncodeUwDnaseHmvecdadPkRep2.narrowPeak
	Blood Vessel	wgEncodeUwDnaseHmvecdbladPkRep1.narrowPeak
	Blood Vessel	wgEncodeUwDnaseHmvecdbladPkRep2.narrowPeak
	Blood Vessel	wgEncodeUwDnaseHmvecdblneoPkRep1.narrowPeak
	Blood Vessel	wgEncodeUwDnaseHmvecdblneoPkRep2.narrowPeak
	Blood Vessel	wgEncodeUwDnaseHmvecdlyadPkRep1.narrowPeak
	Blood Vessel	wgEncodeUwDnaseHmvecdlyadPkRep2.narrowPeak

Blood Vessel	Blood Vessel	wgEncodeUwDnaseHmvecdlyneoPkRep1.narrowPeak
	Blood Vessel	wgEncodeUwDnaseHmvecdlyneoPkRep2.narrowPeak
	Blood Vessel	wgEncodeUwDnaseHmvecdneoPkRep1.narrowPeak
	Blood Vessel	wgEncodeUwDnaseHmvecdneoPkRep2.narrowPeak
	Blood Vessel	wgEncodeUwDnaseHmveclblPkRep1.narrowPeak
	Blood Vessel	wgEncodeUwDnaseHmveclblPkRep2.narrowPeak
	Blood Vessel	wgEncodeUwDnaseHmvecllyPkRep1.narrowPeak
	Blood Vessel	wgEncodeUwDnaseHmvecllyPkRep2.narrowPeak
	Blood Vessel	wgEncodeUwDnaseHpaecPkRep1.narrowPeak
	Blood Vessel	wgEncodeUwDnaseHpafPkRep1.narrowPeak
	Blood Vessel	wgEncodeUwDnaseHpafPkRep2.narrowPeak
	Blood Vessel	wgEncodeUwDnaseHuvecPkRep1V2.narrowPeak
	Blood Vessel	wgEncodeUwDnaseHuvecPkRep2.narrowPeak
Bone	Bone	wgEncodeAwgDnaseDukeOsteoblUniPk.narrowPeak
	Bone	wgEncodeOpenChromDnaseOsteoblPk.narrowPeak
Bone Marrow	Bone Marrow	wgEncodeUwDnaseHs27aPkRep1.narrowPeak
	Bone Marrow	wgEncodeUwDnaseHs5PkRep1.narrowPeak
	Bone Marrow	wgEncodeUwDnaseMscPkRep1.narrowPeak
	Bone Marrow	wgEncodeUwDnaseMscPkRep2.narrowPeak
Brain	Brain	wgEncodeAwgDnaseDukeGlioblaUniPk.narrowPeak
	Brain	wgEncodeAwgDnaseDukeMedulloUniPk.narrowPeak
	Brain	wgEncodeAwgDnaseUwBe2cUniPk.narrowPeak
	Brain	wgEncodeAwgDnaseUwNhaUniPk.narrowPeak
	Brain	wgEncodeAwgDnaseUwSknmcUniPk.narrowPeak
	Brain	wgEncodeAwgDnaseUwSknshraUniPk.narrowPeak
	Brain	wgEncodeOpenChromDnaseGlioblaPk.narrowPeak
	Brain	wgEncodeOpenChromDnaseMedullod341Pk.narrowPeak
	Brain	wgEncodeOpenChromDnaseMedulloPk.narrowPeak
	Brain	wgEncodeOpenChromDnaseSknshPk.narrowPeak
	Brain	wgEncodeUwDnaseBe2cPkRep1.narrowPeak
	Brain	wgEncodeUwDnaseBe2cPkRep2.narrowPeak
	Brain	wgEncodeUwDnaseM059jPkRep1.narrowPeak
	Brain	wgEncodeUwDnaseM059jPkRep2.narrowPeak
	Brain	wgEncodeUwDnaseNhaPkRep1.narrowPeak
	Brain	wgEncodeUwDnaseNhaPkRep2.narrowPeak
	Brain	wgEncodeUwDnaseSknmcPkRep1.narrowPeak
	Brain	wgEncodeUwDnaseSknmcPkRep2.narrowPeak
	Brain	wgEncodeUwDnaseSknshraPkRep1.narrowPeak
	Brain	wgEncodeUwDnaseSknshraPkRep2.narrowPeak
Brain Hippocampus	Brain Hippocampus	wgEncodeAwgDnaseUwHahUniPk.narrowPeak
	Brain Hippocampus	wgEncodeUwDnaseHahPkRep1.narrowPeak
	Brain Hippocampus	wgEncodeUwDnaseHahPkRep2.narrowPeak
Breast	Breast	wgEncodeAwgDnaseDukeMcf7hypoxiaUniPk.narrowPeak
	Breast	wgEncodeAwgDnaseDukeT47dUniPk.narrowPeak
	Breast	wgEncodeAwgDnaseUwdukeHmecUniPk.narrowPeak
	Breast	wgEncodeAwgDnaseUwdukeMcf7UniPk.narrowPeak
	Breast	wgEncodeOpenChromDnaseHmecPk.narrowPeak
	Breast	wgEncodeOpenChromDnaseMcf7CtcfshrnaPk.narrowPeak
	Breast	wgEncodeOpenChromDnaseMcf7HypoxlacconPk.narrowPeak
	Breast	wgEncodeOpenChromDnaseMcf7HypoxlacPk.narrowPeak
	Breast	wgEncodeOpenChromDnaseMcf7Pk.narrowPeak
	Breast	wgEncodeOpenChromDnaseMcf7RandshrnaPk.narrowPeak
	Breast	wgEncodeOpenChromDnaseT47dEst10nm30mPk.narrowPeak
	Breast	wgEncodeOpenChromDnaseT47dPk.narrowPeak
	Breast	wgEncodeUwDnaseHmecPkRep1.narrowPeak
	Breast	wgEncodeUwDnaseHmecPkRep2.narrowPeak
	Breast	wgEncodeUwDnaseMcf7Est100nm1hPkRep1.narrowPeak

Breast	Breast	wgEncodeUwDnaseMcf7Est100nm1hPkRep2.narrowPeak
	Breast	wgEncodeUwDnaseMcf7Estctrl0hPkRep1.narrowPeak
	Breast	wgEncodeUwDnaseMcf7Estctrl0hPkRep2.narrowPeak
	Breast	wgEncodeUwDnaseMcf7PkRep1.narrowPeak
	Breast	wgEncodeUwDnaseMcf7PkRep2.narrowPeak
	Breast	wgEncodeUwDnaseT47dPkRep1.narrowPeak
	Breast	wgEncodeUwDnaseT47dPkRep2.narrowPeak
	Mammary	wgEncodeAwgDnaseUwHmfUniPk.narrowPeak
	Mammary	wgEncodeUwDnaseHmfPkRep1.narrowPeak
	Mammary	wgEncodeUwDnaseHmfPkRep2.narrowPeak
Cerebellar	Cerebellar	wgEncodeAwgDnaseUwHacUniPk.narrowPeak
	Cerebellar	wgEncodeUwDnaseHacPkRep1.narrowPeak
	Cerebellar	wgEncodeUwDnaseHacPkRep2.narrowPeak
	Cerebellum	wgEncodeOpenChromDnaseCerebellumocPk.narrowPeak
Cervix	Cervix	wgEncodeAwgDnaseDukeHelas3ifna4hUniPk.narrowPeak
	Cervix	wgEncodeAwgDnaseUwdukeHelas3UniPk.narrowPeak
	Cervix	wgEncodeOpenChromDnaseHelas3Ifna4hPk.narrowPeak
	Cervix	wgEncodeOpenChromDnaseHelas3Pk.narrowPeak
	Cervix	wgEncodeUwDnaseHelas3PkRep1.narrowPeak
	Cervix	wgEncodeUwDnaseHelas3PkRep2.narrowPeak
Colon	Colon	wgEncodeAwgDnaseUwCaco2UniPk.narrowPeak
	Colon	wgEncodeAwgDnaseUwHct116UniPk.narrowPeak
	Colon	wgEncodeUwDnaseCaco2PkRep1.narrowPeak
	Colon	wgEncodeUwDnaseCaco2PkRep2.narrowPeak
	Colon	wgEncodeUwDnaseHct116PkRep1.narrowPeak
	Colon	wgEncodeUwDnaseHct116PkRep2.narrowPeak
Connective	Connective	wgEncodeAwgDnaseUwHvmfUniPk.narrowPeak
	Connective	wgEncodeUwDnaseHvmfPkRep1.narrowPeak
	Connective	wgEncodeUwDnaseHvmfPkRep2.narrowPeak
Embryonic Lung	Embryonic Lung	wgEncodeAwgDnaseUwWi38tamoxifentamoxifenUniPk.narrowPeak
	Embryonic Lung	wgEncodeAwgDnaseUwWi38UniPk.narrowPeak
	Embryonic Lung	wgEncodeUwDnaseWi38OhtamPkRep1.narrowPeak
	Embryonic Lung	wgEncodeUwDnaseWi38OhtamPkRep2.narrowPeak
	Embryonic Lung	wgEncodeUwDnaseWi38PkRep1.narrowPeak
	Embryonic Lung	wgEncodeUwDnaseWi38PkRep2.narrowPeak
Embryonic Stem Cell	Embryonic Stem Cell	wgEncodeAwgDnaseDukeH9esUniPk.narrowPeak
	Embryonic Stem Cell	wgEncodeAwgDnaseUwdukeH1hescUniPk.narrowPeak
	Embryonic Stem Cell	wgEncodeAwgDnaseUwH7hescUniPk.narrowPeak
	Embryonic Stem Cell	wgEncodeOpenChromDnaseH1hescPk.narrowPeak
	Embryonic Stem Cell	wgEncodeOpenChromDnaseH7esPk.narrowPeak
	Embryonic Stem Cell	wgEncodeOpenChromDnaseH9esPk.narrowPeak
	Embryonic Stem Cell	wgEncodeUwDnaseH1hescPkRep1.narrowPeak
	Embryonic Stem Cell	wgEncodeUwDnaseH7esDiffa14dPkRep1.narrowPeak
	Embryonic Stem Cell	wgEncodeUwDnaseH7esDiffa14dPkRep2.narrowPeak
	Embryonic Stem Cell	wgEncodeUwDnaseH7esDiffa2dPkRep1.narrowPeak
	Embryonic Stem Cell	wgEncodeUwDnaseH7esDiffa5dPkRep1.narrowPeak
	Embryonic Stem Cell	wgEncodeUwDnaseH7esDiffa5dPkRep2.narrowPeak
	Embryonic Stem Cell	wgEncodeUwDnaseH7esDiffa9dPkRep1.narrowPeak
	Embryonic Stem Cell	wgEncodeUwDnaseH7esPkRep1V2.narrowPeak
	Embryonic Stem Cell	wgEncodeUwDnaseH7esPkRep2.narrowPeak
Epithelium	Bronchial Epithelium	wgEncodeUwDnaseNhberaPkRep1.narrowPeak
	Bronchial Epithelium	wgEncodeUwDnaseNhberaPkRep2.narrowPeak
	Epithelium	wgEncodeAwgDnaseDukePhteUniPk.narrowPeak
	Epithelium	wgEncodeAwgDnaseUwdukeA549UniPk.narrowPeak
	Epithelium	wgEncodeAwgDnaseUwHaepicUniPk.narrowPeak
	Epithelium	wgEncodeAwgDnaseUwHcpepicUniPk.narrowPeak
	Epithelium	wgEncodeAwgDnaseUwHeepicUniPk.narrowPeak

Epithelium	Epithelium	wgEncodeAwgDnaseUwHipepicUniPk.narrowPeak
	Epithelium	wgEncodeAwgDnaseUwHnpcepicUniPk.narrowPeak
	Epithelium	wgEncodeAwgDnaseUwHpdlfUniPk.narrowPeak
	Epithelium	wgEncodeAwgDnaseUwHrcepicUniPk.narrowPeak
	Epithelium	wgEncodeAwgDnaseUwHreUniPk.narrowPeak
	Epithelium	wgEncodeAwgDnaseUwHrpepicUniPk.narrowPeak
	Epithelium	wgEncodeAwgDnaseUwRptecUniPk.narrowPeak
	Epithelium	wgEncodeAwgDnaseUwSaecUniPk.narrowPeak
	Epithelium	wgEncodeOpenChromDnaseA549Pk.narrowPeak
	Epithelium	wgEncodeOpenChromDnasePhtePk.narrowPeak
	Epithelium	wgEncodeUwDnaseA549PkRep1.narrowPeak
	Epithelium	wgEncodeUwDnaseA549PkRep2.narrowPeak
	Epithelium	wgEncodeUwDnaseHaePkRep1.narrowPeak
	Epithelium	wgEncodeUwDnaseHaePkRep2.narrowPeak
	Epithelium	wgEncodeUwDnaseHcpePkRep1.narrowPeak
	Epithelium	wgEncodeUwDnaseHcpePkRep2.narrowPeak
	Epithelium	wgEncodeUwDnaseHeePkRep1.narrowPeak
	Epithelium	wgEncodeUwDnaseHeePkRep2.narrowPeak
	Epithelium	wgEncodeUwDnaseHipePkRep1.narrowPeak
	Epithelium	wgEncodeUwDnaseHipePkRep2.narrowPeak
	Epithelium	wgEncodeUwDnaseHnpcePkRep1.narrowPeak
	Epithelium	wgEncodeUwDnaseHnpcePkRep2V2.narrowPeak
	Epithelium	wgEncodeUwDnaseHpdlfPkRep1.narrowPeak
	Epithelium	wgEncodeUwDnaseHpdlfPkRep2.narrowPeak
	Epithelium	wgEncodeUwDnaseHrcePkRep1.narrowPeak
	Epithelium	wgEncodeUwDnaseHrcePkRep2.narrowPeak
	Epithelium	wgEncodeUwDnaseHrePkRep1V2.narrowPeak
	Epithelium	wgEncodeUwDnaseHrePkRep2V2.narrowPeak
	Epithelium	wgEncodeUwDnaseHrpePkRep1V2.narrowPeak
	Epithelium	wgEncodeUwDnaseHrpePkRep2V2.narrowPeak
	Epithelium	wgEncodeUwDnaseRptecPkRep1.narrowPeak
	Epithelium	wgEncodeUwDnaseRptecPkRep2.narrowPeak
	Epithelium	wgEncodeUwDnaseSaecPkRep1.narrowPeak
	Epithelium	wgEncodeUwDnaseSaecPkRep2.narrowPeak
	Luminal Epithelium	wgEncodeOpenChromDnaseEcc1Dm002p1hPk.narrowPeak
	Luminal Epithelium	wgEncodeOpenChromDnaseEcc1Est10nm30mPk.narrowPeak
	Pancreatic Duct	wgEncodeAwgDnaseDukeHpde6e6e7UniPk.narrowPeak
	Pancreatic Duct	wgEncodeOpenChromDnaseHpde6e6e7Pk.narrowPeak
Eye	Eye	wgEncodeAwgDnaseUwHconfUniPk.narrowPeak
	Eye	wgEncodeAwgDnaseUwWerirb1UniPk.narrowPeak
	Eye	wgEncodeUwDnaseHconfPkRep1.narrowPeak
	Eye	wgEncodeUwDnaseHconfPkRep2.narrowPeak
	Eye	wgEncodeUwDnaseWerirb1PkRep1.narrowPeak
	Eye	wgEncodeUwDnaseWerirb1PkRep2.narrowPeak
Fetal Membrane	Fetal Membrane	wgEncodeAwgDnaseDukeChorionUniPk.narrowPeak
	Fetal Membrane	wgEncodeOpenChromDnaseChorionPk.narrowPeak
Fibroblasts	Lung Fibroblast	wgEncodeOpenChromDnaseFibropag08396Pk.narrowPeak
	Skin	wgEncodeAwgDnaseDukeFibroblUniPk.narrowPeak
	Skin	wgEncodeAwgDnaseDukeFibropUniPk.narrowPeak
	Skin	wgEncodeOpenChromDnaseFibroblgm03348LenticonPk.narrowPeak
	Skin	wgEncodeOpenChromDnaseFibroblgm03348LentimyodPk.narrowPeak
	Skin	wgEncodeOpenChromDnaseFibroblgm03348Pk.narrowPeak
	Skin	wgEncodeOpenChromDnaseFibroblPk.narrowPeak
	Skin	wgEncodeOpenChromDnaseFibropPk.narrowPeak
	Skin Fibroblast	wgEncodeOpenChromDnaseFibropag08395Pk.narrowPeak
	Skin Fibroblast	wgEncodeOpenChromDnaseFibropag20443Pk.narrowPeak
Foreskin	Foreskin	wgEncodeAwgDnaseUwHffmycUniPk.narrowPeak

Foreskin	Foreskin	wgEncodeAwgDnaseUwHffUniPk.narrowPeak
	Foreskin	wgEncodeUwDnaseHffmycPkRep1.narrowPeak
	Foreskin	wgEncodeUwDnaseHffmycPkRep2.narrowPeak
	Foreskin	wgEncodeUwDnaseHffPkRep1.narrowPeak
	Foreskin	wgEncodeUwDnaseHffPkRep2.narrowPeak
Frontal Cerebrum	Frontal Cerebrum	wgEncodeOpenChromDnaseCerebrumfrontalocPk.narrowPeak
Frontal Cortex	Frontal Cortex	wgEncodeOpenChromDnaseFrontalcortexocPk.narrowPeak
Gingival	Gingiva	wgEncodeAwgDnaseUwHgfUniPk.narrowPeak
	Gingiva	wgEncodeUwDnaseHgfPkRep1.narrowPeak
	Gingiva	wgEncodeUwDnaseHgfPkRep2.narrowPeak
	Gingival	wgEncodeAwgDnaseUwAg09319UniPk.narrowPeak
	Gingival	wgEncodeUwDnaseAg09319PkRep1V2.narrowPeak
	Gingival	wgEncodeUwDnaseAg09319PkRep2.narrowPeak
Heart	Heart	wgEncodeAwgDnaseUwHcfaaUniPk.narrowPeak
	Heart	wgEncodeAwgDnaseUwHcfUniPk.narrowPeak
	Heart	wgEncodeAwgDnaseUwHcmUniPk.narrowPeak
	Heart	wgEncodeOpenChromDnaseHeartocPk.narrowPeak
	Heart	wgEncodeUwDnaseHcfaaPkRep1.narrowPeak
	Heart	wgEncodeUwDnaseHcfaaPkRep2.narrowPeak
	Heart	wgEncodeUwDnaseHcfPkRep1.narrowPeak
	Heart	wgEncodeUwDnaseHcfPkRep2.narrowPeak
	Heart	wgEncodeUwDnaseHcmPkRep1.narrowPeak
	Heart	wgEncodeUwDnaseHcmPkRep2.narrowPeak
IPS	Induced Pluripotent Cell IPS	wgEncodeOpenChromDnaseIpscwru1Pk.narrowPeak
	Induced Pluripotent Cell IPS	wgEncodeOpenChromDnaseIpsnihi11Pk.narrowPeak
	Induced Pluripotent Cell IPS	wgEncodeOpenChromDnaseIpsnihi7Pk.narrowPeak
	Induced Pluripotent Stem Cell	wgEncodeAwgDnaseDukeIpsUniPk.narrowPeak
	Induced Pluripotent Stem Cell	wgEncodeOpenChromDnaseIpsPk.narrowPeak
Kidney	Kidney	wgEncodeAwgDnaseUwHrgecUniPk.narrowPeak
	Kidney	wgEncodeOpenChromDnaseHek293tPk.narrowPeak
	Kidney	wgEncodeUwDnaseHrgecPkRep1.narrowPeak
	Kidney	wgEncodeUwDnaseHrgecPkRep2.narrowPeak
Liver	Liver	wgEncodeAwgDnaseDuke8988tUniPk.narrowPeak
	Liver	wgEncodeAwgDnaseDukeHepatocytesUniPk.narrowPeak
	Liver	wgEncodeAwgDnaseDukeHuh75UniPk.narrowPeak
	Liver	wgEncodeAwgDnaseDukeHuh7UniPk.narrowPeak
	Liver	wgEncodeAwgDnaseDukeStellateUniPk.narrowPeak
	Liver	wgEncodeAwgDnaseUwdukeHepg2UniPk.narrowPeak
	Liver	wgEncodeOpenChromDnase8988tPk.narrowPeak
	Liver	wgEncodeOpenChromDnaseHepatocytesPk.narrowPeak
	Liver	wgEncodeOpenChromDnaseHepg2Pk.narrowPeak
	Liver	wgEncodeOpenChromDnaseHuh75Pk.narrowPeak
	Liver	wgEncodeOpenChromDnaseHuh7Pk.narrowPeak
	Liver	wgEncodeOpenChromDnaseStellatePk.narrowPeak
	Liver	wgEncodeUwDnaseHepg2PkRep1.narrowPeak
	Liver	wgEncodeUwDnaseHepg2PkRep2.narrowPeak
Lung	Lung	wgEncodeAwgDnaseUwAg04450UniPk.narrowPeak
	Lung	wgEncodeAwgDnaseUwHpfUniPk.narrowPeak
	Lung	wgEncodeAwgDnaseUwNhlfUniPk.narrowPeak
	Lung	wgEncodeOpenChromDnaseImr90Pk.narrowPeak
	Lung	wgEncodeUwDnaseAg04450PkRep1.narrowPeak
	Lung	wgEncodeUwDnaseAg04450PkRep2.narrowPeak
	Lung	wgEncodeUwDnaseHpfPkRep1.narrowPeak
	Lung	wgEncodeUwDnaseHpfPkRep2.narrowPeak
	Lung	wgEncodeUwDnaseNhlfPkRep1.narrowPeak
	Lung	wgEncodeUwDnaseNhlfPkRep2.narrowPeak
Melanoma	Melanoma Cell Line derived from Melanoma Metastasis	wgEncodeOpenChromDnaseMel2183Pk.narrowPeak

Monocytes	Monocytes	wgEncodeAwgDnaseUwMonocytescd14ro01746UniPk.narrowPeak
	Monocytes	wgEncodeOpenChromDnaseMonocd14Pk.narrowPeak
	Monocytes	wgEncodeUwDnaseMonocd14ro1746PkRep2.narrowPeak
Muscle	Muscle	wgEncodeAwgDnaseDukeHsmmembUniPk.narrowPeak
	Muscle	wgEncodeAwgDnaseUwdukeHsmmtubeUniPk.narrowPeak
	Muscle	wgEncodeAwgDnaseUwSkmcUniPk.narrowPeak
	Muscle	wgEncodeOpenChromDnaseHsmmembPk.narrowPeak
	Muscle	wgEncodeOpenChromDnaseHsmmfshdPk.narrowPeak
	Muscle	wgEncodeOpenChromDnaseHsmmtPk.narrowPeak
	Muscle	wgEncodeUwDnaseHsmmtPkRep1.narrowPeak
	Muscle	wgEncodeUwDnaseHsmmtPkRep2.narrowPeak
	Muscle	wgEncodeUwDnaseSkmcPkRep1.narrowPeak
	Muscle	wgEncodeUwDnaseSkmcPkRep2.narrowPeak
	Psoas Muscle	wgEncodeOpenChromDnasePsoasmuscleocPk.narrowPeak
Myometrium	Myometrium	wgEncodeAwgDnaseDukeMyometrUniPk.narrowPeak
	Myometrium	wgEncodeOpenChromDnaseMyometrPk.narrowPeak
Nasal Biopsy	Nasal Biopsy	wgEncodeOpenChromDnaseOlfneurospherePk.narrowPeak
Pancreas	Pancreas	wgEncodeAwgDnaseDukePanisletdUniPk.narrowPeak
	Pancreas	wgEncodeAwgDnaseDukePanisletsUniPk.narrowPeak
	Pancreas	wgEncodeAwgDnaseUwPanc1UniPk.narrowPeak
	Pancreas	wgEncodeOpenChromDnasePanisdPk.narrowPeak
	Pancreas	wgEncodeOpenChromDnasePanisletsPk.narrowPeak
	Pancreas	wgEncodeUwDnasePanc1PkRep1.narrowPeak
	Pancreas	wgEncodeUwDnasePanc1PkRep2.narrowPeak
Prostate	Prostate	wgEncodeAwgDnaseDukeLncapandrogenUniPk.narrowPeak
	Prostate	wgEncodeAwgDnaseDukeRwpe1UniPk.narrowPeak
	Prostate	wgEncodeAwgDnaseUwdukeLncapUniPk.narrowPeak
	Prostate	wgEncodeAwgDnaseUwPrecUniPk.narrowPeak
	Prostate	wgEncodeOpenChromDnaseLncapAndroPk.narrowPeak
	Prostate	wgEncodeOpenChromDnaseLncapPk.narrowPeak
	Prostate	wgEncodeOpenChromDnaseRwpe1Pk.narrowPeak
	Prostate	wgEncodeUwDnaseLncapPkRep1.narrowPeak
	Prostate	wgEncodeUwDnaseLncapPkRep2.narrowPeak
	Prostate	wgEncodeUwDnasePrecPkRep1.narrowPeak
	Prostate	wgEncodeUwDnasePrecPkRep2.narrowPeak
Skeletal	Skeletal Muscle Myoblast	wgEncodeAwgDnaseUwdukeHsmmUniPk.narrowPeak
Muscle	Skeletal Muscle Myoblast	wgEncodeOpenChromDnaseHsmmPk.narrowPeak
Myoblast	Skeletal Muscle Myoblast	wgEncodeUwDnaseHsmmPkRep1.narrowPeak
	Skeletal Muscle Myoblast	wgEncodeUwDnaseHsmmPkRep2.narrowPeak
	Skeletal Muscle Myoblast	wgEncodeUwDnaseLhcnm2Diff4dPkRep1.narrowPeak
	Skeletal Muscle Myoblast	wgEncodeUwDnaseLhcnm2Diff4dPkRep2.narrowPeak
	Skeletal Muscle Myoblast	wgEncodeUwDnaseLhcnm2PkRep1.narrowPeak
	Skeletal Muscle Myoblast	wgEncodeUwDnaseLhcnm2PkRep2.narrowPeak
Skin	Skin	wgEncodeAwgDnaseDukeMelanoUniPk.narrowPeak
	Skin	wgEncodeAwgDnaseDukeProgfibUniPk.narrowPeak
	Skin	wgEncodeAwgDnaseUwAg04449UniPk.narrowPeak
	Skin	wgEncodeAwgDnaseUwAg09309UniPk.narrowPeak
	Skin	wgEncodeAwgDnaseUwAg10803UniPk.narrowPeak
	Skin	wgEncodeAwgDnaseUwBjUniPk.narrowPeak
	Skin	wgEncodeAwgDnaseUwdukeNhekUniPk.narrowPeak
	Skin	wgEncodeAwgDnaseUwNhdfadUniPk.narrowPeak
	Skin	wgEncodeAwgDnaseUwNhdfneoUniPk.narrowPeak
	Skin	wgEncodeOpenChromDnaseColo829Pk.narrowPeak
	Skin	wgEncodeOpenChromDnaseMelanoPk.narrowPeak
	Skin	wgEncodeOpenChromDnaseNhekPk.narrowPeak
	Skin	wgEncodeOpenChromDnaseProgfibPk.narrowPeak
	Skin	wgEncodeUwDnaseAg04449PkRep1.narrowPeak

Skin	Skin	wgEncodeUwDnaseAg04449PkRep2.narrowPeak
	Skin	wgEncodeUwDnaseAg09309PkRep1.narrowPeak
	Skin	wgEncodeUwDnaseAg09309PkRep2.narrowPeak
	Skin	wgEncodeUwDnaseAg10803PkRep1.narrowPeak
	Skin	wgEncodeUwDnaseAg10803PkRep2.narrowPeak
	Skin	wgEncodeUwDnaseBjPkRep1.narrowPeak
	Skin	wgEncodeUwDnaseBjPkRep2.narrowPeak
	Skin	wgEncodeUwDnaseGm04503PkRep1.narrowPeak
	Skin	wgEncodeUwDnaseGm04503PkRep2.narrowPeak
	Skin	wgEncodeUwDnaseGm04504PkRep1.narrowPeak
	Skin	wgEncodeUwDnaseGm04504PkRep2.narrowPeak
	Skin	wgEncodeUwDnaseNhdfadPkRep1.narrowPeak
	Skin	wgEncodeUwDnaseNhdfadPkRep2.narrowPeak
	Skin	wgEncodeUwDnaseNhdfneoPkRep1.narrowPeak
	Skin	wgEncodeUwDnaseNhdfneoPkRep2.narrowPeak
	Skin	wgEncodeUwDnaseNhekPkRep1.narrowPeak
	Skin	wgEncodeUwDnaseNhekPkRep2.narrowPeak
	Skin	wgEncodeUwDnaseRpmi7951PkRep1.narrowPeak
	Skin	wgEncodeUwDnaseRpmi7951PkRep2.narrowPeak
Spinal Cord	Spinal Cord	wgEncodeAwgDnaseUwHaspUniPk.narrowPeak
	Spinal Cord	wgEncodeUwDnaseHaspPkRep1.narrowPeak
	Spinal Cord	wgEncodeUwDnaseHaspPkRep2.narrowPeak
Testis	Testis	wgEncodeAwgDnaseUwNt2d1UniPk.narrowPeak
	Testis	wgEncodeUwDnaseNt2d1PkRep1.narrowPeak
	Testis	wgEncodeUwDnaseNt2d1PkRep2.narrowPeak
Tonsil	Tonsil	wgEncodeOpenChromDnaseGcbcellPk.narrowPeak
	Tonsil	wgEncodeOpenChromDnaseNaivebcellPk.narrowPeak
Urothelium	Urothelium	wgEncodeAwgDnaseDukeUrotheliaUniPk.narrowPeak
	Urothelium	wgEncodeAwgDnaseDukeUrotheliaut189UniPk.narrowPeak
	Urothelium	wgEncodeOpenChromDnaseUrothelPkV2.narrowPeak
	Urothelium	wgEncodeOpenChromDnaseUrothelUt189PkV2.narrowPeak
Uterus	Uterus	wgEncodeAwgDnaseDukeIshikawaestradiolUniPk.narrowPeak
	Uterus	wgEncodeAwgDnaseDukeIshikawatamoxifenUniPk.narrowPeak
	Uterus	wgEncodeOpenChromDnaseIshikawaEst10nm30mPk.narrowPeak
	Uterus	wgEncodeOpenChromDnaseIshikawaTam10030Pk.narrowPeak

Table S3.2: Enrichment of lipid loci in transcription factor binding sites and histone modifications from relevant Tier 1 and Tier 2 cell types

Regulatory Feature	Observed Number of index SNPs in Feature	Expected Number of index SNPs in Feature	Enrichment *P*-value	Annotation
POLR2A	116	59.17	6.23×10^{-24}	
RCOR1	87	41.86	1.75×10^{-16}	
SP1	52	17.29	1.48×10^{-15}	
EP300	82	40.67	1.08×10^{-14}	literature
eGFP JUND	86	43.98	2.20×10^{-14}	
H3K4me3	72	34.27	1.38×10^{-13}	literature
MXI1	62	26.53	2.79×10^{-13}	
MYC	71	33.76	5.75×10^{-13}	
H3K36me3	52	20.55	1.07×10^{-12}	
MYBL2	39	11.69	1.17×10^{-12}	
TBL1XR1	63	29.30	8.89×10^{-12}	lipid gene regulator
H3K9me1	111	70.65	2.11×10^{-11}	literature
SMC3	63	30.21	2.57×10^{-11}	
ARID3A	68	34.44	3.44×10^{-11}	
H3k4me1	74	38.57	4.82×10^{-11}	
H3K9ac	56	25.49	1.25×10^{-10}	
MAZ	70	37.11	3.52×10^{-10}	
BHLHE40	67	34.86	4.42×10^{-10}	
TBP	57	27.07	4.95×10^{-10}	
eGFP GATA2	64	32.64	7.25×10^{-10}	
MAX	63	32.05	7.86×10^{-10}	
JUND	85	52.19	1.04×10^{-9}	
NCOR1	81	45.39	1.22×10^{-9}	lipid gene regulator
FOXA1	46	20.41	5.00×10^{-9}	lipid gene regulator
NFIC	48	21.88	6.26×10^{-9}	
TEAD4	44	19.11	7.09×10^{-9}	
TAL1	49	23.02	1.08×10^{-8}	
CEBPB	90	59.99	1.69×10^{-8}	lipid gene regulator
CCNT2	49	23.51	1.99×10^{-8}	
HDAC2	33	12.05	2.13×10^{-8}	
HNF4G	26	7.82	2.19×10^{-8}	
RFX5	48	23.02	2.20×10^{-8}	
eGFP JUNB	52	25.73	2.39×10^{-8}	literature
RXRA	25	7.50	4.23×10^{-8}	lipid gene regulator
ELF1	40	17.21	4.38×10^{-8}	
JUN	61	34.37	1.08×10^{-7}	
CREB1	45	21.64	1.10×10^{-7}	literature
CHD2	47	23.26	1.65×10^{-7}	
eGFP HDAC8	24	7.62	2.51×10^{-7}	
HMGN3	43	20.66	3.08×10^{-7}	
CUX1	40	18.64	4.04×10^{-7}	
ZNF143	51	27.20	4.49×10^{-7}	
CTCF	75	49.00	7.41×10^{-7}	
ZC3H11A	32	13.68	1.29×10^{-6}	
HNF4A	28	10.93	1.66×10^{-6}	lipid gene regulator
IRF1	49	26.87	1.71×10^{-6}	
YY1	42	21.34	2.22×10^{-6}	literature
TCF7L2	23	8.15	2.28×10^{-6}	literature
USF2	30	12.76	2.98×10^{-6}	
MBD4	16	4.26	3.52×10^{-6}	
ZNF384	47	25.87	3.84×10^{-6}	

Regulatory Feature	Observed Number of index SNPs in Feature	Expected Number of index SNPs in Feature	Enrichment *P*-value	Annotation
SIN3AK20	28	11.66	3.93×10^{-6}	
NFYA	20	6.79	7.28×10^{-6}	
SPI1	37	18.56	8.45×10^{-6}	
BRCA1	22	8.21	9.76×10^{-6}	
RAD21	52	31.14	1.22×10^{-5}	
SREBP1	11	2.29	1.40×10^{-5}	lipid gene regulator
E2F6	35	17.56	1.48×10^{-5}	
HDAC1	22	8.52	1.85×10^{-5}	literature
ZBTB7A	25	10.54	2.06×10^{-5}	
UBTF	32	15.91	3.31×10^{-5}	
HCFC1	38	20.75	4.65×10^{-5}	
TAF1	29	13.98	4.65×10^{-5}	
TCF12	20	7.85	7.83×10^{-5}	
E2F4	23	10.14	8.46×10^{-5}	
CEBPD	16	5.54	8.69×10^{-5}	lipid gene regulator
EGR1	28	13.53	8.85×10^{-5}	
KDM5B	23	10.22	1.02×10^{-4}	
PML	32	17.10	1.42×10^{-4}	
RUNX3	40	23.82	2.09×10^{-4}	
USF1	26	12.74	2.12×10^{-4}	lipid gene regulator
FOS	15	5.37	2.16×10^{-4}	
EBF1	31	16.38	2.26×10^{-4}	
FOXA2	29	15.33	2.86×10^{-4}	lipid gene regulator
eGFP FOS	30	16.04	3.06×10^{-4}	
REST	25	12.46	3.30×10^{-4}	
FOSL2	20	8.83	3.60×10^{-4}	
GTF2F1	24	11.97	4.04×10^{-4}	
CHD1	20	9.14	4.59×10^{-4}	
eGFP NR4A1	11	3.50	6.68×10^{-4}	literature
ATF1	35	21.01	6.95×10^{-4}	
POU2F2	21	10.12	7.41×10^{-4}	
SAP30	17	7.44	7.80×10^{-4}	
CEBPZ	6	1.17	9.52×10^{-4}	literature
NR2F2	20	9.62	1.01×10^{-3}	
PHF8	25	13.65	1.24×10^{-3}	
MAFF	59	43.42	1.30×10^{-3}	
ELK1	20	10.04	1.66×10^{-3}	
MAFK	75	59.30	1.70×10^{-3}	
ATF3	16	7.46	2.48×10^{-3}	literature
SREBP2	2	0.08	2.78×10^{-3}	lipid gene regulator
GATA2	20	10.47	2.84×10^{-3}	
SIN3A	17	8.47	3.37×10^{-3}	
GTF2B	16	7.77	3.38×10^{-3}	
WRNIP1	16	7.77	3.60×10^{-3}	
ETS1	15	7.06	3.66×10^{-3}	
SIX5	9	3.16	3.80×10^{-3}	
KAP1	28	17.39	4.17×10^{-3}	
IRF4	16	8.10	5.70×10^{-3}	
CREBBP	87	70.93	5.90×10^{-3}	lipid gene regulator
ZEB1	8	2.80	6.29×10^{-3}	
GTF3C2	7	2.27	6.98×10^{-3}	
PAX5	22	13.06	8.32×10^{-3}	
GABPA	18	10.15	1.01×10^{-2}	

Regulatory Feature	Observed Number of index SNPs in Feature	Expected Number of index SNPs in Feature	Enrichment *P*-value	Annotation
NR2C2	8	3.08	$1.10x10^{-2}$	literature
NFYB	24	15.13	$1.19x10^{-2}$	
STAT1	10	4.52	$1.36x10^{-2}$	
RBBP5	18	10.59	$1.50x10^{-2}$	
FOSL1	9	4.03	$1.76x10^{-2}$	
GATA1	17	10.02	$2.01x10^{-2}$	
MTA3	15	8.61	$2.25x10^{-2}$	
SMARCA4	7	2.90	$2.39x10^{-2}$	
NRF1	11	5.69	$2.43x10^{-2}$	lipid gene regulator
SIRT6	5	1.67	$2.49x10^{-2}$	lipid gene regulator
ATF2	20	12.81	$2.56x10^{-2}$	literature
STAT2	7	2.92	$2.60x10^{-2}$	
PBX3	8	3.77	$3.30x10^{-2}$	
H3k27me3	46	35.95	$3.44x10^{-2}$	literature
SP2	6	2.54	$4.05x10^{-2}$	
ZBTB33	7	3.23	$4.07x10^{-2}$	
NFE2	5	1.92	$4.19x10^{-2}$	
CTCFL	7	3.29	$4.60x10^{-2}$	
BCLAF1	11	6.32	$4.75x10^{-2}$	
RPC155	3	0.88	$5.83x10^{-2}$	
STAT5A	15	10.13	$7.48x10^{-2}$	
STAT3	7	3.71	$7.70x10^{-2}$	
GRp20	2	0.50	$8.75x10^{-2}$	
THAP1	5	2.46	$9.66x10^{-2}$	
ZNF274	9	5.55	$1.01x10^{-1}$	
MEF2A	13	8.91	$1.03x10^{-1}$	
BACH1	12	8.42	$1.30x10^{-1}$	literature
TAF7	5	2.74	$1.36x10^{-1}$	
IRF3	2	0.67	$1.45x10^{-1}$	literature
BATF	14	10.35	$1.46x10^{-1}$	
RELA	12	8.71	$1.57x10^{-1}$	
ESRRA	2	0.72	$1.59x10^{-1}$	
TCF3	10	7.04	$1.64x10^{-1}$	
EZH2	5	3.06	$1.91x10^{-1}$	literature
BCL3	10	7.33	$1.95x10^{-1}$	
IKZF1	8	6.00	$2.50x10^{-1}$	
MEF2C	6	4.34	$2.63x10^{-1}$	
SMARCB1	3	1.84	$2.79x10^{-1}$	
TRIM28	10	8.17	$2.98x10^{-1}$	literature
NFATC1	9	7.25	$2.99x10^{-1}$	
HSF1	2	1.15	$3.19x10^{-1}$	
FOXM1	15	13.35	$3.54x10^{-1}$	literature
SETDB1	8	6.90	$3.84x10^{-1}$	
RDBP	1	0.49	$3.95x10^{-1}$	
HDAC6	1	0.54	$4.27x10^{-1}$	
SRF	6	5.34	$4.46x10^{-1}$	
ZNF263	3	2.73	$5.18x10^{-1}$	
H3K9me3	116	119.5	$7.73x10^{-1}$	literature

Table S3.3: Primers used in luciferase expression constructs

SPTLC3	rs1321940F	GTGCTCACTGAAACGTGTCT
	rs1321940R	CAGTGCACAATGTCAATATGGA
	rs364585F	CACCTGACCATTTCTCCCCA
	rs364585R	ACGAAACACCCCTGAAGACA
ANGPTL8	rs3810308F	AGAGGAGGCAGAAGTGAAGG
	rs3810308R	CCAGCTCTGAACTCTGGACA
	rs737337F	GGGTAGGGATGTGGAGTGAG
	rs737337R	ATTCCCATTGCCTCTCTGCT
FAM117B	rs11692610F	TAAAAGCCCGAACGAGATGC
	rs11692610R	GGGTTTTGTTGTTGTTGGGC
	rs11694172F	TCCTGGGTTCAAGCAGTTCT
	rs11694172R	ATCCCAAAGGCCTCCAAAGA
SORT1	rs12740374F	ACACATTTTCAGGGGAGCCT
	rs12740374R	AGGAGAGGTGGGGAGATGAT
	rs629301F	TCTCCTCAGTTTTGCCGACT
	rs629301R	CTCTCCCACCGTAGAAGTCC
IRF2BP2	rs526936F	AAAACTAGCTGGGCGTGGTA
	rs526936R	CCCCGAGTAAAACACCCTCT
	rs514230F	CCCCAGACATGAGGACAAGT
	rs514230R	GCAGGCCGGTTTTCTTCTTT
ADH5	rs2602836F	GCCAGCAATGAACAAGTGGA
	rs2602836R	CGCACATGTAACAAACCTGC
	rs1800759F	CTGGCATAGGGGTCACTCAT
	rs1800759R	AATGGGCGATTCTGAGGAGT

CHAPTER IV

Investigating the functional role of structural variation in myocardial infarction risk from whole genome sequencing of a Norwegian population

4.1 Abstract

Structural variation (SV) is a class of genetic variation whose implication in complex disease is currently not well understood. We investigate the role of deletions, duplications, and inversions in risk for heart disease within a cohort of 2,202 Norwegians from The HUNT Study, which includes cases with myocardial infarction (MI) and matched controls. Using complementary approaches for discovering structural variation from whole genome sequencing data, we identify SVs in the Norwegian population and perform genome wide association analyses with myocardial infarction and quantitative lipid traits. We confirm linkage disequilibrium between a deletion on chromosome 2 and a single variant associated with MI at the *WDR12* locus. Structural variants identified by this study can be used for imputation into the larger HUNT cohort for increased power to detect significant associations.

4.2 Introduction

Early-onset myocardial infarction (MI) is a major cause of mortality in the U.S. and throughout the world, with both common and rare genetic mutations contribut-

ing to its multifactorial risk (Mozaffarian et al., 2015). GWAS efforts over the past 8 years have led to discoveries of over 50 single genetic risk variants associated with coronary artery disease (CAD) or MI (McPherson et al., 2007; Samani et al., 2007; Helgadottir et al., 2007; Myocardial Infarction Genetics Consortium et al., 2009; Schunkert et al., 2011; Coronary Artery Disease (C4D) Genetics Consortium, 2011; IBC 50K CAD Consortium, 2011; CARDIoGRAMplusC4D Consortium et al., 2013; CARDIoGRAMplusC4D Consortium et al., 2015). According to the latest meta-analysis involving nearly 185,000 participants, single variants identified by genome wide association studies together explain 12.9 $\pm$ 0.4% of the trait heritability for CAD (CARDIoGRAMplusC4D Consortium et al., 2015). Do et al. (2015) used exome sequencing to find that the burden of rare mutations in *APOA5* and *LDLR* explains 0.14% and 0.24% of the total variance for MI and roughly 0.28% and 0.48% of the heritability, respectively. Many of these risk loci contain variants associated with LDL cholesterol (*LPA*, *APOB*, *SORT1*, *LDLR*, *APOE*, *ABCG5-ABCG8*, and *PCSK9*), HDL cholesterol (*ANKS1A*), and triglycides (*TRIB1* and *APOA5-A4-C3-A1*), suggesting a plausible role of lipid modulation in disease risk (Roberts, 2015). Still, for other disease-associated loci, the risk mechanism remains unclear.

Our current understanding of the functional role of structural variation (SV) in myocardial infarction is in its infancy relative to simpler forms of genetic variation. These balanced or unbalanced copy number changes, typically defined as 50 base pairs to several kb in size, have traditionally been discovered using array-based (McCarroll et al., 2008; Conrad et al., 2010) and clone-based methods (Kidd et al., 2008). For example, Conrad et al. (2010) used array CGH to report a CNV in LD with a variant at the MI-risk locus, *WDR12*. Advances in sequencing technology have prompted the development of methods for discovering and genotyping

structural variants at higher resolution (Mills et al., 2011; Sudmant et al., 2015). Whole-genome sequencing allows for a finer interrogation of the genome to discover structural variants, including low frequency and rare copy number events.

Using whole-genome sequencing in a Norwegian population of 2,202 matched MI cases and controls, we investigate whether structural variation plays a functional role in myocardial infarction risk and regulation of quantitative lipid traits. We hypothesize that there are different frequencies of structural variants in MI cases compared to controls and apply established and complementary SV detection algorithms to discover and genotype deletions, duplications, and inversions. We carry out a genome wide association study framework to test for associations that will implicate structural variation in heart disease risk.

4.3 Methods

4.3.1 Phenotype measurements

The population-based Nord-Trøndelag Health Study (The HUNT Study) is a collaboration between the HUNT Research Centre (Faculty of Medicine, Norwegian University of Science and Technology NTNU), Nord-Trøndelag County Council, Central Norway Health Authority, and the Norwegian Institute of Public Health (Krokstad et al., 2013). A set of 2,202 Norwegian individuals was chosen from the Nord-Trøndelag Health study for whole genome sequencing. Sequenced participants were composed of 1,101 cases with early-onset MI, and 1,101 healthy controls that were one-to-one matched on age, sex, and birth municipality. The earliest-onset cases, primarily from batch 1 (see Section 4.3.2), were defined as an MI event at age $\leq$ 55 years for males and $\leq$ 65 years for females (Figure 4.1). Controls were chosen from cohort participants without self-reported and/or hospital diagnosed MI, MI in first- or second-degree family members, cardiovascular disease, diabetes, or hypertension.

No sequenced individuals had any known first- or second-degree relatives among the others selected for sequencing.

We also collected directly-measured lipid phenotypes including non-fasting plasma HDL cholesterol, triglycerides, and total cholesterol (Table 4.1). LDL cholesterol levels for participants with triglyceride levels <400 mg/dL were estimated using the Friedewald formula, as shown in Equation 4.1 (Friedewald et al., 1972).

$$LDL\text{-}C = TC - HDL\text{-}C - \frac{TG}{5} \tag{4.1}$$

Lipid measurements were collected on the same samples at two time points approximately 10 years apart as included in the HUNT2 (1995-97) and HUNT3 (2006-08) efforts. Residuals estimated from each HUNT stage were averaged for the association analysis (see Section 4.3.4).

4.3.2 Whole-genome sequencing

Illumina-based whole-genome sequencing (∼100 bp reads) of 2,202 samples was performed at the University of Michigan DNA Sequencing Core in 3 batches over a 3-year period. Equal numbers of MI cases and controls in batches 1 (n=602), 2 (n=800), and 3 (n=800) were sequenced with total average coverage of 5.9x, 5.4x, and 4.3x, respectively. A subset of individuals (n=210) was also targeted for exome sequencing, but the targeted sequencing was removed from these samples for the subsequent analysis. Differences in library preparation protocol resulted in varying insert size distributions between samples (Figure 4.2) and across batch (Figure 4.3), with batch 1 samples generally having smaller insert sizes than batch 2 and 3 samples. Insert size standard deviations differ by batch but not by case-control status (Figure 4.4). Sequence alignment was performed using the GotCloud pipeline (`http://genome.sph.umich.edu/wiki/GotCloud`).

4.3.3 Structural variant calling

Deletions were called and genotyped by integrating several technical features of the sequence data as well as population-scale patterns across the 2,202 genomes analyzed (GenomeSTRiP 2.0, Handsaker et al. (2015)). In brief, GenomeSTRiP incorporates information from break-point spanning reads, paired-end sequences, and local variation in read depth coverage to discover deletions with improved sensitivity and specificity relative to other algorithms that use only one or two of these features. To improve upon the power to detect structural variation in a single genome, GenomeSTRiP considers how alleles are shared across multiple genomes and patterns of sequence heterogeneity to accurately determine the state of each variant in every individual genome of the population. Six HUNT individuals with an outlier number of variants by this method were removed from the subsequent association analysis (Figure 4.5).

Deletions (DEL), tandem duplications (DUP), and inversions (INV) were also called per individual whole-genome targeted sample using a read-pair and split-read based method (DELLY, Rausch et al. (2012)). By integrating the paired-end and split-read alignments, DELLY can delineate copy-number variable events as well as balanced rearrangements such as inversions. The human reference genome containing decoy sequence (to remove reads that would otherwise map with low quality in the reference) was downloaded from `ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/technical/reference/phase2_reference_assembly_sequence/`, and telomeric and centromeric regions were excluded in SV calling. Counts of events called by mean sequencing depth show that batch 1 samples with tighter insert size distributions generally called more deletions than samples from the other two batches (Figure 4.6). Structural variants that passed the DELLY quality filter were com-

bined across all 2,202 sequenced samples by merging events that overlapped by 80% reciprocally. Genotypes were combined by assuming that the absence of a variant in the overlapping region implies homozygous reference genotypes.

Finally, we used the HUNT metadata information from GenomeSTRiP to genotype the merged DELLY deletions, as well as those deletions called from the 1000 Genomes Project Phase 3 v5 (1000 Genomes Project Consortium et al., 2012).

We examined whether SVs were in strong linkage disequilibrium (LD) with previously reported CAD- or lipid-associated SNPs in the 2,202 sample HUNT population. Pairwise r^2 was estimated between each SV and the GWAS-reported index SNP using best-guess unphased genotypes.

4.3.4 Association analysis

Association analyses were carried out separately for deletions, duplications, and inversions using PLINK. We performed logistic regression of MI case-control response and genotype predictors with minor allele frequency (MAF) >0.01 (minor allele count >44). We adjusted for covariates age, sex, and batch, as well as the first 10 principle components (PCs) estimated from sequence genotypes of the 2,202 sequenced samples (Equation 4.2).

$$MI\ status = SV\ genotype + birth\ year + sex + batch + PC1\text{–}PC10 \tag{4.2}$$

We also performed association analysis of SV genotypes with quantitative lipid traits HDL-C, LDL-C, TG, and TC. Inverse normalized residuals for each of the four lipid traits were generated separately for each HUNT time point (HUNT2 and HUNT3) with adjustment for birth year and sex. Estimated residuals from the two HUNT stages were then averaged (Figure 4.7) and used as the response in a linear regression (Wald test) with adjustment for covariates MI status, batch, and 10 PCs

(Equation 4.3).

$$\begin{aligned} &\textit{Birth year- and sex-adjusted inverse normalized residuals} = \\ &\textit{SV genotype} + \textit{MI status} + \textit{batch} + \textit{PC}1\text{–}\textit{PC}10 \end{aligned} \tag{4.3}$$

Association analyses in this manner were carried out separately for genotype calls with MAF >0.01 for deletions (from DELLY, GenomeSTRiP, and 1000 Genomes separately), duplications, and inversions that were <1Mb in size.

After checking for cryptic relatedness using genotype information, one individual was found to be contaminated and removed from the association analysis in addition to the six samples with outlier SV calls (see Section 4.3.3). Table 4.2 describes the final counts of individuals used for association with each trait based on QC and phenotype availability.

A schematic diagram of the overall SV analysis pipeline is provided in Figure 4.8.

4.4 Results

We discovered a set of 3,270 deletions (885 2kb-1Mb) with MAF>0.01 in the HUNT population using GenomeSTRiP, as well as 5,564 deletions (1,209 2kb-1Mb), 723 duplications (218 2kb-1Mb), and 493 inversions (183 2kb-1Mb) with MAF>0.01 from DELLY (Table 4.3). We found 252 deletions called in the 1000 Genomes Project Phase 3 v5 that overlapped by 80% reciprocally with those deletions discovered in the HUNT population. Distributions of association P-values for MI status and each quantitative lipid trait are shown in Figures 4.9-4.13. The red P-value distributions in each of these figures represents SVs within 1Mb of any GWAS single variant, while all other variants are shown in black. Known GWAS variants are defined as published variants associated with any of the four lipid traits, published CAD-associated variants, and a set of novel lipid-associated variants identified from an

ongoing exome chip study (Liu and Global Lipids Genetics Consortium, 2014). We did not observe genome-wide significant associations (based on Bonferroni correction for the number of tests) with either MI status or lipids (see top significant results in Tables 4.5-4.7).

Of those 34 CNVs reported by Conrad et al. (2010) with tagged GWAS SNPs, we found 15 nearby CNVs in the HUNT population as well as in the 1000 Genomes Project (Table 4.4). Both DELLY and GenomeSTRiP called a deletion (2:203898933-203904481 and 2:203899034-203904285, respectively) near the Conrad et al. (2010) deletion tagging an MI-associated SNP at the *WDR12* locus. Estimation of linkage disequilibrium between this SNP (rs6725887; chr2:203745885) and the GenomeSTRiP genotypes at this deletion (`DEL_P0227_516`; chr2:203899034-203904285) confirmed their strong linkage disequilibrium (r^2=0.98) (Figure 4.14).

4.5 Discussion

Several conclusions can be drawn from this research including both biological insights and computational lessons. Sequencing experimental protocol can largely affect SV calling, particularly when there is insert size variability between samples. In our study, the 210 samples that were targeted for exome sequencing in addition to whole genome sequencing had systematically smaller insert sizes, which biased deletion calling to smaller events. To eliminate this bias, we excluded the targeted exome from these individual samples and kept only the sequenced whole genomes for SV discovery, genotyping and subsequent analysis.

In addition, the differences in insert size distributions due to changes in library preparation protocol over the sequencing time period resulted in a bias toward calling more small deletions in batch 1 samples (Figure 4.2). Library construction with the

ep*Motion* robotic workstation used gel size selection, which resulted in more consistent insert sizes of the first 576 batch 1 samples. In contrast, the IntegenX library prep protocol used bead size selection for the remaining samples, which resulted in much broader insert size distributions. In paired-end sequencing, fragments are expected to be consistently mapped a particular distance away from each other. A discrepancy in this distance indicates a structural variation between the paired-end tag sequences. For example, a deletion in a sequenced genome will have reads that map further away than expected in the reference genome, since the reference genome will have a DNA fragment that is missing in the sequenced genome. Consequently, variability in size of the sequenced fragments, as produced in the IntegenX-prepped samples, reduces the ability to distinguish small deletion events that deviate from the reference genome. For duplications and inversions however, the numbers of events called were more similar across batch. This is expected since these events were called based on the orientation of mapped reads rather than relying on insert size. We addressed this technical sequencing artifact by adjusting for batch in the association analysis.

Different genotyping approaches give variable results, suggesting the necessity to explore more than one method when studying structural variation. Taking into account the sequence heterogeneity across multiple individuals gives GenomeSTRiP an advantage over other methods, especially when sequencing a sizable cohort. Indeed, we observed more significant deletions by this method than the alternative DELLY approach. This supports the power of harnessing patterns of sequence heterogeneity within a population and integrating paired-end, split-read, and read-depth-based analyses. Sequencing a large number of individuals is a study design that is becoming increasingly more feasible as sequencing costs decline, and we have seen the success-

ful performance of the GenomeSTRiP approach in the 1000 Genomes Project effort (1000 Genomes Project Consortium et al., 2012). In addition to calling deletions, however, the DELLY approach of integrating paired-end and split-read alignments is advantageous in that it can call tandem duplications and balanced rearrangements. Thus, by applying these complementary approaches, we were able to investigate a broad spectrum of genomic rearrangements in the HUNT population.

The choice of algorithm for discovery and genotyping of structural variation may depend on the goals and design of a particular study. For example, in a study for which the primary objective is to discover novel events, taking the union of SV events discovered by both DELLY and GenomeSTRiP would give the most comprehensive set. Where accurate genotyping is of primary concern in a study, genotyping using GenomeSTRiP is the superior choice. To illustrate this, we estimated linkage disequilibrium in the HUNT samples between previously reported (Conrad et al., 2010) trait-associated SNPs and their tagged CNV's. DELLY genotypes consistently resulted in low r^2 estimates (Table 4.4). On the other hand, GenomeSTRiP replicated these LD relationships much more consistently. This suggests that DELLY genotypes are less reliable and perhaps contain an excess of false negatives, calling events as more rare than the truth. Given the insights from this study, thoughtful consideration should be made when choosing a method for structural variation analysis of sequencing data.

Careful study design of matched cases and controls, adjusting for appropriate confounders such as age and sex, and applying the appropriate transformation of quantitative lipid measurements are all critical for identifying true causal associations. Filtering sites based on the quality scores of the respective calling methods is critical to identify a confident set of events and prevent false positive associations.

Our regression results suggest that structural variation does not play a strong role in MI risk and modulating lipid levels. However, the absence of significant large-effect associations for MI and lipids does not suggest an absent role of structural variation in modulating these phenotypes, but rather that we are underpowered to detect them. To increase our power for discovery in the future, structural variants identified here will be imputed into the larger HUNT cohort of 30,000 samples for further study. A comprehensive survey of linkage disequilibrium between structural variation and GWAS-reported single markers did not reveal additional SV-single marker LD relationships. Again, a larger cohort may be needed to identify SVs that tag single markers to suggest novel plausible functional candidates.

4.6 Acknowledgements

We are especially grateful to all the Norwegian volunteers who participated in our study. This work was overseen by Cristen J. Willer and Hyun Min Kang. Sample selection and phenotype collection were performed by Oddgeir L. Holmen and Kristian Hveem. Jin Chen carried out the GotCloud sequence alignment pipeline and He Zhang performed SNP calling and single variant association analyses. Thank you also to Ryan E. Mills and Xuefang Zhao for helpful discussions on this project. EMS was supported by a Rackham Summer Award.

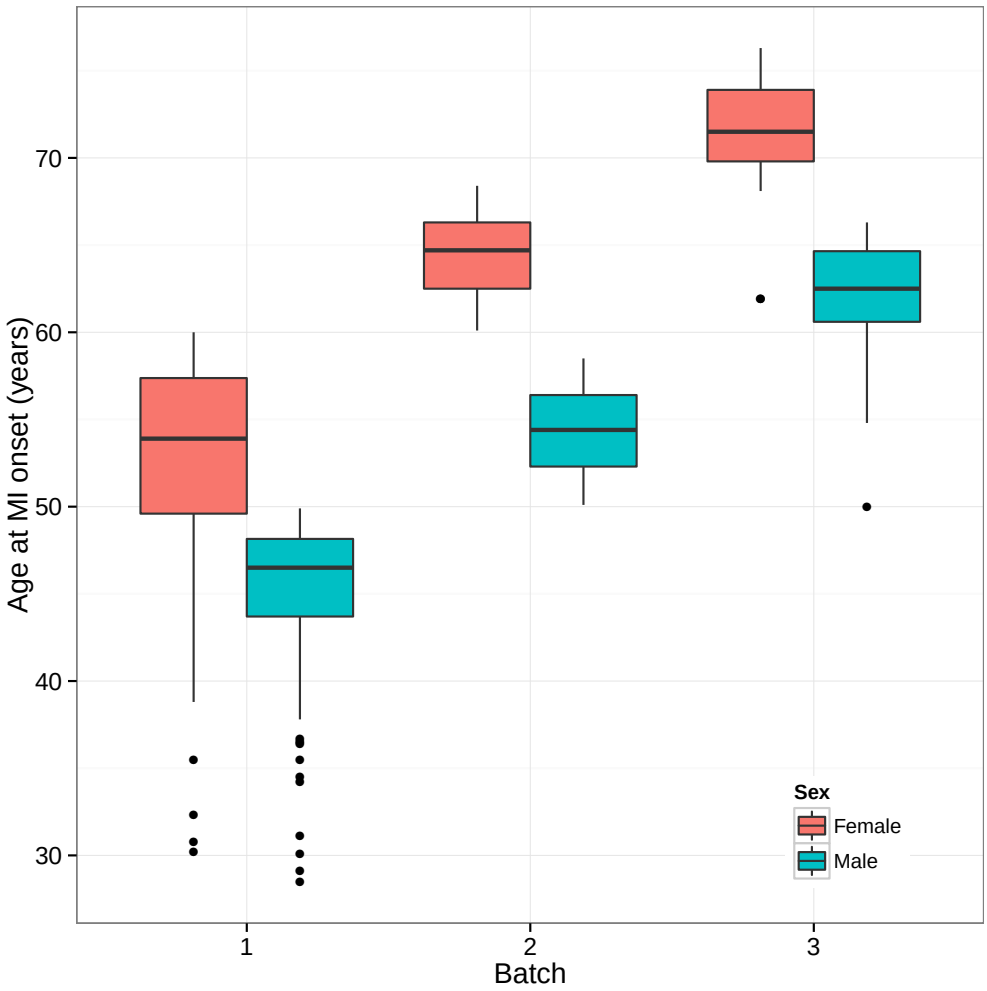


Figure 4.1: Distribution of age at MI onset in 1,101 affected individuals by batch.

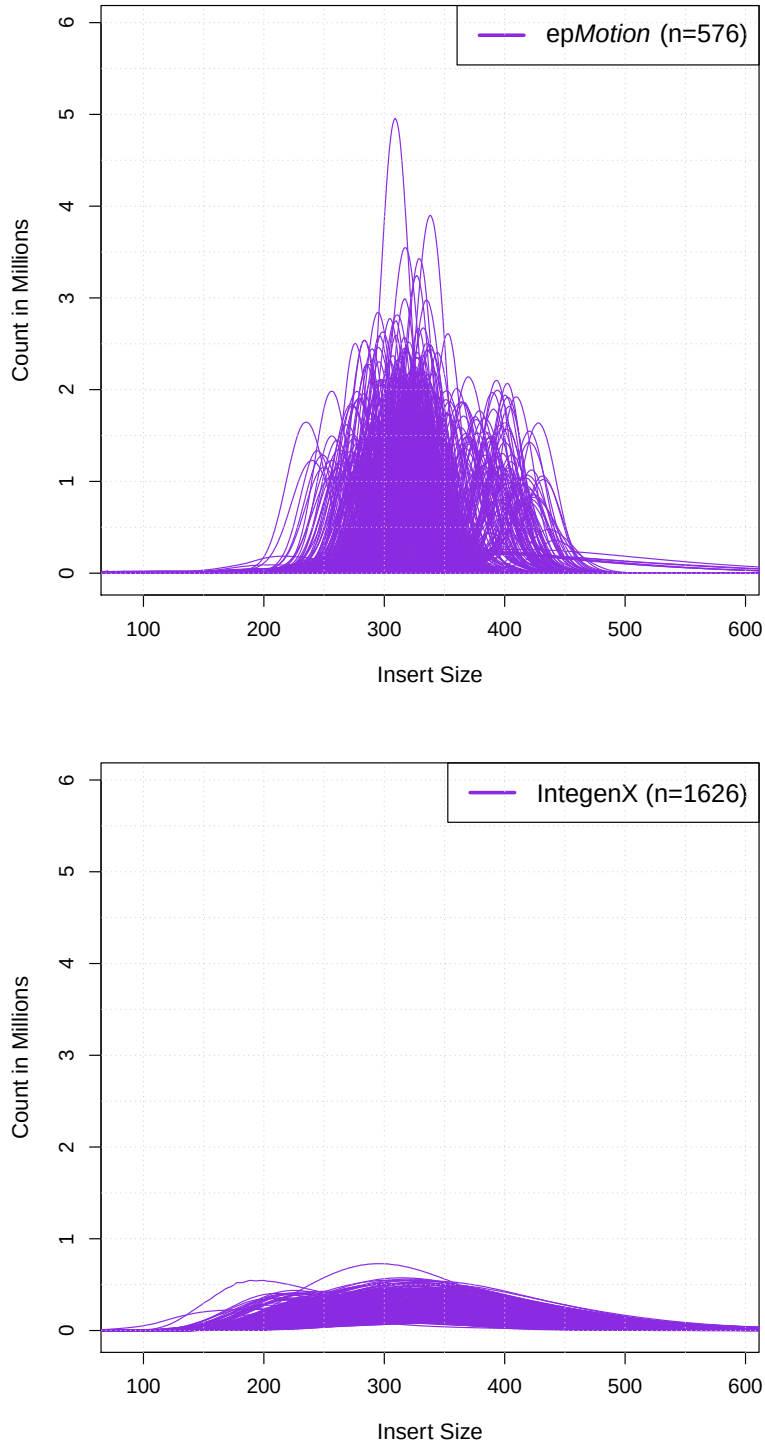


Figure 4.2: Insert size distributions for different library preparation methods. Most batch 1 samples (n=576) were prepped using the ep*Motion* robotic workstation and the remaining samples were prepped using the IntegenX robotic workstation (batch 1, n=26; batch 2, n=800; batch 3, n=800).

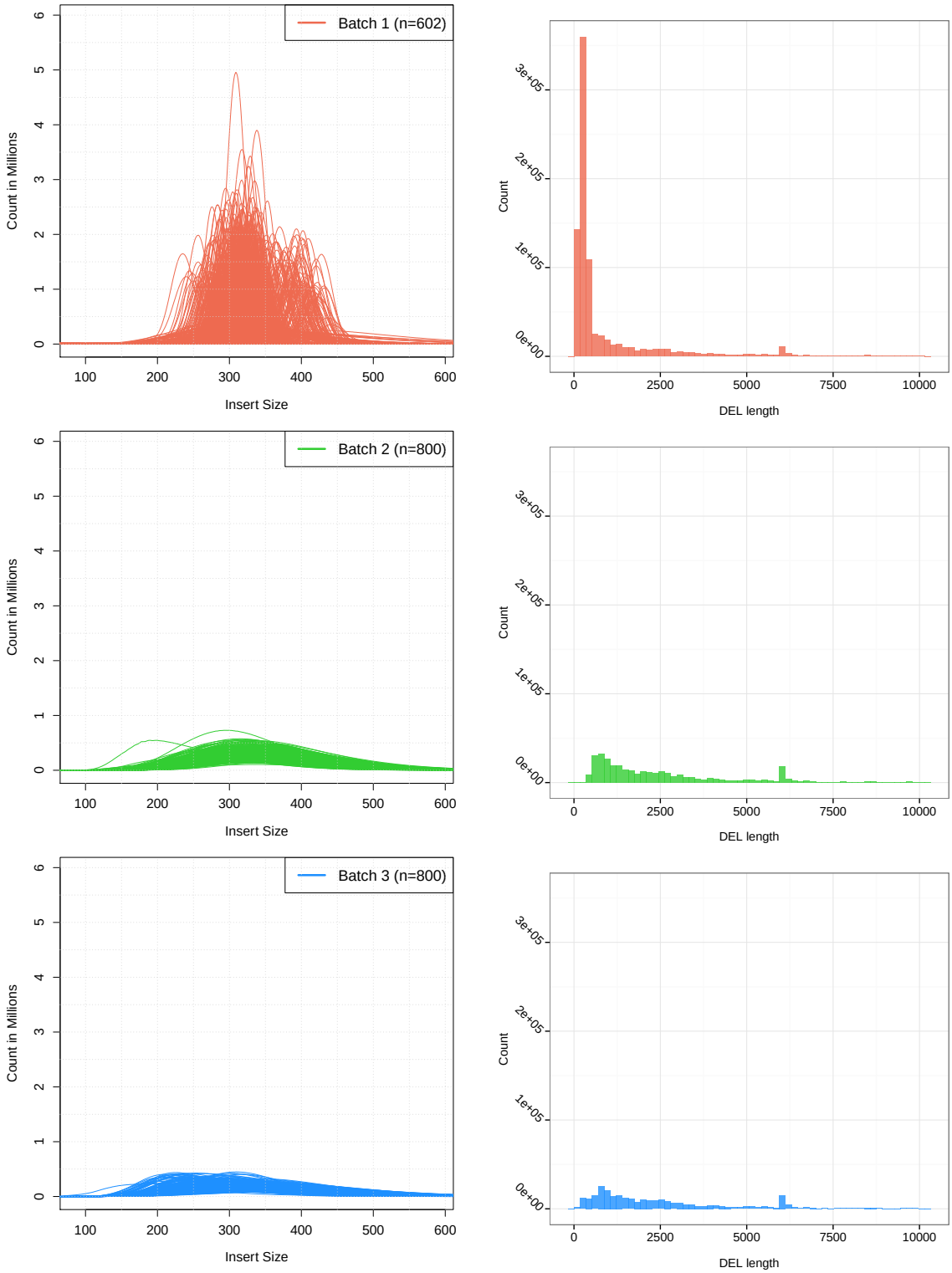


Figure 4.3: Insert size distributions colored by batch. Sequencing data were processed over 3 years in 3 separate library preparation batches 1 (n=602), 2 (n=800), and 3 (n=800).

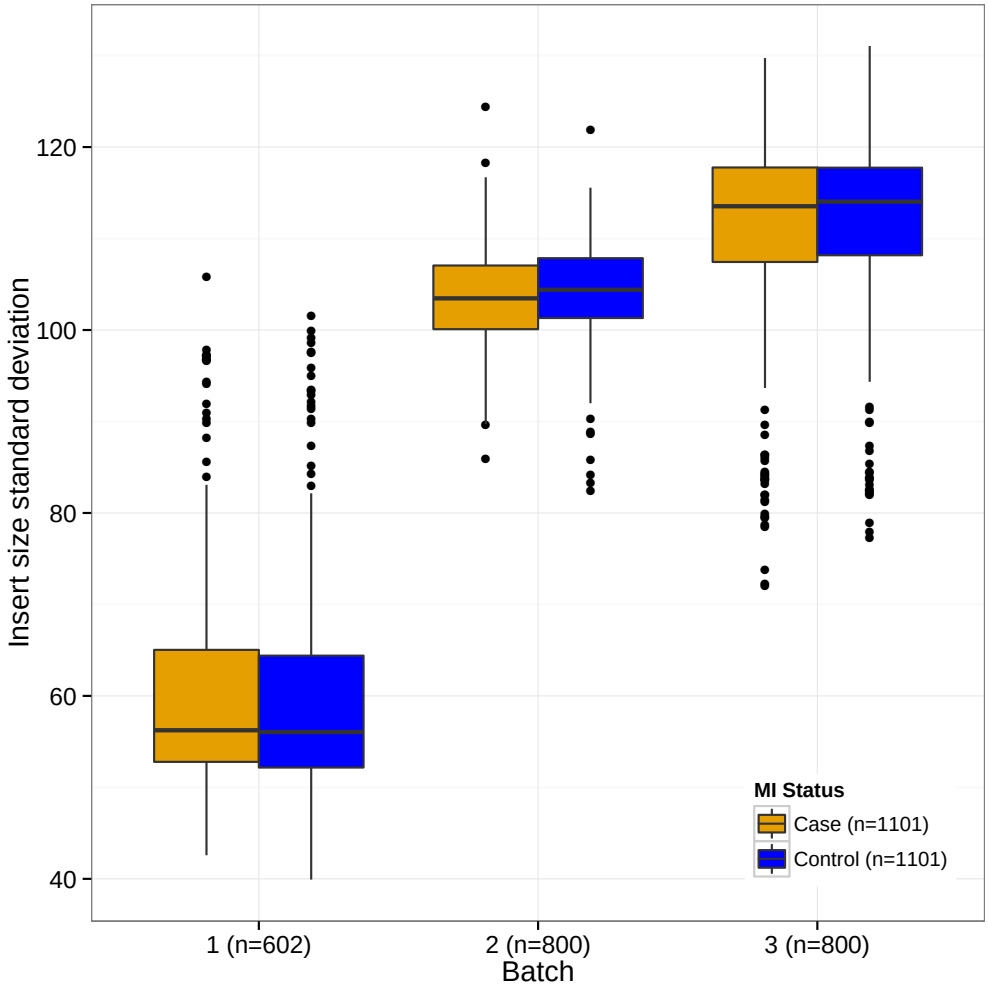


Figure 4.4: Boxplot of insert size standard deviation by batch and disease status.

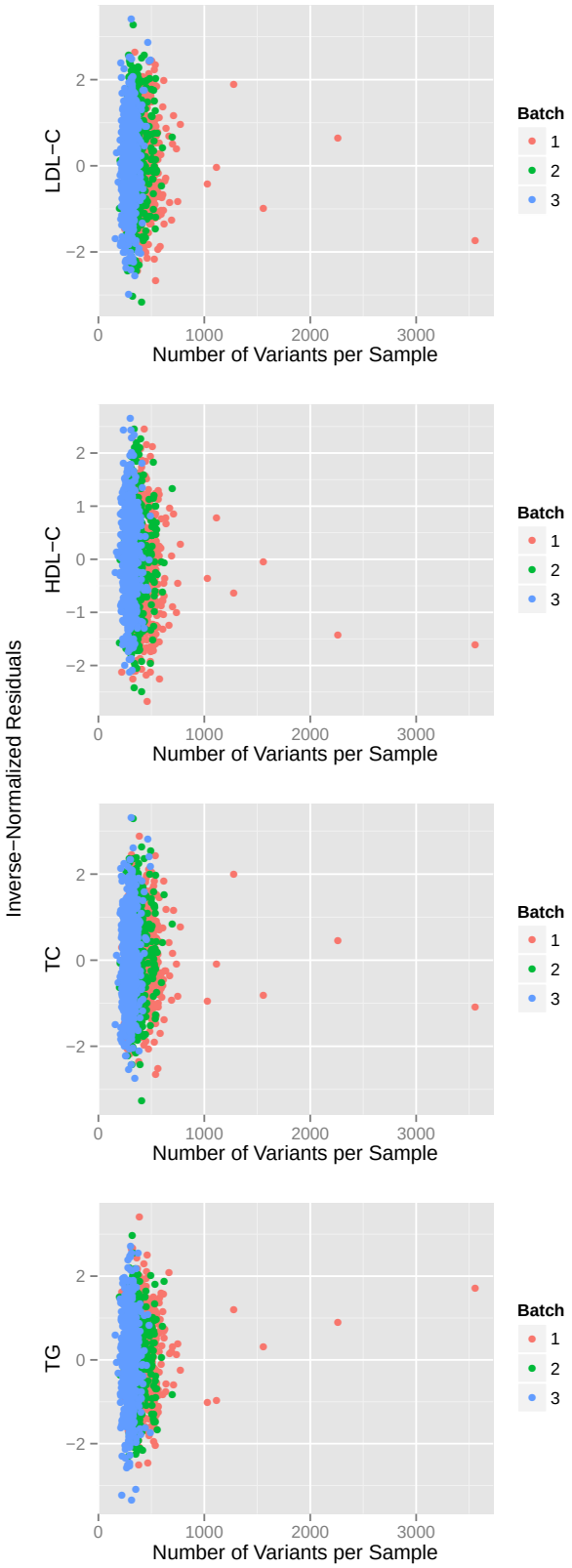


Figure 4.5: Number of structural variants called from GenomeSTRiP by lipid distributions.

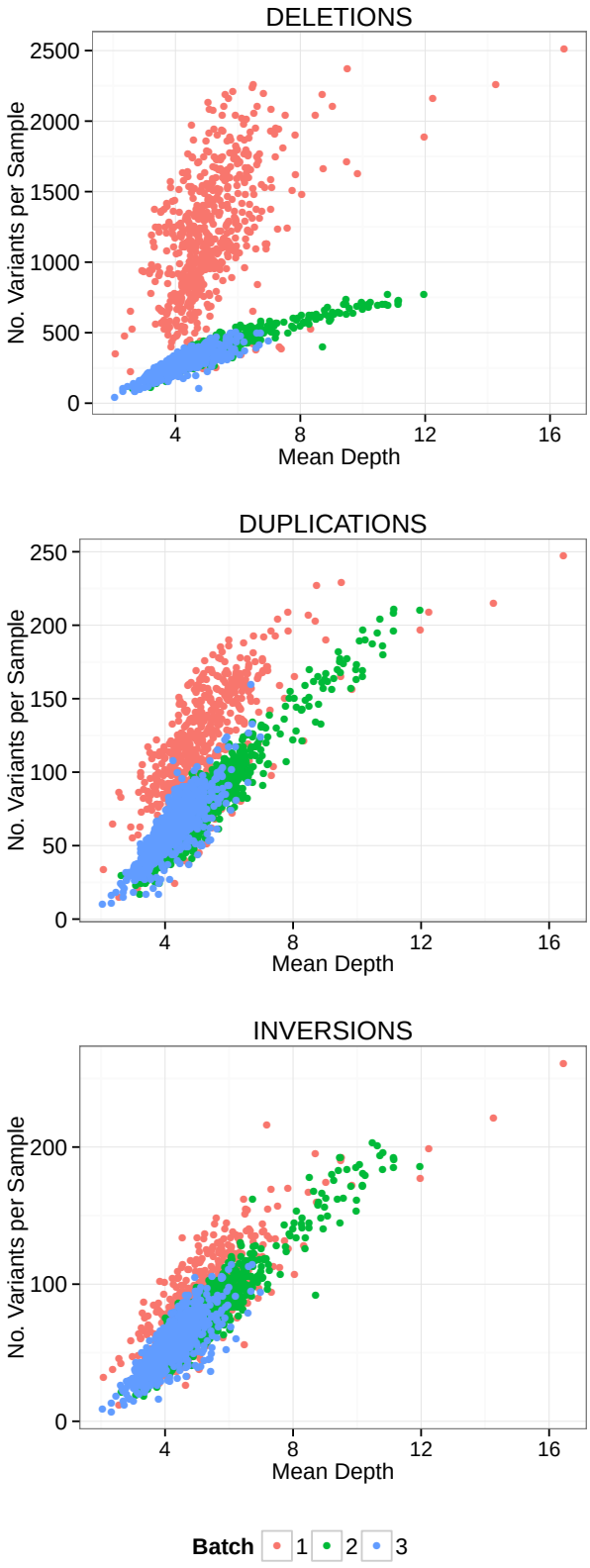


Figure 4.6: Number of structural variants called from DELLY by mean sequencing depth.

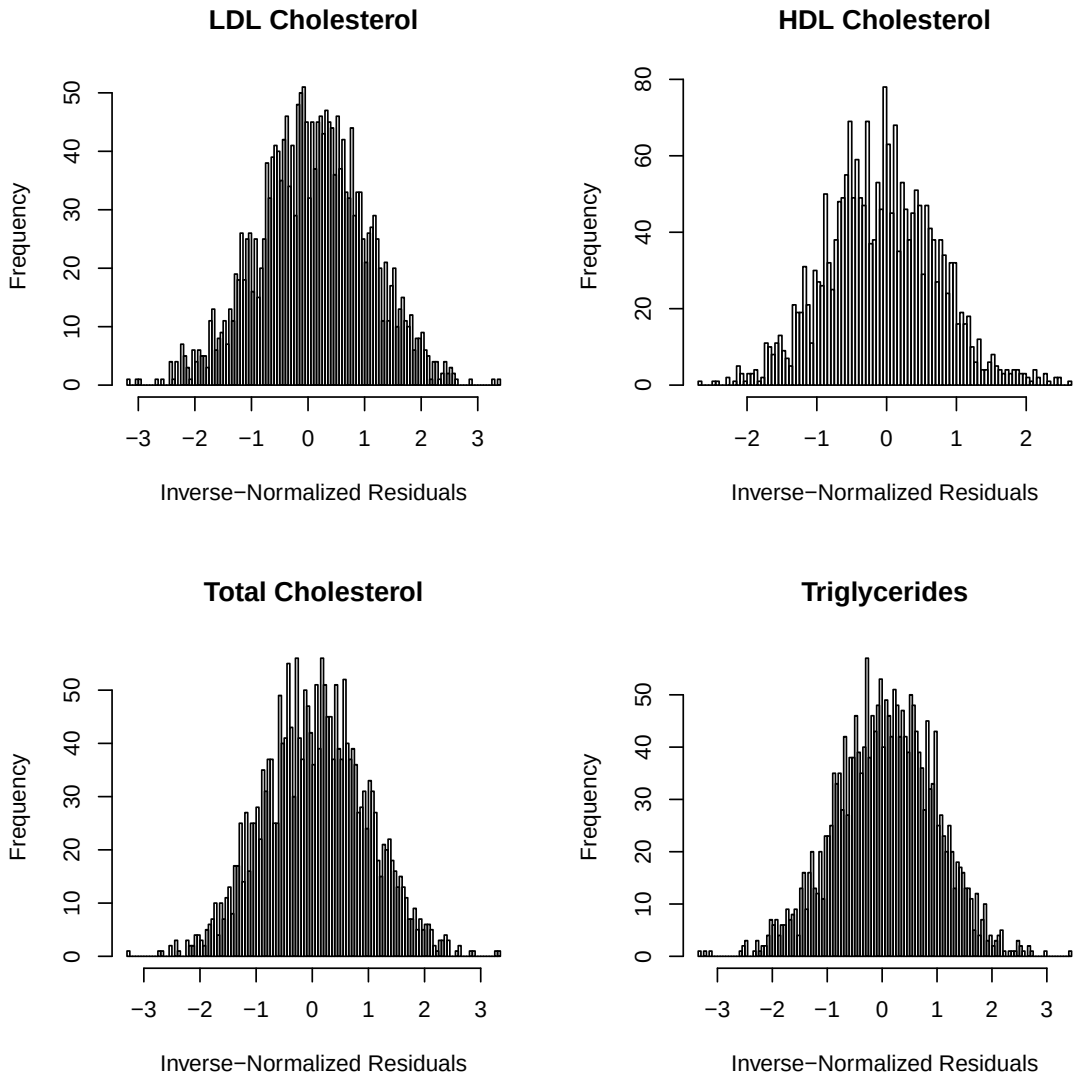


Figure 4.7: Distribution of age- and sex-adjusted residuals for lipids. Residuals were estimated separately for HUNT2 and HUNT3 time points and then averaged.

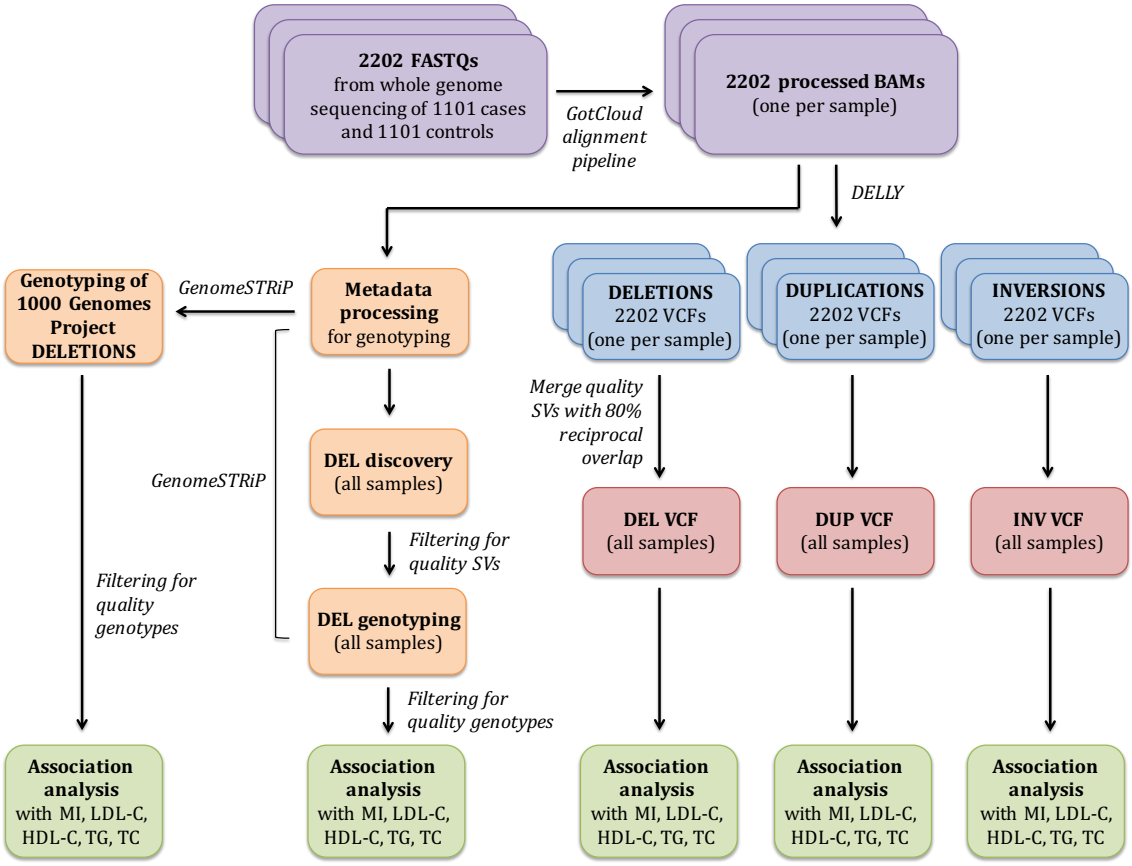
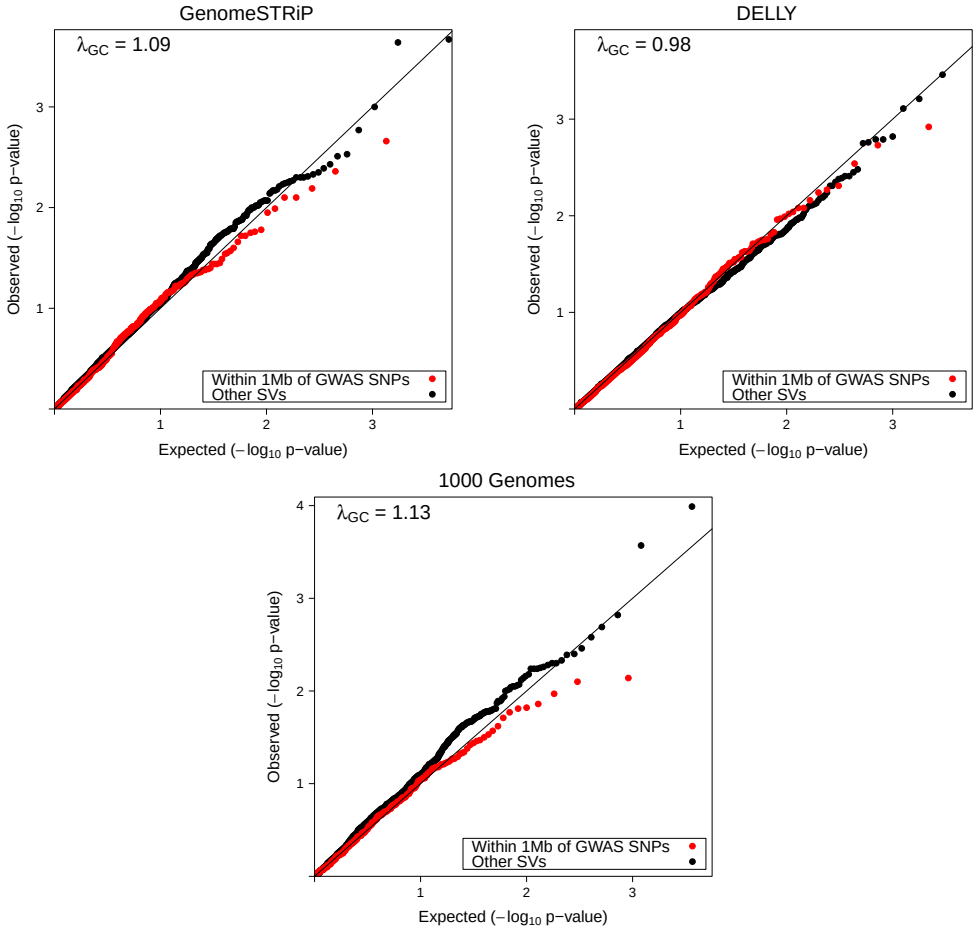


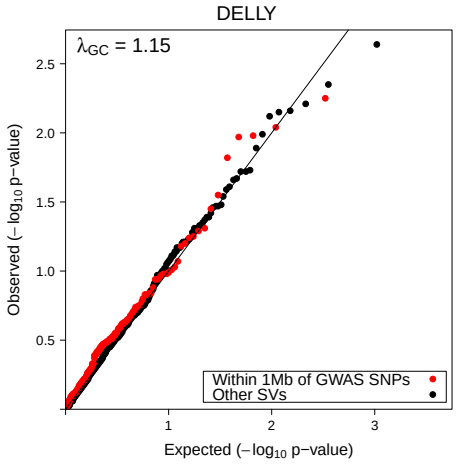
Figure 4.8: Structural variant analysis pipeline.

Figure 4.9: SV association results for MI status. Logistic regression of HUNT genotypes with MI status was carried out separately for **(A)** deletions discovered by GenomeSTRiP, DELLY, and the 1000 Genomes Project; **(B)** duplications discovered by DELLY; and **(C)** inversions discovered by DELLY. QQplots show association *P*-values for structural variants within 1Mb (red) and outside 1Mb (black) of known CAD- or lipid-associated GWAS SNPs.

A. DELETIONS



B. DUPLICATONS



C. INVERSIONS

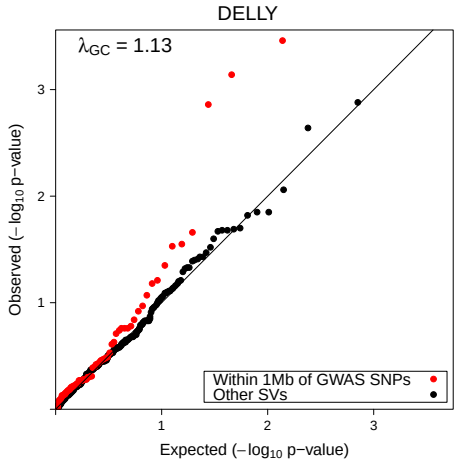
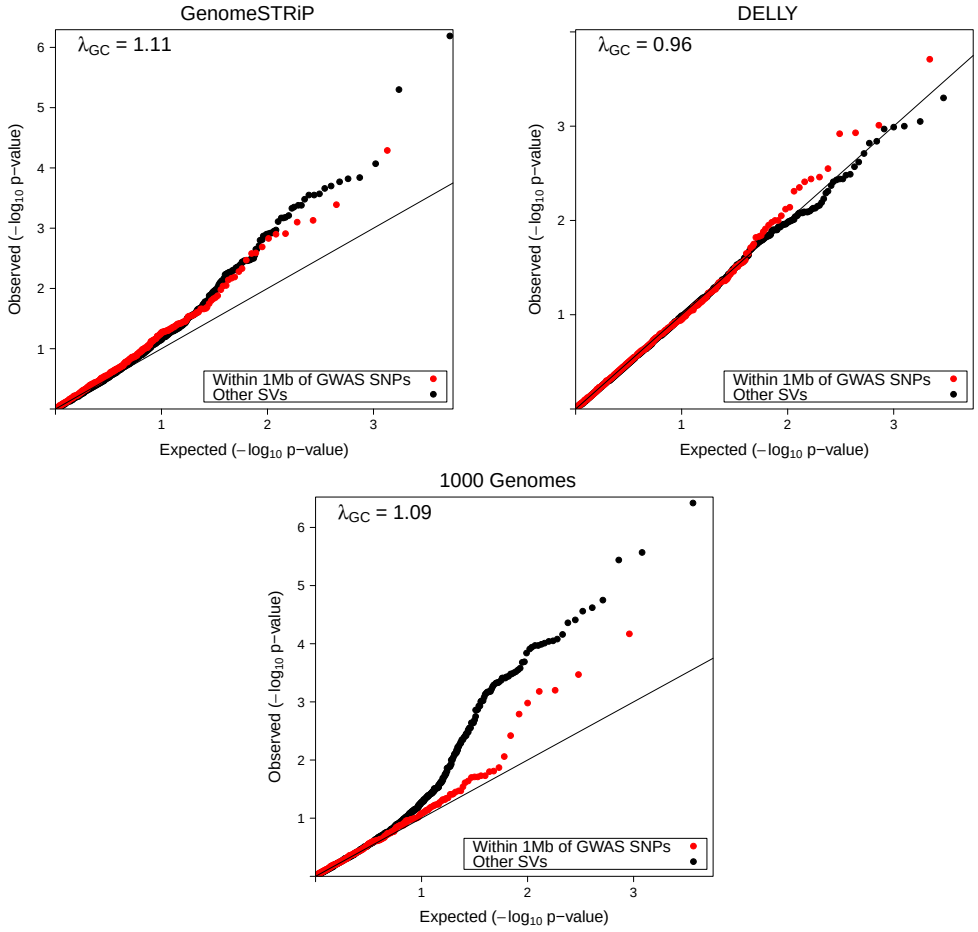
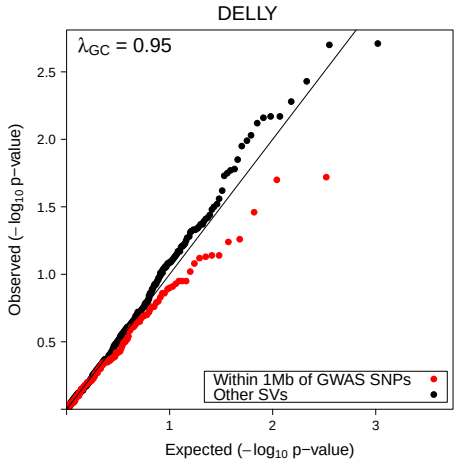


Figure 4.10: SV association results for LDL cholesterol. Linear regression of HUNT genotypes with LDL cholesterol was carried out separately for **(A)** deletions discovered by GenomeSTRiP, DELLY, and the 1000 Genomes Project; **(B)** duplications discovered by DELLY; and **(C)** inversions discovered by DELLY. QQplots show association P-values for structural variants within 1Mb (red) and outside 1Mb (black) of known CAD- or lipid-associated GWAS SNPs.

A. DELETIONS



B. DUPLICATONS



C. INVERSIONS

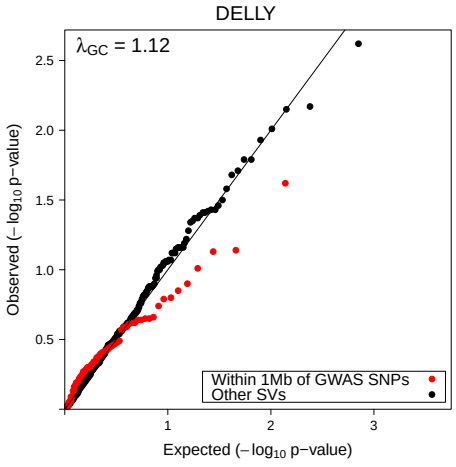
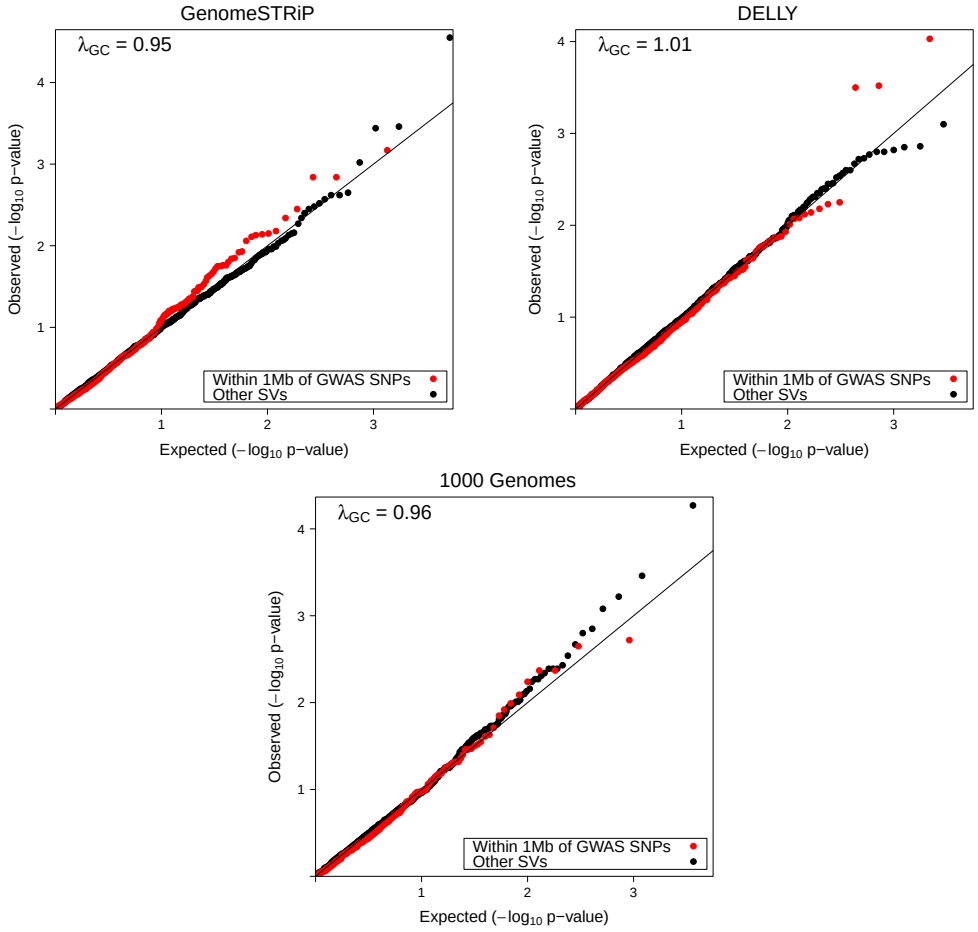
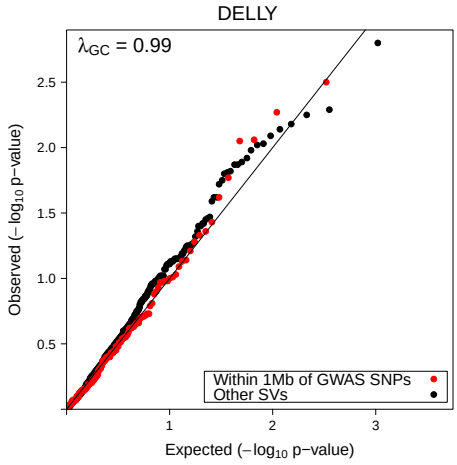


Figure 4.11: SV association results for HDL cholesterol. Linear regression of HUNT genotypes with HDL cholesterol was carried out separately for **(A)** deletions discovered by GenomeSTRiP, DELLY, and the 1000 Genomes Project; **(B)** duplications discovered by DELLY; and **(C)** inversions discovered by DELLY. QQplots show association P-values for structural variants within 1Mb (red) and outside 1Mb (black) of known CAD- or lipid-associated GWAS SNPs.

A. DELETIONS



B. DUPLICATONS



C. INVERSIONS

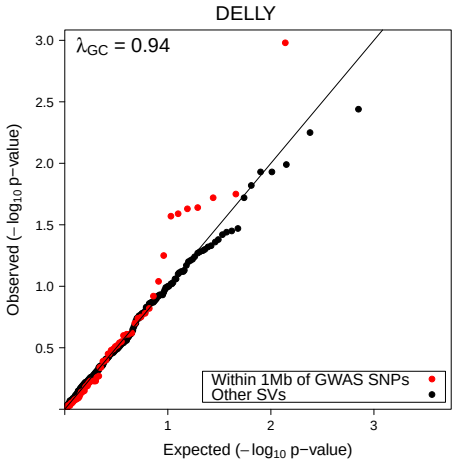
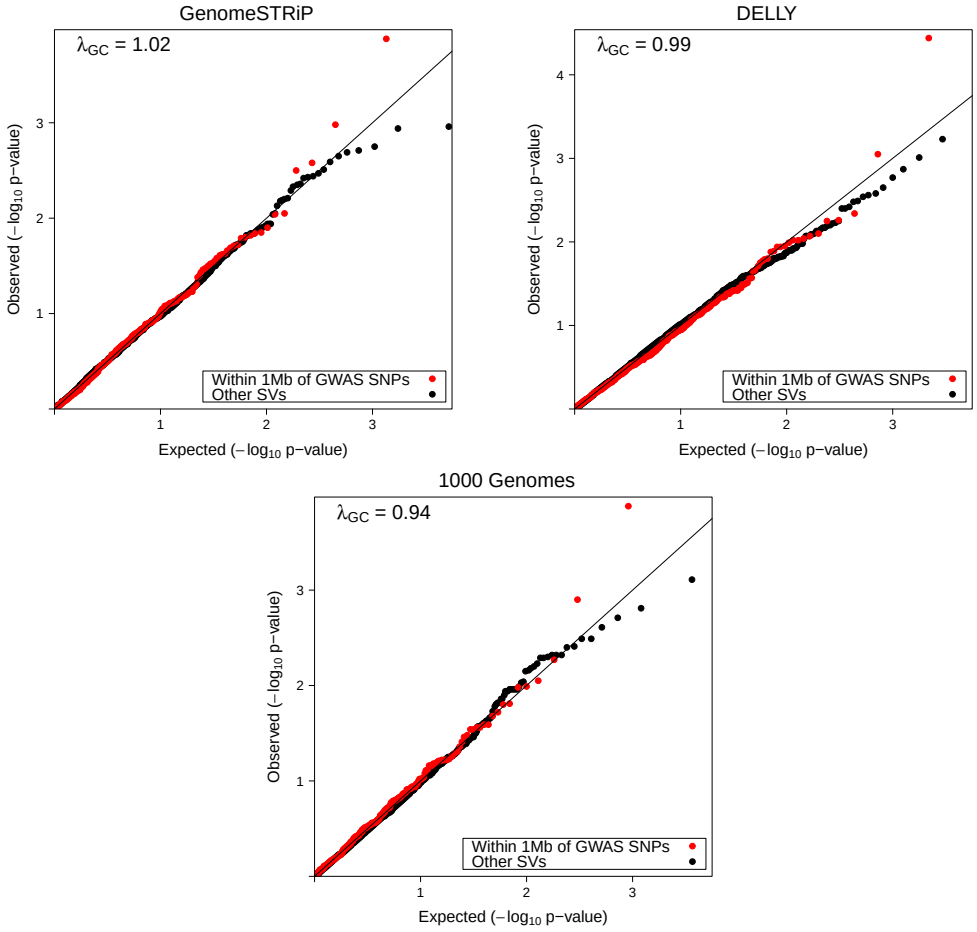
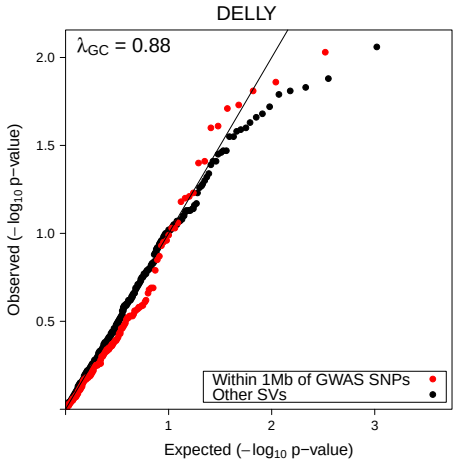


Figure 4.12: SV association results for triglycerides. Linear regression of HUNT genotypes with triglycerides was carried out separately for **(A)** deletions discovered by GenomeSTRiP, DELLY, and the 1000 Genomes Project; **(B)** duplications discovered by DELLY; and **(C)** inversions discovered by DELLY. QQplots show association P-values for structural variants within 1Mb (red) and outside 1Mb (black) of known CAD- or lipid-associated GWAS SNPs.

A. DELETIONS



B. DUPLICATONS



C. INVERSIONS

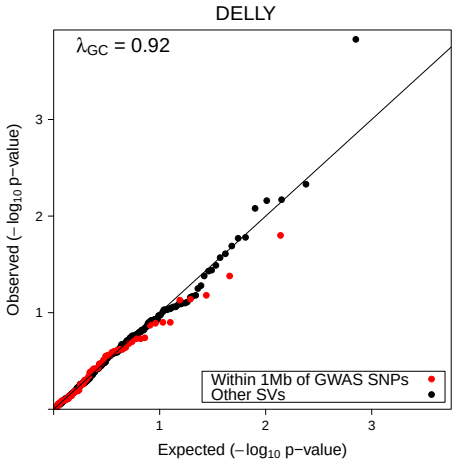
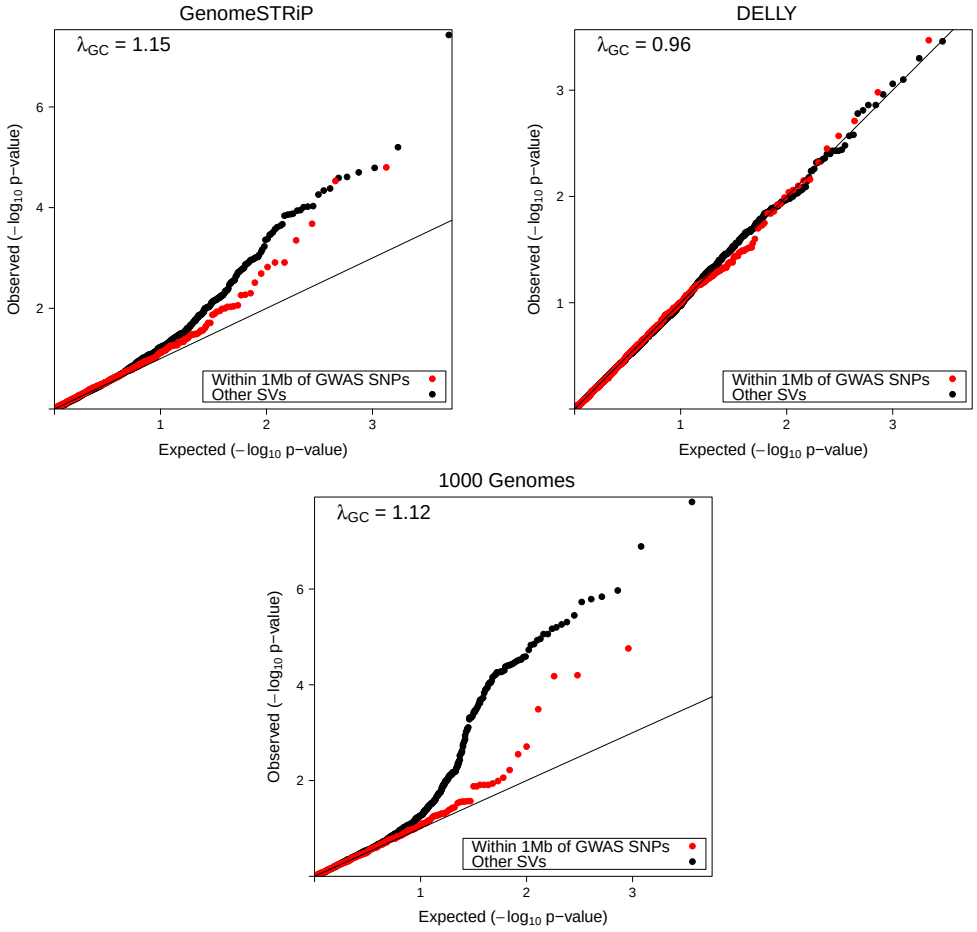


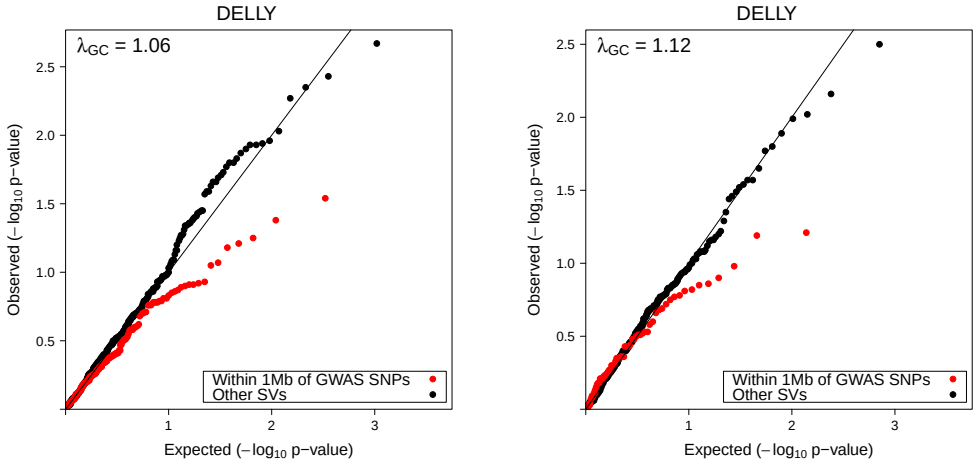
Figure 4.13: SV association results for total cholesterol. Linear regression of HUNT genotypes with total cholesterol was carried out separately for **(A)** deletions discovered by GenomeSTRiP, DELLY, and the 1000 Genomes Project; **(B)** duplications discovered by DELLY; and **(C)** inversions discovered by DELLY. QQplots show association P-values for structural variants within 1Mb (red) and outside 1Mb (black) of known CAD- or lipid-associated GWAS SNPs.

A. DELETIONS



B. DUPLICATONS

C. INVERSIONS



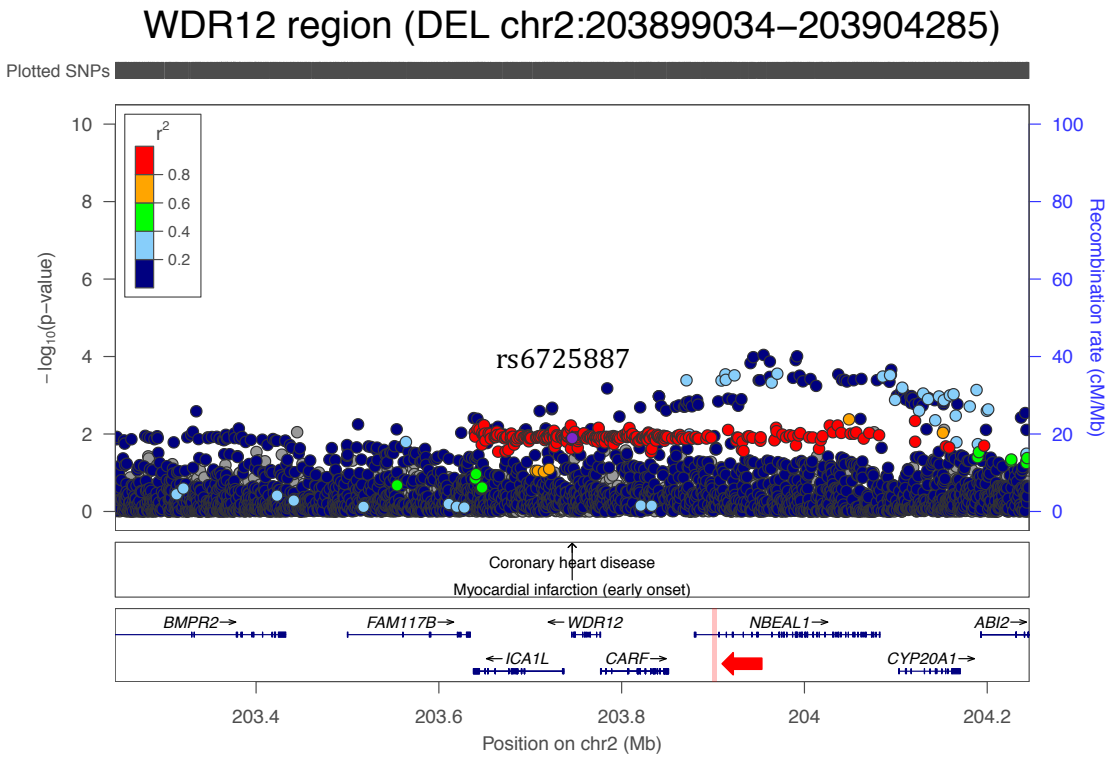


Figure 4.14: HUNT single variant MI association results at the *WDR12* locus. There is a 5,251 base-pair deletion (highlighted in red) in linkage disequilibrium (r^2=0.98) with the previously reported MI-associated index SNP rs6725887 (colored in purple).

Table 4.1: Phenotype descriptive statistics for HUNT sequenced samples

Descriptive statistics for sex					
		Male	Female	Overall	
Sex	Case	756	345	1101	
	Control	756	345	1101	
	Overall	1512	690	2202	
Descriptive statistics for age[a] and quantitative lipid measurements[b]					
		Min	Mean	Median	Max
Age at participation (years)	Case	20.00	54.55	54.00	82.00
	Control	20.00	53.66	53.00	81.00
	Overall	20.00	54.11	54.00	82.00
LDL cholesterol (mg/dL)	Case	21.10	127.30	122.10	345.70
	Control	43.32	148.20	145.90	293.90
	Overall	21.1	138.0	135.4	345.7
HDL cholesterol (mg/dL)	Case	19.3	45.9	42.5	104.2
	Control	23.20	52.72	50.20	115.80
	Overall	19.30	49.31	46.30	115.80
Triglycerides (mg/dL)	Case	35.4	186.4	159.3	1443.0
	Control	35.4	149.3	129.6	902.7
	Overall	35.4	167.6	141.6	1443.0
Total cholesterol (mg/dL)	Case	81.1	209.5	204.6	432.4
	Control	123.6	230.2	227.8	382.2
	Overall	81.1	219.9	216.2	432.4

[a]Age from HUNT2; HUNT3 age was used when HUNT2 age was not available.
[b]Lipid measurements were averaged from HUNT2 and HUNT3;
HUNT3 measurements were used when HUNT2 data were not available.

Table 4.2: Sample sizes for association analysis by trait

Trait	Samples that pass QC	Non-missing phenotypes	Total samples in analysis
MI	2195	2202	2195
LDL-C	2195	2134	2127
HDL-C	2195	2201	2194
TC	2195	2201	2194
TG	2195	2175	2168

Table 4.3: Structural variant counts and size distributions

	SV Type	SV Size	Size Distribution				SV Count
			Min	Median	Mean	Max	
1000 Genomes Project	DEL (<CN0>)	All	204	1933	6345	864000	3106
(EUR AF>0.01)		2kb-1Mb	2001	4643	12020	864000	1502
	DUP (<CN2>)	All	3070	16080	36790	264200	80
		2kb-1Mb	3070	16080	36790	264200	80
	INV (<INV>)	All	281	1191	5772	102600	28
		2kb-1Mb	4120	7486	27890	102600	5
DELLY	DEL	All	160	628	101400	130400000	5564
(HUNT AF>0.01)		2kb-1Mb	2002	4242	16050	957800	1209
	DUP	All	283	957	932200	223500000	723
		2kb-1Mb	2028	12990	77580	997200	218
	INV	All	156	2109	2437000	222200000	493
		2kb-1Mb	2018	31120	128500	957100	183
GenomeSTRiP	DEL	All	202	505	2668	182500	3270
(HUNT AF>0.01)		2kb-1Mb	2003	4719	8333	182500	885

*DEL: deletions; DUP: duplications; INV: inversions; CN0: copy number 0; CN2: copy number 2; AF: allele frequency

Table 4.4: CNVs tagging trait-associated SNPs in Conrad et al. (2010), HUNT and 1000 Genomes Project samples

			Conrad et al. (2010)*			HUNT (DELLY)		1000G Phase3 v5		HUNT (GenomeSTRiP)	
Trait	Reported Gene	SNP	CNV	r^2	Population	CNV	r^2	CNV	r^2	CNV	r^2
Multiple sclerosis	*KIF1B*	rs10492972	1:10482550-10483507	0.92	CEU						
Height	*CATSPER4*	rs11809207	1:26459570-26464632	0.61	CEU	1:26459942-26464932	0.10	1:26460095-26464769	0.57		
Body mass index	*NEGR1*	rs2815752	1:72766282-72811969	0.96	CEU	1:72766156-72812111	0.28	1:72766343-72811815	0.00		
Smoking behavior	*AK002179*	rs7553864	1:87613239-87614258	0.76	CEU						
Psoriasis	*LCE3D,LCE3A*	rs4085613	1:152555610-152590091	0.97	CEU	1:152555535-152587820	0.001	1:152555495-152586932	0.99		
C-reactive protein	*CRP*	rs11265260	1:159648762-159649629	0.62	CHB+JPT	1:159648518-159649798	0.02	1:159648708-159649659	0.08		
QT interval	*NOS1AP*	rs12029454	1:162230745-162231222	0.57	CHB+JPT					1:162230746-162231279	0.12
Myocardial infaction	*WDR12*	rs6725887	2:203899521-203903877	1.00	CEU	2:203898933-203904481	0.17	2:203899045-203904284	0.98	2:203899034-203904285	0.98
Prostate cancer	*CTDSPL*	rs9311171	3:37978470-37986876	1.00	CHB+JPT	3:37978271-37987003	0.14	3:37978345-37986932	0.50	3:37978418-37986927	0.50
Ageing traits	*KCNAB1*	rs3772255	3:156092052-156093564	0.90	CEU	3:156091996-156093857	0.30	3:156092162-156093688	0.90		
Bone mineral density	NR	rs9291683	4:10174154-10234566	0.51	YRI						
Bone mineral density	NR	rs9291683	4:10211321-10234566	0.51	YRI	4:10211066-10234773	0.04	4:10211223-10234580	0.12		
Lung cancer	*CLPTM1L*	rs401681	5:1333043-1333897	0.68	YRI						
Crohn's disease	*IRGM*	rs11747270	5:150177643-150181585	1.00	CEU	5:150177568-150181690	0.21	5:150177636-150181601	0.93	5:150177661-150181600	0.93
Crohn's disease	*IRGM*	rs11747270	5:150203369-150223430	1.00	CEU	5:150203087-150223453	0.20	5:150203162-150223269	1.00	5:150203162-150223263	1.00
Multiple sclerosis	*SGCD*	rs4704970	5:155476656-155495022	0.95	CEU						
Psoriasis	*HLA-C*	rs12191877	6:31276526-31289437	0.79	CEU						
AIDS progression	*HLA-C*	rs10484554	6:31276526-31289437	0.79	CEU						
Multiple sclerosis	*HLA-DRB1*	rs3129934	6:32411907-32779836	0.87	CEU						
Hepatitis B	*HLA-DPB1*	rs9277535	6:33048360-33054740	0.62	CEU						
Hepatitis B	*HLA-DPB1*	rs9277535	6:33051704-33055345	0.67	CEU						
Testicular germ cell tumor	*BAK1*	rs210138	6:33583939-33585879	0.55	CEU						
Crohn's disease	*CCR6*	rs2301436	6:167488131-167489148	0.71	CEU						
Neuroticism	*AK127771*	rs2705293	8:138911640-138912197	0.51	YRI						
Schizophrenia	Intergenic	rs1602565	11:29139377-29140406	0.64	CEU	11:29139247-29140526	0.08	11:29139370-29140327	0.26		
Schizophrenia	Intergenic	rs1602565	11:29139538-29140067	0.61	CEU						
HDL cholesterol	*MADD,FOLH1*	rs7395662	11:48600856-48604301	1.00	CEU	11:48600647-48604511	0.31	11:48600857-48604286	0.96		
Cognitive test performance	Intergenic	rs9300212	12:33715129-33716915	0.84	CEU	12:33714966-33717168	0.32	12:33715107-33716977	0.85	12:33715120-33716969	0.85
Type 2 diabetes	NR	rs1495377	12:71532675-71533665	0.72	CEU						
Height	*DLEU7*	rs3118914	13:51069346-51075130	0.69	CEU	13:51069170-51075273	0.22	13:51069347-51075082	0.99		
Height	*RAB40C*	rs763014	16:661067-663587	0.68	CEU						
QT interval	*LITAF*	rs8049607	16:11684037-11684551	0.88	CHB+JPT						
QT interval	*NDRG4*	rs7188697	16:58673606-58676357	0.61	YRI						
Skin sensitivity to sun	*MC1R*	rs1805007	16:89896098-89898402	0.87	CEU	16:89895927-89898597	0.27	16:89896054-89898405	0.91	16:89895909-89898438	0.73

*CNVs from Table 2 of Conrad et al. (2010); Start and End CNV positions are hg19; NR: no gene reported in original study

Table 4.5: Top significant association results for deletions

Locus	Chr	Start	End	Trait	OR/Effect Size	*P*-value	Source
OR5P2	11	7790133	7792078	MI	0.7857	1.49×10^{-4}	DELLY
AGMO	7	15229569	15230065	MI	0.7224	3.50×10^{-4}	DELLY
UBE2NL	X	143406811	143410134	MI	0.7496	6.18×10^{-4}	DELLY
PFDN4	20	52866065	52867491	MI	0.3224	7.79×10^{-4}	DELLY
WDR47	1	109573112	109575345	MI	0.7800	1.21×10^{-3}	DELLY
CDH6	5	30915523	30916059	MI	0.7668	2.16×10^{-4}	GenomeSTRiP
LONP2	16	48311331	48311679	MI	0.7167	2.30×10^{-4}	GenomeSTRiP
MIR548AE2	14	30817195	30819456	MI	1.9680	9.93×10^{-4}	GenomeSTRiP
G2E3	7	52963199	52964915	MI	0.6465	1.69×10^{-3}	GenomeSTRiP
POM121L12	9	22653155	22653781	MI	0.5957	2.21×10^{-3}	GenomeSTRiP
LINC01239	11	5760086	5762366	MI	0.7813	1.02×10^{-4}	1000G
OR56B1	5	30915526	30916063	MI	0.7702	2.69×10^{-4}	1000G
CDH6	14	30817188	30819455	MI	1.9110	1.50×10^{-3}	1000G
G2E3	1	188539457	188540228	MI	0.7600	2.05×10^{-3}	1000G
BRINP3	19	23465135	23466785	MI	1.2360	2.63×10^{-3}	1000G
IPO5P1	5	86245887	86247320	LDL-C	-0.25	1.19×10^{-4}	DELLY
MIR4280	9	19671646	19673613	LDL-C	-0.26	1.97×10^{-4}	DELLY
SLC24A2	X	124115599	124116298	LDL-C	-0.32	4.98×10^{-4}	DELLY
TENM1	13	85929413	85929777	LDL-C	-0.27	8.91×10^{-4}	DELLY
LINC00351	11	51340884	51363219	LDL-C	-0.09	9.81×10^{-4}	DELLY
OR4A5	4	92472837	92478295	LDL-C	0.64	6.41×10^{-7}	GenomeSTRiP
CCSER1	10	42454653	42457078	LDL-C	0.44	4.97×10^{-6}	GenomeSTRiP
LOC441666	10	91998535	92002030	LDL-C	0.52	5.14×10^{-5}	GenomeSTRiP
RP11-15K3	5	45192227	45194396	LDL-C	0.56	8.49×10^{-5}	GenomeSTRiP
HCN1	12	38335811	38341301	LDL-C	0.53	1.44×10^{-4}	GenomeSTRiP
ALG10B	4	92472835	92478331	LDL-C	0.66	3.78×10^{-7}	1000G
MLLT3	9	20380045	20386726	LDL-C	0.68	2.66×10^{-6}	1000G
MIR1973	4	117499720	117503035	LDL-C	0.62	3.63×10^{-6}	1000G
HCN1	5	46217822	46219362	LDL-C	0.46	1.77×10^{-5}	1000G
RAB9BP1	5	104949722	104955603	HDL-C	0.59	2.42×10^{-5}	1000G
UGT2B7	4	69959686	69960320	HDL-C	-0.15	9.35×10^{-5}	DELLY
TMPRSS11E	4	69373294	69491273	HDL-C	-0.23	3.04×10^{-4}	DELLY
PPP2R5E	14	63887380	63887889	HDL-C	-0.26	3.17×10^{-4}	DELLY
KIRREL3-AS3	11	127456167	127456578	HDL-C	-0.19	6.08×10^{-4}	DELLY
IL1RAPL1	X	29821144	29821508	HDL-C	0.38	7.88×10^{-4}	DELLY
TMEM248	7	66397601	66397942	HDL-C	0.47	2.84×10^{-5}	GenomeSTRiP
MIR5011	18	64959011	64967266	HDL-C	-0.32	3.48×10^{-4}	GenomeSTRiP
P2RY8	X	1667032	1667328	HDL-C	0.23	3.60×10^{-4}	GenomeSTRiP
SCARB1	12	125334384	125334699	HDL-C	-0.08	6.74×10^{-4}	GenomeSTRiP
MYO9A	15	72126376	72126737	HDL-C	0.30	9.54×10^{-4}	GenomeSTRiP
ZBTB11	3	101357527	101359730	HDL-C	-0.13	5.41×10^{-5}	1000G
MIR5011	18	64958984	64967256	HDL-C	-0.32	3.48×10^{-4}	1000G
LOC101928401	7	56652469	56658519	HDL-C	0.09	6.09×10^{-4}	1000G
KIAA1257	3	128672140	128675334	HDL-C	-0.40	8.30×10^{-4}	1000G
SOX5	12	23939446	23940212	HDL-C	0.13	1.43×10^{-3}	1000G
TRIM49B	11	48944428	48945723	TG	-0.12	3.63×10^{-5}	DELLY
TPO	2	1533816	1536862	TG	-0.10	2.70×10^{-4}	DELLY
TSHR	14	81596902	81597479	TG	0.33	5.93×10^{-4}	DELLY
TRIM48	11	54869785	54871917	TG	-0.22	8.97×10^{-4}	DELLY
MIR5007	13	55634075	55634428	TG	-0.25	9.66×10^{-4}	DELLY
ARHGEF18	19	7451063	7454115	TG	-0.47	1.33×10^{-4}	GenomeSTRiP
SLC35F3	1	234318643	234319750	TG	0.14	1.06×10^{-3}	GenomeSTRiP
GPC5	13	93000122	93000324	TG	0.13	1.11×10^{-3}	GenomeSTRiP
PLOD2	3	145644968	145649191	TG	0.09	1.14×10^{-3}	GenomeSTRiP
CACNA1A	19	13301205	13301539	TG	-0.12	1.77×10^{-3}	GenomeSTRiP
ARHGEF18	19	7451039	7454104	TG	-0.47	1.33×10^{-4}	1000G
LOC647859	5	70368390	70391629	TG	-0.11	7.81×10^{-4}	1000G
SLC35F3	1	234318645	234319749	TG	0.14	1.25×10^{-3}	1000G
RASSF2	20	4807411	4808147	TG	-0.24	1.55×10^{-3}	1000G
PLOD2	3	145644966	145649185	TG	0.09	1.94×10^{-3}	1000G
OR4A5	11	51340884	51363219	TC	-0.09	3.40×10^{-4}	DELLY
MIR4280	5	86245887	86247320	TC	-0.22	3.42×10^{-4}	DELLY
TENM1	X	124115599	124116298	TC	-0.31	3.49×10^{-4}	DELLY
LOC100129138	1	105667371	105668654	TC	0.18	4.98×10^{-4}	DELLY
AKAP6	14	32953126	32954531	TC	-0.23	7.87×10^{-4}	DELLY
CCSER1	4	92472837	92478295	TC	0.66	3.68×10^{-8}	GenomeSTRiP
ALG10	12	34709011	34715660	TC	0.36	6.34×10^{-6}	GenomeSTRiP
RP11-15K3	10	91998535	92002030	TC	0.51	1.59×10^{-5}	GenomeSTRiP
LOC441666	10	42454653	42457078	TC	0.39	1.63×10^{-5}	GenomeSTRiP
ALG10	12	34685528	34696065	TC	0.37	1.97×10^{-5}	GenomeSTRiP
CCSER1	4	92472835	92478331	TC	0.68	1.51×10^{-8}	1000G
MLLT3	9	20380045	20386726	TC	0.71	1.30×10^{-7}	1000G
RAB9BP1	5	104949722	104955603	TC	0.63	1.08×10^{-6}	1000G
HCN1	5	46217822	46219362	TC	0.48	1.44×10^{-6}	1000G
LINC00293	8	47216305	47233060	TC	0.56	1.62×10^{-6}	1000G

Table 4.6: Top significant association results for duplications

Locus	Chr	Start	End	Trait	OR/Effect Size	*P*-value	Source
GALNT9	12	132973788	132974587	MI	1.6490	2.27×10^{-3}	DELLY
GALNT9	12	132973707	132974587	MI	1.5790	4.44×10^{-3}	DELLY
PDX1	13	28508684	28509249	MI	1.4260	5.68×10^{-3}	DELLY
DNAH14	1	225133308	225248377	MI	0.6307	6.20×10^{-3}	DELLY
C9orf106	9	132158710	132159317	MI	1.2630	6.99×10^{-3}	DELLY
AKAP11	13	42947269	42947736	LDL-C	0.13	1.95×10^{-3}	DELLY
AKAP11	13	42947268	42947736	LDL-C	0.13	2.01×10^{-3}	DELLY
EPS8L2	11	711724	712153	LDL-C	-0.41	3.73×10^{-3}	DELLY
EPS8L2	11	711722	712153	LDL-C	-0.38	5.19×10^{-3}	DELLY
LINC00473	6	166253236	166253588	LDL-C	-0.27	6.69×10^{-3}	DELLY
COL5A1	9	137576794	137577726	HDL-C	0.10	1.59×10^{-3}	DELLY
HLA-DRB5	6	32469530	32540210	HDL-C	0.06	3.17×10^{-3}	DELLY
COL5A1	9	137576781	137577726	HDL-C	0.09	5.10×10^{-3}	DELLY
ADARB1	21	46575840	46576486	HDL-C	0.07	5.39×10^{-3}	DELLY
CNTNAP5	2	125766561	125768484	HDL-C	0.08	5.65×10^{-3}	DELLY
SORBS1	10	97206785	97208113	TG	-0.09	8.78×10^{-3}	DELLY
TCEA3	1	23718825	23719261	TG	-0.21	9.28×10^{-3}	DELLY
PITRM1	10	3356380	3357458	TG	-0.21	1.31×10^{-2}	DELLY
KLK7	19	51484077	51484767	TG	-0.09	1.39×10^{-2}	DELLY
EPS8L2	11	711724	712153	TG	-0.31	1.48×10^{-2}	DELLY
EPS8L2	11	711724	712153	TC	-0.41	2.15×10^{-3}	DELLY
C3orf38	3	88547135	88547974	TC	-0.23	3.68×10^{-3}	DELLY
LMCD1	3	8601108	8601876	TC	0.18	4.43×10^{-3}	DELLY
EPS8L2	11	711722	712153	TC	-0.36	5.43×10^{-3}	DELLY
LOC644172	17	43655785	44366773	TC	0.20	9.27×10^{-3}	DELLY

Table 4.7: Top significant association results for inversions

Locus	Chr	Start	End	Trait	OR/Effect Size	*P*-value	Source
DPM3	1	155119792	155120021	MI	0.4367	3.47×10^{-4}	DELLY
DPM3	1	155119740	155120021	MI	0.4374	7.16×10^{-4}	DELLY
CCDC129	7	31586877	31590353	MI	1.3310	1.33×10^{-3}	DELLY
DPM3	1	155119793	155120021	MI	0.5049	1.38×10^{-3}	DELLY
ZNF626	19	20801073	20884244	MI	0.8509	2.27×10^{-3}	DELLY
LCMT1	16	25204035	25204734	LDL-C	-0.26	2.41×10^{-3}	DELLY
SLC25A51P1	6	67492324	67492651	LDL-C	-0.26	6.69×10^{-3}	DELLY
RASGRP3	2	33764621	33768033	LDL-C	-0.15	7.13×10^{-3}	DELLY
TPTE2P6	13	25154598	25542722	LDL-C	0.07	9.87×10^{-3}	DELLY
PLEKHB2	2	131886576	131887696	LDL-C	0.10	1.17×10^{-2}	DELLY
KLC2	11	66019004	66020102	HDL-C	0.10	1.04×10^{-3}	DELLY
SFTPA2	10	81316454	81374513	HDL-C	0.12	3.63×10^{-3}	DELLY
MIR3924	10	59257657	59258202	HDL-C	0.05	5.66×10^{-3}	DELLY
AUTS2	7	70420815	70438968	HDL-C	0.05	1.02×10^{-2}	DELLY
RAD51B	14	68907825	69298024	HDL-C	0.16	1.16×10^{-2}	DELLY
NCKAP5L	12	50182514	50183034	TG	0.12	1.49×10^{-4}	DELLY
PRSS35	6	84207259	84610844	TG	-0.28	4.72×10^{-3}	DELLY
ASIC2	17	31683859	31684175	TG	-0.27	6.83×10^{-3}	DELLY
ARID1B	6	157559436	157641398	TG	0.13	6.91×10^{-3}	DELLY
MGLL	3	127496284	127497985	TG	-0.15	8.24×10^{-3}	DELLY
RASGRP3	2	33764621	33768033	TC	-0.15	3.18×10^{-3}	DELLY
MYCN	2	16406391	16407874	TC	-0.10	6.90×10^{-3}	DELLY
LOC644172	17	43663171	44338245	TC	0.15	9.49×10^{-3}	DELLY
SLC25A51P1	6	67492324	67492651	TC	-0.23	1.01×10^{-2}	DELLY
LCMT1	16	25204035	25204734	TC	-0.20	1.28×10^{-2}	DELLY

CHAPTER V

Discussion

5.1 Results Summary

Our collective knowledge of the role of human genetic variation in complex disease has come a long way since the first published genome-wide association studies. As a research community, we've catalogued over 150 common variants and at least 25 loci containing rare variants that influence lipid variability in humans. We've made advances in identifying functional variants at associated loci and recognized the regulatory importance of noncoding variation. In addition, we've been able to leverage genetic tools to answer questions about the clinical implications of lipid-associated variants. Together, these insights provide the groundwork for individualized treatment, diagnosis, and prevention of heart disease. Through this dissertation research, I have advanced our understanding of lipid genetics and developed a tool that has expanded our knowledge of the biological mechanisms underlying noncoding variation associated with lipids as well as other complex traits.

In the manuscript Global Lipids Genetics Consortium et al. (2013), a follow-up study of 100,000 individuals genotyped on Metabochip, we discovered 62 novel genetic loci associated with lipids to contribute to the existing list of known associated loci. Chapter II described the discovery of these loci and several downstream analy-

ses including pathway analyses, investigation of regulation of mRNA expression, and literature review that support the roles of 38 of these loci in regulation of plasma lipids. The mechanistic role of the remaining loci is unknown, leaving considerable opportunity for functional insights from genetic studies in the coming years. Given the non-protein-coding role of so many lipid-associated variants reported by GWAS, I developed a tool in Chapter III to evaluate the enrichment of GWAS variants in tissue-specific chromatin states and regulatory features defined by bioinformatics techniques and new sequencing approaches such as ChIP-seq (Schmidt et al., 2015). Using a data-driven hypothesis, I selected particular variants at a set of five lipid loci as the potential functional variant, and reported experimental luciferase results to confirm my computational predictions. Lastly, in Chapter IV I performed discovery and genotyping of insertions, duplications, and inversions from low-pass whole genome sequencing of nearly 2,000 Norwegian MI-cases and controls. Although we did not have the power to detect significant genome-wide associations with structural variants identified in this dataset, I learned many technical and computational lessons including the importance of accurate sequencing library preparation for CNV calling, and generating an optimal SV analysis pipeline using complementary genotyping approaches.

5.2 Interpreting GWAS: promises and challenges

Despite the strides we have made in understanding lipid genetics, there are still shortcomings to traditional genome-wide association study designs and an incomplete knowledge of the biological mechanisms underlying GWAS-identified signals. Firstly, GWA studies are primarily designed for finding common trait-associated variation, but natural selection has reduced the frequency of high-risk variants in the human

population. Effect sizes of trait-associated variants discovered by GWAS are generally modest (*e.g.* odds ratio <1.5), conferring relatively small modulation in risk. In addition, GWAS variants only explain a fraction of the trait variability, leaving a large proportion of heritability unexplained (*e.g.* ∼90% unexplained heritability for coronary artery disease (CARDIoGRAMplusC4D Consortium et al., 2013)). Finally, some of the largest GWA studies to date typically investigate European-only populations, leaving complex trait genetics in non-Europeans less well understood.

Because rare variants are not captured well by GWAS with imputation, the role of rare variants in complex traits is still largely unknown. More comprehensive scans involving whole-genome or exome sequencing are promising for revealing rare risk variants that may explain more of the missing heritability. For early-onset diseases that are rare and highly-penetrant, the missing heritability will likely be found with extremely low frequency variants of high effect. Although SNP genotyping coupled with imputation is still more cost-effective today than whole-genome sequencing (Yang et al., 2015), rare variants can be difficult to impute. This makes a strong case for sequencing studies in diseases where rare variants are more likely to play a role.

For more common diseases such as CAD however, the remaining missing heritability will likely be found in common variants with small effects. Ongoing efforts in genotyping thousands of unrelated individuals on exome chip are revealing more coding variants with a role in modulation of lipid levels. Meta-analyses of large non-European populations are currently underway, leading us to new discoveries of population-specific variants associated with lipids that may not be significant in Europeans. Fine mapping of lipid-associated loci in ethnically diverse groups will be increasingly important to provide guidance toward identifying the causal variant.

Complex traits and diseases can have variable genetic architectures, making study design and results interpretation challenging. For example, it is possible that not all carriers of an associated risk variant will display manifestation of the trait or disease, suggesting genetic or environmental factors that confer resistance. In addition, some genetic variants may depend on pre-existing environmental contexts, resulting in context-dependent risk variants that don't pass genome wide significance. For heart disease in particular, these factors could include lifestyle elements such as smoking, diet, physical activity, or sudden high-stress events (Peters et al., 2014; Chan et al., 2013). Another caveat of a case-control GWAS design is that disease processes might be active in control individuals, but the clinical symptoms may have not yet manifested when they participate in the study. In this case, having a sufficiently large number of controls or re-evaluating and assigning control individuals at a later stage will improve the study.

An undisputed challenge in complex trait genetics is the interpretation of noncoding variation. In Chapter III, I presented a practical tool for researchers to investigate the biological mechanisms of GWAS signals for any phenotype and provide guidance toward prioritizing the functional variant using epigenomic features. However, there is still progress to be made in methods for refining the association signal to predict functional variants and in understanding the mechanisms by which they act. An improvement over prioritizing variants solely based on the number of overlapping regulatory features could be assigning variants a score based on their likelihood of being functional. This score could involve weighting by effect size or the presence of a nearby motif, or be analogous to the SVM classifier used for filtering variants from sequencing data. In addition, we can use Bayes theorem to determine the likelihood of disease-causing SNPs based on prior probability (Maller et al., 2012).

Integrating other sources of information including functional genomics, chromatin states, evolutionary conservation, and quantitative trait loci to link noncoding sequence with regulation will also be supportive. Burden testing of noncoding variants is another under-developed area that can shed new insight on transcriptional regulation. Without a doubt, the future progress in understanding the noncoding genetic variation implicated in complex disease will rely on the coupling of GWAS findings with cell-type specific sequencing-based functional genomics data.

There is still debate in the genetics community about the clinical implications of various plasma lipids. In particular, genetic studies describing the causal role of triglycerides and HDL cholesterol in heart disease risk have contradicted previous assumptions. In a companion paper published with the results from Global Lipids Genetics Consortium et al. (2013), causality was established between triglycerides and coronary artery disease risk through correlation of effect sizes of trait-associated SNPs (Do et al., 2013). In addition, the study presented by Voight et al. (2012b) concluded no relationship between HDL-C and risk of heart attack. These two important papers have drawn considerable attention from the medical community, and give direction and/or caution to physicians when considering triglycerides or HDL-C in disease risk. Advances over traditional approaches of Mendelian randomization that address the pleiotropy complicating these variants have helped support the causality of triglycerides on heart disease risk, and can perhaps shed light on the relationships of other traits and disorders in the future (Burgess and Thompson, 2015).

Given our limited understanding of pharmacogenetics and the effect that individual genotypic variation plays in drug response, studies of complex trait genetics are extremely relevant. An example illustrating the importance of improved pharmacogenetics understanding is the impaired ability of carriers of *CYP2C19* variants

to metabolize the drug clopidogrel, increasing their risk for heart disease (Kaufman et al., 2015).

5.3 Data integration and bioinformatics challenges

As we progress in our understanding of complex traits, a major theme emerging is data integration. Investigators are increasingly collaborating to build large repositories of high-throughput genomics and epigenomics data for public use. This openness encourages integrative analyses and presents more creative ways to address hypotheses. As illustrated in Chapter III, combining GWAS findings with ENCODE epignomic data can lead to mechanistic insights. Another example of a data-rich repository is the Genotype-Tissue Expression (GTEx) Portal, which contains normalized expression matrices from RNA-seq in a wide range of human tissues (GTEx Consortium, 2015). These data can help answer ongoing questions about the non-coding variation that likely acts through regulation of gene expression. The systems genetics approach, or Genome Wide Network Study as coined by Björkegren et al. (2015), puts emphasis on combining data from intermediate phenotypes such as RNA, proteins, metabolites, and epigenetics in multiple disease-relevant tissues. Together, the data generated by these and many other ongoing efforts will surely help fill in some of the missing knowledge concerning the biological mechanisms underlying trait-associated genetic variation.

The meta-analysis performed in Chapter II is one of many ongoing and future collaborations that will rely heavily on data sharing. As scientists increasingly share their data in public domains, there is a need for more consistent standards in data formats and metadata annotation. Future genetic studies involving common complex diseases that rely on large sample sizes will be particularly affected by the data

sharing and dissemination practices of the larger scientific community. In addition, large-scale studies of human genetic variation require voluntary cooperation from the general public. Efforts such as Genes for Good (`http://genesforgood.sph.umich.edu/`) utilize social media to collect individual genetic data and educate people about the personal benefits and larger scientific contributions resulting from their participation. Data collected in Norway through The HUNT Study (Krokstad et al., 2013), the source from which data were used in Chapter IV, is an excellent model of an extensive collection of volunteer-based personal health data. Health informatics will be increasingly important for optimizing how these large volumes of biomedical data are managed, stored, shared, and interpreted.

On January 30, 2015, President Obama announced an initiative to transform healthcare into the era of big data and personalized medicine. The NIH Precision Medicine Initiative aims to utilize individual risk factors including genetic variability to develop treatment that is tailored to specific patients. The increasing ubiquity of mobile devices that record health-related measures such as heart rate, calorie consumption, and physical activity is revolutionizing the way we can monitor health, reduce risk, and treat disease based on individual lifestyle. Research involving complex trait genetics such as the work presented in this dissertation provides primary foundational knowledge for facilitating the translation from 'bench-to-bedside' and fulfilling this vision of personalized medicine.

BIBLIOGRAPHY

1000 Genomes Project Consortium, Abecasis, G. R., Altshuler, D., Auton, A., Brooks, L. D., Durbin, R. M., Gibbs, R. A., Hurles, M. E. and McVean, G. A. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–73, 2010. 24, 28, 85

1000 Genomes Project Consortium, Abecasis, G. R., Auton, A., Brooks, L. D., DePristo, M. A., Durbin, R. M., Handsaker, R. E., Kang, H. M., Marth, G. T. and McVean, G. A. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, 2012. 13, 135, 139

Alkan, C., Coe, B. P. and Eichler, E. E. Genome structural variation discovery and genotyping. *Nat Rev Genet*, 12(5):363–376, 2011. 12, 14

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M. and Sherlock, G. Gene ontology: tool for the unification of biology. *Nat Genet*, 25(1):25–29, 2000. 26

Asimit, J. L., Day-Williams, A. G., Morris, A. P. and Zeggini, E. ARIEL and AMELIA: Testing for an accumulation of rare variants using next-generation sequencing data. *Human Heredity*, 73(2):84–94, 2012. 4

Asselbergs, F. W., Guo, Y., van Iperen, E. P. A., Sivapalaratnam, S., Tragante, V., Lanktree, M. B., Lange, L. A., Almoguera, B., Appelman, Y. E., Barnard, J., Baumert, J., Beitelshees, A. L., Bhangale, T. R., Chen, Y.-D. I., Gaunt, T. R., Gong, Y., Hopewell, J. C., Johnson, T., Kleber, M. E., Langaee, T. Y., Li, M., Li, Y. R., Liu, K., McDonough, C. W., Meijs, M. F. L., Middelberg, R. P. S., Musunuru, K., Nelson, C. P., O'Connell, J. R., Padmanabhan, S., Pankow, J. S., Pankratz, N., Rafelt, S., Rajagopalan, R., Romaine, S. P. R., Schork, N. J., Shaffer, J., Shen, H., Smith, E. N., Tischfield, S. E., van der Most, P. J., van Vliet-Ostaptchouk, J. V., Verweij, N., Volcik, K. A., Zhang, L., Bailey, K. R., Bailey, K. M., Bauer, F., Boer, J. M. A., Braund, P. S., Burt, A., Burton, P. R., Buxbaum, S. G., Chen, W., Cooper-DeHoff, R. M., Cupples, L. A., deJong, J. S., Delles, C., Duggan, D., Fornage, M., Furlong, C. E., Glazer, N., Gums, J. G., Hastie, C., Holmes, M. V., Illig, T., Kirkland, S. A., Kivimaki, M., Klein, R., Klein, B. E., Kooperberg, C., Kottke-Marchant, K., Kumari, M., LaCroix, A. Z., Mallela, L., Murugesan, G., Ordovas, J., Ouwehand, W. H., Post, W. S., Saxena, R., Scharnagl, H., Schreiner, P. J., Shah, T., Shields, D. C., Shimbo, D., Srinivasan, S. R., Stolk, R. P., Swerdlow, D. I., Taylor Jr., H. A., Topol, E. J., Toskala, E., van Pelt, J. L., van Setten, J., Yusuf, S., Whittaker, J. C., Zwinderman, A. H., Anand, S. S., Balmforth, A. J., Berenson, G. S., Bezzina, C. R., Boehm, B. O., Boerwinkle, E., Casas, J. P., Caulfield, M. J., Clarke, R., Connell, J. M., Cruickshanks, K. J., Davidson, K. W., Day, I. N. M., de Bakker, P. I. W., Doevendans, P. A., Dominiczak, A. F., Hall, A. S., Hartman, C. A., Hengstenberg, C., Hillege, H. L., Hofker, M. H., Humphries,

S. E., Jarvik, G. P., Johnson, J. A., Kaess, B. M., Kathiresan, S., Koenig, W., Lawlor, D. A., März, W., Melander, O., Mitchell, B. D., Montgomery, G. W., Munroe, P. B., Murray, S. S., Newhouse, S. J., Onland-Moret, N. C., Poulter, N., Psaty, B., Redline, S., Rich, S. S., Rotter, J. I., Schunkert, H., Sever, P., Shuldiner, A. R., Silverstein, R. L., Stanton, A., Thorand, B., Trip, M. D., Tsai, M. Y., van der Harst, P., van der Schoot, E., van der Schouw, Y. T., Verschuren, W. M. M., Watkins, H., Wilde, A. A. M., Wolffenbuttel, B. H. R., Whitfield, J. B., Hovingh, G. K., Ballantyne, C. M., Wijmenga, C., Reilly, M. P., Martin, N. G., Wilson, J. G., Rader, D. J., Samani, N. J., Reiner, A. P., Hegele, R. A., Kastelein, J. J. P., Hingorani, A. D., Talmud, P. J., Hakonarson, H., Elbers, C. C., Keating, B. J. and Drenos, F. Large-scale gene-centric meta-analysis across 32 studies identifies multiple lipid loci. *The American Journal of Human Genetics*, 91(5):823–838, 2012. 25

Baigent, C., Keech, A., Kearney, P., Blackwell, L., Buck, G., Pollicino, C., Kirby, A., Sourjina, T., Peto, R., Collins, R., Simes, R. and Cholesterol Treatment Trialists' (CTT) Collaborators. Efficacy and safety of cholesterol-lowering treatment: prospective meta-analysis of data from 90,056 participants in 14 randomised trials of statins. *The Lancet*, 366(9493):1267–1278, 2005. 10

Barter, P. J. and Rye, K. A. Cholesteryl ester transfer protein inhibition as a strategy to reduce cardiovascular risk. *J. Lipid Res.*, 53:1755–1766, 2012. 23

Björkegren, J. L. M., Kovacic, J. C., Dudley, J. T. and Schadt, E. E. Genome-wide significant loci: How important are they?: Systems genetics to understand heritability of coronary artery disease and other common complex disorders. *Journal of the American College of Cardiology*, 65(8):830–845, 2015. 167

Boyle, A. P., Hong, E. L., Hariharan, M., Cheng, Y., Schaub, M. A., Kasowski, M., Karczewski, K. J., Park, J., Hitz, B. C., Weng, S., Cherry, J. M. and Snyder, M. Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res*, 22(9):1790–7, 2012. 83, 93, 106

Burgess, S. and Thompson, S. G. Multivariable mendelian randomization: The use of pleiotropic genetic variants to estimate causal effects. *American Journal of Epidemiology*, 181(4):251–260, 2015. 166

Buyske, S., Wu, Y., Carty, C. L., Cheng, I., Assimes, T. L., Dumitrescu, L., Hindorff, L. A., Mitchell, S., Ambite, J. L., Boerwinkle, E., Buzkova, P., Carlson, C. S., Cochran, B., Duggan, D., Eaton, C. B., Fesinmeyer, M. D., Franceschini, N., Haessler, J., Jenny, N., Kang, H. M., Kooperberg, C., Lin, Y., Le Marchand, L., Matise, T. C., Robinson, J. G., Rodriguez, C., Schumacher, F. R., Voight, B. F., Young, A., Manolio, T. A., Mohlke, K. L., Haiman, C. A., Peters, U., Crawford, D. C. and North, K. E. Evaluation of the metabochip genotyping array in African Americans and implications for fine mapping of GWAS-identified loci: The PAGE study. *PLoS ONE*, 7(4):e35651, 2012. 31

CARDIoGRAMplusC4D Consortium, Deloukas, P., Kanoni, S., Willenborg, C., Farrall, M., Assimes, T. L., Thompson, J. R., Ingelsson, E., Saleheen, D., Erdmann, J., Goldstein, B. A., Stirrups, K., Konig, I. R., Cazier, J.-B., Johansson, A., Hall, A. S., Lee, J.-Y., Willer, C. J., Chambers, J. C., Esko, T., Folkersen, L., Goel, A., Grundberg, E., Havulinna, A. S., Ho, W. K., Hopewell, J. C., Eriksson, N., Kleber, M. E., Kristiansson, K., Lundmark, P., Lyytikainen, L.-P., Rafelt, S., Shungin, D., Strawbridge, R. J., Thorleifsson, G., Tikkanen, E., Van Zuydam, N., Voight, B. F., Waite, L. L., Zhang, W., Ziegler, A., Absher, D., Altshuler, D., Balmforth, A. J., Barroso, I., Braund, P. S., Burgdorf, C., Claudi-Boehm, S., Cox, D., Dimitriou, M., Do, R., Doney, A. S. F., Mokhtari, N. E., Eriksson, P., Fischer, K., Fontanillas, P., Franco-Cereceda, A., Gigante, B., Groop, L., Gustafsson, S., Hager, J., Hallmans, G., Han, B.-G., Hunt, S. E., Kang, H. M., Illig, T., Kessler, T., Knowles, J. W., Kolovou, G., Kuusisto, J., Langenberg, C., Langford, C., Leander, K., Lokki, M.-L., Lundmark, A., McCarthy, M. I., Meisinger, C., Melander, O., Mihailov, E., Maouche, S., Morris, A. D., Muller-Nurasyid, M., Nikus, K., Peden, J. F., Rayner,

N. W., Rasheed, A., Rosinger, S., Rubin, D., Rumpf, M. P., Schafer, A., Sivananthan, M., Song, C., Stewart, A. F. R., Tan, S.-T., Thorgeirsson, G., Schoot, C. E. v. d., Wagner, P. J., Wells, G. A., Wild, P. S., Yang, T.-P., Amouyel, P., Arveiler, D., Basart, H., Boehnke, M., Boerwinkle, E., Brambilla, P., Cambien, F., Cupples, A. L., de Faire, U., Dehghan, A., Diemert, P., Epstein, S. E., Evans, A., Ferrario, M. M., Ferrieres, J., Gauguier, D., Go, A. S., Goodall, A. H., Gudnason, V., Hazen, S. L., Holm, H., Iribarren, C., Jang, Y., Kahonen, M., Kee, F., Kim, H.-S., Klopp, N., Koenig, W., Kratzer, W., Kuulasmaa, K., Laakso, M., Laaksonen, R., Lee, J.-Y., Lind, L., Ouwehand, W. H., Parish, S., Park, J. E., Pedersen, N. L., Peters, A., Quertermous, T., Rader, D. J., Salomaa, V., Schadt, E., Shah, S. H., Sinisalo, J., Stark, K., Stefansson, K., Tregouet, D.-A., Virtamo, J., Wallentin, L., Wareham, N., Zimmermann, M. E., Nieminen, M. S., Hengstenberg, C., Sandhu, M. S., Pastinen, T., Syvanen, A.-C., Hovingh, G. K., Dedoussis, G., Franks, P. W., Lehtimaki, T., Metspalu, A., Zalloua, P. A., Siegbahn, A., Schreiber, S., Ripatti, S., Blankenberg, S. S., Perola, M., Clarke, R., Boehm, B. O., O'Donnell, C., Reilly, M. P., Marz, W., Collins, R., Kathiresan, S., Hamsten, A., Kooner, J. S., Thorsteinsdottir, U., Danesh, J., Palmer, C. N. A., Roberts, R., Watkins, H., Schunkert, H. and Samani, N. J. Large-scale association analysis identifies new risk loci for coronary artery disease. *Nat Genet*, 45(1):25–33, 2013. 9, 21, 131, 164

CARDIoGRAMplusC4D Consortium, Nikpay, M., Goel, A., Won, H. H., Hall, L. M., Willenborg, C., Kanoni, S., Saleheen, D., Kyriakou, T., Nelson, C. P., Hopewell, J. C., Webb, T. R., Zeng, L., Dehghan, A., Alver, M., Armasu, S. M., Auro, K., Bjonnes, A., Chasman, D. I., Chen, S., Ford, I., Franceschini, N., Gieger, C., Grace, C., Gustafsson, S., Huang, J., Hwang, S. J., Kim, Y. K., Kleber, M. E., Lau, K. W., Lu, X., Lu, Y., Lyytikäinen, L. P., Mihailov, E., Morrison, A. C., Pervjakova, N., Qu, L., Rose, L. M., Salfati, E., Saxena, R., Scholz, M., Smith, A. V., Tikkanen, E., Uitterlinden, A., Yang, X., Zhang, W., Zhao, W., de Andrade, M., de Vries, P. S., van Zuydam, N. R., Anand, S. S., Bertram, L., Beutner, F., Dedoussis, G., Frossard, P., Gauguier, D., Goodall, A. H., Gottesman, O., Haber, M., Han, B. G., Huang, J., Jalilzadeh, S., Kessler, T., König, I. R., Lannfelt, L., Lieb, W., Lind, L., Lindgren, C. M., Lokki, M. L., Magnusson, P. K., Mallick, N. H., Mehra, N., Meitinger, T., Memon, F. U., Morris, A. P., Nieminen, M. S., Pedersen, N. L., Peters, A., Rallidis, L. S., Rasheed, A., Samuel, M., Shah, S. H., Sinisalo, J., Stirrups, K. E., Trompet, S., Wang, L., Zaman, K. S., Ardissino, D., Boerwinkle, E., Borecki, I. B., Bottinger, E. P., Buring, J. E., Chambers, J. C., Collins, R., Cupples, L. A., Danesh, J., Demuth, I., Elosua, R., Epstein, S. E., Esko, T., Feitosa, M. F., Franco, O. H., Franzosi, M. G., Granger, C. B., Gu, D., Gudnason, V., Hall, A. S., Hamsten, A., Harris, T. B., Hazen, S. L., Hengstenberg, C., Hofman, A., Ingelsson, E., Iribarren, C., Jukema, J. W., Karhunen, P. J., Kim, B. J., Kooner, J. S., Kullo, I. J., Lehtimäki, T., Loos, R. J., Melander, O., Metspalu, A., März, W., Palmer, C. N., Perola, M., Quertermous, T., Rader, D. J., Ridker, P. M., Ripatti, S., Roberts, R., Salomaa, V., Sanghera, D. K., Schwartz, S. M., Seedorf, U., Stewart, A. F., Stott, D. J., Thiery, J., Zalloua, P. A., O'Donnell, C. J., Reilly, M. P., Assimes, T. L., Thompson, J. R., Erdmann, J., Clarke, R., Watkins, H., Kathiresan, S., McPherson, R., Deloukas, P., Schunkert, H., Samani, N. J. and Farrall, M. A comprehensive 1000 genomes-based genome-wide association meta-analysis of coronary artery disease. *Nat Genet*, 47(10):1121–1130, 2015. 131

Castelli, W. Cholesterol and lipids in the risk of coronary artery disease–the Framingham Heart Study. *Can J Cardiol*, 4(suppl. A):5A–10A, 1988. 23

Chan, C., Elliott, J., Troughton, R., Frampton, C., Smyth, D., Crozier, I. and Bridgman, P. Acute myocardial infarction and stress cardiomyopathy following the christchurch earthquakes. *PLoS ONE*, 8(7):e68504, 2013. 165

Chapman, K., Ferreira, T., Morris, A., Asimit, J. and Zeggini, E. Defining the power limits of genome-wide association scan meta-analyses. *Genetic Epidemiology*, 35(8):781–789, 2011. 4

Chasman, D. I., Paré, G., Mora, S., Hopewell, J. C., Peloso, G., Clarke, R., Cupples, L. A., Hamsten, A., Kathiresan, S., Mälarstig, A., Ordovas, J., Ripatti, S., Parker, A. N., Miletich,

J. P. and Ridker, P. M. Forty-three loci associated with plasma lipoprotein size, concentration, and cholesterol content in genome-wide analysis. *PLoS Genet*, 5(11):e1000730, 2009. 43, 56, 61

Chen, K., Wallis, J. W., McLellan, M. D., Larson, D. E., Kalicki, J. M., Pohl, C. S., McGrath, S. D., Wendl, M. C., Zhang, Q., Locke, D. P., Shi, X., Fulton, R. S., Ley, T. J., Wilson, R. K., Ding, L. and Mardis, E. R. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nat Meth*, 6(9):677–681, 2009. 14

Chen, L. S., Hsu, L., Gamazon, E. R., Cox, N. J. and Nicolae, D. L. An exponential combination procedure for set-based association tests in sequencing studies. *The American Journal of Human Genetics*, 91(6):977–986, 2012. 4

Clarke, R., Emberson, J., Parish, S., Shipley, M., Linksted, P., Sherliker, P., Clark, S., Armitage, J., Fletcher, A. and Collins, R. Cholesterol fractions and apolipoproteins as risk factors for heart disease mortality in older men. *Archives of Internal Medicine*, 167(13):1373–1378, 2007. 33

Claussnitzer, M., Dankel, S. N., Klocke, B., Grallert, H., Glunk, V., Berulava, T., Lee, H., Oskolkov, N., Fadista, J., Ehlers, K., Wahl, S., Hoffmann, C., Qian, K., Rönn, T., Riess, H., Müller-Nurasyid, M., Bretschneider, N., Schroeder, T., Skurk, T., Horsthemke, B., Spieler, D., Klingenspor, M., Seifert, M., Kern, M. J., Mejhert, N., Dahlman, I., Hansson, O., Hauck, S. M., Blüher, M., Arner, P., Groop, L., Illig, T., Suhre, K., Hsu, Y.-H., Mellgren, G., Hauner, H. and Laumen, H. Leveraging cross-species transcription factor binding site patterns: from diabetes risk loci to disease mechanisms. *Cell*, 156(1–2):343–358, 2014. 82, 83

Cohen, J. C., Boerwinkle, E., Mosley, T. H. and Hobbs, H. H. Sequence variations in PCSK9, low LDL, and protection against coronary heart disease. *New England Journal of Medicine*, 354(12):1264–1272, 2006. 10, 18

Cohen, J. C. and Hobbs, H. H. Simple genetics for a complex disease. *Science*, 340(6133):689–690, 2013. 10

Cohen, J. C., Kiss, R. S., Pertsemlidis, A., Marcel, Y. L., McPherson, R. and Hobbs, H. H. Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science*, 305(5685):869–872, 2004. 7

Cohen, J. C., Pertsemlidis, A., Kotowski, I. K., Graham, R., Garcia, C. K. and Hobbs, H. H. Low LDL cholesterol in individuals of African descent resulting from frequent nonsense mutations in PCSK9. *Nat Genet*, 37(2):161–165, 2005. 10, 18

Conrad, D. F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y., Aerts, J., Andrews, T. D., Barnes, C., Campbell, P., Fitzgerald, T., Hu, M., Ihm, C. H., Kristiansson, K., MacArthur, D. G., MacDonald, J. R., Onyiah, I., Pang, A. W. C., Robson, S., Stirrups, K., Valsesia, A., Walter, K., Wei, J., Tyler-Smith, C., Carter, N. P., Lee, C., Scherer, S. W. and Hurles, M. E. Origins and functional impact of copy number variation in the human genome. *Nature*, 464(7289):704–712, 2010. 13, 15, 131, 137, 139, 158

Coronary Artery Disease (C4D) Genetics Consortium. A genome-wide association study in Europeans and South Asians identifies five new loci for coronary artery disease. *Nat Genet*, 43(4):339–44, 2011. 31, 84, 131

Cortes, A. and Brown, M. A. Promise and pitfalls of the Immunochip. *Arthritis Research & Therapy*, 13(1):101–101, 2011. 5

Derkach, A., Lawless, J. F. and Sun, L. Robust and powerful tests for rare variants using Fisher's method to combine evidence of association from two or more complementary tests. *Genetic Epidemiology*, 37(1):110–121, 2013. 4

Devlin, B. and Roeder, K. Genomic control for association studies. *Biometrics*, 55(4):997–1004, 1999. 3, 24, 38

Do, R., Kathiresan, S. and Abecasis, G. R. Exome sequencing and complex disease: practical aspects of rare variant association studies. *Human Molecular Genetics*, 21(R1):R1–R9, 2012. 7

Do, R., Stitziel, N. O., Won, H.-H., Jorgensen, A. B., Duga, S., Angelica Merlini, P., Kiezun, A., Farrall, M., Goel, A., Zuk, O., Guella, I., Asselta, R., Lange, L. A., Peloso, G. M., Auer, P. L., Project, N. E. S., Girelli, D., Martinelli, N., Farlow, D. N., DePristo, M. A., Roberts, R., Stewart, A. F. R., Saleheen, D., Danesh, J., Epstein, S. E., Sivapalaratnam, S., Kees Hovingh, G., Kastelein, J. J., Samani, N. J., Schunkert, H., Erdmann, J., Shah, S. H., Kraus, W. E., Davies, R., Nikpay, M., Johansen, C. T., Wang, J., Hegele, R. A., Hechter, E., Marz, W., Kleber, M. E., Huang, J., Johnson, A. D., Li, M., Burke, G. L., Gross, M., Liu, Y., Assimes, T. L., Heiss, G., Lange, E. M., Folsom, A. R., Taylor, H. A., Olivieri, O., Hamsten, A., Clarke, R., Reilly, D. F., Yin, W., Rivas, M. A., Donnelly, P., Rossouw, J. E., Psaty, B. M., Herrington, D. M., Wilson, J. G., Rich, S. S., Bamshad, M. J., Tracy, R. P., Adrienne Cupples, L., Rader, D. J., Reilly, M. P., Spertus, J. A., Cresci, S., Hartiala, J., Wilson Tang, W. H., Hazen, S. L., Allayee, H., Reiner, A. P., Carlson, C. S., Kooperberg, C., Jackson, R. D., Boerwinkle, E., Lander, E. S., Schwartz, S. M., Siscovick, D. S., McPherson, R., Tybjaerg-Hansen, A., Abecasis, G. R., Watkins, H., Nickerson, D. A., Ardissino, D., Sunyaev, S. R., O'Donnell, C. J., Altshuler, D., Gabriel, S. and Kathiresan, S. Exome sequencing identifies rare LDLR and APOA5 alleles conferring risk for myocardial infarction. *Nature*, 518(7537):102–106, 2015. 131

Do, R., Willer, C. J., Schmidt, E. M., Sengupta, S., Gao, C., Peloso, G. M., Gustafsson, S., Kanoni, S., Ganna, A., Chen, J., Buchkovich, M. L., Mora, S., Beckmann, J. S., Bragg-Gresham, J. L., Chang, H.-Y., Demirkan, A., Den Hertog, H. M., Donnelly, L. A., Ehret, G. B., Esko, T., Feitosa, M. F., Ferreira, T., Fischer, K., Fontanillas, P., Fraser, R. M., Freitag, D. F., Gurdasani, D., Heikkila, K., Hypponen, E., Isaacs, A., Jackson, A. U., Johansson, A., Johnson, T., Kaakinen, M., Kettunen, J., Kleber, M. E., Li, X., Luan, J., Lyytikainen, L.-P., Magnusson, P. K. E., Mangino, M., Mihailov, E., Montasser, M. E., Muller-Nurasyid, M., Nolte, I. M., O'Connell, J. R., Palmer, C. D., Perola, M., Petersen, A.-K., Sanna, S., Saxena, R., Service, S. K., Shah, S., Shungin, D., Sidore, C., Song, C., Strawbridge, R. J., Surakka, I., Tanaka, T., Teslovich, T. M., Thorleifsson, G., Van den Herik, E. G., Voight, B. F., Volcik, K. A., Waite, L. L., Wong, A., Wu, Y., Zhang, W., Absher, D., Asiki, G., Barroso, I., Been, L. F., Bolton, J. L., Bonnycastle, L. L., Brambilla, P., Burnett, M. S., Cesana, G., Dimitriou, M., Doney, A. S. F., Doring, A., Elliott, P., Epstein, S. E., Eyjolfsson, G. I., Gigante, B., Goodarzi, M. O., Grallert, H., Gravito, M. L., Groves, C. J., Hallmans, G., Hartikainen, A.-L., Hayward, C., Hernandez, D., Hicks, A. A., Holm, H., Hung, Y.-J., Illig, T., Jones, M. R., Kaleebu, P., Kastelein, J. J. P., Khaw, K.-T., Kim, E., Klopp, N., Komulainen, P., Kumari, M., Langenberg, C., Lehtimaki, T., Lin, S.-Y., Lindstrom, J., Loos, R. J. F., Mach, F., McArdle, W. L., Meisinger, C., Mitchell, B. D., Muller, G., Nagaraja, R., Narisu, N., Nieminen, T. V. M., Nsubuga, R. N., Olafsson, I., Ong, K. K., Palotie, A., Papamarkou, T., Pomilla, C., Pouta, A., Rader, D. J., Reilly, M. P., Ridker, P. M., Rivadeneira, F., Rudan, I., Ruokonen, A., Samani, N., Scharnagl, H., Seeley, J., Silander, K., Stancakova, A., Stirrups, K., Swift, A. J., Tiret, L., Uitterlinden, A. G., van Pelt, L. J., Vedantam, S., Wainwright, N., Wijmenga, C., Wild, S. H., Willemsen, G., Wilsgaard, T., Wilson, J. F., Young, E. H., Zhao, J. H., Adair, L. S., Arveiler, D., Assimes, T. L., Bandinelli, S., Bennett, F., Bochud, M., Boehm, B. O., Boomsma, D. I., Borecki, I. B., Bornstein, S. R., Bovet, P., Burnier, M., Campbell, H., Chakravarti, A., Chambers, J. C., Chen, Y.-D. I., Collins, F. S., Cooper, R. S., Danesh, J., Dedoussis, G., de Faire, U., Feranil, A. B., Ferrieres, J., Ferrucci, L., Freimer, N. B., Gieger, C., Groop, L. C., Gudnason, V., Gyllensten, U., Hamsten, A., Harris, T. B., Hingorani, A., Hirschhorn, J. N., Hofman, A., Hovingh, G. K., Hsiung, C. A., Humphries, S. E., Hunt, S. C., Hveem, K., Iribarren, C., Jarvelin, M.-R., Jula, A., Kahonen, M., Kaprio, J., Kesaniemi, A., Kivimaki, M., Kooner, J. S., Koudstaal, P. J., Krauss, R. M., Kuh, D., Kuusisto, J., Kyvik, K. O., Laakso, M., Lakka, T. A., Lind, L., Lindgren, C. M., Martin, N. G., Marz, W., McCarthy, M. I., McKenzie, C. A., Meneton, P., Metspalu, A., Moilanen, L., Morris, A. D., Munroe, P. B., Njolstad, I., Pedersen, N. L., Power, C., Pramstaller, P. P., Price, J. F., Psaty, B. M., Quertermous, T., Rauramaa, R., Saleheen, D., Salomaa, V., Sanghera, D. K., Saramies, J., Schwarz, P. E. H., Sheu, W. H.-H., Shuldiner, A. R., Siegbahn, A., Spector, T. D.,

Stefansson, K., Strachan, D. P., Tayo, B. O., Tremoli, E., Tuomilehto, J., Uusitupa, M., van Duijn, C. M., Vollenweider, P., Wallentin, L., Wareham, N. J., Whitfield, J. B., Wolffenbuttel, B. H. R., Altshuler, D., Ordovas, J. M., Boerwinkle, E., Palmer, C. N. A., Thorsteinsdottir, U., Chasman, D. I., Rotter, J. I., Franks, P. W., Ripatti, S., Cupples, L. A., Sandhu, M. S., Rich, S. S., Boehnke, M., Deloukas, P., Mohlke, K. L., Ingelsson, E., Abecasis, G. R., Daly, M. J., Neale, B. M. and Kathiresan, S. Common variants associated with plasma triglycerides and risk for coronary artery disease. *Nat Genet*, 45(11):1345–1352, 2013. 10, 34, 166

Dupuis, J., Langenberg, C., Prokopenko, I., Saxena, R., Soranzo, N., Jackson, A. U., Wheeler, E., Glazer, N. L., Bouatia-Naji, N., Gloyn, A. L., Lindgren, C. M., Magi, R., Morris, A. P., Randall, J., Johnson, T., Elliott, P., Rybin, D., Thorleifsson, G., Steinthorsdottir, V., Henneman, P., Grallert, H., Dehghan, A., Hottenga, J. J., Franklin, C. S., Navarro, P., Song, K., Goel, A., Perry, J. R. B., Egan, J. M., Lajunen, T., Grarup, N., Sparso, T., Doney, A., Voight, B. F., Stringham, H. M., Li, M., Kanoni, S., Shrader, P., Cavalcanti-Proenca, C., Kumari, M., Qi, L., Timpson, N. J., Gieger, C., Zabena, C., Rocheleau, G., Ingelsson, E., An, P., O'Connell, J., Luan, J., Elliott, A., McCarroll, S. A., Payne, F., Roccasecca, R. M., Pattou, F., Sethupathy, P., Ardlie, K., Ariyurek, Y., Balkau, B., Barter, P., Beilby, J. P., Ben-Shlomo, Y., Benediktsson, R., Bennett, A. J., Bergmann, S., Bochud, M., Boerwinkle, E., Bonnefond, A., Bonnycastle, L. L., Borch-Johnsen, K., Bottcher, Y., Brunner, E., Bumpstead, S. J., Charpentier, G., Chen, Y.-D. I., Chines, P., Clarke, R., Coin, L. J. M., Cooper, M. N., Cornelis, M., Crawford, G., Crisponi, L., Day, I. N. M., de Geus, E. J. C., Delplanque, J., Dina, C., Erdos, M. R., Fedson, A. C., Fischer-Rosinsky, A., Forouhi, N. G., Fox, C. S., Frants, R., Franzosi, M. G., Galan, P., Goodarzi, M. O., Graessler, J., Groves, C. J., Grundy, S., Gwilliam, R., Gyllensten, U., Hadjadj, S., Hallmans, G., Hammond, N., Han, X., Hartikainen, A.-L., Hassanali, N., Hayward, C., Heath, S. C., Hercberg, S., Herder, C., Hicks, A. A., Hillman, D. R., Hingorani, A. D., Hofman, A., Hui, J., Hung, J., Isomaa, B., Johnson, P. R. V., Jorgensen, T., Jula, A., Kaakinen, M., Kaprio, J., Kesaniemi, Y. A., Kivimaki, M., Knight, B., Koskinen, S., Kovacs, P., Kyvik, K. O., Lathrop, G. M., Lawlor, D. A., Le Bacquer, O., Lecoeur, C., Li, Y., Lyssenko, V., Mahley, R., Mangino, M., Manning, A. K., Martinez-Larrad, M. T., McAteer, J. B., McCulloch, L. J., McPherson, R., Meisinger, C., Melzer, D., Meyre, D., Mitchell, B. D., Morken, M. A., Mukherjee, S., Naitza, S., Narisu, N., Neville, M. J., Oostra, B. A., Orru, M., Pakyz, R., Palmer, C. N. A., Paolisso, G., Pattaro, C., Pearson, D., Peden, J. F., Pedersen, N. L., Perola, M., Pfeiffer, A. F. H., Pichler, I., Polasek, O., Posthuma, D., Potter, S. C., Pouta, A., Province, M. A., Psaty, B. M., Rathmann, W., Rayner, N. W., Rice, K., Ripatti, S., Rivadeneira, F., Roden, M., Rolandsson, O., Sandbaek, A., Sandhu, M., Sanna, S., Sayer, A. A., Scheet, P., Scott, L. J., Seedorf, U., Sharp, S. J., Shields, B., Sigur[eth]sson, G., Sijbrands, E. J. G., Silveira, A., Simpson, L., Singleton, A., Smith, N. L., Sovio, U., Swift, A., Syddall, H., Syvanen, A.-C., Tanaka, T., Thorand, B., Tichet, J., Tonjes, A., Tuomi, T., Uitterlinden, A. G., van Dijk, K. W., van Hoek, M., Varma, D., Visvikis-Siest, S., Vitart, V., Vogelzangs, N., Waeber, G., Wagner, P. J., Walley, A., Walters, G. B., Ward, K. L., Watkins, H., Weedon, M. N., Wild, S. H., Willemsen, G., Witteman, J. C. M., Yarnell, J. W. G., Zeggini, E., Zelenika, D., Zethelius, B., Zhai, G., Zhao, J. H., Zillikens, M. C., Borecki, I. B., Loos, R. J. F., Meneton, P., Magnusson, P. K. E., Nathan, D. M., Williams, G. H., Hattersley, A. T., Silander, K., Salomaa, V., Smith, G. D., Bornstein, S. R., Schwarz, P., Spranger, J., Karpe, F., Shuldiner, A. R., Cooper, C., Dedoussis, G. V., Serrano-Rios, M., Morris, A. D., Lind, L., Palmer, L. J., Hu, F. B., Franks, P. W., Ebrahim, S., Marmot, M., Kao, W. H. L., Pankow, J. S., Sampson, M. J., Kuusisto, J., Laakso, M., Hansen, T., Pedersen, O., Pramstaller, P. P., Wichmann, H. E., Illig, T., Rudan, I., Wright, A. F., Stumvoll, M., Campbell, H. and Wilson, J. F. New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nat Genet*, 42(2):105–116, 2010. 32

Eisinga, R., Te Grotenhuis, M. and Pelzer, B. Saddlepoint approximations for the sum of independent non-identically distributed binomial random variables. *Statistica Neerlandica*, 67(2):190–201, 2013. ISSN 1467-9574. 97

Emerging Risk Factors Collaboration, Di Angelantonio, E., Sarwar, N., Perry, P., Kaptoge, S., Ray,

K., Thompson, A., Wood, A., Lewington, S., Sattar, N., Packard, C., Collins, R., Thompson, S. and Danesh, J. Major lipids, apolipoproteins, and risk of vascular disease. *JAMA*, 302(18):1993–2000, 2009. 6

ENCODE Project Consortium. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol*, 9(4):e1001046, 2011. 29, 30, 43, 73

ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, 2012. 12, 82

Ernst, J., Kheradpour, P., Mikkelsen, T. S., Shoresh, N., Ward, L. D., Epstein, C. B., Zhang, X., Wang, L., Issner, R., Coyne, M., Ku, M., Durham, T., Kellis, M. and Bernstein, B. E. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, 473(7345):43–9, 2011. 29, 30, 43, 72, 73, 82, 83, 89, 103

Fisher, R., Genetiker, S., Genetician, S., Britain, G. and Geneticien, S. Statistical methods for research workers. *Oliver and Boyd, Edinburgh*, 1970. 4

Fitzgerald, M. L., Moore, K. J. and Freeman, M. W. Nuclear hormone receptors and cholesterol trafficking: the orphans find a new home. 80(5):271–281, 2002. 27

Freathy, R. M., Timpson, N. J., Lawlor, D. A., Pouta, A., Ben-Shlomo, Y., Ruokonen, A., Ebrahim, S., Shields, B., Zeggini, E., Weedon, M. N., Lindgren, C. M., Lango, H., Melzer, D., Ferrucci, L., Paolisso, G., Neville, M. J., Karpe, F., Palmer, C. N. A., Morris, A. D., Elliott, P., Jarvelin, M.-R., Davey Smith, G., McCarthy, M. I., Hattersley, A. T. and Frayling, T. M. Common variation in the FTO gene alters diabetes-related metabolic traits to the extent expected given its effect on BMI. *Diabetes*, 57(5):1419–1426, 2008. 32, 33

Friedewald, W. T., Levy, R. I. and Fredrickson, D. S. Estimation of the concentration of low-density lipoprotein cholesterol in plasma, without use of the preparative ultracentrifuge. *Clinical Chemistry*, 18(6):499–502, 1972. 38, 133

Frikke-Schmidt, R., Nordestgaard, B., Stene, M., Sethi, A., Remaley, A., Schnohr, R., Grande, P. and Tybjaerg-Hansen, A. Association of loss-of-function mutations in the ABCA1 gene with high-density lipoprotein cholesterol levels and risk of ischemic heart disease. *JAMA*, 299(21):2524–2532, 2008. 33

Fu, W., O'Connor, T. D., Jun, G., Kang, H. M., Abecasis, G., Leal, S. M., Gabriel, S., Altshuler, D., Shendure, J., Nickerson, D. A., Bamshad, M. J., NHLBI Exome Sequencing Project and Akey, J. M. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature*, 493(7431):216–220, 2013. 3

Global Lipids Genetics Consortium, Willer, C. J., Schmidt, E. M., Sengupta, S., Peloso, G. M., Gustafsson, S., Kanoni, S., Ganna, A., Chen, J., Buchkovich, M. L., Mora, S., Beckmann, J. S., Bragg-Gresham, J. L., Chang, H.-Y., Demirkan, A., Den Hertog, H. M., Do, R., Donnelly, L. A., Ehret, G. B., Esko, T., Feitosa, M. F., Ferreira, T., Fischer, K., Fontanillas, P., Fraser, R. M., Freitag, D. F., Gurdasani, D., Heikkilä, K., Hyppönen, E., Isaacs, A., Jackson, A. U., Johansson, A., Johnson, T., Kaakinen, M., Kettunen, J., Kleber, M. E., Li, X., Luan, J., Lyytikäinen, L.-P., Magnusson, P. K. E., Mangino, M., Mihailov, E., Montasser, M. E., Müller-Nurasyid, M., Nolte, I. M., O'Connell, J. R., Palmer, C. D., Perola, M., Petersen, A.-K., Sanna, S., Saxena, R., Service, S. K., Shah, S., Shungin, D., Sidore, C., Song, C., Strawbridge, R. J., Surakka, I., Tanaka, T., Teslovich, T. M., Thorleifsson, G., Van den Herik, E. G., Voight, B. F., Volcik, K. A., Waite, L. L., Wong, A., Wu, Y., Zhang, W., Absher, D., Asiki, G., Barroso, I., Been, L. F., Bolton, J. L., Bonnycastle, L. L., Brambilla, P., Burnett, M. S., Cesana, G., Dimitriou, M., Doney, A. S. F., Döring, A., Elliott, P., Epstein, S. E., Eyjolfsson, G. I., Gigante, B., Goodarzi, M. O., Grallert, H., Gravito, M. L., Groves, C. J., Hallmans, G., Hartikainen, A.-L., Hayward, C., Hernandez, D., Hicks, A. A., Holm, H., Hung, Y.-J., Illig, T., Jones, M. R., Kaleebu, P.,

Kastelein, J. J. P., Khaw, K.-T., Kim, E., Klopp, N., Komulainen, P., Kumari, M., Langenberg, C., Lehtimäki, T., Lin, S.-Y., Lindström, J., Loos, R. J. F., Mach, F., McArdle, W. L., Meisinger, C., Mitchell, B. D., Müller, G., Nagaraja, R., Narisu, N., Nieminen, T. V. M., Nsubuga, R. N., Olafsson, I., Ong, K. K., Palotie, A., Papamarkou, T., Pomilla, C., Pouta, A., Rader, D. J., Reilly, M. P., Ridker, P. M., Rivadeneira, F., Rudan, I., Ruokonen, A., Samani, N., Scharnagl, H., Seeley, J., Silander, K., Stancáková, A., Stirrups, K., Swift, A. J., Tiret, L., Uitterlinden, A. G., van Pelt, L. J., Vedantam, S., Wainwright, N., Wijmenga, C., Wild, S. H., Willemsen, G., Wilsgaard, T., Wilson, J. F., Young, E. H., Zhao, J. H., Adair, L. S., Arveiler, D., Assimes, T. L., Bandinelli, S., Bennett, F., Bochud, M., Boehm, B. O., Boomsma, D. I., Borecki, I. B., Bornstein, S. R., Bovet, P., Burnier, M., Campbell, H., Chakravarti, A., Chambers, J. C., Chen, Y.-D. I., Collins, F. S., Cooper, R. S., Danesh, J., Dedoussis, G., de Faire, U., Feranil, A. B., Ferrières, J., Ferrucci, L., Freimer, N. B., Gieger, C., Groop, L. C., Gudnason, V., Gyllensten, U., Hamsten, A., Harris, T. B., Hingorani, A., Hirschhorn, J. N., Hofman, A., Hovingh, G. K., Hsiung, C. A., Humphries, S. E., Hunt, S. C., Hveem, K., Iribarren, C., Järvelin, M.-R., Jula, A., Kähönen, M., Kaprio, J., Kesäniemi, A., Kivimaki, M., Kooner, J. S., Koudstaal, P. J., Krauss, R. M., Kuh, D., Kuusisto, J., Kyvik, K. O., Laakso, M., Lakka, T. A., Lind, L., Lindgren, C. M., Martin, N. G., März, W., McCarthy, M. I., McKenzie, C. A., Meneton, P., Metspalu, A., Moilanen, L., Morris, A. D., Munroe, P. B., Njølstad, I., Pedersen, N. L., Power, C., Pramstaller, P. P., Price, J. F., Psaty, B. M., Quertermous, T., Rauramaa, R., Saleheen, D., Salomaa, V., Sanghera, D. K., Saramies, J., Schwarz, P. E. H., Sheu, W. H.-H., Shuldiner, A. R., Siegbahn, A., Spector, T. D., Stefansson, K., Strachan, D. P., Tayo, B. O., Tremoli, E., Tuomilehto, J., Uusitupa, M., van Duijn, C. M., Vollenweider, P., Wallentin, L., Wareham, N. J., Whitfield, J. B., Wolffenbuttel, B. H. R., Ordovas, J. M., Boerwinkle, E., Palmer, C. N. A., Thorsteinsdottir, U., Chasman, D. I., Rotter, J. I., Franks, P. W., Ripatti, S., Cupples, L. A., Sandhu, M. S., Rich, S. S., Boehnke, M., Deloukas, P., Kathiresan, S., Mohlke, K. L., Ingelsson, E. and Abecasis, G. R. Discovery and refinement of loci associated with lipid levels. *Nat Genet*, 45(11):1274–83, 2013. 6, 9, 20, 21, 22, 44, 84, 87, 110, 162, 166

GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science*, 348(6235):648–660, 2015. 167

Hagberg, C. E., Falkevall, A., Wang, X., Larsson, E., Huusko, J., Nilsson, I., van Meeteren, L. A., Samen, E., Lu, L., Vanwildemeersch, M., Klar, J., Genove, G., Pietras, K., Stone-Elander, S., Claesson-Welsh, L., Yla-Herttuala, S., Lindahl, P. and Eriksson, U. Vascular endothelial growth factor B controls endothelial fatty acid uptake. *Nature*, 464(7290):917–921, 2010. 25

Han, F. and Pan, W. A data-adaptive sum test for disease association with multiple common or rare variants. *Human Heredity*, 70(1):42–54, 2010. 4

Handsaker, R. E., Korn, J. M., Nemesh, J. and McCarroll, S. A. Discovery and genotyping of genome structural polymorphism by sequencing on a population scale. *Nat Genet*, 43(3):269–276, 2011. 14

Handsaker, R. E., Van Doren, V., Berman, J. R., Genovese, G., Kashin, S., Boettger, L. M. and McCarroll, S. A. Large multiallelic copy number variations in humans. *Nat Genet*, 47(3):296–303, 2015. 14, 134

Heid, I. M., Jackson, A. U., Randall, J. C., Winkler, T. W., Qi, L., Steinthorsdottir, V., Thorleifsson, G., Zillikens, M. C., Speliotes, E. K., Magi, R., Workalemahu, T., White, C. C., Bouatia-Naji, N., Harris, T. B., Berndt, S. I., Ingelsson, E., Willer, C. J., Weedon, M. N., Luan, J., Vedantam, S., Esko, T., Kilpelainen, T. O., Kutalik, Z., Li, S., Monda, K. L., Dixon, A. L., Holmes, C. C., Kaplan, L. M., Liang, L., Min, J. L., Moffatt, M. F., Molony, C., Nicholson, G., Schadt, E. E., Zondervan, K. T., Feitosa, M. F., Ferreira, T., Allen, H. L., Weyant, R. J., Wheeler, E., Wood, A. R., Estrada, K., Goddard, M. E., Lettre, G., Mangino, M., Nyholt, D. R., Purcell, S., Smith, A. V., Visscher, P. M., Yang, J., McCarroll, S. A., Nemesh, J., Voight, B. F., Absher, D., Amin, N., Aspelund, T., Coin, L., Glazer, N. L., Hayward, C., Heard-Costa, N. L., Hottenga, J.-J.,

Johansson, A., Johnson, T., Kaakinen, M., Kapur, K., Ketkar, S., Knowles, J. W., Kraft, P., Kraja, A. T., Lamina, C., Leitzmann, M. F., McKnight, B., Morris, A. P., Ong, K. K., Perry, J. R. B., Peters, M. J., Polasek, O., Prokopenko, I., Rayner, N. W., Ripatti, S., Rivadeneira, F., Robertson, N. R., Sanna, S., Sovio, U., Surakka, I., Teumer, A., van Wingerden, S., Vitart, V., Zhao, J. H., Cavalcanti-Proenca, C., Chines, P. S., Fisher, E., Kulzer, J. R., Lecoeur, C., Narisu, N., Sandholt, C., Scott, L. J., Silander, K., Stark, K., Tammesoo, M.-L., Teslovich, T. M., Timpson, N. J., Watanabe, R. M., Welch, R., Chasman, D. I., Cooper, M. N., Jansson, J.-O., Kettunen, J., Lawrence, R. W., Pellikka, N., Perola, M., Vandenput, L., Alavere, H., Almgren, P., Atwood, L. D., Bennett, A. J., Biffar, R., Bonnycastle, L. L., Bornstein, S. R., Buchanan, T. A., Campbell, H., Day, I. N. M., Dei, M., Dorr, M., Elliott, P., Erdos, M. R., Eriksson, J. G., Freimer, N. B., Fu, M., Gaget, S., Geus, E. J. C., Gjesing, A. P., Grallert, H., Graszler, J., Groves, C. J., Guiducci, C., Hartikainen, A.-L., Hassanali, N., Havulinna, A. S., Herzig, K.-H., Hicks, A. A., Hui, J., Igl, W., Jousilahti, P., Jula, A., Kajantie, E., Kinnunen, L., Kolcic, I., Koskinen, S., Kovacs, P., Kroemer, H. K., Krzelj, V., Kuusisto, J., Kvaloy, K., Laitinen, J., Lantieri, O., Lathrop, G. M., Lokki, M.-L., Luben, R. N., Ludwig, B., McArdle, W. L., McCarthy, A., Morken, M. A., Nelis, M., Neville, M. J., Pare, G., Parker, A. N., Peden, J. F., Pichler, I., Pietilainen, K. H., Platou, C. G. P., Pouta, A., Ridderstrale, M., Samani, N. J., Saramies, J., Sinisalo, J., Smit, J. H., Strawbridge, R. J., Stringham, H. M., Swift, A. J., Teder-Laving, M., Thomson, B., Usala, G., van Meurs, J. B. J., van Ommen, G.-J., Vatin, V., Volpato, C. B., Wallaschofski, H., Walters, G. B., Widen, E., Wild, S. H., Willemsen, G., Witte, D. R., Zgaga, L., Zitting, P., Beilby, J. P., James, A. L., Kahonen, M., Lehtimaki, T., Nieminen, M. S., Ohlsson, C., Palmer, L. J., Raitakari, O., Ridker, P. M., Stumvoll, M., Tonjes, A., Viikari, J., Balkau, B., Ben-Shlomo, Y., Bergman, R. N., Boeing, H., Smith, G. D., Ebrahim, S., Froguel, P., Hansen, T., Hengstenberg, C., Hveem, K., Isomaa, B., Jorgensen, T., Karpe, F., Khaw, K.-T., Laakso, M., Lawlor, D. A., Marre, M., Meitinger, T., Metspalu, A., Midthjell, K., Pedersen, O., Salomaa, V., Schwarz, P. E. H., Tuomi, T., Tuomilehto, J., Valle, T. T., Wareham, N. J., Arnold, A. M., Beckmann, J. S., Bergmann, S., Boerwinkle, E., Boomsma, D. I., Caulfield, M. J., Collins, F. S., Eiriksdottir, G., Gudnason, V., Gyllensten, U., Hamsten, A., Hattersley, A. T., Hofman, A., Hu, F. B., Illig, T., Iribarren, C., Jarvelin, M.-R., Kao, W. H. L., Kaprio, J., Launer, L. J., Munroe, P. B., Oostra, B., Penninx, B. W., Pramstaller, P. P., Psaty, B. M., Quertermous, T., Rissanen, A., Rudan, I., Shuldiner, A. R., Soranzo, N., Spector, T. D., Syvanen, A.-C., Uda, M., Uitterlinden, A., Volzke, H., Vollenweider, P., Wilson, J. F., Witteman, J. C., Wright, A. F., Abecasis, G. R., Boehnke, M., Borecki, I. B., Deloukas, P., Frayling, T. M., Groop, L. C., Haritunians, T., Hunter, D. J., Kaplan, R. C., North, K. E., O'Connell, J. R., Peltonen, L., Schlessinger, D., Strachan, D. P., Hirschhorn, J. N., Assimes, T. L., Wichmann, H.-E., Thorsteinsdottir, U., van Duijn, C. M., Stefansson, K., Cupples, L. A., Loos, R. J. F., Barroso, I., McCarthy, M. I., Fox, C. S., Mohlke, K. L. and Lindgren, C. M. Meta-analysis identifies 13 new loci associated with waist-hip ratio and reveals sexual dimorphism in the genetic basis of fat distribution. *Nat Genet*, 42(11):949–960, 2010. 31

Helgadottir, A., Thorleifsson, G., Manolescu, A., Gretarsdottir, S., Blondal, T., Jonasdottir, A., Jonasdottir, A., Sigurdsson, A., Baker, A., Palsson, A., Masson, G., Gudbjartsson, D. F., Magnusson, K. P., Andersen, K., Levey, A. I., Backman, V. M., Matthiasdottir, S., Jonsdottir, T., Palsson, S., Einarsdottir, H., Gunnarsdottir, S., Gylfason, A., Vaccarino, V., Hooper, W. C., Reilly, M. P., Granger, C. B., Austin, H., Rader, D. J., Shah, S. H., Quyyumi, A. A., Gulcher, J. R., Thorgeirsson, G., Thorsteinsdottir, U., Kong, A. and Stefansson, K. A common variant on chromosome 9p21 affects the risk of myocardial infarction. *Science*, 316(5830):1491–1493, 2007. 131

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S. and Manolio, T. A. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences*, 106(23):9362–9367, 2009. 11

Hoffmann, T. J., Marini, N. J. and Witte, J. S. Comprehensive approach to analyzing rare genetic variants. *PLoS ONE*, 5(11):e13584, 2010. 4

Hollox, E. J., Huffmeier, U., Zeeuwen, P. L. J. M., Palla, R., Lascorz, J., Rodijk-Olthuis, D., van de Kerkhof, P. C. M., Traupe, H., de Jongh, G., Heijer, M. d., Reis, A., Armour, J. A. L. and Schalkwijk, J. Psoriasis is associated with increased beta-defensin genomic copy number. *Nat Genet*, 40(1):23–25, 2008. 15

Holmen, O. L., Zhang, H., Fan, Y., Hovelson, D. H., Schmidt, E. M., Zhou, W., Guo, Y., Zhang, J., Langhammer, A., Lochen, M.-L., Ganesh, S. K., Vatten, L., Skorpen, F., Dalen, H., Zhang, J., Pennathur, S., Chen, J., Platou, C., Mathiesen, E. B., Wilsgaard, T., Njolstad, I., Boehnke, M., Chen, Y. E., Abecasis, G. R., Hveem, K. and Willer, C. J. Systematic evaluation of coding variation identifies a candidate causal variant in TM6SF2 influencing total cholesterol and myocardial infarction risk. *Nat Genet*, 46(4):345–351, 2014. 8, 18, 20

Hu, Y.-J., Berndt, S. I., Gustafsson, S., Ganna, A., Hirschhorn, J., North, K. E., Ingelsson, E. and Lin, D.-Y. Meta-analysis of gene-level associations for rare variants based on single-variant statistics. *The American Journal of Human Genetics*, 93(2):236–248, 2013. 5

Iafrate, A. J., Feuk, L., Rivera, M. N., Listewnik, M. L., Donahoe, P. K., Qi, Y., Scherer, S. W. and Lee, C. Detection of large-scale variation in the human genome. *Nat Genet*, 36(9):949–951, 2004. 13

IBC 50K CAD Consortium. Large-scale gene-centric analysis identifies novel variants for coronary artery disease. *PLoS Genet*, 7(9):e1002260, 2011. 131

International Consortium for Blood Pressure Genome-Wide Association Studies, Ehret, G. B., Munroe, P. B., Rice, K. M., Bochud, M., Johnson, A. D., Chasman, D. I., Smith, A. V., Tobin, M. D., Verwoert, G. C., Hwang, S.-J., Pihur, V., Vollenweider, P., O'Reilly, P. F., Amin, N., Bragg-Gresham, J. L., Teumer, A., Glazer, N. L., Launer, L., Zhao, J. H., Aulchenko, Y., Heath, S., Sõber, S., Parsa, A., Luan, J., Arora, P., Dehghan, A., Zhang, F., Lucas, G., Hicks, A. A., Jackson, A. U., Peden, J. F., Tanaka, T., Wild, S. H., Rudan, I., Igl, W., Milaneschi, Y., Parker, A. N., Fava, C., Chambers, J. C., Fox, E. R., Kumari, M., Go, M. J., van der Harst, P., Kao, W. H. L., Sjögren, M., Vinay, D. G., Alexander, M., Tabara, Y., Shaw-Hawkins, S., Whincup, P. H., Liu, Y., Shi, G., Kuusisto, J., Tayo, B., Seielstad, M., Sim, X., Nguyen, K.-D. H., Lehtimäki, T., Matullo, G., Wu, Y., Gaunt, T. R., Onland-Moret, N. C., Cooper, M. N., Platou, C. G. P., Org, E., Hardy, R., Dahgam, S., Palmen, J., Vitart, V., Braund, P. S., Kuznetsova, T., Uiterwaal, C. S. P. M., Adeyemo, A., Palmas, W., Campbell, H., Ludwig, B., Tomaszewski, M., Tzoulaki, I., Palmer, N. D., CARDIoGRAM consortium, CKDGen Consortium, KidneyGen Consortium, EchoGen consortium, CHARGE-HF consortium, Aspelund, T., Garcia, M., Chang, Y.-P. C., O'Connell, J. R., Steinle, N. I., Grobbee, D. E., Arking, D. E., Kardia, S. L., Morrison, A. C., Hernandez, D., Najjar, S., McArdle, W. L., Hadley, D., Brown, M. J., Connell, J. M., Hingorani, A. D., Day, I. N. M., Lawlor, D. A., Beilby, J. P., Lawrence, R. W., Clarke, R., Hopewell, J. C., Ongen, H., Dreisbach, A. W., Li, Y., Young, J. H., Bis, J. C., Kähönen, M., Viikari, J., Adair, L. S., Lee, N. R., Chen, M.-H., Olden, M., Pattaro, C., Bolton, J. A. H., Köttgen, A., Bergmann, S., Mooser, V., Chaturvedi, N., Frayling, T. M., Islam, M., Jafar, T. H., Erdmann, J., Kulkarni, S. R., Bornstein, S. R., Grässler, J., Groop, L., Voight, B. F., Kettunen, J., Howard, P., Taylor, A., Guarrera, S., Ricceri, F., Emilsson, V., Plump, A., Barroso, I., Khaw, K.-T., Weder, A. B., Hunt, S. C., Sun, Y. V., Bergman, R. N., Collins, F. S., Bonnycastle, L. L., Scott, L. J., Stringham, H. M., Peltonen, L., Perola, M., Vartiainen, E., Brand, S.-M., Staessen, J. A., Wang, T. J., Burton, P. R., Soler Artigas, M., Dong, Y., Snieder, H., Wang, X., Zhu, H., Lohman, K. K., Rudock, M. E., Heckbert, S. R., Smith, N. L., Wiggins, K. L., Doumatey, A., Shriner, D., Veldre, G., Viigimaa, M., Kinra, S., Prabhakaran, D., Tripathy, V., Langefeld, C. D., Rosengren, A., Thelle, D. S., Corsi, A. M., Singleton, A., Forrester, T., Hilton, G., McKenzie, C. A., Salako, T., Iwai, N., Kita, Y., Ogihara, T., Ohkubo, T., Okamura, T., Ueshima, H., Umemura, S., Eyheramendy, S., Meitinger, T., Wichmann, H.-E., Cho, Y. S., Kim, H.-L., Lee,

J.-Y., Scott, J., Sehmi, J. S., Zhang, W., Hedblad, B., Nilsson, P., Smith, G. D., Wong, A., Narisu, N., Stančáková, A., Raffel, L. J., Yao, J., Kathiresan, S., O'Donnell, C. J., Schwartz, S. M., Ikram, M. A., Longstreth, Jr, W. T., Mosley, T. H., Seshadri, S., Shrine, N. R. G., Wain, L. V., Morken, M. A., Swift, A. J., Laitinen, J., Prokopenko, I., Zitting, P., Cooper, J. A., Humphries, S. E., Danesh, J., Rasheed, A., Goel, A., Hamsten, A., Watkins, H., Bakker, S. J. L., van Gilst, W. H., Janipalli, C. S., Mani, K. R., Yajnik, C. S., Hofman, A., Mattace-Raso, F. U. S., Oostra, B. A., Demirkan, A., Isaacs, A., Rivadeneira, F., Lakatta, E. G., Orru, M., Scuteri, A., Ala-Korpela, M., Kangas, A. J., Lyytikäinen, L.-P., Soininen, P., Tukiainen, T., Würtz, P., Ong, R. T.-H., Dörr, M., Kroemer, H. K., Völker, U., Völzke, H., Galan, P., Hercberg, S., Lathrop, M., Zelenika, D., Deloukas, P., Mangino, M., Spector, T. D., Zhai, G., Meschia, J. F., Nalls, M. A., Sharma, P., Terzic, J., Kumar, M. V. K., Denniff, M., Zukowska-Szczechowska, E., Wagenknecht, L. E., Fowkes, F. G. R., Charchar, F. J., Schwarz, P. E. H., Hayward, C., Guo, X., Rotimi, C., Bots, M. L., Brand, E., Samani, N. J., Polasek, O., Talmud, P. J., Nyberg, F., Kuh, D., Laan, M., Hveem, K., Palmer, L. J., van der Schouw, Y. T., Casas, J. P., Mohlke, K. L., Vineis, P., Raitakari, O., Ganesh, S. K., Wong, T. Y., Tai, E. S., Cooper, R. S., Laakso, M., Rao, D. C., Harris, T. B., Morris, R. W., Dominiczak, A. F., Kivimaki, M., Marmot, M. G., Miki, T., Saleheen, D., Chandak, G. R., Coresh, J., Navis, G., Salomaa, V., Han, B.-G., Zhu, X., Kooner, J. S., Melander, O., Ridker, P. M., Bandinelli, S., Gyllensten, U. B., Wright, A. F., Wilson, J. F., Ferrucci, L., Farrall, M., Tuomilehto, J., Pramstaller, P. P., Elosua, R., Soranzo, N., Sijbrands, E. J. G., Altshuler, D., Loos, R. J. F., Shuldiner, A. R., Gieger, C., Meneton, P., Uitterlinden, A. G., Wareham, N. J., Gudnason, V., Rotter, J. I., Rettig, R., Uda, M., Strachan, D. P., Witteman, J. C. M., Hartikainen, A.-L., Beckmann, J. S., Boerwinkle, E., Vasan, R. S., Boehnke, M., . Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature*, 478(7367):103–9, 2011. 21, 31, 84

Ionita-Laza, I., Buxbaum, J. D., Laird, N. M. and Lange, C. A new testing strategy to identify rare variants with either risk or protective effect on disease. *PLoS Genet*, 7(2):e1001289, 2011. 4

Kang, H. M., Sul, J. H., Service, S. K., Zaitlen, N. A., Kong, S.-y., Freimer, N. B., Sabatti, C. and Eskin, E. Variance component model to account for sample structure in genome-wide association studies. *Nat Genet*, 42(4):348–354, 2010. 2, 38

Kannel, W. B., Dawber, T. R., Kagan, A., Revotskie, N. and Stokes, J. I. Factors of risk in the development of coronary heart disease—six-year follow-up experience. The Framingham Study. *Annals of Internal Medicine*, 55(1):33–50, 1961. 23

Kaprio, J., Ferrell, R. E., Kottke, B. A., Kamboh, M. I. and Sing, C. F. Effects of polymorphisms in apolipoproteins E, A-IV, and H on quantitative traits related to risk for cardiovascular disease. *Arteriosclerosis, Thrombosis, and Vascular Biology*, 11(5):1330–1348, 1991. 28

Kathiresan, S., Melander, O., Guiducci, C., Surti, A., Burtt, N. P., Rieder, M. J., Cooper, G. M., Roos, C., Voight, B. F., Havulinna, A. S., Wahlstrand, B., Hedner, T., Corella, D., Tai, E. S., Ordovas, J. M., Berglund, G., Vartiainen, E., Jousilahti, P., Hedblad, B., Taskinen, M.-R., Newton-Cheh, C., Salomaa, V., Peltonen, L., Groop, L., Altshuler, D. M. and Orho-Melander, M. Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. *Nat Genet*, 40(2):189–197, 2008. 6

Kathiresan, S., Willer, C. J., Peloso, G. M., Demissie, S., Musunuru, K., Schadt, E. E., Kaplan, L., Bennett, D., Li, Y., Tanaka, T., Voight, B. F., Bonnycastle, L. L., Jackson, A. U., Crawford, G., Surti, A., Guiducci, C., Burtt, N. P., Parish, S., Clarke, R., Zelenika, D., Kubalanza, K. A., Morken, M. A., Scott, L. J., Stringham, H. M., Galan, P., Swift, A. J., Kuusisto, J., Bergman, R. N., Sundvall, J., Laakso, M., Ferrucci, L., Scheet, P., Sanna, S., Uda, M., Yang, Q., Lunetta, K. L., Dupuis, J., de Bakker, P. I. W., O'Donnell, C. J., Chambers, J. C., Kooner, J. S., Hercberg, S., Meneton, P., Lakatta, E. G., Scuteri, A., Schlessinger, D., Tuomilehto, J., Collins, F. S., Groop, L., Altshuler, D., Collins, R., Lathrop, G. M., Melander, O., Salomaa, V., Peltonen, L., Orho-Melander, M., Ordovas, J. M., Boehnke, M., Abecasis, G. R., Mohlke, K. L. and Cupples,

L. A. Common variants at 30 loci contribute to polygenic dyslipidemia. *Nat Genet*, 41(1):56–65, 2009. 6

Kaufman, A. L., Spitz, J., Jacobs, M., Sorrentino, M., Yuen, S., Danahey, K., Saner, D., Klein, T. E., Altman, R. B., Ratain, M. J. and O'Donnell, P. H. Evidence for clinical implementation of pharmacogenomics in cardiac drugs. *Mayo Clinic Proceedings*, 90(6):716–729, 2015. 167

Keating, B. J., Tischfield, S., Murray, S. S., Bhangale, T., Price, T. S., Glessner, J. T., Galver, L., Barrett, J. C., Grant, S. F. A., Farlow, D. N., Chandrupatla, H. R., Hansen, M., Ajmal, S., Papanicolaou, G. J., Guo, Y., Li, M., DerOhannessian, S., de Bakker, P. I. W., Bailey, S. D., Montpetit, A., Edmondson, A. C., Taylor, K., Gai, X., Wang, S. S., Fornage, M., Shaikh, T., Groop, L., Boehnke, M., Hall, A. S., Hattersley, A. T., Frackelton, E., Patterson, N., Chiang, C. W. K., Kim, C. E., Fabsitz, R. R., Ouwehand, W., Price, A. L., Munroe, P., Caulfield, M., Drake, T., Boerwinkle, E., Reich, D., Whitehead, A. S., Cappola, T. P., Samani, N. J., Lusis, A. J., Schadt, E., Wilson, J. G., Koenig, W., McCarthy, M. I., Kathiresan, S., Gabriel, S. B., Hakonarson, H., Anand, S. S., Reilly, M., Engert, J. C., Nickerson, D. A., Rader, D. J., Hirschhorn, J. N. and Fitzgerald, G. A. Concept, design and implementation of a cardiovascular gene-centric 50 k SNP array for large-scale genomic association studies. *PLoS ONE*, 3(10):e3583, 2008. 42

Kichaev, G., Yang, W.-Y., Lindstrom, S., Hormozdiari, F., Eskin, E., Price, A. L., Kraft, P. and Pasaniuc, B. Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS Genet*, 10(10):e1004722, 2014. 82, 83

Kidd, J. M., Cooper, G. M., Donahue, W. F., Hayden, H. S., Sampas, N., Graves, T., Hansen, N., Teague, B., Alkan, C., Antonacci, F., Haugen, E., Zerr, T., Yamada, N. A., Tsang, P., Newman, T. L., Tuzun, E., Cheng, Z., Ebling, H. M., Tusneem, N., David, R., Gillett, W., Phelps, K. A., Weaver, M., Saranga, D., Brand, A., Tao, W., Gustafson, E., McKernan, K., Chen, L., Malig, M., Smith, J. D., Korn, J. M., McCarroll, S. A., Altshuler, D. A., Peiffer, D. A., Dorschner, M., Stamatoyannopoulos, J., Schwartz, D., Nickerson, D. A., Mullikin, J. C., Wilson, R. K., Bruhn, L., Olson, M. V., Kaul, R., Smith, D. R. and Eichler, E. E. Mapping and sequencing of structural variation from eight human genomes. *Nature*, 453(7191):56–64, 2008. 131

Korbel, J. O., Urban, A. E., Affourtit, J. P., Godwin, B., Grubert, F., Simons, J. F., Kim, P. M., Palejev, D., Carriero, N. J., Du, L., Taillon, B. E., Chen, Z., Tanzer, A., Saunders, A. C. E., Chi, J., Yang, F., Carter, N. P., Hurles, M. E., Weissman, S. M., Harkins, T. T., Gerstein, M. B., Egholm, M. and Snyder, M. Paired-end mapping reveals extensive structural variation in the human genome. *Science*, 318(5849):420–426, 2007. 14

Kozlitina, J., Smagris, E., Stender, S., Nordestgaard, B. G., Zhou, H. H., Tybjaerg-Hansen, A., Vogt, T. F., Hobbs, H. H. and Cohen, J. C. Exome-wide association study identifies a TM6SF2 variant that confers susceptibility to nonalcoholic fatty liver disease. *Nat Genet*, 46(4):352–356, 2014. 8

Krokstad, S., Langhammer, A., Hveem, K., Holmen, T., Midthjell, K., Stene, T., Bratberg, G., Heggland, J. and Holmen, J. Cohort Profile: the HUNT Study, Norway. *International Journal of Epidemiology*, 42(4):968–977, 2013. 132, 168

Kryukov, G. V., Shpunt, A., Stamatoyannopoulos, J. A. and Sunyaev, S. R. Power of deep, all-exon resequencing for discovery of human trait genes. *Proceedings of the National Academy of Sciences*, 106(10):3871–3876, 2009. 7

Ladouceur, M., Dastani, Z., Aulchenko, Y. S., Greenwood, C. M. T. and Richards, J. B. The empirical power of rare variant association methods: Results from sanger sequencing in 1,998 individuals. *PLoS Genet*, 8(2):e1002496, 2012. 4

Lange, L. A., Hu, Y., Zhang, H., Xue, C., Schmidt, E. M., Tang, Z.-Z., Bizon, C., Lange, E. M., Smith, J. D., Turner, E. H., Jun, G., Kang, H. M., Peloso, G., Auer, P., Li, K.-p., Flannick, J., Zhang, J., Fuchsberger, C., Gaulton, K., Lindgren, C., Locke, A., Manning, A., Sim, X., Rivas, M. A., Holmen, O. L., Gottesman, O., Lu, Y., Ruderfer, D., Stahl, E. A., Duan, Q., Li, Y., Durda, P., Jiao, S., Isaacs, A., Hofman, A., Bis, J. C., Correa, A., Griswold, M. E., Jakobsdottir, J., Smith, A. V., Schreiner, P. J., Feitosa, M. F., Zhang, Q., Huffman, J. E., Crosby, J., Wassel, C. L., Do, R., Franceschini, N., Martin, L. W., Robinson, J. G., Assimes, T. L., Crosslin, D. R., Rosenthal, E. A., Tsai, M., Rieder, M. J., Farlow, D. N., Folsom, A. R., Lumley, T., Fox, E. R., Carlson, C. S., Peters, U., Jackson, R. D., van Duijn, C. M., Uitterlinden, A., Levy, D., Rotter, J. I., Taylor, H. A., Gudnason Jr., V., Siscovick, D. S., Fornage, M., Borecki, I. B., Hayward, C., Rudan, I., Chen, Y. E., Bottinger, E. P., Loos, R. J. F., Sætrom, P., Hveem, K., Boehnke, M., Groop, L., McCarthy, M., Meitinger, T., Ballantyne, C. M., Gabriel, S. B., O'Donnell, C. J., Post, W. S., North, K. E., Reiner, A. P., Boerwinkle, E., Psaty, B. M., Altshuler, D., Kathiresan, S., Lin, D.-Y., Jarvik, G. P., Cupples, L. A., Kooperberg, C., Wilson, J. G., Nickerson, D. A., Abecasis, G. R., Rich, S. S., Tracy, R. P. and Willer, C. J. Whole-exome sequencing identifies rare and low-frequency coding variants associated with LDL cholesterol. *The American Journal of Human Genetics*, 94(2):233–245, 2014. 8, 19, 20

Lee, S., Abecasis, G. R., Boehnke, M. and Lin, X. Rare-variant association analysis: Study designs and statistical tests. *The American Journal of Human Genetics*, 95(1):5–23, 2014. 3

Lee, S., Teslovich, T. M., Boehnke, M. and Lin, X. General framework for meta-analysis of rare variants in sequencing association studies. *The American Journal of Human Genetics*, 93(1):42–53, 2013. 5

Lee, S., Wu, M. C. and Lin, X. Optimal tests for rare variant effects in sequencing association studies. *Biostatistics*, 13(4):762–775, 2012. 4

Li, B. and Leal, S. M. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *The American Journal of Human Genetics*, 83(3):311–321, 2008. 4

Li, H. Fermikit: assembly-based variant calling for illumina resequencing data. *Bioinformatics*, 31(22):3694–3696, 2015. 14

Li, Y., Willer, C. J., Ding, J., Scheet, P. and Abecasis, G. R. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genetic Epidemiology*, 34(8):816–834, 2010. 5

Lin, D.-Y. and Tang, Z.-Z. A general framework for detecting disease associations with rare variants in sequencing studies. *The American Journal of Human Genetics*, 89(3):354–367, 2011. 4

Lin, D. Y. and Zeng, D. Meta-analysis of genome-wide association studies: no efficiency gain in using individual participant data. *Genetic Epidemiology*, 34(1):60–66, 2010. 4

Liu, D. and Global Lipids Genetics Consortium. Statistical methods for population based studies; (Program Number 199). *Presented at the 64th Annual Meeting of The American Society of Human Genetics*, San Diego, CA, October 20, 2014. 137

Liu, D. J. and Leal, S. M. A novel adaptive method for the analysis of next-generation sequencing data to detect complex trait associations with rare variants due to gene main effects and interactions. *PLoS Genet*, 6(10):e1001156, 2010. 4

Liu, D. J., Peloso, G. M., Zhan, X., Holmen, O. L., Zawistowski, M., Feng, S., Nikpay, M., Auer, P. L., Goel, A., Zhang, H., Peters, U., Farrall, M., Orho-Melander, M., Kooperberg, C., McPherson, R., Watkins, H., Willer, C. J., Hveem, K., Melander, O., Kathiresan, S. and Abecasis, G. R. Meta-analysis of gene-level tests for rare variant association. *Nat Genet*, 46(2):200–204, 2014. 5

Lloyd-Jones, D., Adams, R. J., Brown, T. M., Carnethon, M., Dai, S., De Simone, G., Ferguson, T. B., Ford, E., Furie, K., Gillespie, C., Go, A., Greenlund, K., Haase, N., Hailpern, S., Ho, P. M., Howard, V., Kissela, B., Kittner, S., Lackland, D., Lisabeth, L., Marelli, A., McDermott, M. M., Meigs, J., Mozaffarian, D., Mussolino, M., Nichol, G., Roger, V. L., Rosamond, W., Sacco, R., Sorlie, P., Stafford, R., Thom, T., Wasserthiel-Smoller, S., Wong, N. D., Wylie-Rosett, J. and the American Heart Association Statistics Committee and Stroke Statistics Subcommittee. Heart disease and stroke statistics—2010 update: A report from the American Heart Association. *Circulation*, 121(7):e46–e215, 2010. 23

Lo, K. S., Vadlamudi, S., Fogarty, M. P., Mohlke, K. L. and Lettre, G. Strategies to fine-map genetic associations with lipid levels by combining epigenomic annotations and liver-specific transcription profiles. *Genomics*, 104(2):105–112, 2014. 9, 82, 83

Locke, A. E., Kahali, B., Berndt, S. I., Justice, A. E., Pers, T. H., Day, F. R., Powell, C., Vedantam, S., Buchkovich, M. L., Yang, J., Croteau-Chonka, D. C., Esko, T., Fall, T., Ferreira, T., Gustafsson, S., Kutalik, Z., Luan, J., Mägi, R., Randall, J. C., Winkler, T. W., Wood, A. R., Workalemahu, T., Faul, J. D., Smith, J. A., Hua Zhao, J., Zhao, W., Chen, J., Fehrmann, R., Hedman, Å. K., Karjalainen, J., Schmidt, E. M., Absher, D., Amin, N., Anderson, D., Beekman, M., Bolton, J. L., Bragg-Gresham, J. L., Buyske, S., Demirkan, A., Deng, G., Ehret, G. B., Feenstra, B., Feitosa, M. F., Fischer, K., Goel, A., Gong, J., Jackson, A. U., Kanoni, S., Kleber, M. E., Kristiansson, K., Lim, U., Lotay, V., Mangino, M., Mateo Leach, I., Medina-Gomez, C., Medland, S. E., Nalls, M. A., Palmer, C. D., Pasko, D., Pechlivanis, S., Peters, M. J., Prokopenko, I., Shungin, D., Stančáková, A., Strawbridge, R. J., Ju Sung, Y., Tanaka, T., Teumer, A., Trompet, S., van der Laan, S. W., van Setten, J., Van Vliet-Ostaptchouk, J. V., Wang, Z., Yengo, L., Zhang, W., Isaacs, A., Albrecht, E., Ärnlöv, J., Arscott, G. M., Attwood, A. P., Bandinelli, S., Barrett, A., Bas, I. N., Bellis, C., Bennett, A. J., Berne, C., Blagieva, R., Blüher, M., Böhringer, S., Bonnycastle, L. L., Böttcher, Y., Boyd, H. A., Bruinenberg, M., Caspersen, I. H., Ida Chen, Y.-D., Clarke, R., Daw, E. W., de Craen, A. J. M., Delgado, G., Dimitriou, M., Doney, A. S. F., Eklund, N., Estrada, K., Eury, E., Folkersen, L., Fraser, R. M., Garcia, M. E., Geller, F., Giedraitis, V., Gigante, B., Go, A. S., Golay, A., Goodall, A. H., Gordon, S. D., Gorski, M., Grabe, H.-J., Grallert, H., Grammer, T. B., Gräßler, J., Grönberg, H., Groves, C. J., Gusto, G., Haessler, J., Hall, P., Haller, T., Hallmans, G., Hartman, C. A., Hassinen, M., Hayward, C., Heard-Costa, N. L., Helmer, Q., Hengstenberg, C., Holmen, O., Hottenga, J.-J., James, A. L., Jeff, J. M., Johansson, Å., Jolley, J., Juliusdottir, T., Kinnunen, L., Koenig, W., Koskenvuo, M., Kratzer, W., Laitinen, J., Lamina, C., Leander, K., Lee, N. R., Lichtner, P., Lind, L., Lindström, J., Sin Lo, K., Lobbens, S., Lorbeer, R., Lu, Y., Mach, F., Magnusson, P. K. E., Mahajan, A., McArdle, W. L., McLachlan, S., Menni, C., Merger, S., Mihailov, E., Milani, L., Moayyeri, A., Monda, K. L., Morken, M. A., Mulas, A., Müller, G., Müller-Nurasyid, M., Musk, A. W., Nagaraja, R., Nöthen, M. M., Nolte, I. M., Pilz, S., Rayner, N. W., Renstrom, F., Rettig, R., Ried, J. S., Ripke, S., Robertson, N. R., Rose, L. M., Sanna, S., Scharnagl, H., Scholtens, S., Schumacher, F. R., Scott, W. R., Seufferlein, T., Shi, J., Vernon Smith, A., Smolonska, J., Stanton, A. V., Steinthorsdottir, V., Stirrups, K., Stringham, H. M., Sundström, J., Swertz, M. A., Swift, A. J., Syvänen, A.-C., Tan, S.-T., Tayo, B. O., Thorand, B., Thorleifsson, G., Tyrer, J. P., Uh, H.-W., Vandenput, L., Verhulst, F. C., Vermeulen, S. H., Verweij, N., Vonk, J. M., Waite, L. L., Warren, H. R., Waterworth, D., Weedon, M. N., Wilkens, L. R., Willenborg, C., Wilsgaard, T., Wojczynski, M. K., Wong, A., Wright, A. F., Zhang, Q., LifeLines Cohort Study, Brennan, E. P., Choi, M., Dastani, Z., Drong, A. W., Eriksson, P., Franco-Cereceda, A., Gådin, J. R., Gharavi, A. G., Goddard, M. E., Handsaker, R. E., Huang, J., Karpe, F., Kathiresan, S., Keildson, S., Kiryluk, K., Kubo, M., Lee, J.-Y., Liang, L., Lifton, R. P., Ma, B., McCarroll, S. A., McKnight, A. J., Min, J. L., Moffatt, M. F., Montgomery, G. W., Murabito, J. M., Nicholson, G., Nyholt, D. R., Okada, Y., Perry, J. R. B., Dorajoo, R., Reinmaa, E., Salem, R. M., Sandholm, N., Scott, R. A., Stolk, L., Takahashi, A., Tanaka, T., Van't Hooft, F. M., Vinkhuyzen, A. A. E., Westra, H.-J., Zheng, W., Zondervan, K. T., ADIPOGen Consortium, AGEN-BMI Working Group, CARDIOGRAMplusC4D Consortium, CKDGen Consortium, GLGC, ICBP, MAGIC Investigators, MuTHER Consortium, MIGen Consortium, PAGE Consortium, ReproGen Consortium,

GENIE Consortium, International Endogene Consortium, Heath, A. C., Arveiler, D., Bakker, S. J. L., Beilby, J., Bergman, R. N., Blangero, J., Bovet, P., Campbell, H., Caulfield, M. J., Cesana, G., Chakravarti, A., Chasman, D. I., Chines, P. S., Collins, F. S., Crawford, D. C., Cupples, L. A., Cusi, D., Danesh, J., de Faire, U., den Ruijter, H. M., Dominiczak, A. F., Erbel, R., Erdmann, J., Eriksson, J. G., Farrall, M., Felix, S. B., Ferrannini, E., Ferrières, J., Ford, I., Forouhi, N. G., Forrester, T., Franco, O. H., Gansevoort, R. T., Gejman, P. V., Gieger, C., Gottesman, O., Gudnason, V., Gyllensten, U., Hall, A. S., Harris, T. B., Hattersley, A. T., Hicks, A. A., Hindorff, L. A., Hingorani, A. D., Hofman, A., Homuth, G., Hovingh, G. K., Humphries, S. E., Hunt, S. C., Hyppönen, E., Illig, T., Jacobs, K. B., Jarvelin, M.-R., Jöckel, K.-H., Johansen, B., Jousilahti, P., Jukema, J. W., Jula, A. M., Kaprio,. Genetic studies of body mass index yield new insights for obesity biology. *Nature*, 518(7538):197–206, 2015. 21, 84, 87

Lupski, J. R., de Oca-Luna, R. M., Slaugenhaupt, S., Pentao, L., Guzzetta, V., Trask, B. J., Saucedo-Cardenas, O., Barker, D. F., Killian, J. M., Garcia, C. A., Chakravarti, A. and Patel, P. I. DNA duplication associated with Charcot-Marie-Tooth disease type 1A. *Cell*, 66(2):219–232, 1991. 15

Madsen, B. E. and Browning, S. R. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet*, 5(2):e1000384, 2009. 4

Maller, J. B., McVean, G., Byrnes, J., Vukcevic, D., Palin, K., Su, Z., Howson, J. M. M., Auton, A., Myers, S., Morris, A., Pirinen, M., Brown, M. A., Burton, P. R., Caulfield, M. J., Compston, A., Farrall, M., Hall, A. S., Hattersley, A. T., Hill, A. V. S., Mathew, C. G., Pembrey, M., Satsangi, J., Stratton, M. R., Worthington, J., Craddock, N., Hurles, M., Ouwehand, W., Parkes, M., Rahman, N., Duncanson, A., Todd, J. A., Kwiatkowski, D. P., Samani, N. J., Gough, S. C. L., McCarthy, M. I., Deloukas, P. and Donnelly, P. Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nat Genet*, 44(12):1294–1301, 2012. 165

Marchini, J., Howie, B., Myers, S., McVean, G. and Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet*, 39(7):906–913, 2007. 5

Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., Reynolds, A. P., Sandstrom, R., Qu, H., Brody, J., Shafer, A., Neri, F., Lee, K., Kutyavin, T., Stehling-Sun, S., Johnson, A. K., Canfield, T. K., Giste, E., Diegel, M., Bates, D., Hansen, R. S., Neph, S., Sabo, P. J., Heimfeld, S., Raubitschek, A., Ziegler, S., Cotsapas, C., Sotoodehnia, N., Glass, I., Sunyaev, S. R., Kaul, R. and Stamatoyannopoulos, J. A. Systematic localization of common disease-associated variation in regulatory DNA. *Science*, 337(6099):1190–1195, 2012. 82, 83, 89

McCarroll, S. A., Kuruvilla, F. G., Korn, J. M., Cawley, S., Nemesh, J., Wysoker, A., Shapero, M. H., de Bakker, P. I. W., Maller, J. B., Kirby, A., Elliott, A. L., Parkin, M., Hubbell, E., Webster, T., Mei, R., Veitch, J., Collins, P. J., Handsaker, R., Lincoln, S., Nizzari, M., Blume, J., Jones, K. W., Rava, R., Daly, M. J., Gabriel, S. B. and Altshuler, D. Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat Genet*, 40(10):1166–1174, 2008. 13, 131

McPherson, R., Pertsemlidis, A., Kavaslar, N., Stewart, A., Roberts, R., Cox, D. R., Hinds, D. A., Pennacchio, L. A., Tybjaerg-Hansen, A., Folsom, A. R., Boerwinkle, E., Hobbs, H. H. and Cohen, J. C. A common allele on chromosome 9 associated with coronary heart disease. *Science*, 316(5830):1488–1491, 2007. 131

Mills, R. E., Walter, K., Stewart, C., Handsaker, R. E., Chen, K., Alkan, C., Abyzov, A., Yoon, S. C., Ye, K., Cheetham, R. K., Chinwalla, A., Conrad, D. F., Fu, Y., Grubert, F., Hajirasouliha, I., Hormozdiari, F., Iakoucheva, L. M., Iqbal, Z., Kang, S., Kidd, J. M., Konkel, M. K., Korn, J., Khurana, E., Kural, D., Lam, H. Y. K., Leng, J., Li, R., Li, Y., Lin, C.-Y., Luo, R., Mu, X. J., Nemesh, J., Peckham, H. E., Rausch, T., Scally, A., Shi, X., Stromberg, M. P., Stutz, A. M., Urban, A. E., Walker, J. A., Wu, J., Zhang, Y., Zhang, Z. D., Batzer, M. A., Ding, L., Marth, G. T., McVean, G., Sebat, J., Snyder, M., Wang, J., Ye, K., Eichler, E. E., Gerstein, M. B.,

Hurles, M. E., Lee, C., McCarroll, S. A. and Korbel, J. O. Mapping copy number variation by population-scale genome sequencing. *Nature*, 470(7332):59–65, 2011. 13, 14, 132

Morgenthaler, S. and Thilly, W. G. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: A cohort allelic sums test (CAST). *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis*, 615(1–2):28–56, 2007. 4

Morris, A. P., Voight, B. F., Teslovich, T. M., Ferreira, T., Segrè, A. V., Steinthorsdottir, V., Strawbridge, R. J., Khan, H., Grallert, H., Mahajan, A., Prokopenko, I., Kang, H. M., Dina, C., Esko, T., Fraser, R. M., Kanoni, S., Kumar, A., Lagou, V., Langenberg, C., Luan, J., Lindgren, C. M., Müller-Nurasyid, M., Pechlivanis, S., Rayner, N. W., Scott, L. J., Wiltshire, S., Yengo, L., Kinnunen, L., Rossin, E. J., Raychaudhuri, S., Johnson, A. D., Dimas, A. S., Loos, R. J. F., Vedantam, S., Chen, H., Florez, J. C., Fox, C., Liu, C.-T., Rybin, D., Couper, D. J., Kao, W. H. L., Li, M., Cornelis, M. C., Kraft, P., Sun, Q., van Dam, R. M., Stringham, H. M., Chines, P. S., Fischer, K., Fontanillas, P., Holmen, O. L., Hunt, S. E., Jackson, A. U., Kong, A., Lawrence, R., Meyer, J., Perry, J. R. B., Platou, C. G. P., Potter, S., Rehnberg, E., Robertson, N., Sivapalaratnam, S., Stančáková, A., Stirrups, K., Thorleifsson, G., Tikkanen, E., Wood, A. R., Almgren, P., Atalay, M., Benediktsson, R., Bonnycastle, L. L., Burtt, N., Carey, J., Charpentier, G., Crenshaw, A. T., Doney, A. S. F., Dorkhan, M., Edkins, S., Emilsson, V., Eury, E., Forsen, T., Gertow, K., Gigante, B., Grant, G. B., Groves, C. J., Guiducci, C., Herder, C., Hreidarsson, A. B., Hui, J., James, A., Jonsson, A., Rathmann, W., Klopp, N., Kravic, J., Krjutškov, K., Langford, C., Leander, K., Lindholm, E., Lobbens, S., Männistö, S., Mirza, G., Mühleisen, T. W., Musk, B., Parkin, M., Rallidis, L., Saramies, J., Sennblad, B., Shah, S., Sigursson, G., Silveira, A., Steinbach, G., Thorand, B., Trakalo, J., Veglia, F., Wennauer, R., Winckler, W., Zabaneh, D., Campbell, H., van Duijn, C., Uitterlinden, A. G., Hofman, A., Sijbrands, E., Abecasis, G. R., Owen, K. R., Zeggini, E., Trip, M. D., Forouhi, N. G., Syvänen, A.-C., Eriksson, J. G., Peltonen, L., Nöthen, M. M., Balkau, B., Palmer, C. N. A., Lyssenko, V., Tuomi, T., Isomaa, B., Hunter, D. J., Qi, L., Wellcome Trust Case Control Consortium, Meta-Analyses of Glucose and Insulin-related traits Consortium (MAGIC) Investigators, Genetic Investigation of ANthropometric Traits (GIANT) Consortium, Asian Genetic Epidemiology Network–Type 2 Diabetes (AGEN-T2D) Consortium, South Asian Type 2 Diabetes (SAT2D) Consortium, Shuldiner, A. R., Roden, M., Barroso, I., Wilsgaard, T., Beilby, J., Hovingh, K., Price, J. F., Wilson, J. F., Rauramaa, R., Lakka, T. A., Lind, L., Dedoussis, G., Njølstad, I., Pedersen, N. L., Khaw, K.-T., Wareham, N. J., Keinanen-Kiukaanniemi, S. M., Saaristo, T. E., Korpi-Hyövälti, E., Saltevo, J., Laakso, M., Kuusisto, J., Metspalu, A., Collins, F. S., Mohlke, K. L., Bergman, R. N., Tuomilehto, J., Boehm, B. O., Gieger, C., Hveem, K., Cauchi, S., Froguel, P., Baldassarre, D., Tremoli, E., Humphries, S. E., Saleheen, D., Danesh, J., Ingelsson, E., Ripatti, S., Salomaa, V., Erbel, R., Jöckel, K.-H., Moebus, S., Peters, A., Illig, T., de Faire, U., Hamsten, A., Morris, A. D., Donnelly, P. J., Frayling, T. M., Hattersley, A. T., Boerwinkle, E., Melander, O., Kathiresan, S., Nilsson, P. M., Deloukas, P., Thorsteinsdottir, U., Groop, L. C., Stefansson, K., Hu, F., Pankow, J. S., Dupuis, J., Meigs, J. B., Altshuler, D., Boehnke, M., McCarthy, M. I. and DIAbetes Genetics Replication and Meta-analysis (DIAGRAM) Consortium. Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat Genet*, 44(9):981–90, 2012. 21, 84, 87

Morris, A. P. and Zeggini, E. An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genetic Epidemiology*, 34(2):188–193, 2010. 4

Morrison, A., Voorman, A., Johnson, A., Liu, X., Yu, J., Li, A., Muzny, D., Yu, F., Rice, K., Zhu, C., Bis, J., Heiss, G., O'Donnell, C., Psaty, B., Cupples, L., Gibbs, R., E, B. and the Cohorts for Heart and Aging Research in Genetic Epidemiology (CHARGE) Consortium. Whole-genome sequence-based analysis of high-density lipoprotein cholesterol. *Nat Genet*, 45(8):899–901, 2013. 9

Mozaffarian, D., Benjamin, E. J., Go, A. S., Arnett, D. K., Blaha, M. J., Cushman, M., de Ferranti, S., Després, J.-P., Fullerton, H. J., Howard, V. J., Huffman, M. D., Judd, S. E., Kissela, B. M.,

Lackland, D. T., Lichtman, J. H., Lisabeth, L. D., Liu, S., Mackey, R. H., Matchar, D. B., McGuire, D. K., Mohler, E. R., Moy, C. S., Muntner, P., Mussolino, M. E., Nasir, K., Neumar, R. W., Nichol, G., Palaniappan, L., Pandey, D. K., Reeves, M. J., Rodriguez, C. J., Sorlie, P. D., Stein, J., Towfighi, A., Turan, T. N., Virani, S. S., Willey, J. Z., Woo, D., Yeh, R. W. and Turner, M. B. Heart disease and stroke statistics—2015 update: A report from the American Heart Association. *Circulation*, 131(4):e29–e322, 2015. 6, 131

Musunuru, K., Strong, A., Frank-Kamenetsky, M., Lee, N. E., Ahfeldt, T., Sachs, K. V., Li, X., Li, H., Kuperwasser, N., Ruda, V. M., Pirruccello, J. P., Muchmore, B., Prokunina-Olsson, L., Hall, J. L., Schadt, E. E., Morales, C. R., Lund-Katz, S., Phillips, M. C., Wong, J., Cantley, W., Racie, T., Ejebe, K. G., Orho-Melander, M., Melander, O., Koteliansky, V., Fitzgerald, K., Krauss, R. M., Cowan, C. A., Kathiresan, S. and Rader, D. J. From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature*, 466(7307):714–9, 2010. 7, 23, 31, 33, 91, 92

Myocardial Infarction Genetics Consortium, Kathiresan, S., Voight, B. F., Purcell, S., Musunuru, K., Ardissino, D., Mannucci, P. M., Anand, S., Engert, J. C., Samani, N. J., Schunkert, H., Erdmann, J., Reilly, M. P., Rader, D. J., Morgan, T., Spertus, J. A., Stoll, M., Girelli, D., McKeown, P. P., Patterson, C. C., Siscovick, D. S., O'Donnell, C. J., Elosua, R., Peltonen, L., Salomaa, V., Schwartz, S. M., Melander, O., Altshuler, D., Ardissino, D., Merlini, P. A., Berzuini, C., Bernardinelli, L., Peyvandi, F., Tubaro, M., Celli, P., Ferrario, M., Fetiveau, R., Marziliano, N., Casari, G., Galli, M., Ribichini, F., Rossi, M., Bernardi, F., Zonzin, P., Piazza, A., Mannucci, P. M., Schwartz, S. M., Siscovick, D. S., Yee, J., Friedlander, Y., Elosua, R., Marrugat, J., Lucas, G., Subirana, I., Sala, J., Ramos, R., Kathiresan, S., Meigs, J. B., Williams, G., Nathan, D. M., MacRae, C. A., O'Donnell, C. J., Salomaa, V., Havulinna, A. S., Peltonen, L., Melander, O., Berglund, G., Voight, B. F., Kathiresan, S., Hirschhorn, J. N., Asselta, R., Duga, S., Spreafico, M., Musunuru, K., Daly, M. J., Purcell, S., Voight, B. F., Purcell, S., Nemesh, J., Korn, J. M., McCarroll, S. A., Schwartz, S. M., Yee, J., Kathiresan, S., Lucas, G., Subirana, I., Elosua, R., Surti, A., Guiducci, C., Gianniny, L., Mirel, D., Parkin, M., Burtt, N., Gabriel, S. B., Samani, N. J., Thompson, J. R., Braund, P. S., Wright, B. J., Balmforth, A. J., Ball, S. G., Hall, A., Wellcome Trust Case Control Consortium, Schunkert, H., Erdmann, J., Linsel-Nitschke, P., Lieb, W., Ziegler, A., König, I., Hengstenberg, C., Fischer, M., Stark, K., Grosshennig, A., Preuss, M., Wichmann, H. E., Schreiber, S., Schunkert, H., Samani, N. J., Erdmann, J., Ouwehand, W., Hengstenberg, C., Deloukas, P., Scholz, M., Cambien, F., Reilly, M. P., Li, M., Chen, Z., Wilensky, R., Matthai, W., Qasim, A., Hakonarson, H. H., Devaney, J., Burnett, M. S., Pichard, A. D., Kent, K. M., Satler, L., Lindsay, J. M., Waksman, R., Knouff, C. W., Waterworth, D. M., Walker, M. C., Mooser, V., Epstein, S. E., Rader, D. J., Scheffold, T., Berger, K., Stoll, M., Huge, A., Girelli, D., Martinelli, N., Olivieri, O., Corrocher, R., Morgan, T., Spertus, J. A., McKeown, P., Patterson, C. C., Schunkert, H., Erdmann, E., Linsel-Nitschke, P., Lieb, W., Ziegler, A., König, I. R., Hengstenberg, C., Fischer, M., Stark, K., Grosshennig, A., Preuss, M., Wichmann, H. E., Schreiber, S., Hólm, H., Thorleifsson, G., Thorsteinsdottir, U., Stefansson, K., Engert, J. C., Do, R., Xie, C., Anand, S., Kathiresan, S., Ardissino, D., Mannucci, P. M., Siscovick, D., O'Donnell, C. J., Samani, N. J., Melander, O., Elosua, R., Peltonen, L., Salomaa, V., Schwartz, S. M. and Altshuler, D. Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. *Nat Genet*, 41(3):334–341, 2009. 15, 131

Myocardial Infarction Genetics Consortium Investigators, Stitziel, N. O., Won, H. H., Morrison, A. C., Peloso, G. M., Do, R., Lange, L. A., Fontanillas, P., Gupta, N., Duga, S., Goel, A., Farrall, M., Saleheen, D., Ferrario, P., König, I., Asselta, R., Merlini, P. A., Marziliano, N., Notarangelo, M. F., Schick, U., Auer, P., Assimes, T. L., Reilly, M., Wilensky, R., Rader, D. J., Hovingh, G. K., Meitinger, T., Kessler, T., Kastrati, A., Laugwitz, K. L., Siscovick, D., Rotter, J. I., Hazen, S. L., Tracy, R., Cresci, S., Spertus, J., Jackson, R., Schwartz, S. M., Natarajan, P., Crosby, J., Muzny, D., Ballantyne, C., Rich, S. S., O'Donnell, C. J., Abecasis, G., Sunyaev, S., Nickerson, D. A., Buring, J. E., Ridker, P. M., Chasman, D. I., Austin, E., Ye, Z., Kullo, I. J., Weeke, P. E., Shaffer, C. M., Bastarache, L. A., Denny, J. C., Roden, D. M., Palmer, C.,

Deloukas, P., Lin, D. Y., Tang, Z. Z., Erdmann, J., Schunkert, H., Danesh, J., Marrugat, J., Elosua, R., Ardissino, D., McPherson, R., Watkins, H., Reiner, A. P., Wilson, J. G., Altshuler, D., Gibbs, R. A., Lander, E. S., Boerwinkle, E., Gabriel, S. and Kathiresan, S. Inactivating mutations in NPC1L1 and protection from coronary heart disease. *New England Journal of Medicine*, 371(22):2072–2082, 2014. 10, 19

Neale, B. M., Rivas, M. A., Voight, B. F., Altshuler, D., Devlin, B., Orho-Melander, M., Kathiresan, S., Purcell, S. M., Roeder, K. and Daly, M. J. Testing for an unusual distribution of rare variants. *PLoS Genet*, 7(3):e1001322, 2011. 4

Palmen, J., Smith, A. J. P., Dorfmeister, B., Putt, W., Humphries, S. E. and Talmud, P. J. The functional interaction on in vitro gene expression of APOA5 SNPs, defining haplotype APOA52, and their paradoxical association with plasma triglyceride but not plasma apoAV levels. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, 1782(7–8):447–452, 2008. 31

Pan, W. Asymptotic tests of association with multiple SNPs in linkage disequilibrium. *Genetic Epidemiology*, 33(6):497–507, 2009. 4

Pang, A. W., MacDonald, J. R., Pinto, D., Wei, J., Rafiq, M. A., Conrad, D. F., Park, H., Hurles, M. E., Lee, C., Venter, J. C., Kirkness, E. F., Levy, S., Feuk, L. and Scherer, S. W. Towards a comprehensive structural variation map of an individual human genome. *Genome Biology*, 11(5):R52–R52, 2010. 13

Parker, S. C. J., Stitzel, M. L., Taylor, D. L., Orozco, J. M., Erdos, M. R., Akiyama, J. A., van Bueren, K. L., Chines, P. S., Narisu, N., NISC Comparative Sequencing Program, Black, B. L., Visel, A., Pennacchio, L. A., Collins, F. S., National Institutes of Health Intramural Sequencing Center Comparative Sequencing Program Authors and NISC Comparative Sequencing Program Authors. Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. *Proc Natl Acad Sci U S A*, 110(44):17921–6, 2013. 82, 87

Peloso, G. M., Auer, P. L., Bis, J. C., Voorman, A., Morrison, A. C., Stitziel, N. O., Brody, J. A., Khetarpal, S. A., Crosby, J. R., Fornage, M., Isaacs, A., Jakobsdottir, J., Feitosa, M. F., Davies, G., Huffman, J. E., Manichaikul, A., Davis, B., Lohman, K., Joon, A. Y., Smith, A. V., Grove, M. L., Zanoni, P., Redon, V., Demissie, S., Lawson, K., Peters, U., Carlson, C., Jackson, R. D., Ryckman, K. K., Mackey, R. H., Robinson, J. G., Siscovick, D. S., Schreiner, P. J., Mychaleckyj, J. C., Pankow, J. S., Hofman, A., Uitterlinden, A. G., Harris, T. B., Taylor, K. D., Stafford, J. M., Reynolds, L. M., Marioni, R. E., Dehghan, A., Franco, O. H., Patel, A. P., Lu, Y., Hindy, G., Gottesman, O., Bottinger, E. P., Melander, O., Orho-Melander, M., Loos, R. J. F., Duga, S., Merlini, P. A., Farrall, M., Goel, A., Asselta, R., Girelli, D., Martinelli, N., Shah, S. H., Kraus, W. E., Li, M., Rader, D. J., Reilly, M. P., McPherson, R., Watkins, H., Ardissino, D., Zhang, Q., Wang, J., Tsai, M. Y., Taylor, H. A., Correa, A., Griswold, M. E., Lange, L. A., Starr, J. M., Rudan, I., Eiriksdottir, G., Launer, L. J., Ordovas, J. M., Levy, D., Chen, Y. D. I., Reiner, A. P., Hayward, C., Polasek, O., Deary, I. J., Borecki, I. B., Liu, Y., Gudnason, V., Wilson, J. G., van Duijn, C. M., Kooperberg, C., Rich, S. S., Psaty, B. M., Rotter, J. I., O'Donnell, C. J., Rice, K., Boerwinkle, E., Kathiresan, S. and Cupples, L. A. Association of low-frequency and rare coding-sequence variants with blood lipids and coronary heart disease in 56,000 whites and blacks. *The American Journal of Human Genetics*, 94(2):223–232, 2014. 11, 18

Peters, M. N., Moscona, J. C., Katz, M. J., Deandrade, K. B., Quevedo, H. C., Tiwari, S., Burchett, A. R., Turnage, T. A., Singh, K. Y., Fomunung, E. N., Srivastav, S., Delafontaine, P. and Irimpen, A. M. Natural disasters and myocardial infarction: The six years after hurricane katrina. *Mayo Clinic Proceedings*, 89(4):472–477, 2014. 165

Pickrell, J. K. Joint analysis of functional genomic data and genome-wide association studies of 18 human traits. *The American Journal of Human Genetics*, 94(4):559–573, 2014. 82, 83

Plyte, S. E., Hughes, K., Nikolakaki, E., Pulverer, B. J. and Woodgett, J. R. Glycogen synthase kinase-3: functions in oncogenesis and development. *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*, 1114(2–3):147–162, 1992. 27

Price, A. L., Kryukov, G. V., de Bakker, P. I. W., Purcell, S. M., Staples, J., Wei, L.-J. and Sunyaev, S. R. Pooled association tests for rare variants in exon-resequencing studies. *The American Journal of Human Genetics*, 86(6):832–838, 2010. 4

Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A. and Reich, D. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*, 38(8):904–909, 2006. 2, 38

Pritchard, J. K. and Cox, N. J. The allelic architecture of human disease genes: common disease–common variant...or not? *Human Molecular Genetics*, 11(20):2417–2423, 2002. 7

Rahalkar, A. R. and Hegele, R. A. Monogenic pediatric dyslipidemias: Classification, genetics and clinical spectrum. *Molecular Genetics and Metabolism*, 93(3):282–294, 2008. 23

Rausch, T., Zichner, T., Schlattl, A., Stütz, A. M., Benes, V. and Korbel, J. O. DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, 28(18):i333–i339, 2012. 14, 134

Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M. J., Amin, V., Whitaker, J. W., Schultz, M. D., Ward, L. D., Sarkar, A., Quon, G., Sandstrom, R. S., Eaton, M. L., Wu, Y.-C., Pfenning, A. R., Wang, X., Claussnitzer, M., Liu, Y., Coarfa, C., Harris, R. A., Shoresh, N., Epstein, C. B., Gjoneska, E., Leung, D., Xie, W., Hawkins, R. D., Lister, R., Hong, C., Gascard, P., Mungall, A. J., Moore, R., Chuah, E., Tam, A., Canfield, T. K., Hansen, R. S., Kaul, R., Sabo, P. J., Bansal, M. S., Carles, A., Dixon, J. R., Farh, K.-H., Feizi, S., Karlic, R., Kim, A.-R., Kulkarni, A., Li, D., Lowdon, R., Elliott, G., Mercer, T. R., Neph, S. J., Onuchic, V., Polak, P., Rajagopal, N., Ray, P., Sallari, R. C., Siebenthall, K. T., Sinnott-Armstrong, N. A., Stevens, M., Thurman, R. E., Wu, J., Zhang, B., Zhou, X., Beaudet, A. E., Boyer, L. A., De Jager, P. L., Farnham, P. J., Fisher, S. J., Haussler, D., Jones, S. J. M., Li, W., Marra, M. A., McManus, M. T., Sunyaev, S., Thomson, J. A., Tlsty, T. D., Tsai, L.-H., Wang, W., Waterland, R. A., Zhang, M. Q., Chadwick, L. H., Bernstein, B. E., Costello, J. F., Ecker, J. R., Hirst, M., Meissner, A., Milosavljevic, A., Ren, B., Stamatoyannopoulos, J. A., Wang, T. and Kellis, M. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–330, 2015. 12

Roberts, R. A genetic basis for coronary artery disease. *Trends in Cardiovascular Medicine*, 25(3):171–178, 2015. 131

Rosenthal, E. A., Ranchalis, J., Crosslin, D. R., Burt, A., Brunzell, J. D., Motulsky, A. G., Nickerson, D. A., NHLBI GO Exome Sequencing Project, Wijsman, E. M. and Jarvik, G. P. Joint linkage and association analysis with exome sequence data implicates SLC25A40 in hypertriglyceridemia. *The American Journal of Human Genetics*, 93(6):1035–1045, 2013. 7, 19

Rossin, E. J., Lage, K., Raychaudhuri, S., Xavier, R. J., Tatar, D., Benita, Y., Cotsapas, C., Daly, M. J. and International Inflammatory Bowel Disease Genetics Constortium. Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology. *PLoS Genet*, 7(1):e1001273, 2011. 27

Samani, N. J., Erdmann, J., Hall, A. S., Hengstenberg, C., Mangino, M., Mayer, B., Dixon, R. J., Meitinger, T., Braund, P., Wichmann, H. E., Barrett, J. H., König, I. R., Stevens, S. E., Szymczak, S., Tregouet, D.-A., Iles, M. M., Pahlke, F., Pollard, H., Lieb, W., Cambien, F., Fischer, M., Ouwehand, W., Blankenberg, S., Balmforth, A. J., Baessler, A., Ball, S. G., Strom, T. M., Brænne, I., Gieger, C., Deloukas, P., Tobin, M. D., Ziegler, A., Thompson, J. R. and Schunkert, H. Genomewide association analysis of coronary artery disease. *New England Journal of Medicine*, 357(5):443–453, 2007. 131

Sanna, S., Li, B., Mulas, A., Sidore, C., Kang, H. M., Jackson, A. U., Piras, M. G., Usala, G., Maninchedda, G., Sassu, A., Serra, F., Palmas, M. A., Wood, William H., I., Njølstad, I., Laakso, M., Hveem, K., Tuomilehto, J., Lakka, T. A., Rauramaa, R., Boehnke, M., Cucca, F., Uda, M., Schlessinger, D., Nagaraja, R. and Abecasis, G. R. Fine mapping of five loci associated with low-density lipoprotein cholesterol detects variants that double the explained heritability. *PLoS Genet*, 7(7):e1002198, 2011. 9, 24, 30, 31

Sarria, A. J., Panini, S. R. and Evans, R. M. A functional role for vimentin intermediate filaments in the metabolism of lipoprotein-derived cholesterol in human SW-13 cells. *Journal of Biological Chemistry*, 267(27):19455–19463, 1992. 25

Sarwar, N., Danesh, J., Eiriksdottir, G., Sigurdsson, G., Wareham, N., Bingham, S., Boekholdt, S. M., Khaw, K.-T. and Gudnason, V. Triglycerides and the risk of coronary heart disease: 10,158 incident cases among 262,525 participants in 29 Western prospective studies. *Circulation*, 115(4):450–458, 2007. 10

Schadt, E. E., Molony, C., Chudin, E., Hao, K., Yang, X., Lum, P. Y., Kasarskis, A., Zhang, B., Wang, S., Suver, C., Zhu, J., Millstein, J., Sieberts, S., Lamb, J., GuhaThakurta, D., Derry, J., Storey, J. D., Avila-Campillo, I., Kruger, M. J., Johnson, J. M., Rohl, C. A., van Nas, A., Mehrabian, M., Drake, T. A., Lusis, A. J., Smith, R. C., Guengerich, F. P., Strom, S. C., Schuetz, E., Rushmore, T. H. and Ulrich, R. Mapping the genetic architecture of gene expression in human liver. *PLoS Biol*, 6(5):e107, 2008. 42, 91

Schaub, M. A., Boyle, A. P., Kundaje, A., Batzoglou, S. and Snyder, M. Linking disease associations with regulatory information in the human genome. *Genome Res*, 22(9):1748–59, 2012. 83

Schmidt, E. M. and Willer, C. J. Insights into blood lipids from rare variant discovery. *Current Opinion in Genetics & Development*, 33:25–31, 2015. 1

Schmidt, E. M., Zhang, J., Zhou, W., Chen, J., Mohlke, K. L., Chen, Y. E. and Willer, C. J. GREGOR: evaluating global enrichment of trait-associated variants in epigenomic features using a systematic, data-driven approach. *Bioinformatics*, 31(16):2601–2606, 2015. 9, 81, 163

Schunkert, H., König, I. R., Kathiresan, S., Reilly, M. P., Assimes, T. L., Holm, H., Preuss, M., Stewart, A. F. R., Barbalic, M., Gieger, C., Absher, D., Aherrahrou, Z., Allayee, H., Altshuler, D., Anand, S. S., Andersen, K., Anderson, J. L., Ardissino, D., Ball, S. G., Balmforth, A. J., Barnes, T. A., Becker, D. M., Becker, L. C., Berger, K., Bis, J. C., Boekholdt, S. M., Boerwinkle, E., Braund, P. S., Brown, M. J., Burnett, M. S., Buysschaert, I., Cardiogenics, Carlquist, J. F., Chen, L., Cichon, S., Codd, V., Davies, R. W., Dedoussis, G., Dehghan, A., Demissie, S., Devaney, J. M., Diemert, P., Do, R., Doering, A., Eifert, S., Mokhtari, N. E. E., Ellis, S. G., Elosua, R., Engert, J. C., Epstein, S. E., de Faire, U., Fischer, M., Folsom, A. R., Freyer, J., Gigante, B., Girelli, D., Gretarsdottir, S., Gudnason, V., Gulcher, J. R., Halperin, E., Hammond, N., Hazen, S. L., Hofman, A., Horne, B. D., Illig, T., Iribarren, C., Jones, G. T., Jukema, J. W., Kaiser, M. A., Kaplan, L. M., Kastelein, J. J. P., Khaw, K.-T., Knowles, J. W., Kolovou, G., Kong, A., Laaksonen, R., Lambrechts, D., Leander, K., Lettre, G., Li, M., Lieb, W., Loley, C., Lotery, A. J., Mannucci, P. M., Maouche, S., Martinelli, N., McKeown, P. P., Meisinger, C., Meitinger, T., Melander, O., Merlini, P. A., Mooser, V., Morgan, T., Mühleisen, T. W., Muhlestein, J. B., Münzel, T., Musunuru, K., Nahrstaedt, J., Nelson, C. P., Nöthen, M. M., Olivieri, O., Patel, R. S., Patterson, C. C., Peters, A., Peyvandi, F., Qu, L., Quyyumi, A. A., Rader, D. J., Rallidis, L. S., Rice, C., Rosendaal, F. R., Rubin, D., Salomaa, V., Sampietro, M. L., Sandhu, M. S., Schadt, E., Schäfer, A., Schillert, A., Schreiber, S., Schrezenmeir, J., Schwartz, S. M., Siscovick, D. S., Sivananthan, M., Sivapalaratnam, S., Smith, A., Smith, T. B., Snoep, J. D., Soranzo, N., Spertus, J. A., Stark, K., Stirrups, K., Stoll, M., Tang, W. H. W., Tennstedt, S., Thorgeirsson, G., Thorleifsson, G., Tomaszewski, M., Uitterlinden, A. G., van Rij, A. M., Voight, B. F., Wareham, N. J., Wells, G. A., Wichmann, H.-E., Wild, P. S., Willenborg, C., Witteman, J. C. M., Wright, B. J., Ye, S., Zeller, T., Ziegler, A., Cambien, F., Goodall, A. H., Cupples, L. A., Quertermous,

T., März, W., Hengstenberg, C., Blankenberg, S., Ouwehand, W. H., Hall, A. S., Deloukas, P., Thompson, J. R., Stefansson, K., Roberts, R., Thorsteinsdottir, U., O'Donnell, C. J., McPherson, R., Erdmann, J., CARDIoGRAM Consortium and Samani, N. J. Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nat Genet*, 43(4):333–8, 2011. 31, 84, 87, 131

Scott, R. A., Lagou, V., Welch, R. P., Wheeler, E., Montasser, M. E., Luan, J., Magi, R., Strawbridge, R. J., Rehnberg, E., Gustafsson, S., Kanoni, S., Rasmussen-Torvik, L. J., Yengo, L., Lecoeur, C., Shungin, D., Sanna, S., Sidore, C., Johnson, P. C. D., Jukema, J. W., Johnson, T., Mahajan, A., Verweij, N., Thorleifsson, G., Hottenga, J.-J., Shah, S., Smith, A. V., Sennblad, B., Gieger, C., Salo, P., Perola, M., Timpson, N. J., Evans, D. M., Pourcain, B. S., Wu, Y., Andrews, J. S., Hui, J., Bielak, L. F., Zhao, W., Horikoshi, M., Navarro, P., Isaacs, A., O'Connell, J. R., Stirrups, K., Vitart, V., Hayward, C., Esko, T., Mihailov, E., Fraser, R. M., Fall, T., Voight, B. F., Raychaudhuri, S., Chen, H., Lindgren, C. M., Morris, A. P., Rayner, N. W., Robertson, N., Rybin, D., Liu, C.-T., Beckmann, J. S., Willems, S. M., Chines, P. S., Jackson, A. U., Kang, H. M., Stringham, H. M., Song, K., Tanaka, T., Peden, J. F., Goel, A., Hicks, A. A., An, P., Muller-Nurasyid, M., Franco-Cereceda, A., Folkersen, L., Marullo, L., Jansen, H., Oldehinkel, A. J., Bruinenberg, M., Pankow, J. S., North, K. E., Forouhi, N. G., Loos, R. J. F., Edkins, S., Varga, T. V., Hallmans, G., Oksa, H., Antonella, M., Nagaraja, R., Trompet, S., Ford, I., Bakker, S. J. L., Kong, A., Kumari, M., Gigante, B., Herder, C., Munroe, P. B., Caulfield, M., Antti, J., Mangino, M., Small, K., Miljkovic, I., Liu, Y., Atalay, M., Kiess, W., James, A. L., Rivadeneira, F., Uitterlinden, A. G., Palmer, C. N. A., Doney, A. S. F., Willemsen, G., Smit, J. H., Campbell, S., Polasek, O., Bonnycastle, L. L., Hercberg, S., Dimitriou, M., Bolton, J. L., Fowkes, G. R., Kovacs, P., Lindstrom, J., Zemunik, T., Bandinelli, S., Wild, S. H., Basart, H. V., Rathmann, W., Grallert, H., Maerz, W., Kleber, M. E., Boehm, B. O., Peters, A., Pramstaller, P. P., Province, M. A., Borecki, I. B., Hastie, N. D., Rudan, I., Campbell, H., Watkins, H., Farrall, M., Stumvoll, M., Ferrucci, L., Waterworth, D. M., Bergman, R. N., Collins, F. S., Tuomilehto, J., Watanabe, R. M., de Geus, E. J. C., Penninx, B. W., Hofman, A., Oostra, B. A., Psaty, B. M., Vollenweider, P., Wilson, J. F., Wright, A. F., Hovingh, G. K., Metspalu, A., Uusitupa, M., Magnusson, P. K. E., Kyvik, K. O., Kaprio, J., Price, J. F., Dedoussis, G. V., Deloukas, P., Meneton, P., Lind, L., Boehnke, M., Shuldiner, A. R., van Duijn, C. M., Morris, A. D., Toenjes, A., Peyser, P. A., Beilby, J. P., Korner, A., Kuusisto, J., Laakso, M., Bornstein, S. R., Schwarz, P. E. H., Lakka, T. A., Rauramaa, R., Adair, L. S., Smith, G. D., Spector, T. D., Illig, T., de Faire, U., Hamsten, A., Gudnason, V., Kivimaki, M., Hingorani, A., Keinanen-Kiukaanniemi, S. M., Saaristo, T. E., Boomsma, D. I., Stefansson, K., van der Harst, P., Dupuis, J., Pedersen, N. L., Sattar, N., Harris, T. B., Cucca, F., Ripatti, S., Salomaa, V., Mohlke, K. L., Balkau, B., Froguel, P., Pouta, A., Jarvelin, M.-R., Wareham, N. J., Bouatia-Naji, N., McCarthy, M. I., Franks, P. W., Meigs, J. B., Teslovich, T. M., Florez, J. C., Langenberg, C., Ingelsson, E., Prokopenko, I. and Barroso, I. Large-scale association analyses identify new loci influencing glycemic traits and provide insight into the underlying biological pathways. *Nat Genet*, 44(9):991–1005, 2012. 21

Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., Walsh, T., Yamrom, B., Yoon, S., Krasnitz, A., Kendall, J., Leotta, A., Pai, D., Zhang, R., Lee, Y.-H., Hicks, J., Spence, S. J., Lee, A. T., Puura, K., Lehtimäki, T., Ledbetter, D., Gregersen, P. K., Bregman, J., Sutcliffe, J. S., Jobanputra, V., Chung, W., Warburton, D., King, M.-C., Skuse, D., Geschwind, D. H., Gilliam, T. C., Ye, K. and Wigler, M. Strong association of de novo copy number mutations with autism. *Science*, 316(5823):445–449, 2007. 15

Segrè, A. V., Groop, L., Mootha, V. K., Daly, M. J., Altshuler, D., DIAGRAM Consortium and MAGIC investigators. Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits. *PLoS Genet*, 6(8):e1001058, 2010. 26, 41

Soccio, R. E., Chen, E. R., Rajapurkar, S. R., Safabakhsh, P., Marinis, J. M., Dispirito, J. R., Emmett, M. J., Briggs, E. R., Fang, B., Everett, L. J., Lim, H.-W., Won, K.-J., Steger, D. J.,

Wu, Y., Civelek, M., Voight, B. F. and Lazar, M. A. Genetic variation determines PPARγ function and anti-diabetic drug response in vivo. *Cell*, 162(1):33–44, 2015. 12

Speliotes, E. K., Willer, C. J., Berndt, S. I., Monda, K. L., Thorleifsson, G., Jackson, A. U., Allen, H. L., Lindgren, C. M., Luan, J., Magi, R., Randall, J. C., Vedantam, S., Winkler, T. W., Qi, L., Workalemahu, T., Heid, I. M., Steinthorsdottir, V., Stringham, H. M., Weedon, M. N., Wheeler, E., Wood, A. R., Ferreira, T., Weyant, R. J., Segre, A. V., Estrada, K., Liang, L., Nemesh, J., Park, J.-H., Gustafsson, S., Kilpelainen, T. O., Yang, J., Bouatia-Naji, N., Esko, T., Feitosa, M. F., Kutalik, Z., Mangino, M., Raychaudhuri, S., Scherag, A., Smith, A. V., Welch, R., Zhao, J. H., Aben, K. K., Absher, D. M., Amin, N., Dixon, A. L., Fisher, E., Glazer, N. L., Goddard, M. E., Heard-Costa, N. L., Hoesel, V., Hottenga, J.-J., Johansson, A., Johnson, T., Ketkar, S., Lamina, C., Li, S., Moffatt, M. F., Myers, R. H., Narisu, N., Perry, J. R. B., Peters, M. J., Preuss, M., Ripatti, S., Rivadeneira, F., Sandholt, C., Scott, L. J., Timpson, N. J., Tyrer, J. P., van Wingerden, S., Watanabe, R. M., White, C. C., Wiklund, F., Barlassina, C., Chasman, D. I., Cooper, M. N., Jansson, J.-O., Lawrence, R. W., Pellikka, N., Prokopenko, I., Shi, J., Thiering, E., Alavere, H., Alibrandi, M. T. S., Almgren, P., Arnold, A. M., Aspelund, T., Atwood, L. D., Balkau, B., Balmforth, A. J., Bennett, A. J., Ben-Shlomo, Y., Bergman, R. N., Bergmann, S., Biebermann, H., Blakemore, A. I. F., Boes, T., Bonnycastle, L. L., Bornstein, S. R., Brown, M. J., Buchanan, T. A., Busonero, F., Campbell, H., Cappuccio, F. P., Cavalcanti-Proenca, C., Chen, Y.-D. I., Chen, C.-M., Chines, P. S., Clarke, R., Coin, L., Connell, J., Day, I. N. M., Heijer, M. d., Duan, J., Ebrahim, S., Elliott, P., Elosua, R., Eiriksdottir, G., Erdos, M. R., Eriksson, J. G., Facheris, M. F., Felix, S. B., Fischer-Posovszky, P., Folsom, A. R., Friedrich, N., Freimer, N. B., Fu, M., Gaget, S., Gejman, P. V., Geus, E. J. C., Gieger, C., Gjesing, A. P., Goel, A., Goyette, P., Grallert, H., Graszler, J., Greenawalt, D. M., Groves, C. J., Gudnason, V., Guiducci, C., Hartikainen, A.-L., Hassanali, N., Hall, A. S., Havulinna, A. S., Hayward, C., Heath, A. C., Hengstenberg, C., Hicks, A. A., Hinney, A., Hofman, A., Homuth, G., Hui, J., Igl, W., Iribarren, C., Isomaa, B., Jacobs, K. B., Jarick, I., Jewell, E., John, U., Jorgensen, T., Jousilahti, P., Jula, A., Kaakinen, M., Kajantie, E., Kaplan, L. M., Kathiresan, S., Kettunen, J., Kinnunen, L., Knowles, J. W., Kolcic, I., Konig, I. R., Koskinen, S., Kovacs, P., Kuusisto, J., Kraft, P., Kvaloy, K., Laitinen, J., Lantieri, O., Lanzani, C., Launer, L. J., Lecoeur, C., Lehtimaki, T., Lettre, G., Liu, J., Lokki, M.-L., Lorentzon, M., Luben, R. N., Ludwig, B., Manunta, P., Marek, D., Marre, M., Martin, N. G., McArdle, W. L., McCarthy, A., McKnight, B., Meitinger, T., Melander, O., Meyre, D., Midthjell, K., Montgomery, G. W., Morken, M. A., Morris, A. P., Mulic, R., Ngwa, J. S., Nelis, M., Neville, M. J., Nyholt, D. R., O'Donnell, C. J., O'Rahilly, S., Ong, K. K., Oostra, B., Pare, G., Parker, A. N., Perola, M., Pichler, I., Pietilainen, K. H., Platou, C. G. P., Polasek, O., Pouta, A., Rafelt, S., Raitakari, O., Rayner, N. W., Ridderstrale, M., Rief, W., Ruokonen, A., Robertson, N. R., Rzehak, P., Salomaa, V., Sanders, A. R., Sandhu, M. S., Sanna, S., Saramies, J., Savolainen, M. J., Scherag, S., Schipf, S., Schreiber, S., Schunkert, H., Silander, K., Sinisalo, J., Siscovick, D. S., Smit, J. H., Soranzo, N., Sovio, U., Stephens, J., Surakka, I., Swift, A. J., Tammesoo, M.-L., Tardif, J.-C., Teder-Laving, M., Teslovich, T. M., Thompson, J. R., Thomson, B., Tonjes, A., Tuomi, T., van Meurs, J. B. J., van Ommen, G.-J., Vatin, V., Viikari, J., Visvikis-Siest, S., Vitart, V., Vogel, C. I. G., Voight, B. F., Waite, L. L., Wallaschofski, H., Walters, G. B., Widen, E., Wiegand, S., Wild, S. H., Willemsen, G., Witte, D. R., Witteman, J. C., Xu, J., Zhang, Q., Zgaga, L., Ziegler, A., Zitting, P., Beilby, J. P., Farooqi, I. S., Hebebrand, J., Huikuri, H. V., James, A. L., Kahonen, M., Levinson, D. F., Macciardi, F., Nieminen, M. S., Ohlsson, C., Palmer, L. J., Ridker, P. M., Stumvoll, M., Beckmann, J. S., Boeing, H., Boerwinkle, E., Boomsma, D. I., Caulfield, M. J., Chanock, S. J., Collins, F. S., Cupples, L. A., Smith, G. D., Erdmann, J., Froguel, P., Gronberg, H., Gyllensten, U., Hall, P., Hansen, T., Harris, T. B., Hattersley, A. T., Hayes, R. B., Heinrich, J., Hu, F. B., Hveem, K., Illig, T., Jarvelin, M.-R., Kaprio, J., Karpe, F., Khaw, K.-T., Kiemeney, L. A., Krude, H., Laakso, M., Lawlor, D. A., Metspalu, A., Munroe, P. B., Ouwehand, W. H., Pedersen, O., Penninx, B. W., Peters, A., Pramstaller, P. P., Quertermous, T., Reinehr, T., Rissanen, A., Rudan, I., Samani, N. J., Schwarz, P. E. H., Shuldiner, A. R., Spector, T. D., Tuomilehto, J., Uda, M., Uitterlinden, A., Valle, T. T., Wabitsch, M., Waeber, G., Wareham, N. J., Watkins,

H., Wilson, J. F., Wright, A. F., Zillikens, M. C., Chatterjee, N., McCarroll, S. A., Purcell, S., Schadt, E. E., Vis. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat Genet*, 42(11):937–948, 2010. 31

Stouffer, S. A., Suchman, E. A., DeVinney, L. C., Star, S. A. and Williams, R. M. Adjustment during army life. *Princeton University Press, Princeton, NJ*, 1949. 5, 39

Stranger, B. E., Forrest, M. S., Dunning, M., Ingle, C. E., Beazley, C., Thorne, N., Redon, R., Bird, C. P., de Grassi, A., Lee, C., Tyler-Smith, C., Carter, N., Scherer, S. W., Tavaré, S., Deloukas, P., Hurles, M. E. and Dermitzakis, E. T. Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science*, 315(5813):848–853, 2007. 13

Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Hsi-Yang Fritz, M., Konkel, M. K., Malhotra, A., Stutz, A. M., Shi, X., Paolo Casale, F., Chen, J., Hormozdiari, F., Dayama, G., Chen, K., Malig, M., Chaisson, M. J. P., Walter, K., Meiers, S., Kashin, S., Garrison, E., Auton, A., Lam, H. Y. K., Jasmine Mu, X., Alkan, C., Antaki, D., Bae, T., Cerveira, E., Chines, P., Chong, Z., Clarke, L., Dal, E., Ding, L., Emery, S., Fan, X., Gujral, M., Kahveci, F., Kidd, J. M., Kong, Y., Lameijer, E.-W., McCarthy, S., Flicek, P., Gibbs, R. A., Marth, G., Mason, C. E., Menelaou, A., Muzny, D. M., Nelson, B. J., Noor, A., Parrish, N. F., Pendleton, M., Quitadamo, A., Raeder, B., Schadt, E. E., Romanovitch, M., Schlattl, A., Sebra, R., Shabalin, A. A., Untergasser, A., Walker, J. A., Wang, M., Yu, F., Zhang, C., Zhang, J., Zheng-Bradley, X., Zhou, W., Zichner, T., Sebat, J., Batzer, M. A., McCarroll, S. A., Consortium, T. . G. P., Mills, R. E., Gerstein, M. B., Bashir, A., Stegle, O., Devine, S. E., Lee, C., Eichler, E. E. and Korbel, J. O. An integrated map of structural variation in 2,504 human genomes. *Nature*, 526(7571):75–81, 2015. 132

Sun, J., Zheng, Y. and Hsu, L. A unified mixed-effects model for rare-variant association in sequencing studies. *Genetic Epidemiology*, 37(4):334–344, 2013. 4

Surakka, I., Horikoshi, M., Magi, R., Sarin, A.-P., Mahajan, A., Lagou, V., Marullo, L., Ferreira, T., Miraglio, B., Timonen, S., Kettunen, J., Pirinen, M., Karjalainen, J., Thorleifsson, G., Hagg, S., Hottenga, J.-J., Isaacs, A., Ladenvall, C., Beekman, M., Esko, T., Ried, J. S., Nelson, C. P., Willenborg, C., Gustafsson, S., Westra, H.-J., Blades, M., de Craen, A. J. M., de Geus, E. J., Deelen, J., Grallert, H., Hamsten, A., Havulinna, A. S., Hengstenberg, C., Houwing-Duistermaat, J. J., Hypponen, E., Karssen, L. C., Lehtimaki, T., Lyssenko, V., Magnusson, P. K. E., Mihailov, E., Muller-Nurasyid, M., Mpindi, J.-P., Pedersen, N. L., Penninx, B. W. J. H., Perola, M., Pers, T. H., Peters, A., Rung, J., Smit, J. H., Steinthorsdottir, V., Tobin, M. D., Tsernikova, N., van Leeuwen, E. M., Viikari, J. S., Willems, S. M., Willemsen, G., Schunkert, H., Erdmann, J., Samani, N. J., Kaprio, J., Lind, L., Gieger, C., Metspalu, A., Slagboom, P. E., Groop, L., van Duijn, C. M., Eriksson, J. G., Jula, A., Salomaa, V., Boomsma, D. I., Power, C., Raitakari, O. T., Ingelsson, E., Jarvelin, M.-R., Thorsteinsdottir, U., Franke, L., Ikonen, E., Kallioniemi, O., Pietiainen, V., Lindgren, C. M., Stefansson, K., Palotie, A., McCarthy, M. I., Morris, A. P., Prokopenko, I., Ripatti, S. and Consortium, E. The impact of low-frequency and rare variants on lipid levels. *Nat Genet*, 47(6):589–597, 2015. 8, 18, 19

Tachmazidou, I., Dedoussis, G., Southam, L., Farmaki, A.-E., Ritchie, G. R. S., Xifara, D. K., Matchan, A., Hatzikotoulas, K., Rayner, N. W., Chen, Y., Pollin, T. I., O'Connell, J. R., Yerges-Armstrong, L. M., Kiagiadaki, C., Panoutsopoulou, K., Schwartzentruber, J., Moutsianas, L., consortium, U., Tsafantakis, E., Tyler-Smith, C., McVean, G., Xue, Y. and Zeggini, E. A rare functional cardioprotective APOC3 variant has risen in frequency in distinct population isolates. *Nat Commun*, 4, 2013. 18

Teslovich, T. M., Musunuru, K., Smith, A. V., Edmondson, A. C., Stylianou, I. M., Koseki, M., Pirruccello, J. P., Ripatti, S., Chasman, D. I., Willer, C. J., Johansen, C. T., Fouchier, S. W., Isaacs, A., Peloso, G. M., Barbalic, M., Ricketts, S. L., Bis, J. C., Aulchenko, Y. S., Thorleifsson, G., Feitosa, M. F., Chambers, J., Orho-Melander, M., Melander, O., Johnson, T., Li, X., Guo,

X., Li, M., Shin Cho, Y., Jin Go, M., Jin Kim, Y., Lee, J.-Y., Park, T., Kim, K., Sim, X., Twee-Hee Ong, R., Croteau-Chonka, D. C., Lange, L. A., Smith, J. D., Song, K., Hua Zhao, J., Yuan, X., Luan, J., Lamina, C., Ziegler, A., Zhang, W., Zee, R. Y. L., Wright, A. F., Witteman, J. C. M., Wilson, J. F., Willemsen, G., Wichmann, H.-E., Whitfield, J. B., Waterworth, D. M., Wareham, N. J., Waeber, G., Vollenweider, P., Voight, B. F., Vitart, V., Uitterlinden, A. G., Uda, M., Tuomilehto, J., Thompson, J. R., Tanaka, T., Surakka, I., Stringham, H. M., Spector, T. D., Soranzo, N., Smit, J. H., Sinisalo, J., Silander, K., Sijbrands, E. J. G., Scuteri, A., Scott, J., Schlessinger, D., Sanna, S., Salomaa, V., Saharinen, J., Sabatti, C., Ruokonen, A., Rudan, I., Rose, L. M., Roberts, R., Rieder, M., Psaty, B. M., Pramstaller, P. P., Pichler, I., Perola, M., Penninx, B. W. J. H., Pedersen, N. L., Pattaro, C., Parker, A. N., Pare, G., Oostra, B. A., O'Donnell, C. J., Nieminen, M. S., Nickerson, D. A., Montgomery, G. W., Meitinger, T., McPherson, R., McCarthy, M. I., McArdle, W., Masson, D., Martin, N. G., Marroni, F., Mangino, M., Magnusson, P. K. E., Lucas, G., Luben, R., Loos, R. J. F., Lokki, M.-L., Lettre, G., Langenberg, C., Launer, L. J., Lakatta, E. G., Laaksonen, R., Kyvik, K. O., Kronenberg, F., König, I. R., Khaw, K.-T., Kaprio, J., Kaplan, L. M., Johansson, A., Jarvelin, M.-R., Janssens, A. C. J. W., Ingelsson, E., Igl, W., Kees Hovingh, G., Hottenga, J.-J., Hofman, A., Hicks, A. A., Hengstenberg, C., Heid, I. M., Hayward, C., Havulinna, A. S., Hastie, N. D., Harris, T. B., Haritunians, T., Hall, A. S., Gyllensten, U., Guiducci, C., Groop, L. C., Gonzalez, E., Gieger, C., Freimer, N. B., Ferrucci, L., Erdmann, J., Elliott, P., Ejebe, K. G., Döring, A., Dominiczak, A. F., Demissie, S., Deloukas, P., de Geus, E. J. C., de Faire, U., Crawford, G., Collins, F. S., Chen, Y.-d. I., Caulfield, M. J., Campbell, H., Burtt, N. P., Bonnycastle, L. L., Boomsma, D. I., Boekholdt, S. M., Bergman, R. N., Barroso, I., Bandinelli, S., Ballantyne, C. M., Assimes, T. L., Quertermous, T., Altshuler, D., Seielstad, M., Wong, T. Y., Tai, E.-S., Feranil, A. B., Kuzawa, C. W., Adair, L. S., Taylor, Jr, H. A., Borecki, I. B., Gabriel, S. B., Wilson, J. G., Holm, H., Thorsteinsdottir, U., Gudnason, V., Krauss, R. M., Mohlke, K. L., Ordovas, J. M., Munroe, P. B., Kooner, J. S., Tall, A. R., Hegele, R. A., Kastelein, J. J. P., Schadt, E. E., Rotter, J. I., Boerwinkle, E., Strachan, D. P., Mooser, V., Stefansson, K., Reilly, M. P., Samani, N. J., Schunkert, H., Cupples, L. A., Sandhu, M. S., Ridker, P. M., Rader, D. J., van Duijn, C. M., Peltonen, L., Abecasis, G. R., Boehnke, M. and Kathiresan, S. Biological, clinical and population relevance of 95 loci for blood lipids. *Nature*, 466(7307):707–13, 2010. 6, 7, 23, 24, 26, 30, 38, 40, 42, 106

TG and HDL Working Group of the Exome Sequencing Project, National Heart, Lung, and Blood Institute, Crosby, J., Peloso, G. M., Auer, P. L., Crosslin, D. R., Stitziel, N. O., Lange, L. A., Lu, Y., Tang, Z. Z., Zhang, H., Hindy, G., Masca, N., Stirrups, K., Kanoni, S., Do, R., Jun, G., Hu, Y., Kang, H. M., Xue, C., Goel, A., Farrall, M., Duga, S., Merlini, P. A., Asselta, R., Girelli, D., Olivieri, O., Martinelli, N., Yin, W., Reilly, D., Speliotes, E., Fox, C. S., Hveem, K., Holmen, O. L., Nikpay, M., Farlow, D. N., Assimes, T. L., Franceschini, N., Robinson, J., North, K. E., Martin, L. W., DePristo, M., Gupta, N., Escher, S. A., Jansson, J. H., Van Zuydam, N., Palmer, C. N., Wareham, N., Koch, W., Meitinger, T., Peters, A., Lieb, W., Erbel, R., Konig, I. R., Kruppa, J., Degenhardt, F., Gottesman, O., Bottinger, E. P., O'Donnell, C. J., Psaty, B. M., Ballantyne, C. M., Abecasis, G., Ordovas, J. M., Melander, O., Watkins, H., Orho-Melander, M., Ardissino, D., Loos, R. J., McPherson, R., Willer, C. J., Erdmann, J., Hall, A. S., Samani, N. J., Deloukas, P., Schunkert, H., Wilson, J. G., Kooperberg, C., Rich, S. S., Tracy, R. P., Lin, D. Y., Altshuler, D., Gabriel, S., Nickerson, D. A., Jarvik, G. P., Cupples, L. A., Reiner, A. P., Boerwinkle, E. and Kathiresan, S. Loss-of-function mutations in APOC3, triglycerides, and coronary disease. *New England Journal of Medicine*, 371(1):22–31, 2014. 10, 18, 19

Thurman, R. E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M. T., Haugen, E., Sheffield, N. C., Stergachis, A. B., Wang, H., Vernot, B., Garg, K., John, S., Sandstrom, R., Bates, D., Boatman, L., Canfield, T. K., Diegel, M., Dunn, D., Ebersol, A. K., Frum, T., Giste, E., Johnson, A. K., Johnson, E. M., Kutyavin, T., Lajoie, B., Lee, B.-K., Lee, K., London, D., Lotakis, D., Neph, S., Neri, F., Nguyen, E. D., Qu, H., Reynolds, A. P., Roach, V., Safi, A., Sanchez, M. E., Sanyal, A., Shafer, A., Simon, J. M., Song, L., Vong, S., Weaver, M., Yan, Y., Zhang, Z., Zhang, Z., Lenhard, B., Tewari, M., Dorschner, M. O., Hansen, R. S., Navas, P. A., Stamatoyannopoulos,

G., Iyer, V. R., Lieb, J. D., Sunyaev, S. R., Akey, J. M., Sabo, P. J., Kaul, R., Furey, T. S., Dekker, J., Crawford, G. E. and Stamatoyannopoulos, J. A. The accessible chromatin landscape of the human genome. *Nature*, 489(7414):75–82, 2012. 82, 83

Timpson, N. J., Walter, K., Min, J. L., Tachmazidou, I., Malerba, G., Shin, S.-Y., Chen, L., Futema, M., Southam, L., Iotchkova, V., Cocca, M., Huang, J., Memari, Y., McCarthy, S., Danecek, P., Muddyman, D., Mangino, M., Menni, C., Perry, J. R. B., Ring, S. M., Gaye, A., Dedoussis, G., Farmaki, A.-E., Burton, P., Talmud, P. J., Gambaro, G., Spector, T. D., Smith, G. D., Durbin, R., Richards, J. B., Humphries, S. E., Zeggini, E., Soranzo, N. and UK1OK Consortium members. A rare variant in APOC3 is associated with plasma triglyceride and VLDL levels in Europeans. *Nat Commun*, 5, 2014. 8, 18, 20

Toker, A. and Cantley, L. C. Signalling through the lipid products of phosphoinositide-3-OH kinase. *Nature*, 387(6634):673–676, 1997. 27

Trynka, G., Sandor, C., Han, B., Xu, H., Stranger, B. E., Liu, X. S. and Raychaudhuri, S. Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nat Genet*, 45(2):124–30, 2013. 82, 83

Voight, B. F., Kang, H. M., Ding, J., Palmer, C. D., Sidore, C., Chines, P. S., Burtt, N. P., Fuchsberger, C., Li, Y., Erdmann, J., Frayling, T. M., Heid, I. M., Jackson, A. U., Johnson, T., Kilpeläinen, T. O., Lindgren, C. M., Morris, A. P., Prokopenko, I., Randall, J. C., Saxena, R., Soranzo, N., Speliotes, E. K., Teslovich, T. M., Wheeler, E., Maguire, J., Parkin, M., Potter, S., Rayner, N. W., Robertson, N., Stirrups, K., Winckler, W., Sanna, S., Mulas, A., Nagaraja, R., Cucca, F., Barroso, I., Deloukas, P., Loos, R. J. F., Kathiresan, S., Munroe, P. B., Newton-Cheh, C., Pfeufer, A., Samani, N. J., Schunkert, H., Hirschhorn, J. N., Altshuler, D., McCarthy, M. I., Abecasis, G. R. and Boehnke, M. The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. *PLoS Genet*, 8(8):e1002793, 2012a. 5, 24, 38

Voight, B. F., Peloso, G. M., Orho-Melander, M., Frikke-Schmidt, R., Barbalic, M., Jensen, M. K., Hindy, G., Hólm, H., Ding, E. L., Johnson, T., Schunkert, H., Samani, N. J., Clarke, R., Hopewell, J. C., Thompson, J. F., Li, M., Thorleifsson, G., Newton-Cheh, C., Musunuru, K., Pirruccello, J. P., Saleheen, D., Chen, L., Stewart, A. F., Schillert, A., Thorsteinsdottir, U., Thorgeirsson, G., Anand, S., Engert, J. C., Morgan, T., Spertus, J., Stoll, M., Berger, K., Martinelli, N., Girelli, D., McKeown, P. P., Patterson, C. C., Epstein, S. E., Devaney, J., Burnett, M.-S., Mooser, V., Ripatti, S., Surakka, I., Nieminen, M. S., Sinisalo, J., Lokki, M.-L., Perola, M., Havulinna, A., de Faire, U., Gigante, B., Ingelsson, E., Zeller, T., Wild, P., de Bakker, P. I. W., Klungel, O. H., Maitland-van der Zee, A.-H., Peters, B. J. M., de Boer, A., Grobbee, D. E., Kamphuisen, P. W., Deneer, V. H. M., Elbers, C. C., Onland-Moret, N. C., Hofker, M. H., Wijmenga, C., Verschuren, W. M., Boer, J. M., van der Schouw, Y. T., Rasheed, A., Frossard, P., Demissie, S., Willer, C., Do, R., Ordovas, J. M., Abecasis, G. R., Boehnke, M., Mohlke, K. L., Daly, M. J., Guiducci, C., Burtt, N. P., Surti, A., Gonzalez, E., Purcell, S., Gabriel, S., Marrugat, J., Peden, J., Erdmann, J., Diemert, P., Willenborg, C., König, I. R., Fischer, M., Hengstenberg, C., Ziegler, A., Buysschaert, I., Lambrechts, D., Van de Werf, F., Fox, K. A., El Mokhtari, N. E., Rubin, D., Schrezenmeir, J., Schreiber, S., Schäfer, A., Danesh, J., Blankenberg, S., Roberts, R., McPherson, R., Watkins, H., Hall, A. S., Overvad, K., Rimm, E., Boerwinkle, E., Tybjaerg-Hansen, A., Cupples, L. A., Reilly, M. P., Melander, O., Mannucci, P. M., Ardissino, D., Siscovick, D., Elosua, R., Stefansson, K., O'Donnell, C. J., Salomaa, V., Rader, D. J., Peltonen, L., Schwartz, S. M., Altshuler, D. and Kathiresan, S. Plasma HDL cholesterol and risk of myocardial infarction: a mendelian randomisation study. *The Lancet*, 380(9841):572–580, 2012b. 10, 33, 166

Voight, B. F., Scott, L. J., Steinthorsdottir, V., Morris, A. P., Dina, C., Welch, R. P., Zeggini, E., Huth, C., Aulchenko, Y. S., Thorleifsson, G., McCulloch, L. J., Ferreira, T., Grallert, H., Amin, N., Wu, G., Willer, C. J., Raychaudhuri, S., McCarroll, S. A., Langenberg, C., Hofmann, O. M., Dupuis, J., Qi, L., Segre, A. V., van Hoek, M., Navarro, P., Ardlie, K., Balkau, B.,

Benediktsson, R., Bennett, A. J., Blagieva, R., Boerwinkle, E., Bonnycastle, L. L., Bostrom, K. B., Bravenboer, B., Bumpstead, S., Burtt, N. P., Charpentier, G., Chines, P. S., Cornelis, M., Couper, D. J., Crawford, G., Doney, A. S. F., Elliott, K. S., Elliott, A. L., Erdos, M. R., Fox, C. S., Franklin, C. S., Ganser, M., Gieger, C., Grarup, N., Green, T., Griffin, S., Groves, C. J., Guiducci, C., Hadjadj, S., Hassanali, N., Herder, C., Isomaa, B., Jackson, A. U., Johnson, P. R. V., Jorgensen, T., Kao, W. H. L., Klopp, N., Kong, A., Kraft, P., Kuusisto, J., Lauritzen, T., Li, M., Lieverse, A., Lindgren, C. M., Lyssenko, V., Marre, M., Meitinger, T., Midthjell, K., Morken, M. A., Narisu, N., Nilsson, P., Owen, K. R., Payne, F., Perry, J. R. B., Petersen, A.-K., Platou, C., Proenca, C., Prokopenko, I., Rathmann, W., Rayner, N. W., Robertson, N. R., Rocheleau, G., Roden, M., Sampson, M. J., Saxena, R., Shields, B. M., Shrader, P., Sigurdsson, G., Sparso, T., Strassburger, K., Stringham, H. M., Sun, Q., Swift, A. J., Thorand, B., Tichet, J., Tuomi, T., van Dam, R. M., van Haeften, T. W., van Herpt, T., van Vliet-Ostaptchouk, J. V., Walters, G. B., Weedon, M. N., Wijmenga, C., Witteman, J., Bergman, R. N., Cauchi, S., Collins, F. S., Gloyn, A. L., Gyllensten, U., Hansen, T., Hide, W. A., Hitman, G. A., Hofman, A., Hunter, D. J., Hveem, K., Laakso, M., Mohlke, K. L., Morris, A. D., Palmer, C. N. A., Pramstaller, P. P., Rudan, I., Sijbrands, E., Stein, L. D., Tuomilehto, J., Uitterlinden, A., Walker, M., Wareham, N. J., Watanabe, R. M., Abecasis, G. R., Boehm, B. O., Campbell, H., Daly, M. J., Hattersley, A. T., Hu, F. B., Meigs, J. B., Pankow, J. S., Pedersen, O., Wichmann, H.-E., Barroso, I., Florez, J. C., Frayling, T. M., Groop, L., Sladek, R., Thorsteinsdottir, U., Wilson, J. F., Illig, T., Froguel, P., van Duijn, C. M., Stefansson, K., Altshuler, D., Boehnke, M. and McCarthy, M. I. Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nat Genet*, 42(7):579–589, 2010. 31

Ward, L. D. and Kellis, M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res*, 40(Database issue):D930–4, 2012. 83

Weiss, L. A., Pan, L., Abney, M. and Ober, C. The sex-specific genetic architecture of quantitative traits in humans. *Nat Genet*, 38(2):218–222, 2006. 6

Welch, C. L., Xia, Y. R., Shechter, I., Farese, R., Mehrabian, M., Mehdizadeh, S., Warden, C. H. and Lusis, A. J. Genetic regulation of cholesterol homeostasis: chromosomal organization of candidate genes. *Journal of Lipid Research*, 37(7):1406–1421, 1996. 25

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L. and Parkinson, H. The NHGRI GWAS catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res*, 42(Database issue):D1001–6, 2014. 2, 6, 82

Willer, C. J., Li, Y. and Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics*, 26(17):2190–2191, 2010. 5, 39

Willer, C. J., Sanna, S., Jackson, A. U., Scuteri, A., Bonnycastle, L. L., Clarke, R., Heath, S. C., Timpson, N. J., Najjar, S. S., Stringham, H. M., Strait, J., Duren, W. L., Maschio, A., Busonero, F., Mulas, A., Albai, G., Swift, A. J., Morken, M. A., Narisu, N., Bennett, D., Parish, S., Shen, H., Galan, P., Meneton, P., Hercberg, S., Zelenika, D., Chen, W.-M., Li, Y., Scott, L. J., Scheet, P. A., Sundvall, J., Watanabe, R. M., Nagaraja, R., Ebrahim, S., Lawlor, D. A., Ben-Shlomo, Y., Davey-Smith, G., Shuldiner, A. R., Collins, R., Bergman, R. N., Uda, M., Tuomilehto, J., Cao, A., Collins, F. S., Lakatta, E., Lathrop, G. M., Boehnke, M., Schlessinger, D., Mohlke, K. L. and Abecasis, G. R. Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nat Genet*, 40(2):161–169, 2008. 6, 33

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M. and Lin, X. Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics*, 89(1):82–93, 2011. 4

Wu, Y., Waite, L. L., Jackson, A. U., Sheu, W. H.-H., Buyske, S., Absher, D., Arnett, D. K., Boerwinkle, E., Bonnycastle, L. L., Carty, C. L., Cheng, I., Cochran, B., Croteau-Chonka, D. C.,

Dumitrescu, L., Eaton, C. B., Franceschini, N., Guo, X., Henderson, B. E., Hindorff, L. A., Kim, E., Kinnunen, L., Komulainen, P., Lee, W.-J., Le Marchand, L., Lin, Y., Lindström, J., Lingaas-Holmen, O., Mitchell, S. L., Narisu, N., Robinson, J. G., Schumacher, F., Stančáková, A., Sundvall, J., Sung, Y.-J., Swift, A. J., Wang, W.-C., Wilkens, L., Wilsgaard, T., Young, A. M., Adair, L. S., Ballantyne, C. M., Bůžková, P., Chakravarti, A., Collins, F. S., Duggan, D., Feranil, A. B., Ho, L.-T., Hung, Y.-J., Hunt, S. C., Hveem, K., Juang, J.-M. J., Kesäniemi, A. Y., Kuusisto, J., Laakso, M., Lakka, T. A., Lee, I.-T., Leppert, M. F., Matise, T. C., Moilanen, L., Njølstad, I., Peters, U., Quertermous, T., Rauramaa, R., Rotter, J. I., Saramies, J., Tuomilehto, J., Uusitupa, M., Wang, T.-D., Boehnke, M., Haiman, C. A., Chen, Y.-D. I., Kooperberg, C., Assimes, T. L., Crawford, D. C., Hsiung, C. A., North, K. E. and Mohlke, K. L. Trans-ethnic fine-mapping of lipid loci identifies population-specific signals and allelic heterogeneity that increases the trait variance explained. *PLoS Genet*, 9(3):e1003379, 2013. 9

Yang, J., Bakshi, A., Zhu, Z., Hemani, G., Vinkhuyzen, A. A. E., Lee, S. H., Robinson, M. R., Perry, J. R. B., Nolte, I. M., van Vliet-Ostaptchouk, J. V., Snieder, H., Study, T. L. C., Esko, T., Milani, L., Magi, R., Metspalu, A., Hamsten, A., Magnusson, P. K. E., Pedersen, N. L., Ingelsson, E., Soranzo, N., Keller, M. C., Wray, N. R., Goddard, M. E. and Visscher, P. M. Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat Genet*, 47(10):1114–1120, 2015. 164

Yang, Y., Chung, E. K., Wu, Y. L., Savelli, S. L., Nagaraja, H. N., Zhou, B., Hebert, M., Jones, K. N., Shu, Y., Kitzmiller, K., Blanchong, C. A., McBride, K. L., Higgins, G. C., Rennebohm, R. M., Rice, R. R., Hackshaw, K. V., Roubey, R. A. S., Grossman, J. M., Tsao, B. P., Birmingham, D. J., Rovin, B. H., Hebert, L. A. and Yung Yu, C. Gene copy-number variation and associated polymorphisms of complement component c4 in human systemic lupus erythematosus (SLE): Low copy number is a risk factor for and high copy number is a protective factor against SLE susceptibility in European Americans. *The American Journal of Human Genetics*, 80(6):1037–1054, 2007. 15

Ye, K., Schulz, M. H., Long, Q., Apweiler, R. and Ning, Z. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*, 25(21):2865–2871, 2009. 14

Yoon, S., Xuan, Z., Makarov, V., Ye, K. and Sebat, J. Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Research*, 19(9):1586–1592, 2009. 14

Zerbino, D. R. and Birney, E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, 18(5):821–829, 2008. 14

Zuk, O., Schaffner, S. F., Samocha, K., Do, R., Hechter, E., Kathiresan, S., Daly, M. J., Neale, B. M., Sunyaev, S. R. and Lander, E. S. Searching for missing heritability: Designing rare variant association studies. *Proceedings of the National Academy of Sciences*, 111(4):E455–E464, 2014. 3