Improved Analysis of Large Genetic Association Studies Using Summary Statistics

by

Sebanti Sengupta

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in the University of Michigan
2018

Doctoral Committee:

Professor Gonçalo Abecasis, Chair
Professor Michael Boehnke
Associate Professor Xiaoquan William Wen
Associate Professor Cristen Willer

Sebanti Sengupta

sebanti@umich.edu

ORCID iD: 0000-0002-9618-3600

To Ma, Baba and Dada

Acknowledgements

I am honored to have been a part of the University of Michigan Biostatistics department. I am grateful to my thesis advisor, Gonçalo Abecasis, for his guidance and mentorship. His insight into the field of statistical genetics remains an inspiration and this thesis would not have been possible without him.

I greatly appreciate the support and direction I received from the other members of my dissertation committee: Cristen Willer, Xiaoquan William Wen, and Michael Boehnke. Thank you, Cristen, for the tremendous guidance you provided for the GLGC project. Thank you, William, for your patient direction on the enrichment project. Thank you, Mike, for all the advice on the meta-analysis overlap project.

I am especially thankful for the numerous opportunities I received to attend conferences and collaborate with experts in the field. It has been a pleasure to have worked with so many wonderful people in the Abecasis Lab and the Center for Statistical Genetics. I would like to especially thank Ellen Schmidt for the wonderful collaboration on the GLGC project. I am grateful to many people including Adrian Tan, Ryan Welch, Jin Chen, Alan Kwong, Daniel Taliun, Sarah Gagliano and Wei Zhou for helping me with bits of codes or bouncing off ideas.

I'd be remiss if I didn't mention the wonderful Bengali community here in Ann Arbor that made me feel at home. Thank you to all the wonderful ISI seniors and juniors, especially Sumanta Bose, who made sure I didn't have any worries about moving so far away. Thank you, Bhramar Mukherjee, for all the food, 'adda' and care. I am grateful for the great friends I found in Sreyoshi Das, Shrijita Bhattacharya and Raka Mandal. Most of all, I am thankful to Pramita

Bagchi for all the support and encouragement and for always being ready to listen to me complain.

Finally, I want to thank my family without whose support this would not have been possible. I am grateful to my parents for all their love and their faith in me. I am thankful to my brother who has always taken a keen interest in my research and to my nephew who always makes me smile. My grandmother, sister-in-law, aunts and cousins have all shown their love and encouragement in so many different ways and I am very thankful for that.

Table of Contents

Acknowledgements iii

List of Tables viii

List of Figures xii

Abstract xiv

Chapter 1: Introduction 1

An Overview of Meta-Analysis Methods for GWAS 2
Prioritizing Variants for Follow-up 4
Overlapping Samples in Meta-Analysis Studies 6
Thesis Overview 7

Chapter 2 : Discovery and Refinement of Loci Associated with Lipid Levels 9

Introduction 9
Results 10
Novel loci associated with blood lipid levels 10
Overlap of genetic discoveries and prior knowledge 11
Pathway analyses 12
Protein-protein interactions 13
Regulation of gene expression by associated variants 13
Coding variation 14
Overlap between association signals and regulators of transcription in liver 14
Initial fine-mapping of 65 lipid-associated loci 15
Association of lipid loci with metabolic and cardiovascular traits 17
Association of individual lipids with coronary artery disease 18
Evidence for additional loci, not yet reaching genome-wide significance 19
Discussion 20
Online Methods 29
Supplementary Tables 35
Supplementary Figures 71

Supplementary Note: Candidate Genes at Novel Loci 90
Chapter 3: Prioritizing Functional Variants in Genetic Association Studies 97
Introduction 97
Methods 99
Modeling the Association 100
Linking the Annotations 101
Conditional Expectation 102
EM Algorithm 102
Defining Loci and Modifying the Prior 103
Prioritizing Variants for Follow-up 106
Loci with Multiple Causal Variants 107
Application to Multiple Traits 107
Simulation 108
Real Data Application 109
Results 111
Simulation 111
Real Data Application 114
Discussion 117
Supplementary Methods 130
Algorithm 130
Modeling the Association 130
Simpler Approach: Focusing on Loci that Reach Genome-wide Significance 131
An Alternative Bayesian Approach 133
Chapter 4: Correcting for Sample Overlap in GWAS Meta-Analysis Using Summary Statistics 141
Introduction 141
Material and Methods 142
Standard Meta-Analysis Method 142
Meta-Analysis Correcting for Sample Overlap 143
Simulation Set-up 149
Artificially Creating Overlapping Datasets based on GWAS Data 149

Results 150
Simulation Results 150
Artificially Created Overlapping Datasets: Quantitative Trait 150
Artificially Created Overlapping Datasets: Case-Control Study 151
Discussion 152
Chapter 5: Conclusion 163
Prioritizing Variants for Follow-up Studies 165
Meta-Analysis of Studies with Sample Overlap 166
In Summary 168
Bibliography 169

List of Tables

Table 2-A Novel Loci Primarily Associated with HDL Cholesterol Obtained from Joint GWAS and Metabochip Meta-analysis 25

Table 2-B Novel Loci Primarily Associated with LDL Cholesterol Obtained from Joint GWAS and Metabochip Meta-analysis 26

Table 2-C Novel Loci Primarily Associated with Total Cholesterol Obtained from Joint GWAS and Metabochip Meta-analysis 27

Table 2-D Novel Loci Primarily Associated with Triglycerides Obtained from Joint GWAS and Metabochip Meta-analysis 28

Supplementary Table S2.1: Phenotypic Summary of Samples with Metabochip Genotype Results 35

Supplementary Table S2.2: Biological Candidate Genes at Novel Loci based on Literature Search, Nonsynonymous Variants, Gene Expression Levels (eQTLs) and Pathway Analysis 38

Supplementary Table S2.3: Summary of Joint Meta-Analysis Association Results for 95 Previously Discovered Lipid Loci 42

Supplementary Table S2.4: Overlap of Novel Loci and Literature 45

Supplementary Table S2.5: Pathways that Show Enrichment of Genes at Novel Loci by MAGENTA analysis 48

Supplementary Table S2.5-A: Pathways that Show Enrichment of Genes at Novel HDL Associated Loci by MAGENTA analysis 49

Supplementary Table S2.5-B: Pathways that Show Enrichment of Genes at Novel LDL Associated Loci by MAGENTA analysis 50
Supplementary Table S2.5-C: Pathways that Show Enrichment of Genes at Novel Total Cholesterol Associated Loci by MAGENTA analysis 51
Supplementary Table S2.5-D: Pathways that Show Enrichment of Genes at Novel Triglyceride Associated Loci by MAGENTA analysis 52
Supplementary Table S2.6: Overlap Between eQTL Loci and New Lipid Associated Loci 53
Supplementary Table S2.7: Nonsynonymous Variants in Linkage Disequilibrium with Index SNPs at Novel Loci 54
Supplementary Table S2.8: Overlap of SNPs at Known and Novel Lipid Loci with Chromatin States in 9 Different Cell Types 55
Supplementary Table S2.9: Overlap with Chromatin States, Histone Marks and Transcription Factor ChIP-Seq in HepG2 Cells 56
Supplementary Table S2.10: Overlap of Regulatory Features and Associated SNPs at Novel Lipid Loci 57
Supplementary Table S2.11: Fine-Mapping Results in Different Ancestries 61
Supplementary Table S2.12: Novel and Known Lipid Loci Associated with BMI, CAD, DBP, SBP, Fasting Glucose, T2D, and WHR adj BMI 63
Supplementary Table S12.12-A: Novel and Known Lipid Loci with BMI P-value < 0.05 from GIANT* 63
Supplementary Table S12.12-B: Novel and Known Lipid Loci with CAD P-value < 0.05 from CARDIOGRAM+C4D Meta-analysis* 64

Supplementary Table S2.12-C: Novel and Known Lipid Loci with DBP P-value < 0.05 from ICBP** 65
Supplementary Table S2.12-D: Novel and Known Lipid Loci with SBP P-value < 0.05 from ICBP* 66
Supplementary Table S2.12-E: Novel and Known Lipid Loci with Fasting Glucose P-value < 0.05 from MAGIC* 67
Supplementary Table S2.12-F: Novel and Known Lipid Loci with T2D P-value < 0.05 from DIAGRAM* 68
Supplementary Table S2.12-G: Novel and Known Lipid Loci with WHR adj BMI P-value < 0.05 from GIANT* 69
Supplementary Table S2.13: Overlap of Lipid Subfractions in Framingham with Novel and Known Lipid Associated Loci (P<1.4x10-5) 70
Table 3.1: Example of a Contingency Table if Causal Variants are Known 99
Table 3.2: Estimated Confidence Interval Lengths for the Enrichment Parameter 120
Table 3.3: Decrease in Credible Set Size using Enrichment Parameter 120
Table 3.4: Estimated enrichment parameter in AMD data for different genomic features 121
Table 3.5: Credible Set Sizes Based on Different Genomic Features for AMD Data 121
Table 3.6: Some Highlighted Variants in the Associated Loci for AMD Data 122
Table 3.7: Estimated Enrichment Parameter for Nonsynonymous Variants in Publicly Available Datasets 123
Table 3.8: Highlighted Variants in the Enrichment Analysis for eQTL SNPs in MGI Data 124
Table 3.9: Traits where Nonsynonymous Variants are found to be Significantly Enriched in UK Biobank Data 125

Table 3.10: Traits where eQTL SNPs are found to be Significantly Enriched in UK Biobank Data 126

Table 4.1: Overlap estimate and confidence interval when trait is independent of genotypes... 154

Table 4.2: Estimated Sample Overlap for Artificially Created Overlapping Studies for HDL-cholesterol 155

Table 4.3: Estimated Sample Overlap for Artificially Created Overlapping Studies for Type 2 Diabetes..... 156

List of Figures

Figure 2.1 Overlap between loci associated with different lipid traits 24

Supplementary Figure S2.1: Study Design 71

Supplementary Figure S2.2: QQ Plots of Metabochip Meta-Analysis P-value Distributions 72

Supplementary Figure S2.3: Manhattan Plots of Lipid-specific Association Results 74

Supplementary Figure S2.4: Effect Size vs. Allele Frequency at Lipid Associated Loci 76

Supplementary Figure S2.5: Direct Protein-Protein Interactions from Dapple Analysis 77

Supplementary Figure S2.6: Lipid vs. CAD Effect Sizes 78

Supplementary Figure S2.7: Association with Lipid Subfractions 79

Figure 3.1: Power Curves for Different Sample Sizes 127

Figure 3.2: Empirical Mean and SD of Enrichment Parameter Estimate 127

Figure 3.3: Coverage Probability 128

Figure 3.4: Power Comparison with fGWAS 128

Figure 3.5: Locuszoom plot for region around chr19:6718387 for AMD 129

Supplementary Figure S3.1: Phenotypes for MGI Data 139

Supplementary Figure S3.2: Outline of Method to Estimate Enrichment of eQTL's in MGI Data 140

Figure 4.1: Outline of procedure to meta-analyze correcting for overlap 157

Figure 4.2: Example Illustrating the Need for Stratification Based on Sample Size 158

Figure 4.3 : Example of Sample Size Distribution in a Large Meta-Analysis 158

Figure 4.4: Performance of Meta-Analysis Correcting for Overlap in HDL-Cholesterol..... 159

Figure 4.5: Outliers in Meta-Analysis Correcting for Overlap in HDL-Cholesterol..... 159

Figure 4.6: Performance of Meta-Analysis Correcting for Overlap in Type 2 Diabetes..... 160

Figure 4.7: Outliers in Meta-Analysis Correcting for Overlap in Type 2 Diabetes 160

Supplementary Figure S4.1: Creating Overlapping Datasets for HDL-Cholesterol..... 161

Supplementary Figure S4.2: Creating Overlapping Datasets for Type 2 Diabetes 161

Supplementary Figure S4.3: Effect of Not Stratifying by Sample Size 162

Abstract

Genome-wide association studies, which examine millions of genetic variants in thousands of individuals, have identified many complex trait associated loci. As sample sizes increase, particularly through meta-analysis, the number of disease associated loci has increased rapidly. The objective of this dissertation is to demonstrate the advantages of combining data across studies using summary statistics and to demonstrate methods that use publicly available information, such as functional annotation of the genome, to gain further insight into the genetics of human disease.

In the first project, we analyze data from 188,578 individuals using genome-wide and custom genotyping arrays to identify new loci and refine known loci for lipid traits low-density lipoprotein (LDL) cholesterol, high-density lipoprotein (HDL) cholesterol, triglycerides, and total cholesterol. We identify and annotate 157 loci associated with lipid levels at $P < 5x10^{-8}$, 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 lipids 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 illustrate the value of genetic data from individuals of diverse ancestries and provide insights into biological mechanisms regulating blood lipids to guide future genetic, biological, and therapeutic research.

In the second project, we propose that causal variants for a trait may share certain genomic features. Importantly, we show that when these genomic features can be identified, we can use

them to help pinpoint likely causal variants among many trait associated variants. We develop a model that identifies genomic features enriched among the associated loci and uses this information to prioritize likely functional variants in each locus leading to narrower sets of variants for follow-up. Our models work for both quantitative and case-control data and can be used with summary statistics, making it convenient to incorporate in ongoing meta-analysis of genome-wide association studies that can include 100,000s of individuals.

In the third project, we consider meta-analysis where studies may have overlapping sets of participants. In such scenarios, meta-analysis methods that do not account for overlap will perform poorly and have inflated Type I error. We propose a method to identify participant overlap between GWAS using only summary statistics, estimate the degree of overlap, and correctly meta-analyze studies taking into account the overlap. Our method builds upon and extends previous methods that allow meta-analysis of GWAS studies with known overlap proportions. We illustrate our method using simulations and artificially created overlapping samples using real GWAS data.

Chapter 1: Introduction

Conducting genome-wide association studies (GWAS) is an established way to identify genetic factors contributing to human traits and diseases. Typically, millions of single nucleotide polymorphisms (SNPs) are genotyped or imputed in large cohorts, and then each SNP is tested for association with the trait of interest (McCarthy et al. 2008; Hindorff et al. 2009). Developments in high throughput genotyping technology have led to decreases in the cost of conducting a GWAS, enabling increased sample sizes and the study of increasingly diverse traits. In recent years, GWAS have reproducibly identified thousands of variants associated with phenotypes as diverse as psychiatric diseases (Cross-Disorder Group of the Psychiatric Genomics Consortium 2013; Neale et al. 2010; Duncan et al. 2017), type 2 diabetes (Morris et al. 2012), and food taste preferences (Pirastu et al. 2016), leading to discoveries about genes and pathways involved in common diseases and complex traits, providing new insights about their biology and disease etiology (Visscher et al. 2012).

Contrary to early expectations, however, the variants identified by GWAS typically explain only a modest portion of risk for most complex diseases, as the genetic effects due to common alleles are typically quite small (Ioannidis et al 2006; Manolio et al 2009). Larger samples facilitate detection of small effect variants and have enabled gradual increases in the proportion of risk that can be explained (Moonesinghe et al. 2008; Burton et al. 2009; Chapman et al. 2011). A common strategy to enable these large sample sizes is to perform a meta-analysis – the statistical synthesis of information from multiple independent studies (Evangelou and

Ioannidis 2013). Currently, most meta-analyses are conducted using summary statistics of the association results on a per variant basis (see next section for details), which reduces logistical burdens of pooling data and mitigates risks associated with sharing individual level data (Solovieff et al. 2013). Most genetic risk variants discovered recently have come from large-scale meta-analyses of GWAS (Zeggini and Ioannidis 2009; Panagiotou et al. 2013).

With this increasing popularity of meta-analysis of GWAS, many consortia have been formed to investigate the genetics of various complex traits and diseases (Seminara et al. 2007) such as type 2 diabetes (DIAGRAM) (Morris et al. 2012), lipids (GLGC) (Global Lipids Genetics Consortium 2013), anthropomorphic traits like height and BMI (GIANT) (Wen et al. 2012; Wood et al. 2014), and various psychiatric disorders (PGC) like ADHD and Schizophrenia (Demontis et al. 2017; Ripke et al. 2013). Many of these consortia make their summary statistics publicly available, so that newer studies can use their results in a meta-analysis, thus leveraging these large-scale efforts to power new discoveries or to refine known loci.

An Overview of Meta-Analysis Methods for GWAS

In a typical meta-analysis, individual level data are analyzed locally, and summary level results are shared with the coordinating meta-analysis team (Zeggini and Ioannidis 2009). The summary results may be odds ratios, standardized effect sizes or other metrics, with a measure of their uncertainty such as variance or p-values. Additionally, marker information such as alleles and allele frequencies are provided.

Several models can be used for meta-analysis, and can be broadly categorized into three groupings: fixed effects, random effects, or Bayesian. Most models involve taking some weighted combination of estimators from individual studies. Fisher's approach combines *P*-values directly by taking the mean of $-\log(P)$ across studies, but fails to take the direction of

the effect into account. A similar approach is based on Z-scores, calculating the average Z-scores across studies, weighing them proportionally to the sample sizes (Stouffer et al. 1949). This approach does consider the direction of the effect and is useful in cases where only *P*-values are available (Cooper et al. 2009).

Fixed effects meta-analysis assumes that the true effect of each risk allele is the same in all the datasets. A common method is to use inverse variance weighting for the effect sizes (Willer et al. 2010). This also requires that the trait be measured on the same scale for each study, with the same units and transformations so that the effect sizes are comparable across studies. This is the most powerful of the approaches; though there may be some questions regarding the assumption of fixed effects (Kavvoura and Ioannidis, 2008).

In contrast, random effects models do not assume that the true effect sizes are constant across studies, which is a more realistic assumption especially if the samples are of mixed ancestry. While random effects models such as the DerSimonian and Laird estimator (1986) may be more appropriate, they generally have limited power and are not used in discovery efforts. Newer random effects models have been proposed that attempt to increase the power by making simplifying assumptions -- such as the assumption that there is no heterogeneity in genetic effects under the null hypothesis – and thus focusing analyses on the most interesting portions of the parameter space (Han and Eskin 2011). Another proposed alternative is to use a subset-based approach that uses a fixed effects model on a subset of the studies (Bhattacharjee et al. 2012). While these methods increase power when their specific assumptions are met, the fixed-effects model outperforms them when analyzing a single trait across homogeneous populations.

Bayesian methods have also been used for meta-analysis (Burton et al. 2007), but they are less common as they depend on assumptions made about the prior distributions of parameters, and can be computationally intensive to implement genome-wide.

Regardless of which method is chosen, once the meta-analysis is conducted, the most significant variants are investigated. Generally, variants with *P*-values less than a genome-wide significant cut-off (typically $p < 5 \times 10^{-8}$) are flagged. Other secondary analyses such as conditional analysis may be conducted to gain insight into loci associated with the trait of interest.

Prioritizing Variants for Follow-up

Once a GWAS or meta-analysis is completed, it is important to sort through the flagged SNPs to examine interesting variants for follow-up studies or secondary analyses. Variants and genes can be prioritized for follow-up based on criteria such as literature review for biological plausibility (Minelli et al. 2013), evidence from other GWAS of the same or related traits, pathway analysis, regulation of mRNA expression levels, and presence of protein-altering variants. "Causal" variants are defined as the functional genetic variants that influence the risk for disease and explain the observed association. However, all variants in high linkage disequilibrium may show association with the trait or disease, and hence, it can be difficult to identify a specific causal variant. Fine mapping involves refining lists of potentially causal variants in regions with dense coverage where all the variants are genotyped or imputed with high quality genotypes. One of the goals of GWAS is to identify genes that are potentially causal for the trait of interest and thus fine mapping the associated loci is important for understanding the biological mechanisms involved in the trait under study. Variants in strong linkage disequilibrium with a causal SNP show strong association signals and thus the most significantly

associated variant is often not the causal one (Schaub et al. 2012; Faye et al. 2013). Fine mapping can be done through iterative conditional analysis or joint analysis at associated loci (Yang et al. 2012), ridge regression based methods (Malo et al. 2008), and Bayesian modeling (Maller et al. 2012; Hormozdiari et al. 2014). These methods focus on using the association results, that is, they do not consider biological plausibility, and generally require individual level data at the fine mapping loci.

Large-scale initiatives such as The Encyclopedia of DNA Elements (ENCODE Project Consortium 2012) and the Roadmap Epigenomics Project (Bernstein et al. 2010) provide detailed maps of regulatory regions for more than 80% of the human genome. Functional annotations such as transcription factor binding sites and expression quantitative trait loci tend to be enriched in complex trait associated loci (Veyrieras et al. 2008; Gaffney et al. 2012; Trynka and Raychaudhuri 2013; Karczewski et al. 2013). A systematic investigation into the enrichment of these characteristics among associated loci can lend insight for future functional studies.

With advances in sequencing technology, it is feasible to obtain the sequencing data or impute all common variants in associated regions with high quality. Thus, it is plausible to assume that the causal variant exists in the data (Chen et al. 2015) and attempt to narrow down a list of potentially causal variants by systematically modeling genomic features that they may share (Pickrell 2014; Kichaev et al. 2014). For example, a plausible assumption may be that causal variants for traits tend to be non-synonymous variants, which alter an amino acid in a protein-coding sequence, potentially resulting in a functionally impaired protein. There exist methods to integrate diverse annotations into one measure such as the CADD score which prioritizes functional, deleterious and pathogenic variants across many functional categories

(Kircher et al. 2014). Quantifying the enrichment of such genomic features can help systematically prioritize variants for follow-up.

Overlapping Samples in Meta-Analysis Studies

To date, more than 2,500 GWAS and meta-analyses have been published (MacArthur et al. 2017). For certain traits, several independent or partially overlapping consortia may exist (Evangelou and Ioannidis 2013), and meta-analyses combining data from the different consortia as well as using future studies can greatly increase the power to detect weak signals. However, a basic assumption in the meta-analysis methods discussed above is that the studies are independent; that is, the samples analyzed the studies are independent with no overlap of participants. If there is overlap between studies, using these methods may result in to inflated Type I error and, hence, increased false signals (Lin and Sullivan 2009).

Moreover, due to GWAS requirements for large sample sizes, sometimes controls are shared among various studies. For example, many psychiatric GWAS have used controls ascertained and sampled by P.V. Gejman (Shi et al. 2009) and many case-control studies use publicly available genotype data for large sets of population-based controls such as WTCCC (Burton et al. 2007). Additionally, same cohorts may contribute to different meta-analysis efforts. For example, for Type 2 Diabetes (http://www.type2diabetesgenetics.org/), data from FUSION was used in both GoT2D GWAS (Fuchsberger et al. 2016) and 70KforT2D GWAS (Bonàs-Guarch et al. 2017).

If studies have overlapping samples, covariance between studies needs to be accounted for when analyzing them together through meta-analysis. Methods exist for this purpose when individual level data are available, or if the number of overlapping samples is known (Lin and

Sullivan 2009). However, precise sample overlap numbers are not always known, and it is difficult to obtain individual level data to determine the overlap.

Thesis Overview

The scope of this dissertation is to show diverse uses of summary level genetic association data to gain insight into the genetics of diseases and complex traits. I first conduct a large-scale meta-analysis and demonstrate the challenges inherent in trying to interpret the results. I then show how publicly available data, such as genomic features, can be used to prioritize variants for follow-up using summary statistics from published GWAS or meta-analyses. I finally develop a method to allow for meta-analysis when studies have potentially overlapping samples without requiring individual level data.

In the second chapter, we conducted a large-scale meta-analysis of 188,578 individuals for the lipid traits of low-density lipoprotein cholesterol (LDL-C), high-density lipoprotein cholesterol (HDL-C), triglycerides and total cholesterol. We identified sixty-two novel loci in addition to the 95 loci already known in literature. There are 240 genes within 100 kb of the 62 novel loci, which is a daunting challenge for future functional studies. We prioritized variants based on literature review, pathway analysis, protein altering variants and overlap with regulators of transcription in liver and tested enrichment using different tests including permutation based ones. We found lipid associated loci to be strongly associated with Coronary Artery Disease (CAD), Type 2 Diabetes (T2D), Body Mass Index (BMI) and blood pressure. Additionally, we investigated fine mapped loci for different ancestries and identified several loci where the fine mapped signal was clearly different from the known GWAS signal.

In the third chapter, I assume that causal variants are likely to share certain genomic features and model and estimate the enrichment of the genomic feature among trait associated

loci. If the enrichment parameter is found to be statistically significant, we use it to weight the association signals to generate posterior probabilities of each variant being causal. I make an assumption that every locus has at most one causal variant and propose two different approaches: (i) a Bayesian approach that divides the entire genome into loci with each locus having some probability of being causal and there being 0 or 1 causal variant at each locus, and (ii) assuming that variants outside the associated loci cannot be causal and thus considering only associated loci and assuming that there is exactly 1 causal variant at each locus. While the Bayesian approach may be more realistic, it requires summary level data across the whole genome whereas the second approach only requires summary level data for the associated loci and the number of variants with and without the genomic feature of interest outside the associated loci. The second approach is thus computationally faster. After testing the enrichment parameter for significance, we calculate the posterior probability of each variant being causal and use that to construct 95% credible sets of potentially causal variants.

In the fourth chapter, I propose a method to use summary level data to estimate the overlap proportions between a pair of studies. While independent studies are likely to have positively correlated Z-scores due to trait associated loci showing similar effect sizes, truncating the Z-scores and estimating the correlation leads to accurate estimates of overlap proportion. Thus, my method identifies if there is overlap between a pair of studies, estimates it and corrects for it to conduct a meta-analysis where covariance due to overlap is accounted for.

Chapter 2 : Discovery and Refinement of Loci Associated with Lipid Levels[1]

Introduction

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

Here, we report on studies of naturally occurring variation in 188,578 European-ancestry individuals and 7,898 non-European ancestry individuals. Our analyses identify 157 loci associated with lipid levels at $P < 5 \times 10^{-8}$, 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

[1] This work was published in Nature Genetics as Discovery and Refinement of Loci Associated with Lipid Levels (2013) 45(11), pp.1274-1283. I was joint first author and contributed in performing the meta-analysis, bioinformatics analyses, preparing the tables, figures and supplementary information as well as discussing the analysis, results, interpretation and presentation of the results. The full list of authors is at the end of the chapter as well as in the supplementary information.

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.

Results

Novel loci associated with blood lipid levels

We examined subjects of European ancestry, including 94,595 individuals from 23 studies genotyped with GWAS arrays (Teslovich et al. 2010) and 93,982 individuals from 37 studies genotyped with the Metabochip array (Voight et al. 2012) (**Supplementary Table S2.1** and **Supplementary Fig. S2.1**). 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 (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 key variants were directly genotyped, rather than imputed.

We excluded individuals known to be on lipid lowering medications and evaluated the additive effects 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 only a minor impact on our results (**Supplementary Fig. S2.2**). After genomic control correction, 157 loci associated with blood lipid levels were identified ($P < 5x10^{-8}$), including 62 new loci (**Tables 2A-D, Figure 2.1, Supplementary Tables S2.2 and S2.3**). Loci were >1 Mb apart and nearly independent ($r^2 < 0.10$). Of the 62 novel loci, 24 demonstrated the strongest evidence of association with HDL

cholesterol, 15 with LDL cholesterol, 8 with triglyceride levels, and 15 with total cholesterol (**Supplementary Fig. S2.3**). Several of these loci were validated by a similar extension based on GLGC GWAS results (Asselbergs et al. 2012).

The effects of newly identified loci were generally smaller than in earlier GWAS (**Supplementary Fig. S2.4**). For the 62 newly identified variants, trait variance explained in the Framingham offspring were 1.6% for HDL cholesterol, 2.1% for triglycerides, 2.4% for LDL cholesterol, and 2.6% for total cholesterol.

Overlap of genetic discoveries and prior 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 OMIM for occurrences of these gene names and their aliases in the context of relevant keywords. After manual curation, we identified at least one strong candidate in 32 of the 62 loci (52%) (**Supplementary Table S2.4**). For the remaining 30 loci, we found no literature support for the role of a nearby gene on 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*. For 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 levels.

Multiple types of evidence supported several literature candidates (**Supplementary Table S2.2**). For example, *VLDLR* is categorized by Gene Ontology (Ashburner et al. 2000) in the retinoid X nuclear receptor (RXR) activation pathway, which also includes genes (*APOB,*

APOE, CYP7A1, APOA1, HNF1A, HNF4A) in previously implicated loci (Teslovich et al. 2010). However, since 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 linkage disequilibrium proxies. To approximate hand-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, $P = 8x10^{-8}$).

Pathway analyses

We performed a gene-set enrichment analysis, using MAGENTA (Segre et al. 2010), to evaluate over-representation of biological pathways among associated loci. Across the 157 loci, MAGENTA identified 71 enriched pathways. These pathways included at least one gene in 20 of our newly identified loci (**Supplementary Table S2.5**). Examples include *DAGLB* (connected to previously associated loci by genes in the triglyceride lipase activity pathway), *INSIG2* (connected by the cholesterol and steroid metabolic process pathways), *AKR1C4* (connected by the steroid metabolic process and bile acid biosynthesis pathways), *VLDLR* (connected by the retinoic X receptor activation and lipid transport pathways, among others), *PPARA*, *ABCB11,* and *UGT1A1* (three genes assigned to pathways implicated in activation of nuclear hormone receptors, which play an important role in lipid metabolism through the transcriptional regulation of genes in sterol metabolic pathways (Fitzgerald et al. 2001)). Among the 16 loci where literature review and pathway analysis both suggested a candidate, the predictions overlapped 14 times (**Supplementary Table S2.2**; by chance, we expect 6.6 overlapping predictions,

$P = 1x10^{-5}$).

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 (10 interactions, $P = 0.0002$), HDL (8 interactions, $P = 0.002$), and total cholesterol (6 interactions, $P = 0.017$), but not for triglycerides (2 interactions, $P = 0.27$) (**Supplementary Fig. S2.5**). 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 literature and pathway analyses (such as those connecting *VLDLR, APOE, APOB, CETP*, and *LPL*). Among novel loci, we identified a link between *AKT1* and *GSK3B*. *GSK3B* has been shown to play 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 blood lipid levels for these two genes.

Regulation of gene expression by associated variants

Many complex trait associated variants act through the regulation of gene expression. We examined whether our 62 novel variants were associated with expression levels of nearby genes in liver, omental fat, or subcutaneous fat. Fifteen were associated with expression of a nearby transcript with $P < 5x10^{-8}$ (**Supplementary Table S2.6**) and, in seven, the lipid-associated variant was in strong disequilibrium with the strongest expression-quantitative trait locus (eQTL) for the region ($r^2 > 0.8$). In three of these loci, literature search 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.

Coding variation

In some loci where previous coding variant association studies were inconclusive, we now find convincing evidence of association, demonstrating the benefits of the large sample sizes achievable by collaboration. For example, in the *APOH* locus (Kaprio et al. 1991), our most strongly associated variant is rs1801689 (*APOH* C325G, $P = 1x10^{-11}$ for LDL cholesterol). Overall, at 15 of the 62 new loci, there is at least one nonsynonymous variant within 100kb and in strong (r^2>0.8) linkage disequilibrium with the index SNP (**Supplementary Table S2.7)** (18 loci with no restrictions on distance). This ~30% overlap between associated loci and coding variation is similar to that in other complex traits (The 1000 Genomes Project 2010). Unexpectedly, in the 11 loci where a candidate was suggested by literature review and by coding variation, the two coincided seven times ($P = 0.03$ compared to expected chance overlap of 3.8 times); thus, agreement between literature and coding variation was less significant than for eQTL and pathway analysis or protein-protein interactions.

Overlap between association signals and regulators of transcription in liver

Despite our efforts, 18 of the 62 new loci remain without prioritized candidate genes. The liver is an important hub of lipid biosynthesis and there is evidence that lipid loci might be associated with changes in gene regulation in liver cells (Ernst et al. 2011). Using ENCODE data (Ernst et al. 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, distance to the nearest gene, and number of SNPs in $r^2 > 0.8$ (described in

Methods). In HepG2 cells, lipid-associated SNPs were enriched in eight of the 15 functional chromatin states defined by Ernst *et al.* (The ENCODE Project Consortium 2011) ($P < 1x10^{-5}$; **Supplementary Table S2.8**). The strongest enrichment was in regions with "strong enhancer activity" (3.7-fold enrichment, $P = 2x10^{-25}$; **Supplementary Table S2.9**). In the other eight cell types examined by Ernst *et al*., 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 loci and functional marks in HepG2 cells in more detail (**Supplementary Table S2.9**). Notable regulatory elements showing significant overlap with lipid loci included histone marks associated with active regulatory regions (H3K27ac, $P = 3x10^{-20}$; H3K9ac, $P = 3x10^{-22}$), promoters (H3K4me3, $P = 2x10^{-15}$, H3K4me2, $P = 8x10^{-12}$), transcribed regions (H3K36me3, $P = 4x10^{-14}$), indicators of open chromatin (FAIRE, $P = 5x10^{-9}$; DNase, $P = 2x10^{-4}$), and regions that interact with transcription factors HNF4A ($P = 6x10^{-10}$) and CEBP/B ($P = 1x10^{-5}$). Overall, 56 of our 62 new loci contained at least one SNP that overlaps a functional mark (The ENCODE Project Consortium 2011) and/or chromatin state (Ernst et al. 2011) highlighted in **Supplementary Table S2.9,** including all but 3 of the loci where no candidates were suggested by literature review or analyses of pathways, coding variation, or gene expression (**Supplementary Table S2.10**).

Initial fine-mapping of 65 lipid-associated loci

Previous fine-mapping of five LDL-associated lipid loci found that variants showing the strongest association were often substantially different in frequency and effect size from those identified in 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 based on our previous study[4] and 5 nominated for fine-mapping because of association to other traits.

For each of these loci, we identified the most strongly associated Metabochip variant and evaluated whether it (a) reached genome-wide significant evidence for association (to avoid chance fluctuations in regions where the signal was relatively weak) and (b) 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 (**Supplementary Table S2.11**). The two largest differences were at the loci near *PCSK9* (top GWAS variant with minor allele frequency f = 0.24 and $P = 9x10^{-24}$; 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 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 R46L variant in *PCSK9*, the variants showing strongest association in fine-mapped loci all had minor allele frequency > .05.

We also attempted fine-mapping in African (N=3,263), East Asian (N=1,771), and South Asian (N=4,901) ancestry samples. Despite comparatively small samples, ancestry-specific analyses identified SNPs clearly distinct from the original GWAS variant in five loci (**Supplementary Table S2.11**). These were: *APOE*, consistent with European ancestry analyses above; three loci where differences in linkage disequilibrium between populations enabled fine-

mapping in African (*SORT1*, *LDLR*) or East Asian (*APOA5*) ancestry samples; and *CETP*, where an African-specific variant was present. For *CETP*, *SORT1,* and *APOA5,* results are consistent with other fine-mapping and functional studies (Musunuru et al. 2010; Buyske et al. 2012; Palmen et al. 2012).

Association of lipid loci with metabolic and cardiovascular traits

To evaluate the role of the 157 loci identified here on related traits, we evaluated the most strongly associated SNPs for each locus in genetic studies of coronary artery disease (CAD, N=114,590 including 37,653 cases) (Schunkert et al. 2011; The Coronary Artery Disease (C4D) Consortium 2010), 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) (International Consortium for Blood Pressure Genome-Wide Association Studies 2011), and fasting glucose (N=46,186 non-diabetics) (Dupuis et al. 2010). We observed an excess of SNPs nominally associated ($P < 0.05$) with all these traits: a 5.1 fold excess for CAD (40 nominally significant loci, $P = 2\text{x}10^{-19}$), a 4.1 fold excess for BMI (32 loci, $P = 1\text{x}10^{-11}$), 3.7 fold excesses for DBP (29 loci, $P = 1\text{x}10^{-9}$), a 3.4 fold excess for WHR (27 loci, $P = 1\text{x}10^{-9}$), a 2.5 fold excess for SBP (20 loci, $P = 1\text{x}10^{-4}$), a 2.3 fold excess for T2D (18 loci, $P = 0.001$), and a 2.2 fold excess for fasting glucose (17 loci, $P = 3\text{x}10^{-3}$) (**Supplementary Table S2.12**). Interestingly, among the novel loci, we observed greater overlap with BMI, SBP, and DBP (9 overlapping loci each) than with CAD (8 overlapping loci). Among new loci, the two SNPs showing strongest association to CAD map near *RBM5* (rs2013208, $P_{HDL} = 9\text{x}10^{-12}$, $P_{CAD} = 7\text{x}10^{-5}$) and *CMTM6* (rs7640978, $P_{LDL} = 1\text{x}10^{-8}$, $P_{CAD} = 4\text{x}10^{-4}$).

We tested whether the LDL-, total cholesterol- or triglyceride- increasing allele, or HDL-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 2A-D**. 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, BMI, SBP, and DBP), and near *MIR581* (associated with HDL cholesterol, BMI, T2D, and DBP). In some cases, like *FTO*, a strong association with BMI or another phenotype generates weaker association signals for other metabolic traits (Freathy et al. 2008). In other cases, like *SORT1*, a primary effect on lipid levels may mediate secondary association with other traits, like CAD (Musunuru et al. 2010).

Association of individual lipids with coronary artery disease

Epidemiological studies consistently show 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 and CAD is clear, whereas the results for HDL cholesterol levels are more equivocal (Willer et al. 2008; Voight et al. 2012; Frikke-Schmidt et al. 2008). In our data, trait increasing alleles at the loci showing strongest association with LDL cholesterol (31 loci), triglycerides (30 loci), or total cholesterol (38 loci) were associated with increased risk of CAD ($P = 2x10^{-12}$, $P = 2x10^{-16}$, and $P = 0.006$). Conversely, trait decreasing alleles at loci showing the strongest association with HDL cholesterol (64 loci), were associated with increased CAD risk

with P = 0.02. When we focused on loci uniquely associated with LDL cholesterol (12 loci where P > .05 for other lipids), triglycerides (6 loci), or HDL cholesterol (14 loci), only the LDL association remained significant (P = 0.03).

To better explore how associations with individual lipid levels 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 loci (CAD results were not available for 8 SNPs) was correlated with LDL (Pearson r=0.74, $P = 7x10^{-6}$) and triglyceride (Pearson r=0.46, P = 0.02) effect sizes, but not HDL effect sizes (Pearson r=$-9x10^{-4}$, P = 0.99; **Supplementary Fig. S2.6**). Since most variants affect multiple lipid fractions (**Figure 2.1**), dissecting the relationship between lipid level and CAD effects requires multivariate analysis. In a companion manuscript, we use multivariate analysis and detailed examination of triglyceride associated loci to show that increased LDL and triglyceride levels, but not HDL, appear causally related to CAD risk.

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

To evaluate evidence for loci not yet reaching genome-wide significance, we compared direction of effect in GWAS and Metabochip analyses of non-overlapping samples, outside the 157 genome-wide significant loci. Among independent variants ($r^2 < 0.1$) with $P < 0.1$ in the GWAS-only analysis, a significant excess were concordant in direction of effect for HDL (62.9% in 1,847 SNPs, $P < 10^{-16}$), LDL (58.6% of 1,730 SNPs, $P < 10^{-16}$), triglyceride levels (59.1% of 1,783 SNPs, $P < 10^{-16}$), and total cholesterol (61.0% of 1,904 SNPs, $P < 10^{-16}$), suggesting many additional loci to be discovered in future studies.

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 (See **Tables 2A-D** and **Figure 2.1**). All but one of the loci identified here include protein-coding genes within 100 kb of the SNP showing strongest association. While 38 of the 62 new loci include genes whose role in blood lipid levels is supported by literature review or analysis of curated pathway databases, the remainder includes only genes whose role on blood lipid levels has not been documented.

In total, there are 240 genes within 100 kb of one 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 **Supplementary Table S2.2**). While we found significant overlap, different sources of prioritization sometimes disagreed. This result suggests that truly understanding causality will be very challenging. The **Supplementary Note** includes an interpreted digest of genes highlighted by our study. Clearly, a range of approaches will be needed to follow-up these findings. To illustrate possibilities, consider U. S. Patent Application #20,090,036,394 disclosing that, in the mouse, knockout of *Gpr146* modifies blood lipid levels. Here, we show that variants near the human homologue of this gene, *GPR146*, are associated with 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 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 prior GWAS signal in 12 instances. In three of these 12 instances, fine-mapping was enabled by analysis of a few thousand African or East Asian ancestry individuals, whereas in the remaining instances, fine-mapping was possible through examination of nearly 100,000 European ancestry samples. A more detailed fine-mapping exercise, including imputation of variants from emerging very large reference panels, may help refine the location of additional signals.

Lipid-associated loci were strongly associated with CAD, T2D, BMI, SBP, and DBP. In univariate analyses, we found that impact on LDL and triglycerides all predicted association with CAD, but HDL did not. In a more detailed multivariate investigation, a companion manuscript shows that our data is consistent with the hypothesis that both LDL and triglycerides, but not HDL, are causally related to CAD risk. HDL, LDL, and triglycerides levels summarize aggregate levels of different lipid particles, each with potentially distinct consequences for CAD risk. We evaluated association of our loci with lipid subfractions in 2,900 individuals from the Framingham Heart Study (**Supplementary Table S2.13, Supplementary Fig. S2.7**) and with sphingolipids, which are components of lipid membranes in cells, in 4,034 individuals from five samples of European ancestry (Demirkan et al. 2012) (**Supplementary Table S2.14**). The results suggest HDL-associated variants can have a markedly different impact on these sub-phenotypes.

For example, among HDL loci, variants near *LIPC* were strongly associated with plasmalogen levels ($P < 10^{-40}$), variants near *ABCA1* were associated with sphingomyelin levels ($P < 10^{-5}$), and variants near *CETP* – which show the strongest association with HDL cholesterol overall – were associated with neither of these. Detailed genetic dissection of these sub-phenotypes in larger samples, could lead to functional groupings of HDL-associated variants that reconcile the results of genetic studies (which show no clear connection between HDL cholesterol-associated variants and CAD risk) and epidemiologic studies (which show clear association between plasma HDL 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.

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/lipids2013/

ACKNOWLEDGEMENTS

We especially thank the >196,000 volunteers who participated in our study. Detailed acknowledgement of funding sources is provided in the supplementary online material.

AUTHOR CONTRIBUTIONS

Writing and Analysis Group

G.R.A., M.B., L.A.C., P.D., P.W.F., S.K., K.L.M., E.I., G.M.P., S.S.R., S.R., M.S.S., E.M.S., S.S., C.J.W. (Lead). E.M.S. and S.S. performed meta-analysis and E.M.S., S.S., G.M.P., M.B., J.C., S.G., A.G., and S. K. performed bioinformatics analyses. E.M.S. and S.S. prepared the tables, figures and supplementary material. C.J.W. led the analysis and bioinformatics efforts. E.I. and K.M. led the biological interpretation of results. C.J.W. and G.R.A. wrote the manuscript. All analysis and writing group authors extensively discussed the analysis, results, interpretation and presentation of results.

All authors contributed to the research and reviewed the manuscript.

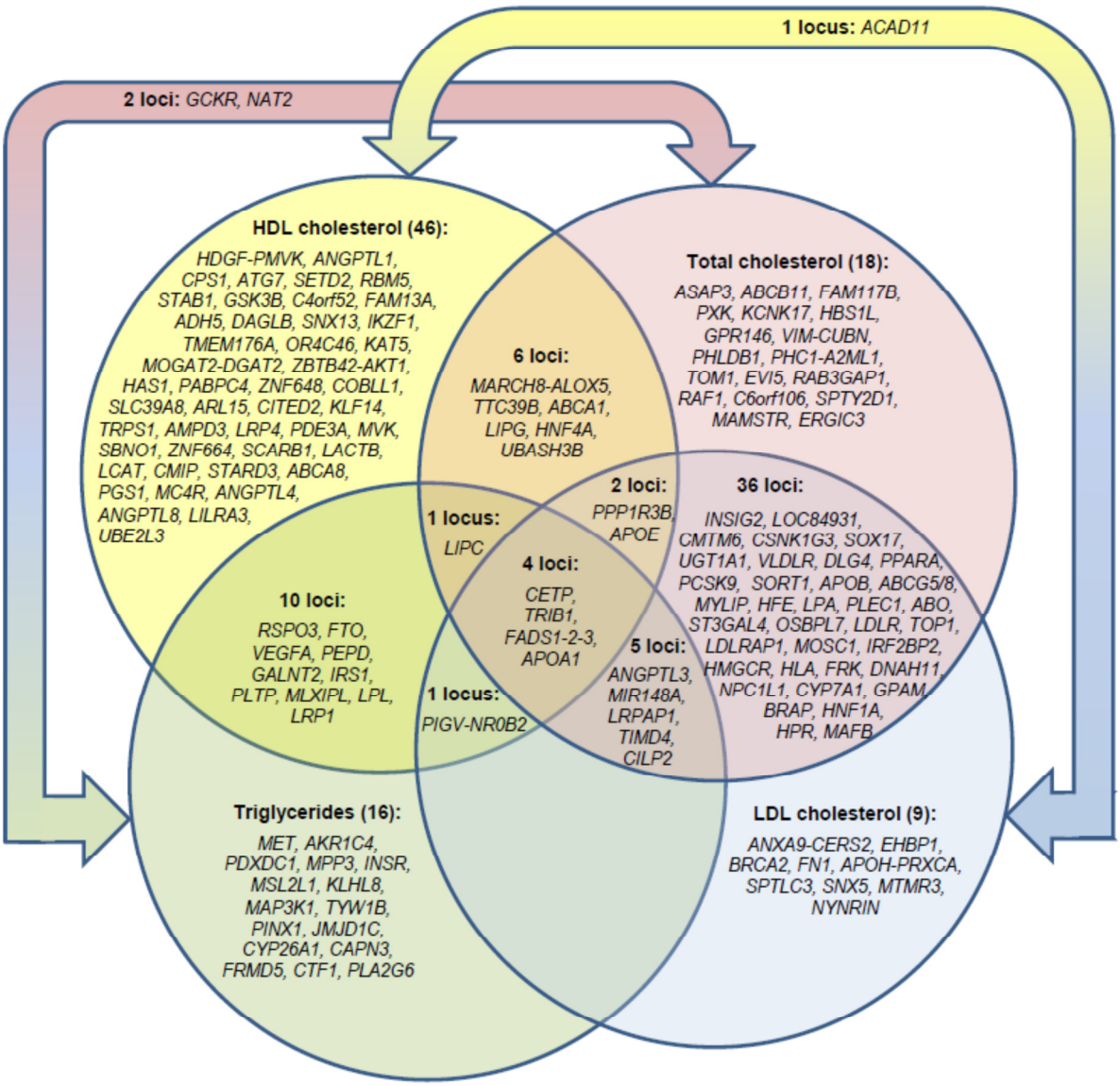


Figure 2.1 Overlap between loci associated with different lipid traits

This Venn diagram illustrates the number of loci that show association with multiple lipid traits. The number of loci primarily associated with only one trait is listed in parentheses after the trait name and the locus name is listed below in italics. Loci that show association with two or more traits are shown in the appropriate section.

Table 2-A Novel Loci Primarily Associated with HDL Cholesterol Obtained 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 (in 1000s)	Joint *P*-value
PIGV-NR0B2	rs12748152	1	27.14	HDL, LDL, TG	.09	T/C	-.051/.050/.037	187/173/178	$1x10^{-15}/3x10^{-12}/1x10^{-9}$
HDGF-PMVK	rs12145743	1	156.70	HDL	.34	G/T	.020	181	$2x10^{-8}$
ANGPTL1	rs4650994	1	178.52	HDL	.49	G/A	.021	187	$7x10^{-9}$
CPS1	rs1047891	2	211.54	HDL	.33	A/C	-.027	182	$9x10^{-10}$
ATG7	rs2606736	3	11.40	HDL	.39	C/T	.025	129	$5x10^{-8}$
SETD2	rs2290547	3	47.06	HDL	.20	A/G	-.030	187	$4x10^{-9}$
RBM5	rs2013208	3	50.13	HDL	.50	T/C	.025	170	$9x10^{-12}$
STAB1	rs13326165	3	52.53	HDL	.21	A/G	.029	187	$9x10^{-11}$
GSK3B	rs6805251	3	119.56	HDL	.39	T/C	.020	186	$1x10^{-8}$
C4orf52	rs10019888	4	26.06	HDL	.18	G/A	-.027	187	$5x10^{-8}$
FAM13A	rs3822072	4	89.74	HDL	.46	A/G	-.025	187	$4x10^{-12}$
ADH5	rs2602836	4	100.01	HDL	.44	A/G	.019	187	$5x10^{-8}$
RSPO3	rs1936800	6	127.44	HDL, TG[a]	.49	C/T	.020/-.020	187/168	$3x10^{-10}/3x10^{-8}$
DAGLB	rs702485	7	6.45	HDL	.45	G/A	.024	187	$7x10^{-12}$
SNX13	rs4142995	7	17.92	HDL	.38	T/G	-.026	165	$9x10^{-12}$
IKZF1	rs4917014	7	50.31	HDL	.32	G/T	.022	187	$1x10^{-8}$
TMEM176A	rs17173637	7	150.53	HDL	.12	C/T	-.036	184	$2x10^{-8}$
MARCH8-ALOX5	rs970548	10	46.01	HDL, TC	.26	C/A	.026/-.026	187/187	$2x10^{-10}/8x10^{-9}$
OR4C46	rs11246602	11	51.51	HDL	.15	C/T	.034	176	$2x10^{-10}$
KAT5	rs12801636	11	65.39	HDL	.23	A/G	.024	187	$3x10^{-8}$
MOGAT2-DGAT2	rs499974	11	75.46	HDL	.19	A/C	-.026	187	$1x10^{-8}$
ZBTB42-AKT1	rs4983559	14	105.28	HDL	.40	G/A	.020	184	$1x10^{-8}$
FTO	rs1121980	16	53.81	HDL, TG	.43	A/G	-.020/-.021	186/155	$7x10^{-9}/3x10^{-8}$
HAS1	rs17695224	19	52.32	HDL	.26	A/G	-.029	185	$2x10^{-13}$

Chr, chromosome; MAF, minor allele frequency; A1, minor allele; A2, major allele. 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. At one locus, the secondary trait was most strongly associated with a different SNP: [a]rs719726 (within 1Mb of rs1936800, $r^2 = 0.74$).

Table 2-B Novel Loci Primarily Associated with LDL Cholesterol Obtained 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 (in 1000s)	Joint *P*-value
ANXA9-CERS2	rs267733	1	150.96	LDL	.16	G/A	-.033	165	$5x10^{-9}$
EHBP1	rs2710642	2	63.15	LDL	.35	G/A	-.024	173	$6x10^{-9}$
INSIG2	rs10490626	2	118.84	LDL, TC[b]	.08	A/G	-.051/.042	173/184	$2x10^{-12}/6x10^{-9}$
LOC84931	rs2030746	2	121.31	LDL, TC	.40	T/C	.021/.020	173/187	$9x10^{-9}/4x10^{-8}$
FN1	rs1250229	2	216.30	LDL	.27	T/C	-.024	173	$3x10^{-8}$
CMTM6	rs7640978	3	32.53	LDL, TC	.09	T/C	-.039/-.038	172/186	$1x10^{-8}$
ACAD11	rs17404153	3	132.16	LDL, HDL[c]	.14	T/G	-.034/.028	172/187	$2x10^{-9}/5x10^{-9}$
CSNK1G3	rs4530754	5	122.86	LDL, TC	.46	G/A	-.028/-.023	173/187	$4x10^{-12}/2x10^{-9}$
MIR148A	rs4722551	7	25.99	LDL, TG[d], TC	.20	C/T	.039/.029/.023	173/187/178	$4x10^{-14}/9x10^{-11}/7.0x10^{-9}$
SOX17	rs10102164	8	55.42	LDL, TC	.21	A/G	.032/.030	173/187	$4x10^{-11}/5x10^{-11}$
BRCA2	rs4942486	13	32.95	LDL	.48	T/C	.024	172	$2x10^{-11}$
APOH-PRXCA	rs1801689	17	64.21	LDL	.04	C/A	.103	111	$1x10^{-11}$
SPTLC3	rs364585	20	12.96	LDL	.38	A/G	-.025	172	$4x10^{-10}$
SNX5	rs2328223	20	17.85	LDL	.21	C/A	.03	171	$6x10^{-9}$
MTMR3	rs5763662	22	30.38	LDL	.04	T/C	.077	163	$1x10^{-8}$

Chr, chromosome; MAF, minor allele frequency; A1, minor allele; A2, major allele. 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. At three loci, secondary traits were most strongly associated with different SNPs. [b]rs17526895 (within 1Mb of rs10490626, $r^2 = 0.98$); [c]rs13076253 (within 1Mb of rs17404153, $r^2 = 0.00$); [d]rs4719841 (within 1Mb of rs4722551, $r^2 = 0.10$).

Table 2-C Novel Loci Primarily Associated with Total Cholesterol Obtained 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 (in 1000s)	Joint *P*-value
ASAP3	rs1077514	1	23.77	TC	.15	C/T	-0.03	184	$6x10^{-9}$
ABCB11	rs2287623	2	169.83	TC	.41	G/A	0.027	184	$4x10^{-12}$
FAM117B	rs11694172	2	203.53	TC	.25	G/A	0.028	187	$2x10^{-9}$
UGT1A1	rs11563251	2	234.68	TC, LDL	.12	T/C	0.037/0.034	187/173	$1x10^{-9}/5x10^{-8}$
PXK	rs13315871	3	58.38	TC	.10	A/G	-0.036	187	$4x10^{-8}$
KCNK17	rs2758886	6	39.25	TC	.30	A/G	0.023	187	$3x10^{-8}$
HBS1L	rs9376090	6	135.41	TC	.28	T/C	-0.025	187	$3x10^{-9}$
GPR146	rs1997243	7	1.08	TC	.16	G/A	0.033	183	$3x10^{-10}$
VLDLR	rs3780181	9	2.64	TC, LDL	.08	G/A	-0.044/-0.044	186/172	$7x10^{-10}/2x10^{-9}$
VIM-CUBN	rs10904908	10	17.26	TC	.43	G/A	0.025	187	$3x10^{-11}$
PHLDB1	rs11603023	11	118.49	TC	.42	T/C	0.022	187	$1x10^{-8}$
PHC1-A2ML1	rs4883201	12	9.08	TC	.12	G/A	-0.035	187	$2x10^{-9}$
DLG4	rs314253	17	7.09	TC, LDL	.37	C/T	-0.023/-0.024	184/170	$3x10^{-10}/3x10^{-10}$
TOM1	rs138777	22	35.71	TC	.36	A/G	0.021	185	$5x10^{-8}$
PPARA	rs4253772	22	46.63	TC, LDL[e]	.11	T/C	0.032/-0.031	185/171	$1x10^{-8}/3x10^{-8}$

Chr, chromosome; MAF, minor allele frequency; A1, minor allele; A2, major allele. 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. At one locus, the secondary trait was most strongly associated with a different SNP: [e]rs4253776 (within 1Mb of rs4253772, $r^2 = 0.95$).

Table 2-D Novel Loci Primarily Associated with Triglycerides Obtained 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 (in 1000s)	Joint *P*-value
LRPAP1	rs6831256	4	3.47	TG, TC[f],LDL[f]	.42	G/A	0.026/-0.022/-0.025	177/173/187	$2x10^{-12}/1x10^{-10}/2x10^{-8}$
VEGFA	rs998584	6	43.76	TG, HDL	.49	A/C	0.029/-0.026	175/184	$3x10^{-15}/2x10^{-11}$
MET	rs38855	7	116.36	TG	.47	G/A	-0.019	178	$2x10^{-8}$
AKR1C4	rs1832007	10	5.25	TG	.18	G/A	-0.033	178	$2x10^{-12}$
PDXDC1	rs3198697	16	15.13	TG	.43	T/C	-0.020	176	$2x10^{-8}$
MPP3	rs8077889	17	41.88	TG	.22	C/A	0.025	176	$1x10^{-8}$
INSR	rs7248104	19	7.22	TG	.42	A/G	-0.022	176	$5x10^{-10}$
PEPD	rs731839	19	33.90	TG, HDL	.35	G/A	0.022/-0.022	176/185	$3x10^{-9}/3x10^{-9}$

Chr, chromosome; MAF, minor allele frequency; A1, minor allele; A2, major allele. 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. At one locus, secondary traits were most strongly associated with a different SNP: [f]rs6818397 (within 1 Mb of rs6831256, $r^2 = 0.18$).

Online Methods

Samples studied: We collected summary statistics for Metabochip SNPs from 45 studies. Among these, 37 studies consisted primarily of individuals of European ancestry (see **Supplementary Table 1** and **Supplementary Note** for details), including both population-based studies and case-control studies of CAD and T2D. Another 8 studies consisted primarily of individuals with non-European ancestry: two studies of South Asian descent, AIDHS/SDS (N=1,516) and PROMIS (N=3,385); two studies of East Asian descent, CLHNS (N=1,771) and TAI-CHI (N=7044); and five studies of recent African ancestry, MRC/UVRI GPC (N=1,687) from Uganda, SEY (N=426) from the Caribbean, and FBPP (N=1,614, TG results unavailable), GXE (N=397), and SPT (N=838) from the United States (more details in **Supplementary Table 1** and **Supplementary Note**).

Genotyping: We genotyped 196,710 genetic variants prioritized on the basis of prior GWAS for cardiovascular and metabolic phenotypes using the Illumina iSelect Metabochip (Voight et al. 2012) 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 for LDL cholesterol, 5,056 for triglycerides, and 938 for total cholesterol. These independent SNPs represent most loci with $P < .005$ in our original GWAS for HDL cholesterol, LDL cholesterol and triglycerides and with $P < .0005$ 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 based on 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).

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 10 studies (24% of total study individuals) and 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 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 individuals); all studies were adjusted using genomic control prior to meta-analysis (Devlin and Roeder 2012). In studies ascertained on diabetes or CVD status, cases and controls were analyzed separately (**Supplementary Table 1**). All meta-analyses were limited to a single ancestral group (e.g. European only).

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 individuals), EMMAX (9 samples, 14% of individuals), Merlin (4 samples, 9% of individuals), GENABEL (1 sample, 3% of individuals), and MMAP (1 sample, 1% of individuals) (**Supplementary Table 1**).

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 > 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 < $5x10^{-8}$ in the combined GWAS and Metabochip meta-analysis and were >1 Mb away from the nearest previously described lipid locus and other novel loci. We used only European samples for the discovery of novel genome-wide significant loci. The non-European samples were meta-analyzed and examined only for fine-mapping analyses.

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.

Proportion of trait variance explained: We estimated the increase in trait variance explained by novel loci in the Framingham cohort (N=7,132) using three models for each trait-residual: 1) lead and secondary SNPs from the previously published loci (Teslovich et al. 2010) and 2) previously published lipid loci plus newly reported loci; and 3) newly reported loci. We regressed lipid residuals on these sets of SNPs using the lme kinship package in R.

Initial automated review of the published literature: An initial list of candidates within each locus was generated with Snipper (http://csg.sph.umich.edu/boehnke/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.

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 SNPs ($P > 0.10$ for any of the 4 lipid traits) and created bins based on 3 statistics: minor allele frequency, 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 one SNP for each permuted list.

Pathway analyses: To investigate if lipid-associated variants overlapped previously annotated pathways, we used gene set enrichment analysis (GSEA), as implemented in MAGENTA (Segre 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, 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 their gene association score, 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 by 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 for GSEA *P*-values below $1x10^{-4}$.

We used the Disease Association Protein–Protein Link Evaluator package (DAPPLE; http://www.broadinstitute.org/mpg/dapple/dapple.php) to examine evidence for protein-protein interaction networks connecting genes across different lipid loci. This analysis included the 62 novel loci as well as the 95 previously known loci; we focus our discussion on pathways that included one or more genes from novel loci.

Cis-expression quantitative trait locus analysis: To determine whether lipid-associated SNPs might act as *cis*-regulators of nearby genes, we examined association with 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 the 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).

Functional annotation of associated variants: We attempted to identify lipid-associated SNPs that fall 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 linkage disequilibrium ($r^2 > 0.8$ from 1000 Genomes or HapMap) with each index SNP. We compared the position of the index SNPs and their proxies to previously described functional marks (Ernst et al. 2011; The ENCODE Project Consortium

2011). To assess the expected overlap with functional marks, we created 100,000 permuted sets of non-associated SNPs (see above) 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.

Association with lipid subfractions: Lipoprotein fractions for Women's Genome Health Study (WGHS) samples (N = 23170) were measured using the LipoProtein-II assay (Liposcience Inc. Raleigh, NC) and Framingham Heart Study Offspring samples (N = 2900) were measured with the LipoProtein-I assay (Liposcience Inc. Raleigh, NC) (Chasman et al. 2009). Additional information on sub-fraction measurements can be found in **Supplementary Fig. 7**. Log transformations were used for non-normalized traits. All models were adjusted for age, sex, and PCs. The genetic association analysis of WGHS used SNP genotypes imputed from the HapMap r22 CEU reference panel using MACH. 16,730 out of 23,170 WGHS participants were fasting for 8 hours prior to blood draw (72.2%).

Supplementary Tables

Supplementary Table S2.1: Phenotypic Summary of Samples with Metabochip Genotype Results

Short study name	N	% Female	Mean age (SD)	Mean HDL (SD) mg/dL	Mean LDL (SD) mg/dL	Mean TC (SD) mg/dL	Mean TG (SD) mg/dL	Excluded individuals on lipid-lowering medication	LDL-C estimated using Friedewald (F) or measured (M)	Fasting > 8 hrs (F) or non-fasting (NF) blood draw	Adjustment for population structure with PCA	Analysis software used	Study Reference (PMID)
ADVANCE	505	40.6	65.7 (2.9)	54.6 (16.5)	128.3 (29.9)	209.7 (34.5)	137.7 (87.6)	Yes	F	F	Yes	PLINK	18443000
AIDHS/SDS [a]	1516	47.2	53.0 (12.1)	38.8 (13.8)	108.9 (38.7)	181.5 (49.2)	181.5 (116.7)	Yes	F	F	Yes	PLINK	18598350
AMC-PAS	304	25.0	43.0 (5.4)	44.2 (14.8)	148.9 (46)	234.2 (59.9)	167.1 (112.9)	Yes	F	F	No	PLINK	19164808
AMISH	1081	50.0	46.7 (15.1)	54.2 (11.6)	135.5 (38.7)	209 (42.6)	74.4 (44.3)	Yes	F	F	No	MMAP	17261661
BC58	2136	57.8	45.0 (0)	60 (15.1)	134.7 (35.6)	230.3 (41.8)	185.2 (130.2)	Yes	M	NF	No	SNPTest	16155052
CLHNS [a]	1771	47.3	21.5 (0.3)	42.3 (11.2)	94 (29.2)	157.2 (36.6)	105.5 (65.6)	Yes	F	F	Yes	PLINK	20507864
D2D 2007 (T2D)	287	43.6	62.3 (7.8)	50.8 (12.4)	132.4 (34.5)	213.2 (41.2)	151.1 (90.8)	Yes	F	F	Yes	EMMAX	20459722
D2D 2007 (controls)	1821	56.3	58.3 (8.3)	57.4 (13.4)	138 (31.3)	218 (35.9)	114.6 (70.6)	Yes	F	F	Yes	EMMAX	20459722
deCODE	15612	62.8	60.9 (17.1)	56 (17.9)	135.2 (39.9)	217.3 (43.8)	134.7 (81.8)	Yes	F	F	No	SNPTEST	17478679
DIAGEN (T2D)	439	50.3	66.0 (11.9)	47.5 (16)	114 (38.4)	198.7 (48.2)	199.6 (176.4)	No	F	F	Yes	EMMAX	16801592
DIAGEN (controls)	1093	56.7	62.4 (15.2)	59.3 (17.9)	127.7 (38.4)	207.2 (46.4)	135 (170.3)	No	F	F	Yes	EMMAX	16801592
DILGOM	3738	58.1	51.6 (13.5)	56.1 (13.5)	122.3 (32.1)	202 (36.8)	104.5 (62.9)	Yes	M	F	No	PLINK	19959603
DPS (T2D)	85	63.5	55.1 (6.4)	44 (11.7)	134.4 (32.8)	212.2 (37.7)	172.3 (85.5)	Yes	F	F	Yes	EMMAX	11333990
DPS (controls)	362	69.6	55.2 (7.3)	47.9 (11.1)	141.5 (31.2)	218.3 (34.2)	147.2 (61.8)	Yes	F	F	Yes	EMMAX	11333990
DRAGON (TAICHI) [a]	1052	41.7	62.9 (14.6)	43.3 (16.3)	101.9 (40.6)	174.5 (47.8)	146.6 (100.1)	Yes	F	F	Yes	PLINK	18632180
DR'S EXTRA (T2D)	121	50.4	68.7 (5.8)	58.3 (17.4)	112.1 (33.6)	181.1 (36.4)	142.6 (73.3)	No	M	F	Yes	EMMAX	21186108
DR'S EXTRA (controls)	1174	53.8	66.4 (5.3)	66.8 (18.6)	125.6 (32.2)	198.7 (35.9)	115.5 (58.9)	No	M	F	Yes	EMMAX	21186108
EAS	733	53.0	64.4 (5.7)	55.7 (13.2)	206.7 (47.2)	274.4 (51.1)	461.6 (152.4)	Yes	F	F	Yes	PLINK	1917239
EGCUT	1240	53.5	64.2 (10.9)	53.8 (27.9)	136.6 (44.1)	211.3 (46.8)	161.3 (111.6)	Yes	M	F	Yes	SNPTest	19424496
Ely	1602	53.6	61.1 (9.2)	56.5 (15.5)	135.5 (36.8)	215.6 (41)	128.5 (74.4)	No	F	F	No	PLINK	7712700
EPIC-CAD cases (EPIC-Norfolk CAD set)	1529	35.1	65.2 (7.9)	49.9 (14.3)	164.1 (40.2)	250.8 (46.4)	192.3 (108.1)	No	F	NF	No	PLINK	10466767
EPIC-T2D cases (EPIC-Norfolk T2D set)	700	40.3	62.2 (8.3)	46.4 (12.8)	153.3 (40.2)	245.4 (48.4)	243.7 (140.9)	No	F	NF	No	PLINK	14693662

EPIC-T2D controls (EPIC-Norfolk T2D set)	994	56.8	59.4 (9.4)	55 (16.3)	154 (38.3)	239.2 (43)	156.8 (89.5)	No	F	NF	No	PLINK	14693662
FBPP [a]	1614	64.0	43.9 (12.5)	53.6 (15.8)	118.4 (37.8)	193.2 (43.5)	NA	Yes	F	F	No	Merlin	11799070
Fenland	3186	53.3	46.9 (7.1)	58.8 (15.1)	129.3 (34.8)	207 (39.5)	103.7 (71.8)	No	F	F	No	PLINK	
FINCAVAS	1201	44.4	58.5 (12.6)	54.6 (17.8)	113 (33.3)	191.2 (38.7)	119.6 (65.6)	Yes	F	F	Yes	SNPTest	16515696
FRISCII	2963	30.6	66.2 (9.7)	46.7 (15.2)	145.7 (38.2)	228.1 (43.1)	185.7 (119.5)	Yes	F	F	No	GenABEL	10892758
FUSION2 (T2D)	843	43.3	59.5 (8.5)	51.1 (14.8)	123.1 (36.2)	217.6 (47.3)	163.3 (92.8)	Yes	F	F	Yes	EMMAX	17463248
FUSION2 (controls)	1880	44.7	56.2 (8.2)	58.1 (15.8)	137.9 (34.9)	225.7 (42.5)	114.8 (68.5)	Yes	F	F	Yes	EMMAX	17463248
GLACIER	5764[b]	61.4 [b]	54.2 (8.0) [b]	56.3 (13.4)	173 (45.3)	232.5 (47.4)	142.6 (47.4)	Yes	F	F	Yes	PLINK	20870969
Go-DARTs	6759	46.1	57.8 (10.3)	55 (17)	120.7 (35.2)	216.7 (46.8)	177.2 (126.7)	No	F	F	No	PLINK	17429603
GXE [a]	397	76.8	39.8 (8.2)	50.9 (12)	146.1 (41.6)	204.3 (43.7)	84.7 (50.1)	Yes	F	F	Yes	PLINK	21347282
HALST (TAICHI) [a]	2375	49.4	68.9 (8.3)	52.6 (13.5)	120.0 (31.8)	197.7 (36.8)	123.1 (76.2)	Yes	F	F	Yes	PLINK	
HUNT (T2D)	588	49.5	69.3 (11.3)	45.8 (14.6)	151.9 (42)	241.7 (51.8)	238.7 (163.4)	No	F	NF	Yes	EMMAX	22879362
HUNT (controls)	784	49.0	66.3 (14.4)	50.9 (14.3)	163.4 (44.1)	249.7 (48.3)	183.8 (101.1)	No	F	NF	Yes	EMMAX	22879362
IMPROVE	1769	50.0	64.4 (5.3)	48.8 (14.7)	148.6 (36.4)	224.1 (42.2)	139.1 (96.6)	Yes	F	F	Yes	PLINK	19952003
KORA F3	2816	52.1	56.3 (12.8)	59.2 (18.2)	129.6 (32.1)	220.2 (39.5)	163.9 (124.9)	Yes	M	NF	No	PLINK	16032514
KORA F4	2678	53.1	54.5 (13.1)	56.5 (14.7)	138.5 (34.8)	218.3 (39.5)	122.3 (86.8)	Yes	M	F	No	PLINK	16032514
MRC/UVRIGPC[ad]	1687	56.7	35.0 (19.1)	39.9 (14.3)	78.6 (29)	138.9 (36.8)	104.5 (55.8)	No	M	NF	No	EMMAX	
LURIC (cases)	983	26.4	64.9 (9.7)	38.5 (10.7)	124.5 (33.4)	201.2 (37.3)	170 (99.9)	Yes	M	F	No	PLINK	11258203
LURIC (controls)	523	46.1	57.9 (12.5)	42.8 (11.5)	121.2 (29.4)	198.8 (35)	154.5 (100.9)	Yes	M	F	No	PLINK	11258203
MDC	2125	51.3	57.4 (6.0)	53.9 (13.9)	160 (36.6)	235.5 (40.2)	109.2 (50.5)	No	F	F	Yes	PLINK	8429286
METSIM (T2D)	634	0.0	59.7 (6.8)	52.1 (16.1)	134.6 (36.2)	212.8 (44.4)	167 (111.8)	Yes	M	F	Yes	EMMAX	19223598
METSIM (controls)	829	0.0	53.7 (5)	57.3 (14.7)	138.4 (29.9)	215.3 (34)	120.1 (76.2)	Yes	M	F	Yes	EMMAX	19223598
NFBC86	4164	52.0	16.0 (0.4)	54.6 (11.2)	87.1 (22.4)	164.9 (30.2)	73.5 (36.3)	No	F	F	No	PLINK	
NSHD	941	52.8	53.0	63.5 (19.7)	136.2 (38.3)	237.2 (43)	206.4 (154.2)	No	F	NF	No	PLINK	16204333
PIVUS	854	51.0	70.0 (0.2)	58.1 (15.5)	135.5 (31)	216.7 (38.7)	106.3 (44.3)	Yes	F	F	Yes	PLINK	18489581
PROMIS [a]	3385	18.0	52.5 (9.9)	35.5 (9.9)	122.3 (43.7)	192.7 (50.7)	210.2 (128.8)	No	M	NF	Yes	PLINK	19404752
SAPPHIRe (TAICHI) [a]	251	49.4	54.6 (10.5)	44.2 (12.6)	127.1 (38.2)	200.4 (43.4)	143.0 (80.0)	Yes	F	F	Yes	PLINK	22839215
SardiNIA	5378	56.8	43.2 (17.4)	64.3 (14.9)	127.1 (35.5)	208.4 (42.6)	86.9 (68.3)	Yes	F	F	No	Merlin	16934002
SCARFSHEEP	2973	0.3	58.3 (7.2)	45.7 (13.9)	156.3 (37.5)	230.7 (42.2)	148 (101)	Yes	F	F	No	PLINK	
SEY [a]	426	54.7	48.7 (14.1)	48.3 (13.1)	141.4 (44.1)	213.3 (49.1)	117.8 (80.1)	Yes	F	F	No	Merlin	15610228

SPT [a]	838[c]	61.8[c]	46.7 (0.5)[c]	48.8 (13)	133.7 (41.3)	192.8 (42.1)	89.1 (53.8)	Yes	F	F	Yes	PLINK	9103091
STR	2543	57.0	75.0 (10.2)	54.2 (15.5)	150.9 (42.6)	243.8 (50.3)	150.6 (79.7)	Yes	F	F	No	Merlin	19606474
TACT (TAICHI) [a]	173	31.2	64.1 (10.7)	34.3 (8.0)	103.5 (24.3)	180.1 (36.8)	130.0 (74.2)	Yes	F	F	Yes	PLINK	19050055
TCAD (TAICHI) [a]	2284	23.3	65.8 (11.7)	44.0 (12.5)	112.3 (40.7)	182.6 (42.4)	133.2 (84.7)	Yes	F	F	Yes	PLINK	17967444
TCAGEN (TAICHI) [a]	383	34.7	64.3 (13.3)	45.5 (17.7)	112.1 (37.0)	185.5 (44.6)	166.9 (131.6)	Yes	F	F	Yes	PLINK	21184753
THISEAS	929	50.7	58.6 (13.5)	52.7 (15.7)	134.5 (38.1)	211.9 (42.6)	127.1 (77.9)	Yes	F	F	No	PLINK	20167083
TROMSO (T2D)	710	50.4	60.0 (12.5)	51.2 (14.8)	168.4 (42.2)	260.5 (46.9)	223.3 (141.5)	No	F	NF	Yes	EMMAX	21422063
TROMSO (controls)	711	50.2	60.0 (12.5)	59.5 (16.4)	166.9 (43.7)	254.9 (48.3)	145.3 (89.9)	No	F	NF	Yes	EMMAX	21422063
TUDR (TAICHI) [a]	669	45.7	64.6 (12.1)	42.0 (15.0)	105.5 (40.5)	178.2 (52.7)	151.5 (97.6)	Yes	F	F	Yes	PLINK	18632180
ULSAM	1113	0.0	71.0 (0.6)	50.3 (11.6)	150.9 (34.8)	224.5 (38.7)	124 (70.9)	Yes	F	F	Yes	PLINK	16030278
WHII	3212	23.0	48.9 (6.0)	53.4 (28.3)	160.2 (78.6)	249.2 (43.3)	129.4 (101.9)	Yes	F	F	No	PLINK	15576467

[a] Studies of non-European ancestry
[b] GLACIER sample sizes differ by trait: TC 5,764, HDL 3,052, LDL 2,034, TG 3,365; %Female: TC 61.1, HDL 61.4, LDL 59.4, TG 59.6; mean age (SD): TC 49.5 (8.7), HDL 53.3 (8.4), LDL 54.2 (8.0), TG 50.9 (8.5)
[c] SPT sample sizes differ by trait: TC 826, HDL 757, LDL 691, TG 838; % Female: TC 60.7, HDL 61.8, LDL 60.9, TG 61.3; mean age (SD): TC 46.7 (0.5), HDL 46.4 (0.5), LDL 46.5 (0.5), TG 46.7 (0.5)
[d] MRC/UVRI GPC is a GWAS cohort from which ~19,800 Metabochipfine-mapping SNPs were used in analysis

Supplementary Table S2.2: Biological Candidate Genes at Novel Loci based on Literature Search, Nonsynonymous Variants, Gene Expression Levels (eQTLs) and Pathway Analysis

Locus	Lead SNP	Chr	hg19 Position (Mb)	Traits GWS	Nearest Gene	Nearest Gene (kb)	No. of Genes within 100kb	Literature Candidate	Gene with Nonsynonymous SNP (r^2>0.8)	eQTL Gene (P<5x10^{-8})	Pathway Analysis
Loci Primarily Associated with HDL Cholesterol											
PIGV-NR0B2	rs12748152	1	27.14	HDL, LDL, TG	*PIGV*	13.5	7	*PIGV, NR0B2*	*NUDC*, C1orf172*, NR0B2*		*NR0B2*
*HDGF-PMVK**	rs12145743	1	156.70	HDL	*RRNAD1*	0	10	*HDGF, CRABP2*	*HDGF*		
*ANGPTL1**	rs4650994	1	178.52	HDL	*C1orf220*	0	3				
CPS1	rs1047891	2	211.54	HDL	*CPS1*	0	2		*CPS1*		*CPS1*
ATG7	rs2606736	3	11.40	HDL	*ATG7*	0	2				
SETD2	rs2290547	3	47.06	HDL	*SETD2*	0	4		*NBEAL2*		
RBM5	rs2013208	3	50.13	HDL	*RBM5*	0	4		*MST1R**	*RBM5*	
STAB1	rs13326165	3	52.53	HDL	*STAB1*	0	10	*STAB1, NISCH*	*NISCH*		
GSK3B	rs6805251	3	119.56	HDL	*GSK3B*	0	3	*GSK3B, NR1I2*			*GSK3B*
*C4orf52**	rs10019888	4	26.06	HDL	*C4orf52**	131.5	0				
FAM13A	rs3822072	4	89.74	HDL	*FAM13A*	0	2				
ADH5	rs2602836	4	100.01	HDL	*ADH5*	4.9	4			*ADH5*	
RSPO3	rs1936800	6	127.44	HDL, TG	*RSPO3*	4	1				
DAGLB	rs702485	7	6.42	HDL	*DAGLB*	0	5	*DAGLB*		*DAGLB*	*DAGLB*
SNX13	rs4142995	7	17.92	HDL	*SNX13*	0	1	*SNX13*			
IKZF1	rs4917014	7	50.31	HDL	*IKZF1*	0	1	*IKZF1*			
TMEM176A	rs17173637	7	150.53	HDL	*ABP1*	20.1	5			*TMEM176A*	
MARCH8-ALOX5	rs970548	10	46.01	HDL, TC	*MARCH8*	0	3	*ALOX5*	*MARCH8*		
OR4C46	rs11246602	11	51.51	HDL	*OR4C46*	3.2	2		*OR5W2*, OR5D13*, OR5AS1**		
KAT5	rs12801636	11	65.39	HDL	*PCNXL3*	0	12	*KAT5*			
MOGAT2-DGAT2	rs499974	11	75.46	HDL	*MOGAT2*	12.7	4	*MOGAT2, DGAT2*			
ZBTB42-AKT1	rs4983559	14	105.28	HDL	*ZBTB42*	6.2	7	*AKT1*			*AKT1*
FTO	rs1121980	16	53.81	HDL, TG	*FTO*	0	2				
HAS1	rs17695224	19	52.32	HDL	*FPR3*	0	6	*HAS1*			

Supplementary Table S2.2 (continued)

Locus	Lead SNP	Chr	hg19 Position (Mb)	Traits GWS	Nearest Gene	Nearest Gene (kb)	No. of Genes within 100kb	Literature Candidate	Gene with Nonsynonymous SNP (r^2>0.8)	eQTL Gene (P<5x10^{-8})	Pathway Analysis
Loci Primarily Associated with LDL Cholesterol											
ANXA9-CERS2	rs267733	1	150.96	LDL	*ANXA9*	0	10	*CERS2*	*ANXA9*		*ANXA9*
EHBP1	rs2710642	2	63.15	LDL	*EHBP1*	0	1	*EHBP1*			
INSIG2	rs10490626	2	118.84	LDL, TC	*INSIG2*	10.2	2	*INSIG2*	*CCDC93*		*INSIG2*
LOC84931	rs2030746	2	121.31	LDL, TC	*LOC84931*	85.6	1				
FN1	rs1250229	2	216.30	LDL	*FN1*	3.6	2	*FN1*	*FN1*		
CMTM6	rs7640978	3	32.53	LDL, TC	*CMTM6*	0	3		*DYNC1LI1*		
ACAD11	rs17404153	3	132.16	LDL, HDL	*DNAJC13*	0	2		*ACAD11**		
CSNK1G3	rs4530754	5	122.86	LDL, TC	*CSNK1G3*	0	2				
MIR148A	rs4722551	7	25.99	LDL, TC, TG	*MIR148A*	2.2	1				
SOX17	rs10102164	8	55.42	LDL, TC	*SOX17*	48.2	1				
BRCA2	rs4942486	13	32.95	LDL	*BRCA2*	0	5				*BRCA2*
APOH-PRXCA	rs1801689	17	64.21	LDL	*APOH*	0	3	*APOH, PRKCA*	*APOH*		*APOH*
SPTLC3	rs364585	20	12.96	LDL	*SPTLC3*	26.9	1	*SPTLC3*		*SPTLC3*	
SNX5	rs2328223	20	17.85	LDL	*SNX5*	76.3	2	*SNX5*			
MTMR3	rs5763662	22	30.38	LDL	*MTMR3*	0	2				

Supplementary Table S2.2 (continued)

Locus	Lead SNP	Chr	hg19 Position (Mb)	Traits GWS	Nearest Gene	Nearest Gene (kb)	No. of Genes within 100kb	Literature Candidate	Gene with Nonsynonymous SNP (r^2>0.8)	eQTL Gene (P<5x10^{-8})	Pathway Analysis
Loci Primarily Associated with Total Cholesterol											
ASAP3	rs1077514	1	23.77	TC	*ASAP3*	0	6				
ABCB11	rs2287623	2	169.83	TC	*ABCB11*	0	4	*ABCB11*	*ABCB11*		*ABCB11*
FAM117B	rs11694172	2	203.53	TC	*FAM117B*	0	2				
UGT1A1	rs11563251	2	234.68	TC, LDL	*UGT1A1*	0	12	*UGT1A1/3/4/5 UGT1A6/7/8/9*			*UGT1A1*
PXK	rs13315871	3	58.38	TC	*PXK*	0	4	*PXK*		*PXK*	
KCNK17	rs2758886	6	39.25	TC	*KCNK17*	15.9	4				
HBS1L	rs9376090	6	135.41	TC	*HBS1L*	35.2	2				
GPR146	rs1997243	7	1.08	TC	*C7orf50*	0	7		*GPR146*	*GPR146*	
VLDLR	rs3780181	9	2.64	TC, LDL	*VLDLR*	0	3	*VLDLR*			*VLDLR*
VIM-CUBN	rs10904908	10	17.26	TC	*VIM*	10.0	3	*VIM, CUBN*			*CUBN*
PHLDB1	rs11603023	11	118.49	TC	*PHLDB1*	0	7				
PHC1-A2ML1	rs4883201	12	9.08	TC	*PHC1*	0	4	*A2ML1*			
DLG4	rs314253	17	7.09	TC, LDL	*DLG4*	1.6	13	*ACADVL, CTDNEP1, SLC2A4*			*DLG4*
TOM1	rs138777	22	35.71	TC	*TOM1*	0	4	*HMOX1*	*HMGXB4*		
PPARA	rs4253772	22	46.63	TC, LDL	*PPARA*	0	6	*PPARA*			*PPARA*

Supplementary Table S2.2 (continued)

Locus	Lead SNP	Chr	hg19 Position (Mb)	Traits GWS	Nearest Gene	Nearest Gene (kb)	No. of Genes within 100kb	Literature Candidate	Gene with Nonsynonymous SNP (r^2>0.8)	eQTL Gene (P<5x10^{-8})	Pathway Analysis
Loci Primarily Associated with Triglycerides											
LRPAP1	rs6831256	4	3.47	TG, LDL, TC	*DOK7*	0	4	*LRPAP1*			*LRPAP1*
VEGFA	rs998584	6	43.76	TG, HDL	*VEGFA*	3.7	1	*VEGFA*			*VEGFA*
MET	rs38855	7	116.36	TG	*MET*	0	1				
AKR1C4	rs1832007	10	5.25	TG	*AKR1C4*	0	2	*AKR1C4*	*AKR1C4*		*AKR1C4*
PDXDC1	rs3198697	16	15.13	TG	*PDXDC1*	0	4				
MPP3	rs8077889	17	41.88	TG	*MPP3*	0	6				*MPP3*
INSR	rs7248104	19	7.22	TG	*INSR*	0	1				*INSR*
PEPD	rs731839	19	33.90	TG, HDL	*PEPD*	0	2	*CEBPG*			

Supplementary Table S2.2 summarizes results of our search for candidates at each locus. The locus label includes a gene used to refer to the locus throughout the text. Except for loci labeled * (*PMVK*, *ANGPTL1* and *C4orf52*) the locus label always refers to a gene within 100kb of the SNP with strongest association; in these three cases, the gene selected as the locus label was judged to be an especially worthy candidate >100kb or no genes within 100kb of the lead SNP were available. The columns labeled literature candidate, non-synonymous SNP, eQTL and pathway analysis candidate indicate genes flagged in our various searches for candidate genes, further detailed in the text and in supplementary tables. Genes with a non-synonymous SNP in disequilibrium with the lead SNP for the locus but more than 100kb away are also labeled *.

Supplementary Table S2.3: Summary of Joint Meta-Analysis Association Results for 95 Previously Discovered Lipid Loci

Nearest gene	MarkerName	Chr	hg19 Position (Mb)	Primary trait, Secondary trait(s)	MAF	Alleles minor/ major	Effect	Joint N (in 1000s)	Joint *P*-value
Loci Primarily Associated with HDL Cholesterol									
PABPC4	rs4660293	1	40.03	HDL	.24	G/A	-.035	187	$3x10^{-18}$
ZNF648	rs1689800	1	182.17	HDL	.35	G/A	-.034	187	$5x10^{-20}$
GALNT2	rs4846914	1	230.30	HDL,TG	.41	G/A	-.048/.040	187/178	$4x10^{-41}/7x10^{-31}$
COBLL1	rs12328675	2	165.54	HDL	.13	C/T	.045	187	$2x10^{-15}$
IRS1	rs2972146	2	227.10	HDL,TG	.37	G/T	.032/-.028	184/175	$2x10^{-17}/3x10^{-15}$
SLC39A8	rs13107325	4	103.19	HDL	.08	T/C	-.071	179	$1x10^{-15}$
ARL15	rs6450176	5	53.30	HDL	.26	A/G	-.025	187	$7x10^{-10}$
CITED2	rs605066	6	139.83	HDL	.42	C/T	-.028	94	$3x10^{-8}$
KLF14	rs4731702	7	130.43	HDL	.49	T/C	.029	187	$5x10^{-17}$
PPP1R3B	rs9987289	8	9.18	HDL,TC,LDL	.10	A/G	-.082/-.084/-.071	169/174/160	$2x10^{-41}/2x10^{-36}/9x10^{-24}$
TRPS1	rs2293889	8	116.60	HDL	.41	T/G	-.031	180102	$4x10^{-17}$
TTC39B	rs581080	9	15.31	HDL,TC	.21	G/C	-.042/-.038	187/187	$1x10^{-19}/1x10^{-13}$
ABCA1	rs1883025	9	107.66	HDL,TC	.25	T/C	-.07/-.067	186/187	$2x10^{-65}/6x10^{-53}$
AMPD3	rs2923084	11	10.39	HDL	.18	G/A	-.026	187	$5x10^{-8}$
LRP4	rs3136441	11	46.74	HDL	.18	C/T	.054	187	$7x10^{-29}$
PDE3A	rs7134375	12	20.47	HDL	.43	A/C	.021	187	$1x10^{-8}$
MVK	rs7134594	12	110.00	HDL	.48	C/T	-.035	94	$2x10^{-13}$
SBNO1	rs4759375	12	123.80	HDL	.08	T/C	.056	94	$3x10^{-8}$
ZNF664	rs4765127	12	124.46	HDL	.35	T/G	.032/-.029	94/91	$8x10^{-10}/2x10^{-8}$
SCARB1	rs838880	12	125.26	HDL	.34	C/T	.048	173	$6x10^{-32}$
LIPC	rs1532085	15	58.68	HDL,TC,TG	.40	A/G	.107/.054/.031	185/186/176	$1x10^{-188}/7x10^{-47}/2x10^{-18}$
LACTB	rs2652834	15	63.40	HDL	.21	A/G	-.028	186	$4x10^{-11}$
CETP	rs3764261	16	56.99	HDL,LDL,TC,TG	.32	A/C	.241/-.053/.050/-.040	178/165/177/169	$1x10^{-769}/2x10^{-34}/4x10^{-31}/2x10^{-25}$
LCAT	rs16942887	16	67.93	HDL	.14	A/G	.083	186	$8x10^{-54}$
CMIP	rs2925979	16	81.53	HDL	.31	T/C	-.035	186	$1x10^{-19}$
STARD3	rs11869286	17	37.81	HDL	.35	G/C	-.032	178	$3x10^{-17}$
ABCA8	rs4148008	17	66.88	HDL	.33	G/C	-.028	166	$1x10^{-12}$
PGS1	rs4129767	17	76.40	HDL	.48	G/A	-.024	185	$2x10^{-11}$
LIPG	rs7241918	18	47.16	HDL,TC	.19	G/T	-.09/-.058	93/93	$1x10^{-44}/4x10^{-18}$
MC4R	rs12967135	18	57.85	HDL	.25	A/G	-.026	154	$4x10^{-8}$
ANGPTL4	rs7255436	19	8.43	HDL	.47	C/A	-.032	93	$2x10^{-8}$
ANGPTL8	rs737337	19	11.35	HDL	.11	C/T	-.056	185	$5x10^{-17}$
LILRA3	rs386000	19	54.79	HDL	.26	C/G	.048	165	$3x10^{-23}$
HNF4A	rs1800961	20	43.04	HDL,TC	.05	T/C	-.127/-.106	158/156	$2x10^{-34}/1x10^{-24}$
PLTP	rs6065906	20	44.55	HDL,TG	.19	C/T	-.059/.053	186/176	$5x10^{-40}/2x10^{-34}$
UBE2L3	rs181362	22	21.93	HDL	.23	T/C	-.038	178	$4x10^{-18}$

Supplementary Table S2.3 (continued)

Loci Primarily Associated with LDL Cholesterol									
PCSK9	rs2479409	1	55.50	LDL,TC	.32	G/A	.064/.054	173/187	$3x10^{-50}/2x10^{-39}$
SORT1	rs629301	1	109.82	LDL,TC	.24	G/T	-.167/-.134	143/156	$5x10^{-241}/2x10^{-170}$
APOB	rs1367117	2	21.26	LDL,TC	.32	A/G	.119/.100	173/187	$1x10^{-182}/3x10^{-139}$
ABCG5/8	rs4299376	2	44.07	LDL,TC	.31	G/T	.081/.079	145/158	$4x10^{-72}/3x10^{-73}$
MYLIP	rs3757354	6	16.13	LDL,TC	.24	T/C	-.038/-.035	173/187	$2x10^{-17}/2x10^{-15}$
HFE	rs1800562	6	26.09	LDL,TC	.07	A/G	-.062/-.056	171/185	$8x10^{-14}/2x10^{-12}$
LPA	rs1564348	6	160.58	LDL,TC	.18	C/T	.048/.049	173/187	$3x10^{-21}/3x10^{-23}$
PLEC1	rs11136341	8	145.04	LDL,TC	.40	G/A	.045/.038	83/87	$7x10^{-12}/6x10^{-9}$
ABO	rs9411489	9	136.155	LDL,TC	.21	T/C	.077/.069	119/130	$2x10^{-41}/3x10^{-35}$
ST3GAL4	rs11220462	11	126.24	LDL,TC	.14	A/G	.059/.047	145/157	$7x10^{-21}/6x10^{-15}$
NYNRIN	rs8017377	14	24.88	LDL	.46	A/G	.030	173	$3x10^{-15}$
OSBPL7	rs7206971	17	45.43	LDL,TC	.49	A/G	.029/.030	81/85	$3x10^{-7}/1x10^{-7}$
LDLR	rs6511720	19	11.20	LDL,TC	.12	T/G	-.221/-.185	171/185	$4x10^{-262}/5x10^{-202}$
APOE	rs4420638	19	45.42	LDL,TC,HDL	.19	G/A	.225/.197/-.067	93/104/100	$2x10^{-178}/1x10^{-149}/2x10^{-21}$
TOP1	rs6029526	20	39.67	LDL,TC	.47	A/T	.044/.040	88/93	$5x10^{-18}/1x10^{-16}$
Loci Primarily Associated with Total Cholesterol									
LDLRAP1	rs12027135	1	25.78	TC,LDL	.46	A/T	-.027/-.030	178/165	$5x10^{-12}/2x10^{-14}$
EVI5	rs7515577	1	93.01	TC	.23	C/A	-.037	95	$2x10^{-8}$
MOSC1	rs2642442	1	220.97	TC,LDL	.33	C/T	-.035/-.036	111/102	$3x10^{-11}/5x10^{-11}$
IRF2BP2	rs514230	1	234.86	TC,LDL	.48	A/T	-.039/-.036	95/90	$5x10^{-14}/9x10^{-12}$
RAB3GAP1	rs7570971	2	135.84	TC	.35	A/C	.030	185	$1x10^{-13}$
RAF1	rs2290159	3	12.63	TC	.23	C/G	-.037	94	$2x10^{-9}$
HMGCR	rs12916	5	74.66	TC,LDL	.40	C/T	.068/.073	183/168	$5x10^{-74}/8x10^{-78}$
TIMD4	rs6882076	5	156.39	TC,TG,LDL	.36	T/C	-.051/-.029/-.046	187/178/173	$5x10^{-41}/2x10^{-15}/3x10^{-31}$
HLA	rs3177928	6	32.41	TC,LDL	.17	A/G	.048/.045	180/166	$1x10^{-21}/3x10^{-17}$
C6orf106	rs2814982	6	34.55	TC	.12	T/C	-.044	187	$4x10^{-15}$
FRK	rs9488822	6	116.31	TC,LDL	.36	T/A	.034/.031	95/90	$1x10^{-9}/2x10^{-7}$
DNAH11	rs12670798	7	21.61	TC,LDL	.25	C/T	.036/.034	187/173	$1x10^{-16}/5x10^{-14}$
NPC1L1	rs2072183	7	44.58	TC,LDL	.29	C/G	.036/.039	184/170	$4x10^{-15}/7x10^{-16}$
CYP7A1	rs2081687	8	59.39	TC,LDL	.36	T/C	.038/.031	95/90	$9x10^{-12}/1x10^{-7}$
GPAM	rs2255141	10	113.93	TC,LDL	.30	A/G	.031/.030	187/173	$7x10^{-16}/1x10^{-13}$
SPTY2D1	rs10128711	11	18.63	TC	.30	T/C	-.031	157	$1x10^{-11}$
UBASH3B	rs7941030	11	122.52	TC,HDL	.39	C/T	.028/.027	187/187	$2x10^{-14}/1x10^{-14}$
BRAP	rs11065987	12	112.07	TC,LDL	.41	G/A	-.031/-.027	187/173	$2x10^{-16}/1x10^{-11}$
HNF1A	rs1169288	12	121.42	TC,LDL	.34	C/A	.032/.038	176/163	$4x10^{-17}/6x10^{-21}$
HPR	rs2000999	16	72.11	TC,LDL	.20	A/G	.062/.065	186/172	$7x10^{-41}/4x10^{-41}$
CILP2	rs10401969	19	19.41	TC,TG,LDL	.09	C/T	-.137/-.121/-.118	186/176/171	$4x10^{-77}/1x10^{-69}/3x10^{-54}$
FLJ36070	rs492602	19	49.21	TC	.47	G/A	.031	184	$1x10^{-16}$
ERGIC3	rs2277862	20	34.15	TC	.15	T/C	-.035	186	$5x10^{-11}$
MAFB	rs2902940	20	39.09	TC,LDL	.30	G/A	-.024/-.027	186/172	$9x10^{-10}/2x10^{-11}$

Supplementary Table S2.3 (continued)

Loci Primarily Associated with Triglycerides									
ANGPTL3	rs2131925	1	63.03	TG,LDL,TC	.34	G/T	-.066/-.049/-.075	178/173/187	$3x10^{-74}/3x10^{-32}/4x10^{-80}$
GCKR	rs1260326	2	27.73	TG,TC	.39	T/C	.115/.051	178/187	$2x10^{-239}/3x10^{-42}$
MSL2L1	rs645040	3	135.93	TG	.23	G/T	-.029	178	$2x10^{-12}$
KLHL8	rs442177	4	88.03	TG	.42	G/T	-.031	178	$1x10^{-18}$
MAP3K1	rs9686661	5	55.86	TG	.20	T/C	.038	177	$3x10^{-16}$
TYW1B	rs13238203	7	72.13	TG	.04	T/C	-.059	102	$3x10^{-6}$
MLXIPL	rs17145738	7	72.98	TG,HDL	.13	T/C	-.115/.041	176/185	$9x10^{-99}/5x10^{-13}$
PINX1	rs11776767	8	10.68	TG	.37	C/G	.022	177	$3x10^{-11}$
NAT2	rs1495741	8	18.27	TG,TC	.26	G/A	.040/.032	88/92	$3x10^{-12}/3x10^{-8}$
LPL	rs12678919	8	19.84	TG,HDL	.13	G/A	-.170/.155	178/187	$2x10^{-199}/1x10^{-149}$
TRIB1	rs2954029	8	126.49	TG,TC,LDL,HDL	.47	T/A	-.076/-.062/-.056/.040	178/187/173/187	$1x10^{-107}/2x10^{-65}/2x10^{-50}/3x10^{-29}$
JMJD1C	rs10761731	10	65.03	TG	.44	T/A	-.031	91	$8x10^{-12}$
CYP26A1	rs2068888	10	94.84	TG	.45	A/G	-.024	178	$2x10^{-11}$
FADS1-2-3	rs174546	11	61.57	TG,LDL,TC,HDL	.36	T/C	.045/-.051/-.048/-.039	178/173/187/187	$7x10^{-38}/2x10^{-39}/3x10^{-37}/8x10^{-28}$
APOA1	rs964184	11	116.65	TG,TC,HDL,LDL	.84	C/G	-.234/-.121/.106/-.086	91/95/94/90	$7x10^{-224}/3x10^{-55}/6x10^{-48}/2x10^{-26}$
LRP1	rs11613352	12	57.79	TG,HDL	.26	T/C	-.028/.028	178/187	$9x10^{-14}/2x10^{-13}$
CAPN3	rs2412710	15	42.68	TG	.04	A/G	.099	154	$2x10^{-11}$
FRMD5	rs2929282	15	44.25	TG	.07	T/A	.072	84	$2x10^{-9}$
CTF1	rs11649653	16	30.92	TG	.40	G/C	-.027	90	$2x10^{-7}$
PLA2G6	rs5756931	22	38.55	TG	.40	C/T	-.020	174	$3x10^{-8}$

Supplementary Table S2.4: Overlap of Novel Loci and Literature

Locus	Lead SNP	Chr	hg19 Position (Mb)	Traits GWS	Literature Candidate	Complete Gene Name	Reference
Loci Primarily Associated with HDL Cholesterol							
PIGV-NR0B2	rs12748152	1	27.14	HDL, LDL, TG	*PIGV*	*phosphatidylinositol glycan anchor biosynthesis, class V*	PMID 20802478 PMID 15623507
PIGV-NR0B2	rs12748152	1	27.14	HDL, LDL, TG	*NR0B2*	*nuclear receptor subfamily 0, group B, member 2*	PMID 22577560 PMID 20375098
HDGF-PMVK	rs12145743	1	156.70	HDL	*HDGF*	*hepatoma-derived growth factor*	PMID 14635185
HDGF-PMVK	rs12145743	1	156.70	HDL	*CRABP2*	*cellular retinoic acid binding protein 2*	PMID 17484622
ANGPTL1	rs4650994	1	178.52	HDL			
CPS1	rs1047891	2	211.54	HDL			
ATG7	rs2606736	3	11.40	HDL			
SETD2	rs2290547	3	47.06	HDL			
RBM5	rs2013208	3	50.13	HDL			
STAB1	rs13326165	3	52.53	HDL	*STAB1*	*stabilin 1*	PMID 21480214 PMID 19726632 PMID 21030611
STAB1	rs13326165	3	52.53	HDL	*NISCH*	*nischarin*	PMID 21484668
GSK3B	rs6805251	3	119.56	HDL	*GSK3B*	*glycogen synthase kinase 3 beta*	PMID 21334395 PMID 21328461
GSK3B	rs6805251	3	119.56	HDL	*NR1I2*	*nuclear receptor subfamily 1, group I, member 2*	PMID 21295138
C4orf52	rs10019888	4	26.06	HDL			
FAM13A	rs3822072	4	89.74	HDL			
ADH5	rs2602836	4	100.01	HDL			
RSPO3	rs1936800	6	127.44	HDL, TG			
DAGLB	rs702485	7	6.45	HDL	*DAGLB*	*diacylglycerol lipase, beta*	PMID 21949825
SNX13	rs4142995	7	17.92	HDL	*SNX13*	*sorting nexin 13*	PMID 12461558
IKZF1	rs4917014	7	50.31	HDL	*IKZF1*	*IKAROS family zinc finger 1 (Ikaros)*	PMID 18483254
TMEM176A	rs17173637	7	150.53	HDL			
MARCH8-ALOX5	rs970548	10	46.01	HDL, TC	*ALOX5*	*arachidonate 5-lipoxygenase*	PMID 22293202
OR4C46	rs11246602	11	51.51	HDL			
KAT5	rs12801636	11	65.39	HDL	*KAT5*	*K(lysine) acetyltransferase 5*	PMID 18096664 PMID 17996965
MOGAT2-DGAT2	rs499974	11	75.46	HDL	*MOGAT2*	*monoacylglycerol O-acyltransferase 2*	PMID 21734185 PMID 14966132

Supplementary Table S2.4 (continued)

MOGAT2-DGAT2	rs499974	11	75.46	HDL	*DGAT2*	*diacylglycerol O-acyltransferase 2*	PMID 22493088 PMID 21317108 PMID 22155452
ZBTB42-AKT1	rs4983559	14	105.28	HDL	*AKT1*	*v-akt murine thymoma viral oncogene homolog 1*	PMID 18054314 PMID 20054340
FTO	rs1121980	16	53.81	HDL, TG			
HAS1	rs17695224	19	52.32	HDL	*HAS1*	*hyaluronan synthase 1*	PMID 9933623
Loci Primarily Associated with LDL Cholesterol							
ANXA9-CERS2	rs267733	1	150.96	LDL	*CERS2*	*ceramide synthase 2*	PMID 20940143 PMID 20110363 PMID 19801672
EHBP1	rs2710642	2	63.15	LDL	*EHBP1*	*EH domain binding protein 1*	PMID 21332221
INSIG2	rs10490626	2	118.84	LDL, TC	*INSIG2*	*insulin induced gene 2*	PMID 22143767 PMID 20817058 PMID 20090767
LOC84931	rs2030746	2	121.31	LDL, TC			
FN1	rs1250229	2	216.30	LDL	*FN1*	*fibronectin 1*	PMID 16150826
CMTM6	rs7640978	3	32.53	LDL, TC			
ACAD11	rs17404153	3	132.16	LDL, HDL			
CSNK1G3	rs4530754	5	122.86	LDL, TC			
MIR148A	rs4722551	7	25.99	LDL, TC, TG			
SOX17	rs10102164	8	55.42	LDL, TC			
BRCA2	rs4942486	13	32.95	LDL			
APOH-PRXCA	rs1801689	17	64.21	LDL	*APOH*	*apolipoprotein H*	PMID 12740481
APOH-PRXCA	rs1801689	17	64.21	LDL	*PRKCA*	*protein kinase C, alpha*	PMID 20692055 PMID 12952980
SPTLC3	rs364585	20	12.96	LDL	*SPTLC3*	*serine palmitoyltransferase, long chain base subunit 3*	PMID 19648650
SNX5	rs2328223	20	17.85	LDL	*SNX5*	*sorting nexin 5*	PMID 15561769
MTMR3	rs5763662	22	30.38	LDL			

Supplementary Table S2.4 (continued)

Loci Primarily Associated with Total Cholesterol							
ASAP3	rs1077514	1	23.77	TC			
ABCB11	rs2287623	2	169.83	TC	*ABCB11*	*ATP-binding cassette, sub-family B (MDR/TAP), member 11*	PMID 21726512 PMID 19228692
FAM117B	rs11694172	2	203.53	TC			
UGT1A1	rs11563251	2	234.68	TC, LDL	*UGT1A1/3/4/5/6/7/8/9/20*	*UDP glucuronosyltransferase 1 family, polypeptide A1*	PMID 17908920
PXK	rs13315871	3	58.38	TC	*PXK*	*PX domain containing serine/threonine kinase*	PMID 20086096 PMID 17178602
KCNK17	rs2758886	6	39.25	TC			
HBS1L	rs9376090	6	135.41	TC			
GPR146	rs1997243	7	1.08	TC			
VLDLR	rs3780181	9	2.64	TC, LDL	*VLDLR*	*very low density lipoprotein receptor*	PMID 8827514
VIM-CUBN	rs10904908	10	17.26	TC	*VIM*	*vimentin*	PMID 22535769 PMID 7706405 PMID 1527066
VIM-CUBN	rs10904908	10	17.26	TC	*CUBN*	*cubilin*	PMID 10371504
PHLDB1	rs11603023	11	118.49	TC			
PHC1-A2ML1	rs4883201	12	9.08	TC	*A2ML1*	*alpha-2-macroglobulin-like 1*	PMID 18648652
DLG4	rs314253	17	7.09	TC, LDL	*ACADVL*	*acyl-CoA dehydrogenase, very long chain*	PMID 19889959
DLG4	rs314253	17	7.09	TC, LDL	*CTDNEP1*	*CTD nuclear envelope phosphatase 1*	PMID 22134922
DLG4	rs314253	17	7.09	TC, LDL	*SLC2A4*	*solute carrier family 2, member 4*	PMID 16096283
TOM1	rs138777	22	35.71	TC	*HMOX1*	*hemeoxygenase (decycling) 1*	PMID 22004613
PPARA	rs4253772	22	46.63	TC, LDL	*PPARA*	*peroxisome proliferator-activated receptor alpha*	PMID 21540177 PMID 21487230
Loci Primarily Associated with Triglycerides							
LRPAP1	rs6831256	4	3.47	TG, LDL, TC	*LRPAP1*	*low density lipoprotein receptor-related protein associated protein 1*	PMID 16973241
VEGFA	rs998584	6	43.76	TG, HDL	*VEGFA*	*vascular endothelial growth factor A*	PMID 21348596 PMID 18789802
MET	rs38855	7	116.36	TG			
AKR1C4	rs1832007	10	5.25	TG	*AKR1C4*	*aldo-ketoreductase family 1, member C4*	PMID 18024509
PDXDC1	rs3198697	16	15.13	TG			
MPP3	rs8077889	17	41.88	TG			
INSR	rs7248104	19	7.22	TG			
PEPD	rs731839	19	33.90	TG, HDL	*CEBPG*	*CCAAT/enhancer binding protein (C/EBP), gamma*	PMID 12177065

Supplementary Table S2.5: Pathways that Show Enrichment of Genes at Novel Loci by MAGENTA analysis

*Supplementary Tables 5A-D.*Database and Gene Set define the source of the gene set with evidence for enrichment; Effective Gene Set Size, the number of genes in a pathway independently assigned a score, after clustering nearby genes and excluding genes in regions with no SNP data; Expected Number of Hits, the number of genes expected to have a score in the top 5% of all scores given the gene set size; Observed Number of Hits, the number of genes observed in the top 5% of all gene scores; FDR *P*-value, the false discovery rate incurred by rejecting the null for this gene set and all others with more extreme enrichment using all GWAS+Metabochip results; Genome-wide Significant Genes, genes in the pathway labeled as hits by MAGENTA (Known and Novel refer to association evidence reported by this study); Other Enriched Genes, genes with scores in the top 5% of all gene scores but that do not reach genome-wide significance. We show here significant pathways (FDR p< .05) which contain at least one gene from one of the 62 Novel loci.

Supplementary Table S2.5-A: Pathways that Show Enrichment of Genes at Novel HDL Associated Loci by MAGENTA analysis

Database	Gene Set	Effective Gene Set Size	No. of Expected Genes (>95% Cutoff)	No. of Observed Genes (>95% Cutoff)	FDR *P*-value	Genome-wide Significant Genes		
						Novel	Known	Other Enriched Genes
HDL Cholesterol								
Ingenuity	FXR RXR activation	54	3	13	$6.0x10^{-4}$	*NR0B2*	*SCARB1,LIPC,HNF4A,PLTP, APOB,APOE,MLXIPL,APOA1*	*APOC3,NR1H3,APOC2,NR1I2*
GOTERM	Cholesterol metabolic process	52	3	11	$1.3x10^{-2}$	*NR0B2*	*CETP,LCAT,STARD3,APOB, APOE,APOA1*	*ABCA1,APOA4,APOC3,APOC1*
Ingenuity	LPS IL-1 mediated inhibition of RXR function	52	3	9	$2.0x10^{-2}$	*NR0B2*	*ABCA1,SCARB1,LIPC,CETP, PLTP,APOE*	*APOC2,NR1I2*
KEGG	Neurotrophin signaling	116	6	16	$2.1x10^{-2}$	*AKT1,GSK3 B*	*SORT1*	*RAC1,RPS6KA1,SH2B3,NFKB1, NTRK1,MAP2K7,RELA,PLCG2, PRKCD,PTPN11,MAP2K2,TP53, MAPK10*
Ingenuity	PXR RXR activation	45	2	8	$2.4x10^{-2}$	*NR0B2*	*HNF4A*	*RELA,NR1I2,ABCB9,GSTM1, INSR,CPT1A*
KEGG	Adipocytokine signaling	63	3	11	$2.5x10^{-2}$	*AKT1*		*AGRP,TRADD,SOCS3,ACSL5, NFKB1,RELA,PTPN11,CHUK, CPT1A,MAPK10*
GOTERM	Triglyceride lipase activity	15	1	5	$2.5x10^{-2}$	*DAGLB*	*LIPC,LIPG,LPL*	*DAGLA*
Ingenuity	NFKB signaling	39	2	7	$2.9x10^{-2}$	*GSK3B*		*RELB,CD40,NFKB1,RELA, PLCG2,CHUK*
Ingenuity	PPARaRXRa activation	50	3	8	$3.4x10^{-2}$	*NR0B2*	*ABCA1,LPL,APOA1*	*CKAP5,MED1,NCOA6,INSR*
GOTERM	Enzyme binding	108	5	16	$3.8x10^{-2}$	*AKT1*	*UBE2L3,SORT1,APOB, APOA1*	*APOA5,RAC1,CSF3,CD40, PRKCD,PLAUR,DNM2,CBX1, TP53,MIZF,HMGA1*
GOTERM	Phospholipid binding	47	2	9	$4.2x10^{-2}$	*CPS1*	*ABCA1,APOB,APOE,APOA1*	*APOA5,LYPLA3,APOC3, MAP1LC3A*

Supplementary Table S2.5-B: Pathways that Show Enrichment of Genes at Novel LDL Associated Loci by MAGENTA analysis

Database	Gene Set	Effective Gene Set Size	No. of Expected Genes (>95% Cutoff)	No. of Observed Genes (>95% Cutoff)	FDR *P*-value	Genome-wide Significant Genes		
						Novel	Known	Other Enriched Genes
LDL Cholesterol								
GOTERM	Cholesterol metabolic process	53	3	18	$< 3.3 \times 10^{-5}$	*NR0B2,INSIG2, VLDLR,CUBN*	*ABCA1,CETP,PCSK9,APOB, LDLR,APOE,LDLRAP1,HNF1A, ANGPTL3,APOA1*	*APOC1,APOA4,APOC3,PPARD*
Ingenuity	FXR RXR activation	54	3	16	$< 3.3 \times 10^{-5}$	*NR0B2,PPARA, VLDLR, ABCB11*	*HNF4A,APOB,APOE,CYP7A1,H NF1A,APOA1*	*PPARG,APOC2,ABCG5,ABCG8,APO C3,MTTP*
GOTERM	Lipoprotein metabolic process	15	1	7	1.7×10^{-4}	*PPARA*	*PCSK9,NPC1L1*	*APOC1,APOA5,APOA4,APOC3*
GOTERM	Lipid transport	61	3	14	1.6×10^{-3}	*VLDLR*	*CETP,APOB,LPA,LDLR,APOE,A POA1*	*APOC1,APOC2,APOC4,COL4A3BP, APOA5,APOA4,APOC3*
Ingenuity	LPS I-1 mediated inhibition of RXR function	53	3	11	2.2×10^{-3}	*NR0B2,PPARA, ABCB11*	*ABCA1,CETP,APOE,CYP7A1*	*APOC2,ABCG5,ABCG8,LY96*
GOTERM	Low-density lipoprotein receptor binding	11	1	5	2.9×10^{-3}	*LRPAP1*	*PCSK9,APOB,APOE*	*APOA5*
GOTERM	Negative regulation of macrophage derived foam cell differentiation	12	1	5	3.4×10^{-3}	*PPARA*	*ABCA1,CETP*	*ITGB3,PPARG*
Ingenuity	Hepatic cholestasis	58	3	11	3.4×10^{-3}	*NR0B2,PPARA, ABCB11*	*CETP,HNF4A,CYP7A1, HNF1A*	*TIRAP,MAP3K4,LY96,NR1H4*
GOTERM	Steroid metabolic process	67	3	13	6.5×10^{-3}	*INSIG2,VLDLR, CUBN*	*ABCA1,CETP,PCSK9,APOB, OSBPL7,LDLR,LDLRAP1, NPC1L1,CYP7A1*	*SORL1*
Ingenuity	PPAR Signaling	18	1	5	7.9×10^{-3}	*NR0B2,PPARA*	*RAF1*	*PPARG,PPARD*
GOTERM	Phosphatidylserine binding	10	1	4	1.3×10^{-2}	*ANXA9*	*SCARB1*	*CPNE1,TRIM72*
Ingenuity	PXR RXR activation	46	2	8	1.7×10^{-2}	*NR0B2,PPARA, ABCB11, UGT1A1*	*HNF4A,CYP7A1*	*GSTM1,UGT1A9*
GOTERM	Lipoprotein transport	10	1	4	1.7×10^{-2}	*CUBN*	*APOB*	*PPARG,MTTP*
GOTERM	Steroid hormone receptor activity	45	2	9	3.3×10^{-2}	*NR0B2,PPARA*	*HNF4A*	*PPARG,PPARD,NR1H4,RARB, NR4A3,THRA*
GOTERM	Organ regeneration	24	1	6	3.5×10^{-2}	*APOH*		*PPARG,ATIC,GAS6,NR4A3,LIF*
GOTERM	Receptor-mediated endocytosis	39	2	8	3.6×10^{-2}	*CUBN*	*HFE,APOE*	*IGF2R,ASGR1,M6PR,SORL1, ARHGAP27*

Supplementary Table S2.5-C: Pathways that Show Enrichment of Genes at Novel Total Cholesterol Associated Loci by MAGENTA analysis

Database	Gene Set	Effective Gene Set Size	No. of Expected Genes (>95% Cutoff)	No. of Observed Genes (>95% Cutoff)	FDR *P*-value	Genome-wide Significant Genes		
						Novel	Known	Other Enriched Genes
					Total Cholesterol			
GOTERM	Cholesterol metabolic process	52	3	18	$< 3.3 \times 10^{-5}$	*CUBN,INSIG2,VLDLR*	*ABCA1,CETP,LCAT,PCSK9,APOB,LDLR,APOE,LDLRAP1,HNF1A,ANGPTL3,APOA1*	*APOC1,APOA4,APOC3,PPARD*
Ingenuity	FXR RXR activation	54	3	18	$< 3.3 \times 10^{-5}$	*ABCB11,PPARA,VLDLR,NR0B2*	*SCARB1,LIPC,HNF4A,APOB,APOE,CYP7A1,HNF1A,APOA1*	*APOC2,ABCG5,ABCG8,APOC3,PPARG,SDC1*
Ingenuity	LPS IL-1 Mediated Inhibition of RXR Function	53	3	14	6.7×10^{-5}	*ABCB11,PPARA,NR0B2*	*ABCA1,SCARB1,LIPC,CETP,APOE,CYP7A1*	*APOC2,ABCG5,ABCG8,LY96,ABCB9*
GOTERM	Low-density lipoprotein receptor binding	11	1	6	1.6×10^{-4}	*LRPAP1*	*PCSK9,APOB,APOE*	*APOA5,SNX17*
Ingenuity	Hepatic cholestasis	56	3	12	7.5×10^{-4}	*ABCB11,PPARA,NR0B2*	*CETP,HNF4A,CYP7A1,HNF1A*	*LY96,NR1I2,MAP3K4,NR1H4,HSD3B7*
Ingenuity	PXR RXR activation	46	2	10	1.1×10^{-3}	*ABCB11,PPARA,UGT1A1,NR0B2*	*HNF4A,CYP7A1*	*GSTM1,UGT1A9,ABCB9,NR1I2*
GOTERM	Negative regulation of macrophage-derived foam cell differentiation	12	1	5	4.2×10^{-3}	*PPARA*	*ABCA1,CETP*	*PPARG,ITGB3*
Ingenuity	PPAR signaling	18	1	5	7.7×10^{-3}	*PPARA,NR0B2*	*RAF1*	*PPARG,PPARD*
GOTERM	Steroid metabolic process	68	3	13	9.8×10^{-3}	*CUBN,INSIG2,VLDLR*	*ABCA1,CETP,LCAT,PCSK9,APOB,LDLR,LDLRAP1,NPC1L1,CYP7A1*	*NR1I2*
GOTERM	Gamma-tubulin binding	10	1	4	1.6×10^{-2}	*BRCA2*		*SPATC1,MARK4,BLOC1S2*
GOTERM	Phosphatidylserine binding	10	1	4	2.0×10^{-2}	*ANXA9*	*SCARB1*	*CPNE1,TRIM72*
Ingenuity	NRF2-mediated oxidative stress Response	50	3	8	3.3×10^{-2}	*GSK3B*	*SCARB1,RAF1*	*HERPUD1,KEAP1,ERP29,HMOX1,FTH1*
Ingenuity	Axonal guidance signaling	65	3	9	3.3×10^{-2}	*GSK3B*	*RAF1*	*SDCBP,VASP,PTPN11,CXCR4,ARHGEF15,GDF7,ERBB2*
Ingenuity	Neuregulin signaling	25	1	5	3.4×10^{-2}	*DLG4*	*RAF1*	*PTPN11,RPS6,GRB7*
Ingenuity	Estrogen receptor signaling	30	2	5	4.4×10^{-2}	*NR0B2*	*RAF1*	*SMARCA4,CARM1,PELP1*
Ingenuity	PPARaRXRa activation	50	3	7	4.5×10^{-2}	*PPARA,NR0B2*	*ABCA1,RAF1,APOA1*	*MED24,MED1*
GOTERM	Receptor-mediated endocytosis	39	2	8	4.7×10^{-2}	*CUBN*	*APOE*	*IGF2R,ASGR1,M6PR,CXCL16,PLD2,SNX17*

Supplementary Table S2.5-D: Pathways that Show Enrichment of Genes at Novel Triglyceride Associated Loci by MAGENTA analysis

Database	Gene Set	Effective Gene Set Size	No. of Expected Genes (>95% Cutoff)	No. of Observed Genes (>95% Cutoff)	FDR *P*-value	Genome-wide Significant Genes		
						Novel	Known	Other Enriched Genes
Triglycerides								
Ingenuity	FXR RXR activation	54	3	15	$< 3.3 \times 10^{-5}$	*NR0B2*	*SCARB1,LIPC,PLTP,APOB, APOE,CYP7A1,MLXIPL ,APOA1*	*APOC2,APOC3,PPARG,CYP27 A1,NR1H3,SLCO1B1*
GOTERM	Low-density lipoprotein receptor binding	11	1	5	2.3×10^{-3}	*LRPAP1*	*APOB,APOE*	*SNX17,APOA5*
KEGG	Primary bile acid biosynthesis	16	1	5	1.0×10^{-2}	*AKR1C4*	*CYP7A1*	*HSD3B7,CYP27A1,HSD17B4*
GOTERM	Cholesterol metabolic process	52	3	11	1.2×10^{-2}	*NR0B2*	*ABCA1,CETP,APOB, APOE,ANGPTL3,APOA1*	*APOA4,APOC1,APOC3,CYP27 A1*
Ingenuity	PPAR signaling	18	1	5	1.4×10^{-2}	*INSR,NR0B2*	*RAF1*	*PPARG,NR1H3*
GOTERM	Cell surface	184	9	23	3.0×10^{-2}	*VEGFA ,LRPAP1,MPP3*	*SCARB1*	*PVRL2,EDG4,BACE1,STX4,BC AM,STRC,FLT3LG,MPP2,TME M102,PCSK6,DSCAML1 ,HSPB1,CD6,C9orf127,BMPR2, IGF2R ,ITGAL,SDC1,HFE2*
Ingenuity	LPS IL-1 mediated inhibition of RXR function	51	3	8	5.0×10^{-2}	*NR0B2*	*ABCA1,LIPC,CETP,PLTP, APOE,CYP7A1*	*APOC2*

Supplementary Table S2.6: 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									
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									
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, omentalfat, 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.

Supplementary Table S2.7: Nonsynonymous Variants in Linkage Disequilibrium with Index SNPs at Novel Loci

Lead SNP	Chr	hg19 Position (Mb)	Lead Trait	Nonsynonymous SNP	r^2	Gene with Nonsynonymous SNP	Amino Acid Change	PolyPhen-2 Classifier[a]
rs12748152	1	27.14	HDL	rs17360994	1.00	*C1orf172*	Gln100Arg	0.20
				rs7545442	.90	*NUDC*	Thr68Met	NA
				rs6659176	1.00	*NR0B2*	Gly171Ala	0.99
rs12145743	1	156.70	HDL	rs4399146	1.00	*HDGF*	Pro201Leu	0.00
rs1047891	2	211.54	HDL	rs1047891	--	*CPS1*	Thr1412Asn	0.01
rs2290547	3	47.06	HDL	rs2305637	.94	*NBEAL2*	Ser2054Phe	0.99
rs2013208	3	50.13	HDL	rs2230590	.89	*MST1R*	Gln523Arg	0.00
				rs1062633	.93	*MST1R*	Arg1335Gly	0.00
rs13326165	3	52.53	HDL	rs887515	.85	*NISCH*	Ala1056Val	0.00
rs970548	10	46.01	HDL	rs2291429	.95	*MARCH8*	Leu269Trp	NA
				rs2291428	.95	*MARCH8*	Phe277Leu	NA
rs11246602	11	55.20	HDL	rs12419022	.97	*OR5W2*	His65Arg	0.01
				rs11230983	.97	*OR5D13*	Arg124His	0.02
				rs12224086	.94	*OR5AS1*	Arg122Leu	0.90
rs267733	1	150.96	LDL	rs267733	--	*ANXA9*	Asp166Gly	0.99
rs10490626	2	118.84	LDL	rs17512204	1.00	*CCDC93*	Pro228Leu	0.01
rs1250229	2	216.30	LDL	rs1250259	1.00	*FN1*	Gln15Leu	0.00
rs7640978	3	32.53	LDL	rs2303857	.91	*DYNC1LI1*	Gln277Arg	0.02
rs17404153	3	132.16	LDL	rs41272321	.85	*ACAD11*	Lys414Thr	NA
rs1801689	17	64.21	LDL	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

[a]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 PolyPhenwebservice at: http://genetics.bwh.harvard.edu/pph2/bgi.shtml

Supplementary Table S2.8: 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 (of 13 tested, $P< 1x10^{-5}$)	Chromatin States* Showing Excess Overlap with Lipid Loci ($P< 1x10^{-5}$)
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 J et al. Nature 473, 43-9, 2011) based on hidden Markov models 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 was compared to matched sets of HapMap SNPs (see Methods)*

Supplementary Table S2.9: 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. (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)						
HNF4A	38	16.2	$6x10^{-10}$	14	7.1	$6x10^{-3}$
CEBP/B	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 J et al. Nature **473**, 43-9, 2011) based on hidden Markov models 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, PLoS Biol. **9**:e1001046, 2011). SNPs in high linkage disequilibrium (r^2> .8 in 1000 Genomes Project European ancestry samples) with known or novel lipid loci were compared to matched sets of HapMap SNPs (see **Methods**). The table lists only regulatory elements that exhibited a significant excess overlap ($P< 1x10^{-3}$ to account for 31 HepG2 regulatory elements tested).

Supplementary Table S2.10: Overlap of Regulatory Features and Associated SNPs at Novel Lipid Loci

		Hidden Markov model-defined regulatory domains from histone methylation marks									Histone methylation marks					Markers of open chromatin		Transcription factor binding (ChIP-Seq)			
Locus	**Lead SNP**	**Strong Enhancer (HMM4)**	**Weak Enhancer (HMM7)**	**WeakTxn (HMM11)**	**Strong Enhancer (HMM5)**	**Txn Elongation (HMM10)**	**Weak Enhancer (HMM6)**	**Active Promoter (HMM1)**	**Weak Promoter (HMM2)**	**Txn Transition (HMM9)**	**H3k9ac**	**H3k27ac**	**H3k4me3**	**H3k36me3**	**H3k4me2**	**FAIRE**	**DNase**	**Hnf4A (Forskolin)**	**Cebpb (Forskolin)**	**CTCF**	**Hsf1**
Loci Primarily Associated with HDL Cholesterol																					
PIGV-NR0B2	rs12748152	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
HDGF-PMVK	rs12145743	x		x	x	x	x		x		x	x	x	x	x	x	x	x	x		x
ANGPTL1	rs4650994	x	x	x	x	x					x	x	x	x	x	x	x	x	x		
CPS1	rs1047891			x										x							
ATG7	rs2606736		x	x		x		x			x		x	x	x	x	x	x	x	x	
SETD2	rs2290547			x			x			x	x	x		x	x						
RBM5	rs2013208		x	x		x	x	x	x	x	x	x	x	x	x	x	x			x	
STAB1	rs13326165						x							x							
GSK3B	rs6805251		x	x			x			x				x	x			x		x	
C4orf52	rs10019888																			x	
FAM13A	rs3822072	x	x	x		x	x	x		x	x	x	x	x	x	x				x	
ADH5	rs2602836		x	x					x		x	x	x	x	x	x	x	x	x		
RSPO3	rs1936800												x		x					x	
DAGLB	rs702485			x			x			x	x	x		x	x					x	
SNX13	rs4142995			x										x							
IKZF1	rs4917014																				
TMEM176A	rs17173637	x									x	x	x		x						
MARCH8-ALOX5	rs970548		x	x		x	x				x	x	x	x	x						
OR4C46	rs11246602			x							x	x	x	x	x					x	
KAT5	rs12801636	x		x			x	x			x	x	x	x	x					x	
MOGAT2-DGAT2	rs499974		x			x		x			x		x		x	x			x		
ZBTB42-AKT1	rs4983559			x											x						
FTO	rs1121980	x	x	x	x	x	x				x	x	x	x	x		x				
HAS1	rs17695224																				

Supplementary Table S2.10 (continued)

Loci Primarily Associated with LDL Cholesterol																													
ANXA9-CERS2	rs267733				x														x										
EHBP1	rs2710642	x			x				x				x	x	X	x	x	x					x						
INSIG2	rs10490626	x			x	x	x	x	x	x	x	x	x	X	x	x	x	x				x		x		x			
LOC84931	rs2030746	x	x	x			x		x			x	X	x		x										x			
FN1	rs1250229	x	x	x			x		x	x	x	x	X	x	x	x													
CMTM6	rs7640978		x	x					x			x	X	x	x	x													
ACAD11	rs17404153		x	x				x	x	x	x	x	X	x	x	x													
CSNK1G3	rs4530754			x						x		x	X	x	x	x	x												
MIR148A	rs4722551									x		x	X	x		x													
SOX17	rs10102164																												
BRCA2	rs4942486			x				x											x										
APOH-PRXCA	rs1801689									x	x	x	X	x	x	x													
SPTLC3	rs364585	x		x								x	X	x		x				x									
SNX5	rs2328223	x										x	X	x		x							x		x				
MTMR3	rs5763662	x	x	x	x	x	x				x	x	X	x	x	x	x						x				x		

Supplementary Table S2.10 (continued)

		Loci Primarily Associated with Total Cholesterol																						
ASAP3	rs1077514	x	x	x	x		x			x	x	X	x	x	x									
ABCB11	rs2287623			x																				
FAM117B	rs11694172		x	x			x		x	x	x	X	x	x	x	x			x					
UGT1A1	rs11563251																							
PXK	rs13315871	x	x	x		x	x	x	x		x	X	x	x	x	x			x	x	x			
KCNK17	rs2758886																							
HBS1L	rs9376090																							
GPR146	rs1997243	x	x	x	x	x	x	x	x	x	x	X	x	x	x	x	x	x	x	x	x	x		
VLDLR	rs3780181			x										x		x								
VIM-CUBN	rs10904908																				x			
PHLDB1	rs11603023		x	x		x	x				x	X		x	x									
PHC1-A2ML1	rs4883201													x										
DLG4	rs314253			x		x		x			x		x		x						x			
TOM1	rs138777	x	x	x	x		x	x	x	x	x	X	x	x	x	x	x			x				
PPARA	rs4253772		x	x		x	x							x	x									

Supplementary Table S2.10 (continued)

Loci Primarily Associated with Triglycerides																															
LRPAP1	rs6831256							x												x								x		x	
VEGFA	rs998584																			x						x					
MET	rs38855				x				x										x												
AKR1C4	rs1832007			x		x								x	X			x		x		x								x	
PDXDC1	rs3198697								x										x												
MPP3	rs8077889	x		x		x		x				x		x	X		x		x		x		x				x		x		
INSR	rs7248104											x							x		x										
PEPD	rs731839	x		x				x						x	X		x				x		x				x		x		x

Supplementary Table S2.10 annotates overlap (denoted as x) between regulatory features and either the index SNP or a variant in high linkage disequilibrium (r^2>0.8) with the index SNP. Regulatory features were obtained from Ernst J et al. Nature **473**, 43-9, 2011, and the ENCODE Project (ENCODE Project Consortium, PLoS Biol. **9**:e1001046, 2011). The corresponding BED file available on the UCSC Genome Browser for each regulatory feature is listed below.

Strong Enhancer (HMM4): wgEncodeBroadHmmHepg2HMM.bed.4_Strong_Enhancer.bed
Weak Enhancer (HMM7): wgEncodeBroadHmmHepg2HMM.bed.7_Weak_Enhancer.bed
WeakTxn (HMM11): wgEncodeBroadHmmHepg2HMM.bed.11_Weak_Txn.bed
Strong Enhancer (HMM5): wgEncodeBroadHmmHepg2HMM5_Strong_Enhancer.bed
Txn Elongation (HMM10): wgEncodeBroadHmmHepg2HMM.bed.10_Txn_Elongation.bed
Weak Enhancer (HMM6): wgEncodeBroadHmmHepg2HMM.bed.6_Weak_Enhancer.bed
Active Promoter (HMM1): wgEncodeBroadHmmHepg2HMM1_Active_Promoter.bed
Weak Promoter (HMM2): wgEncodeBroadHmmHepg2HMM.bed.2_Weak_Promoter.bed
Txn Transition (HMM9): wgEncodeBroadHmmHepg2HMM.bed.9_Txn_Transition.bed
H3k9ac: wgEncodeBroadChipSeqPeaksHepg2H3k9ac.bed
H3k27ac: wgEncodeBroadChipSeqPeaksHepg2H3k27ac.bed
H3k4me3: wgEncodeBroadChipSeqPeaksHepg2H3k4me3.bed
H3k36me3: wgEncodeBroadChipSeqPeaksHepg2H3k36me3.bed
H3k4me2: wgEncodeBroadChipSeqPeaksHepg2H3k4me2.bed
FAIRE: wgEncodeUncFAIRESeqPeaksHepg2V3.bed
DNase: wgEncodeUwDnaseSeqPeaksRep1Hepg2.bed
Hnf4a-Forskln: wgEncodeYaleChIPseqPeaksHepg2Hnf4aForskln.narrowPeak
Cebpb-Forskln: wgEncodeYaleChIPseqPeaksHepg2CebpbForskln.narrowPeak
CTCF: wgEncodeBroadChipSeqPeaksHepg2Ctcf
HSF1: wgEncodeYaleChIPseqPeaksHepg2Hsf1Forskln.narrowPeak

Supplementary Table S2.11: Fine-Mapping Results in Different Ancestries

Chr	Fine Mapping Interval (hg19 Mb)	Locus Name	Top GWAS SNP	# LD Proxies in Europe	Estimates from GWAS Samples for Top GWAS SNP				Estimates from Ancestry-specific Metabochip Samples for Top GWAS SNP				Top Metabochip SNP	# LD Proxies	EUR r^2 with GWAS SNP	Other r^2 with GWAS SNP	Estimates from Ancestry-specific Metabochip Samples for Top MC SNP			
					P	N	% Var	Freq	*P*	N	% Var	Freq					*P*	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 (□2)	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

Supplementary Table S2.11 (continued)

Chr	Fine Mapping Interval (hg19 Mb)	Locus Name	Top GWAS SNP	# LD Proxies in Europe	Estimates from GWAS Samples for Top GWAS SNP				Estimates from Ancestry-specific Metabochip Samples for Top GWAS SNP				Top Metabochip SNP	# LD Proxies	EUR r^2 with GWA SNP	Other r^2 with GWA SNP	Estimates from Ancestry-specific Metabochip Samples for Top MC SNP			
					P	N	% Var	Freq	*P*	N	% Var	Freq					*P*	N	% Var	Freq
Triglycerides																				
East Asian																				
11	116.53-116.67	*APOA5-A4-C3-A1*	rs2160669	20	$3x10^{-128}$	91,013	0.96	0.9	$3x10^{-27}$	8,743	1.37	0.79	rs651821	16	0.85	0.76	$2x10^{-55}$	8,743	2.83	0.73

Supplementary Table S2.11: Locus labels are from Teslovich *et al.* (2010). # LD Proxies in Europe, the number of SNPs r^2> 0.7 with GWAS SNP in 1000 Genomes European Ancestry samples; # Ancestry-Specific LD Proxies, number of SNPs r^2> 0.7 with top Metabochip SNP in the relevant ancestry group; EUR r^2, LD between top GWAS SNP and top Metabochip SNP in European ancestry samples; Other r^2, LD between top GWAS SNP and top Metabochip SNP in the relevant ancestry group.

Supplementary Table S2.12: Novel and Known Lipid Loci Associated with BMI, CAD, DBP, SBP, Fasting Glucose, T2D, and WHR adj BMI

In silico Association Results (*P*< .05) at Lipid Associated Loci for A. Body Mass Index (BMI), B. Coronary Artery Disease (CAD), C. Diastolic Blood Pressure (DBP), D. Systolic Blood Pressure (SBP), E. Fasting Glucose (FG), F. Type 2 Diabetes (T2D), and G. Waist-Hip Ratio adjusted for BMI (WHRadjBMI)

Supplementary Table S12.12-A: Novel and Known Lipid Loci with BMI P-value < 0.05 from GIANT*

Locus	SNP	Chr	hg19 Pos (Mb)	Type	Trait	A1/A2	Lipid Direction	Lipid N	Lipid *P*-value	BMI Direction	BMI N	BMI *P*-value
FTO	rs1121980	16	53.81	novel	HDL	A/G	-	185,524	$6.8x10^{-9}$	+	123,845	$1.8x10^{-57}$
MC4R	rs12967135	18	57.85	known	HDL	A/G	-	153,533	$3.6x10^{-8}$	+	123,864	$5.3x10^{-22}$
SLC39A8	rs13107325	4	103.19	known	HDL	T/C	-	179,316	$1.1x10^{-15}$	+	123,348	$1.4x10^{-7}$
ARL15	rs6450176	5	53.3	known	HDL	A/G	-	187,132	$6.9x10^{-10}$	-	123,861	$7.7x10^{-5}$
BRAP	rs11065987	12	112.07	known	TC	A/G	+	187,309	$2.1x10^{-16}$	+	123,855	$1.2x10^{-4}$
HMGCR	rs12916	5	74.66	known	TC	T/C	-	182,530	$4.6x10^{-74}$	+	123,863	$1.5x10^{-4}$
UBASH3B	rs7941030	11	122.52	known	TC	T/C	-	187,106	$2.4x10^{-14}$	-	123,819	$6.6x10^{-4}$
JMJD1C	rs10761731	10	65.03	known	TG	A/T	+	91,013	$8.4x10^{-12}$	+	123,863	$9.9x10^{-4}$
RBM5	rs2013208	3	50.13	novel	HDL	T/C	+	169,708	$8.9x10^{-12}$	-	123,864	$1.4x10^{-3}$
ZNF664	rs4765127	12	124.46	known	HDL	T/G	+	94,198	$7.8x10^{-10}$	+	123,737	$1.7x10^{-3}$
RAB3GAP1	rs7570971	2	135.84	known	TC	A/C	+	184,956	$1.2x10^{-13}$	-	123,850	$3.3x10^{-3}$
HPR	rs2000999	16	72.11	known	TC	A/G	+	185,692	$6.8x10^{-41}$	+	123,673	$4.9x10^{-3}$
PDE3A	rs7134375	12	20.47	known	HDL	A/C	+	187,088	$1.1x10^{-8}$	+	123,830	$4.9x10^{-3}$
PEPD	rs731839	19	33.9	novel	TG	A/G	-	176,161	$2.7x10^{-9}$	+	123,854	$5.2x10^{-3}$
PGS1	rs4129767	17	76.4	known	HDL	A/G	+	185,469	$2.1x10^{-11}$	-	123,798	$6.1x10^{-3}$
IRS1	rs2972146	2	227.1	known	HDL	T/G	-	184,044	$1.9x10^{-17}$	-	123,855	$7.2x10^{-3}$
TOP1	rs6029526	20	39.67	known	LDL	A/T	+	88,433	$4.8x10^{-18}$	-	123,862	$7.2x10^{-3}$
FRMD5	rs2929282	15	44.25	known	TG	A/T	-	83,616	$2.0x10^{-9}$	+	122,284	$1.1x10^{-2}$
ZBTB42-AKT1	rs4983559	14	105.28	novel	HDL	A/G	-	183,672	$9.6x10^{-9}$	+	119,958	$1.6x10^{-2}$
KCNK17	rs2758886	6	39.25	novel	TC	A/G	+	187,266	$3.0x10^{-8}$	-	123,863	$1.6x10^{-2}$
LRP1	rs11613352	12	57.79	known	TG	T/C	-	177,799	$9.4x10^{-14}$	+	123,865	$1.9x10^{-2}$
EHBP1	rs2710642	2	63.15	novel	LDL	A/G	+	172,994	$6.1x10^{-9}$	+	123,853	$2.1x10^{-2}$
TRPS1	rs2293889	8	116.6	known	HDL	T/G	-	180,102	$4.3x10^{-17}$	+	123,863	$2.1x10^{-2}$
C6orf106	rs2814982	6	34.55	known	TC	T/C	-	187,263	$3.7x10^{-15}$	+	123,848	$3.1x10^{-2}$
VEGFA	rs998584	6	43.76	novel	TG	A/C	+	174,573	$3.4x10^{-15}$	-	119,481	$3.3x10^{-2}$
COBLL1	rs12328675	2	165.54	known	HDL	T/C	-	187,092	$2.1x10^{-15}$	-	123,856	$3.4x10^{-2}$
CTF1	rs11649653	16	30.92	known	TG	C/G	+	89,449	$1.6x10^{-7}$	+	123,819	$3.6x10^{-2}$
PDXDC1	rs3198697	16	15.13	novel	TG	T/C	-	175,934	$2.2x10^{-8}$	+	123,669	$3.7x10^{-2}$
SETD2	rs2290547	3	47.06	novel	HDL	A/G	-	187,142	$3.7x10^{-9}$	-	118,647	$3.8x10^{-2}$
UBE2L3	rs181362	22	21.93	known	HDL	T/C	-	178,283	$4.3x10^{-18}$	+	123,910	$4.7x10^{-2}$
LRP4	rs3136441	11	46.74	known	HDL	T/C	-	186,975	$6.8x10^{-29}$	+	123,866	$4.7x10^{-2}$
FLJ36070	rs492602	19	49.21	known	TC	A/G	-	184,180	$1.1x10^{-16}$	+	120,451	$4.9x10^{-2}$

**Speliotes EK et al. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. Nat Genet. 2010;42, 937-948*

Supplementary Table S12.12-B: Novel and Known Lipid Loci with CAD *P*-value < 0.05 from CARDIOGRAM+C4D Meta-analysis*

Locus	SNP	Chr	hg19 Pos (Mb)	Type	Trait	A1/A2	Lipid Direction	Lipid N	Lipid *P*-value	CAD Direction	CAD N	CAD *P*-value
APOA1	rs964184	11	116.65	known	TG	C/G	-	90,991	$6.6x10^{-224}$	-	110,492	$4.8x10^{-11}$
SORT1	rs629301	1	109.82	known	LDL	T/G	+	142,643	$5.4x10^{-241}$	+	82,222	$6.1x10^{-10}$
BRAP	rs11065987	12	112.07	known	TC	A/G	+	187,308	$2.1x10^{-16}$	-	73,578	$2.4x10^{-9}$
LDLR	rs6511720	19	11.2	known	LDL	T/G	-	170,607	$3.9x10^{-262}$	-	86,870	$1.2x10^{-7}$
ABCG5/8	rs4299376	2	44.07	known	LDL	T/G	-	144,861	$3.9x10^{-72}$	-	106,016	$3.7x10^{-7}$
HNF1A	rs1169288	12	121.42	known	TC	A/C	-	175,774	$3.9x10^{-17}$	-	80,633	$3.5x10^{-6}$
NAT2	rs1495741	8	18.27	known	TG	A/G	-	87,977	$2.7x10^{-12}$	-	109,804	$1.2x10^{-5}$
TRIB1	rs2954029	8	126.49	known	TG	A/T	+	177,729	$1.0x10^{-107}$	+	81,977	$2.8x10^{-5}$
LPL	rs12678919	8	19.84	known	TG	A/G	+	177,749	$1.8x10^{-199}$	+	111,065	$4.7x10^{-5}$
RBM5	rs2013208	3	50.13	novel	HDL	T/C	+	169,708	$8.9x10^{-12}$	-	82,470	$7.0x10^{-5}$
LPA	rs1564348	6	160.58	known	LDL	T/C	-	172,988	$2.8x10^{-21}$	-	108,431	$1.8x10^{-4}$
APOE	rs4420638	19	45.42	known	LDL	A/G	-	93,103	$1.5x10^{-178}$	-	36,066	$2.1x10^{-4}$
CILP2	rs10401969	19	19.41	known	TC	T/C	+	185,666	$4.1x10^{-77}$	+	81,644	$2.4x10^{-4}$
IRS1	rs2972146	2	227.1	known	HDL	T/G	-	184,044	$1.9x10^{-17}$	+	82,540	$3.8x10^{-4}$
CMTM6	rs7640978	3	32.53	novel	LDL	T/C	-	172,227	$9.8x10^{-9}$	-	81,843	$4.1x10^{-4}$
C6orf106	rs2814982	6	34.55	known	TC	T/C	-	187,262	$3.7x10^{-15}$	+	99,096	$1.6x10^{-3}$
ACAD1	rs17404153	3	132.16	novel	LDL	T/G	-	172,898	$1.8x10^{-9}$	-	83,225	$1.8x10^{-3}$
CETP	rs3764261	16	56.99	known	HDL	A/C	+	177,533	$1.4x10^{-769}$	-	83,626	$2.2x10^{-3}$
FRMD5	rs2929282	15	44.25	known	TG	A/T	-	83,616	$2.0x10^{-9}$	-	81,446	$2.8x10^{-3}$
MAP3K1	rs9686661	5	55.86	known	TG	T/C	+	177,050	$2.5x10^{-16}$	+	81,234	$3.2x10^{-3}$
KLF14	rs4731702	7	130.43	known	HDL	T/C	+	187,085	$4.8x10^{-17}$	-	99,195	$3.2x10^{-3}$
ZNF664	rs4765127	12	124.46	known	HDL	T/G	+	94,198	$7.8x10^{-10}$	-	83,532	$3.6x10^{-3}$
SPTY2D1	rs10128711	11	18.63	known	TC	T/C	-	157,199	$1.1x10^{-11}$	-	80,934	$3.9x10^{-3}$
CAPN3	rs2412710	15	42.68	known	TG	A/G	+	153,909	$1.7x10^{-11}$	+	79,267	$5.3x10^{-3}$
HMGCR	rs12916	5	74.66	known	TC	T/C	-	182,529	$4.6x10^{-74}$	-	81,050	$5.3x10^{-3}$
CYP26A1	rs2068888	10	94.84	known	TG	A/G	-	177,712	$1.7x10^{-11}$	-	83,627	$7.2x10^{-3}$
ST3GAL4	rs11220462	11	126.24	known	LDL	A/G	+	145,030	$6.6x10^{-21}$	+	109,031	$7.4x10^{-3}$
VEGFA	rs998584	6	43.76	novel	TG	A/C	+	174,573	$3.4x10^{-15}$	+	66,823	$9.0x10^{-3}$
PCSK9	rs2479409	1	55.5	known	LDL	A/G	-	172,970	$2.5x10^{-50}$	-	83,207	$1.1x10^{-2}$
PINX1	rs11776767	8	10.68	known	TG	C/G	+	177,360	$2.9x10^{-11}$	-	81,760	$1.2x10^{-2}$
CITED2	rs605066	6	139.83	known	HDL	T/C	+	94,311	$2.8x10^{-8}$	-	81,709	$1.5x10^{-2}$
ABCA8	rs4148008	17	66.88	known	HDL	C/G	+	165,732	$1.1x10^{-12}$	-	96,645	$2.0x10^{-2}$
HBS1L	rs9376090	6	135.41	novel	TC	T/C	+	187,263	$2.6x10^{-9}$	+	81,664	$2.1x10^{-2}$
APOB	rs1367117	2	21.26	known	LDL	A/G	+	173,007	$9.5x10^{-183}$	+	79,823	$2.3x10^{-2}$
IKZF1	rs4917014	7	50.31	novel	HDL	T/G	-	186,868	$1.0x10^{-8}$	+	111,434	$3.4x10^{-2}$
KAT5	rs12801636	11	65.39	novel	HDL	A/G	+	187,099	$3.2x10^{-8}$	-	74,817	$3.8x10^{-2}$
HPR	rs2000999	16	72.11	known	TC	A/G	+	185,692	$6.8x10^{-41}$	+	97,651	$4.1x10^{-2}$
GALNT2	rs4846914	1	230.3	known	HDL	A/G	+	186,995	$3.5x10^{-41}$	-	84,068	$4.1x10^{-2}$
ASAP3	rs1077514	1	23.77	novel	TC	T/C	+	184,079	$6.4x10^{-9}$	+	84,078	$4.4x10^{-2}$
KLHL8	rs442177	4	88.03	known	TG	T/G	+	177,798	$1.3x10^{-18}$	+	82,034	$4.6x10^{-2}$

Supplementary Table S2.12B:

**Schunkert H et al. Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. Nat Genet. 2011;43(4):333-8*

**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. 2011; 43(4):339-44*

Supplementary Table S2.12-C: Novel and Known Lipid Loci with DBP *P*-value < 0.05 from ICBP**

Locus	SNP	Chr	hg19 Pos (Mb)	Type	Trait	A1/A2	Lipid Direction	Lipid N	Lipid *P*-value	DBP Direction	DBP N	DBP *P*-value
BRAP	rs11065987	12	112.07	known	TC	A/G	+	187,309	$2.1x10^{-16}$	-	62,481	$3.4x10^{-12}$
SLC39A8	rs13107325	4	103.19	known	HDL	T/C	-	179,316	$1.1x10^{-15}$	-	58,926	$7.5x10^{-7}$
VEGFA	rs998584	6	43.76	novel	TG	A/C	+	174,573	$3.4x10^{-15}$	+	49,589	$1.6x10^{-4}$
HFE	rs1800562	6	26.09	known	LDL	A/G	-	171,209	$8.3x10^{-14}$	+	65,399	$3.2x10^{-4}$
CITED2	rs605066	6	139.83	known	HDL	T/C	+	94,311	$2.8x10^{-8}$	-	68,145	$9.9x10^{-4}$
LACTB	rs2652834	15	63.4	known	HDL	A/G	-	185,613	$3.6x10^{-11}$	+	61,977	$1.4x10^{-3}$
PABPC4	rs4660293	1	40.03	known	HDL	A/G	+	187,027	$2.9x10^{-18}$	-	69,815	$1.8x10^{-3}$
LOC55908	rs737337	19	11.35	known	HDL	T/C	+	185,432	$4.6x10^{-17}$	-	61,569	$2.1x10^{-3}$
PDE3A	rs7134375	12	20.47	known	HDL	A/C	+	187,088	$1.1x10^{-8}$	-	63,231	$2.3x10^{-3}$
FAM13A	rs3822072	4	89.74	novel	HDL	A/G	-	187,115	$4.1x10^{-12}$	+	66,600	$2.8x10^{-3}$
FADS1-2-3	rs174546	11	61.57	known	TG	T/C	+	177,785	$7.4x10^{-38}$	+	69,718	$6.9x10^{-3}$
RSPO3	rs1936800	6	127.44	novel	HDL	T/C	-	187,111	$3.1x10^{-10}$	-	67,494	$7.3x10^{-3}$
JMJD1C	rs10761731	10	65.03	known	TG	A/T	+	91,013	$8.4x10^{-12}$	+	68,336	$8.3x10^{-3}$
PINX1	rs11776767	8	10.68	known	TG	C/G	+	177,360	$2.9x10^{-11}$	-	68,201	$1.2x10^{-2}$
TOM1	rs138777	22	35.71	novel	TC	A/G	+	185,274	$4.7x10^{-8}$	+	67,303	$1.4x10^{-2}$
KAT5	rs12801636	11	65.39	novel	HDL	A/G	+	187,099	$3.2x10^{-8}$	-	62,171	$1.7x10^{-2}$
KCNK17	rs2758886	6	39.25	novel	TC	A/G	+	187,266	$3.0x10^{-8}$	+	69,242	$1.9x10^{-2}$
ABCA1	rs1883025	9	107.66	known	HDL	T/C	-	186,365	$1.5x10^{-65}$	+	61,161	$1.9x10^{-2}$
FTO	rs1121980	16	53.81	novel	HDL	A/G	-	185,524	$6.8x10^{-9}$	-	67,121	$2.7x10^{-2}$
MAFB	rs2902940	20	39.09	known	TC	A/G	+	185,716	$8.8x10^{-10}$	+	67,497	$2.7x10^{-2}$
SBNO1	rs4759375	12	123.8	known	HDL	T/C	+	94,311	$3.0x10^{-8}$	+	62,022	$2.9x10^{-2}$
APOE	rs4420638	19	45.42	known	LDL	A/G	-	93,103	$1.5x10^{-178}$	+	43,118	$3.2x10^{-2}$
ARL15	rs6450176	5	53.3	known	HDL	A/G	-	187,132	$6.9x10^{-10}$	+	65,297	$3.3x10^{-2}$
KLF14	rs4731702	7	130.43	known	HDL	T/C	+	187,085	$4.8x10^{-17}$	-	68,636	$3.5x10^{-2}$
PEPD	rs731839	19	33.9	novel	TG	A/G	-	176,161	$2.7x10^{-9}$	-	62,641	$3.9x10^{-2}$
CYP26A1	rs2068888	10	94.84	known	TG	A/G	-	177,712	$1.7x10^{-11}$	-	56,303	$3.9x10^{-2}$
MTMR3	rs5763662	22	30.38	novel	LDL	T/C	+	162,777	$1.2x10^{-8}$	+	58,243	$4.4x10^{-2}$
TYW1B	rs13238203	7	72.13	known	TG	T/C	-	101,951	$3.1x10^{-6}$	-	34,202	$4.6x10^{-2}$
CMIP	rs2925979	16	81.53	known	HDL	T/C	-	185,553	$1.3x10^{-19}$	+	65,526	$4.7x10^{-2}$

***International Consortium for Blood Pressure Genome-Wide Association Studies, Ehret GB et al. Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. Nature. 2011;478(7367):103-9*

Supplementary Table S2.12-D: Novel and Known Lipid Loci with SBP P-value < 0.05 from ICBP*

Locus	SNP	Chr	hg19 Pos (Mb)	Type	Trait	A1/A2	Lipid Direction	Lipid N	Lipid *P*-value	SBP Direction	SBP N	SBP *P*-value
BRAP	rs11065987	12	112.07	known	TC	A/G	+	187,309	$2.1x10^{-16}$	-	62,444	$2.1x10^{-8}$
SLC39A8	rs13107325	4	103.19	known	HDL	T/C	-	179,316	$1.1x10^{-15}$	-	58,910	$2.6x10^{-7}$
VEGFA	rs998584	6	43.76	novel	TG	A/C	+	174,573	$3.4x10^{-15}$	+	49,585	$9.3x10^{-5}$
CITED2	rs605066	6	139.83	known	HDL	T/C	+	94,311	$2.8x10^{-8}$	-	68,154	$1.1x10^{-3}$
LACTB	rs2652834	15	63.4	known	HDL	A/G	-	185,613	$3.6x10^{-11}$	+	61,931	$2.6x10^{-3}$
KCNK17	rs2758886	6	39.25	novel	TC	A/G	+	187,266	$3.0x10^{-8}$	+	69,249	$2.9x10^{-3}$
PABPC4	rs4660293	1	40.03	known	HDL	A/G	+	187,027	$2.9x10^{-18}$	-	69,821	$3.0x10^{-3}$
KAT5	rs12801636	11	65.39	novel	HDL	A/G	+	187,099	$3.2x10^{-8}$	-	62,173	$3.7x10^{-3}$
KLF14	rs4731702	7	130.43	known	HDL	T/C	+	187,085	$4.8x10^{-17}$	-	68,646	$7.0x10^{-3}$
MTMR3	rs5763662	22	30.38	novel	LDL	T/C	+	162,777	$1.2x10^{-8}$	+	58,275	$1.8x10^{-2}$
DAGLB	rs702485	7	6.45	novel	HDL	A/G	-	186,974	$6.5x10^{-12}$	+	67,622	$2.2x10^{-2}$
RSPO3	rs1936800	6	127.44	novel	HDL	T/C	-	187,111	$3.1x10^{-10}$	-	67,485	$2.5x10^{-2}$
PLEC1	rs11136341	8	145.04	known	LDL	A/G	-	82,810	$7.1x10^{-12}$	+	45,602	$2.7x10^{-2}$
TOM1	rs138777	22	35.71	novel	TC	A/G	+	185,274	$4.7x10^{-8}$	+	67,285	$3.0x10^{-2}$
CSNK1G3	rs4530754	5	122.86	novel	LDL	A/G	+	173,003	$3.6x10^{-12}$	-	69,174	$3.3x10^{-2}$
HFE	rs1800562	6	26.09	known	LDL	A/G	-	171,209	$8.3x10^{-14}$	+	65,402	$3.3x10^{-2}$
MVK	rs7134594	12	110	known	HDL	T/C	+	94,311	$1.8x10^{-13}$	+	69,719	$3.9x10^{-2}$
PEPD	rs731839	19	33.9	novel	TG	A/G	-	176,161	$2.7x10^{-9}$	-	62,643	$4.1x10^{-2}$
LOC55908	rs737337	19	11.35	known	HDL	T/C	+	185,432	$4.6x10^{-17}$	-	61,587	$4.1x10^{-2}$
PDE3A	rs7134375	12	20.47	known	HDL	A/C	+	187,088	$1.1x10^{-8}$	-	63,215	$4.3x10^{-2}$

**International Consortium for Blood Pressure Genome-Wide Association Studies, Ehret GB et al. Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. Nature. 2011;478(7367):103-9*

Supplementary Table S2.12-E: Novel and Known Lipid Loci with Fasting Glucose P-value < 0.05 from MAGIC*

Locus	SNP	Chr	hg19 Pos (Mb)	Type	Trait	A1/A2	Lipid Direction	Lipid N	Lipid *P*-value	FG Effect	FG *P*-value
GCKR	rs1260326	2	27.73	known	TG	T/C	+	177,765	$2.3x10^{-239}$	-0.027	$4.3x10^{-13}$
FADS1-2-3	rs174546	11	61.57	known	TG	T/C	+	177,785	$7.4x10^{-38}$	-0.021	$2.7x10^{-8}$
PPP1R3B	rs9987289	8	9.18	known	HDL	A/G	-	169,235	$2.0x10^{-41}$	0.028	$7.5x10^{-6}$
HBS1L	rs9376090	6	135.41	novel	TC	T/C	+	187,263	$2.6x10^{-9}$	0.014	$1.1x10^{-3}$
DNAH11	rs12670798	7	21.61	known	TC	T/C	-	187,287	$9.5x10^{-17}$	0.014	$1.5x10^{-3}$
TRPS1	rs2293889	8	116.6	known	HDL	T/G	-	180,102	$4.3x10^{-17}$	0.011	$2.3x10^{-3}$
TOM1	rs138777	22	35.71	novel	TC	A/G	+	185,274	$4.7x10^{-8}$	0.012	$3.1x10^{-3}$
LIPC	rs1532085	15	58.68	known	HDL	A/G	+	185,482	$1.2x10^{-188}$	-0.011	$4.9x10^{-3}$
LRP1	rs11613352	12	57.79	known	TG	T/C	-	177,799	$9.4x10^{-14}$	-0.012	$5.8x10^{-3}$
INSR	rs7248104	19	7.22	novel	TG	A/G	-	176,083	$5.1x10^{-10}$	0.0085	$2.1x10^{-2}$
NPC1L1	rs2072183	7	44.58	known	TC	C/G	+	183,969	$4.2x10^{-15}$	-0.012	$2.1x10^{-2}$
ABCA1	rs1883025	9	107.66	known	HDL	T/C	-	186,365	$1.5x10^{-65}$	0.01	$2.2x10^{-2}$
APOB	rs1367117	2	21.26	known	LDL	A/G	+	173,007	$9.5x10^{-183}$	-0.009	$2.8x10^{-2}$
UGT1A1	rs11563251	2	234.68	novel	TC	T/C	+	187,107	$1.3x10^{-9}$	0.014	$3.1x10^{-2}$
STARD3	rs11869286	17	37.81	known	HDL	C/G	+	177,918	$2.7x10^{-17}$	-0.0078	$4.1x10^{-2}$
MVK	rs7134594	12	110	known	HDL	T/C	+	94,311	$1.7x10^{-13}$	0.0075	$4.4x10^{-2}$
PABPC4	rs4660293	1	40.03	known	HDL	A/G	+	187,027	$2.9x10^{-18}$	-0.0087	$4.5x10^{-2}$

**Data on glycaemic traits have been contributed by MAGIC investigators and have been downloaded from www.magicinvestigators.org; Dupuis J et al. New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. Nat Genet. 2010;42:105-16*

Supplementary Table S2.12-F: Novel and Known Lipid Loci with T2D P-value < 0.05 from DIAGRAM*

Locus	SNP	chr	hg19 Pos (Mb)	Type	Trait	A1/A2	Lipid Direction	Lipid N	Lipid *P*-value	T2D N	T2D OR (95% CI)	T2D *P*-value
FTO	rs1121980	16	53.81	novel	HDL	A/G	-	185,524	$6.8x10^{-9}$	22,570	1.12 (1.07,1.26)	$1.4x10^{-7}$
KLF14	rs4731702	7	130.43	known	HDL	T/C	+	187,085	$4.8x10^{-17}$	22,570	-1.10 (-1.06,-1.15)	$2.1x10^{-6}$
APOE	rs4420638	19	45.42	known	LDL	A/G	-	93,103	$1.5x10^{-178}$	18,352	1.15 (1.07,1.23)	$5.4x10^{-5}$
IRS1	rs2972146	2	227.1	known	HDL	T/G	-	184,044	$1.9x10^{-17}$	22,570	1.09 (1.04,1.13)	$9.0x10^{-5}$
ARL15	rs6450176	5	53.3	known	HDL	A/G	-	187,132	$6.9x10^{-10}$	22,570	1.09 (1.04,1.14)	$4.0x10^{-4}$
MAP3K1	rs9686661	5	55.86	known	TG	T/C	+	177,050	$2.5x10^{-16}$	22,570	1.09 (1.03,1.14)	$1.7x10^{-3}$
CILP2	rs10401969	19	19.41	known	TC	T/C	+	185,667	$4.1x10^{-77}$	22,570	-1.14 (-1.04,-1.24)	$3.1x10^{-3}$
HNF1A	rs1169288	12	121.42	known	TC	A/C	-	175,774	$3.9x10^{-17}$	22,570	-1.06 (-1.02,-1.11)	$4.7x10^{-3}$
CMIP	rs2925979	16	81.53	known	HDL	T/C	-	185,553	$1.3x10^{-19}$	21,198	1.07 (1.02,1.12)	$4.9x10^{-3}$
NPC1L1	rs2072183	7	44.58	known	TC	C/G	+	183,969	$4.2x10^{-15}$	17,302	-1.10 (-1.03,-1.18)	$5.0x10^{-3}$
COBLL1	rs12328675	2	165.54	known	HDL	T/C	-	187,092	$2.1x10^{-15}$	22,570	1.08 (1.02,1.16)	$1.2x10^{-2}$
ABO	rs9411489	9	136.155	known	LDL	T/C	+	119,312	$1.8x10^{-41}$	21,520	1.07 (1.01,1.13)	$1.5x10^{-2}$
VEGFA	rs998584	6	43.76	novel	TG	A/C	+	174,573	$3.4x10^{-15}$	17,302	1.07 (1.01,1.13)	$1.8x10^{-2}$
GPAM	rs2255141	10	113.93	known	TC	A/G	+	187,266	$6.5x10^{-16}$	22,570	-1.05 (-1.01,-1.10)	$2.1x10^{-2}$
FADS1-2-3	rs174546	11	61.57	known	TG	T/C	+	177,785	$7.4x10^{-38}$	22,570	-1.04 (-1.01,-1.09)	$2.6x10^{-2}$
MC4R	rs12967135	18	57.85	known	HDL	A/G	-	153,533	$3.6x10^{-8}$	22,570	1.05 (1.01,1.10)	$2.9x10^{-2}$
LIPC	rs1532085	15	58.68	known	HDL	A/G	+	185,482	$1.2x10^{-188}$	22,570	-1.05 (-1.00,-1.09)	$2.9x10^{-2}$
HNF4A	rs1800961	20	43.04	known	HDL	T/C	-	157,871	$1.6x10^{-34}$	13,971	1.14 (1.00,1.30)	$4.7x10^{-2}$

**Voight BF et al. Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. Nat Genet. 2010;42:579–589*

Supplementary Table S2.12-G: Novel and Known Lipid Loci with WHR adj BMI P-value < 0.05 from GIANT*

Locus	SNP	Chr	hg19 Pos (Mb)	Type	Trait	A1/A2	Lipid Direction	Lipid N	Lipid *P*-value	WHR Direction	WHR N	WHR *P*-value
RSPO3	rs1936800	6	127.44	novel	HDL	T/C	-	187,111	$3.1x10^{-10}$	+	77,164	$5.0x10^{-14}$
VEGFA	rs998584	6	43.76	novel	TG	A/C	+	174,573	$3.4x10^{-15}$	+	72,804	$3.2x10^{-9}$
ZNF664	rs4765127	12	124.46	known	HDL	T/G	+	94,198	$7.8x10^{-10}$	-	77,048	$1.8x10^{-5}$
COBLL1	rs12328675	2	165.54	known	HDL	T/C	-	187,092	$2.1x10^{-15}$	+	77,160	$2.8x10^{-5}$
C4orf52	rs10019888	4	26.06	novel	HDL	A/G	+	187,077	$4.9x10^{-8}$	-	77,165	$5.1x10^{-5}$
MAP3K1	rs9686661	5	55.86	known	TG	T/C	+	177,050	$2.5x10^{-16}$	+	77,164	$8.0x10^{-5}$
CITED2	rs605066	6	139.83	known	HDL	T/C	+	94,311	$2.8x10^{-8}$	-	77,164	$1.3x10^{-4}$
TOM1	rs138777	22	35.71	novel	TC	A/G	+	185,274	$4.7x10^{-8}$	+	77,218	$2.8x10^{-4}$
GCKR	rs1260326	2	27.73	known	TG	T/C	+	177,765	$2.3x10^{-239}$	+	77,128	$3.4x10^{-4}$
FAM13A	rs3822072	4	89.74	novel	HDL	A/G	-	187,115	$4.1x10^{-12}$	+	77,163	$3.5x10^{-4}$
FN1	rs1250229	2	216.3	novel	LDL	T/C	-	173,032	$3.1x10^{-8}$	+	77,155	$6.6x10^{-4}$
APOE	rs4420638	19	45.42	known	LDL	A/G	-	93,103	$1.5x10^{-178}$	+	69,832	$8.5x10^{-4}$
STAB1	rs13326165	3	52.53	novel	HDL	A/G	+	187,134	$9.0x10^{-11}$	-	77,168	$1.0x10^{-3}$
CILP2	rs10401969	19	19.41	known	TC	T/C	+	185,666	$4.1x10^{-77}$	-	77,160	$2.5x10^{-3}$
ERGIC3	rs2277862	20	34.15	known	TC	T/C	-	185,738	$5.3x10^{-11}$	-	77,165	$9.9x10^{-3}$
TOP1	rs6029526	20	39.67	known	LDL	A/T	+	88,433	$4.8x10^{-18}$	+	77,165	$1.0x10^{-2}$
KCNK17	rs2758886	6	39.25	novel	TC	A/G	+	187,266	$3.0x10^{-8}$	+	77,167	$1.2x10^{-2}$
CMIP	rs2925979	16	81.53	known	HDL	T/C	-	185,553	$1.3x10^{-19}$	+	77,164	$1.5x10^{-2}$
ARL15	rs6450176	5	53.3	known	HDL	A/G	-	187,131	$6.9x10^{-10}$	-	77,165	$1.8x10^{-2}$
ACAD1	rs17404153	3	132.16	novel	LDL	T/G	-	172,898	$1.8x10^{-9}$	-	77,166	$1.9x10^{-2}$
PPP1R3B	rs9987289	8	9.18	known	HDL	A/G	-	169,234	$1.9x10^{-41}$	+	77,170	$2.0x10^{-2}$
MYLIP	rs3757354	6	16.13	known	LDL	T/C	-	172,986	$2.1x10^{-17}$	-	72,863	$2.6x10^{-2}$
HBS1L	rs9376090	6	135.41	novel	TC	T/C	+	187,263	$2.6x10^{-9}$	-	77,165	$3.3x10^{-2}$
ANXA9-CERS2	rs267733	1	150.96	novel	LDL	A/G	+	164,562	$5.3x10^{-9}$	-	77,162	$4.1x10^{-2}$
LRPAP1	rs6831256	4	3.47	novel	TG	A/G	-	177,494	$1.6x10^{-12}$	-	77,141	$4.1x10^{-2}$
NAT2	rs1495741	8	18.27	known	TG	A/G	-	87,977	$2.7x10^{-12}$	+	77,166	$4.6x10^{-2}$
TTC39B	rs581080	9	15.31	known	HDL	C/G	+	186,937	$1.0x10^{-19}$	+	77,165	$4.9x10^{-2}$

**Heid IM et al. Meta-analysis identifies 13 new loci associated with waist-hip ratio and reveals sexual dimorphism in the genetic basis of fat distribution. Nat Genet 2010;42, 949-960.*

Supplementary Table S2.13: Overlap of Lipid Subfractions in Framingham with Novel and Known Lipid Associated Loci (P<1.4x10-5)

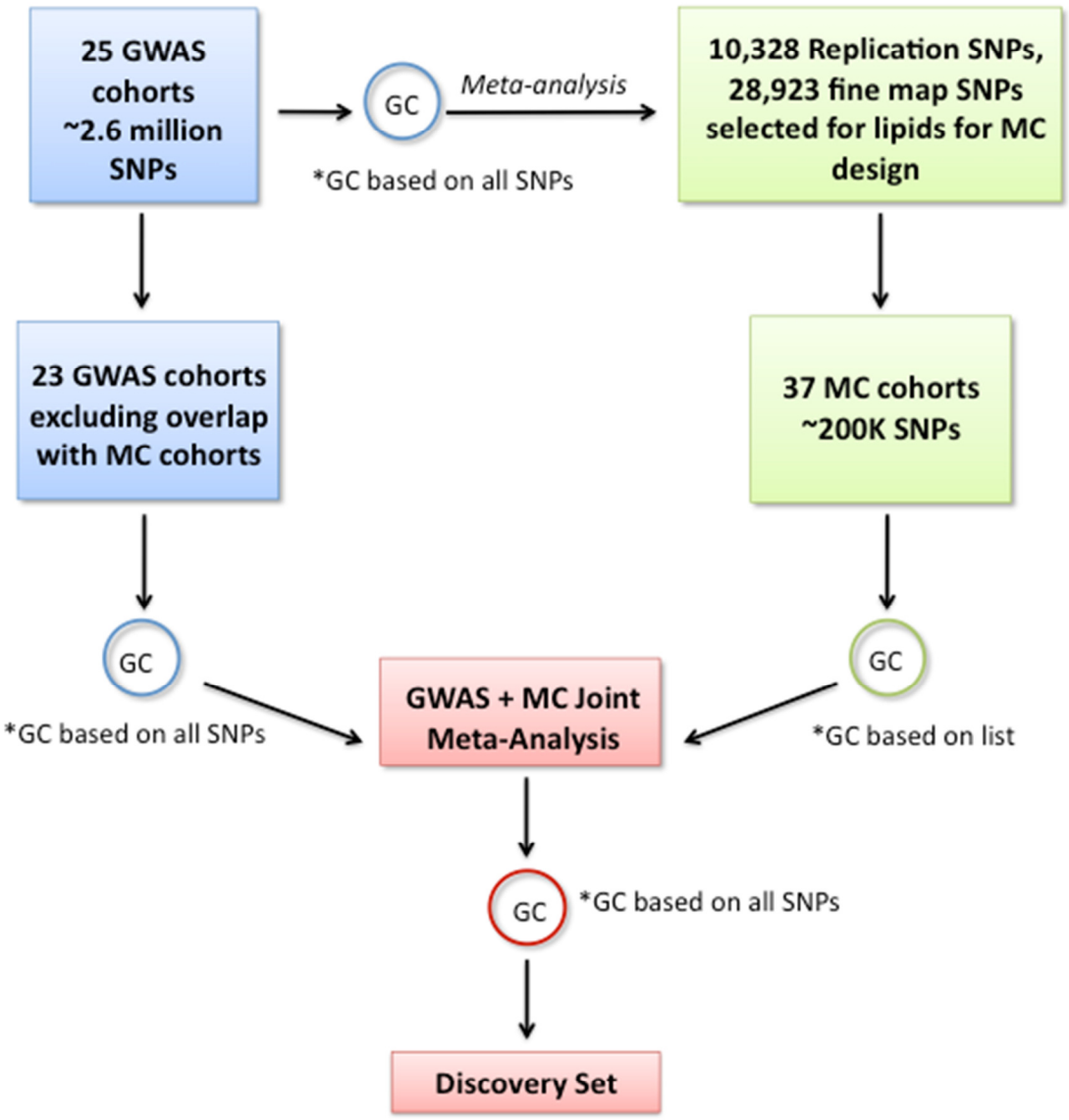
Locus	SNP	Lipid Subfraction Trait*	A1/A2	N	MAF	Beta	*P*-value	Novel Lipid Locus	Lipid *P*-value
Overlap of Lipid Subfractions with HDL Loci									
LIPC	rs1532085	HDL2 cholesterol subfraction	A/G	2,900	0.38	0.13	$2x10^{-06}$	N	$1x10^{-188}$
LIPC	rs1532085	HDL size	A/G	2,742	0.38	0.17	$4x10^{-09}$	N	$1x10^{-188}$
LIPC	rs1532085	Large particles of HDL	A/G	2,742	0.38	0.16	$6x10^{-08}$	N	$1x10^{-188}$
CETP	rs3764261	Intermediate density lipoprotein	A/C	2,742	0.31	-0.16	$9x10^{-08}$	N	$1x10^{-769}$
CETP	rs3764261	HDL2 cholesterol subfraction	A/C	2,900	0.31	0.18	$1x10^{-09}$	N	$1x10^{-769}$
CETP	rs3764261	LDL size	A/C	2,742	0.31	0.17	$7x10^{-08}$	N	$1x10^{-769}$
CETP	rs3764261	Large particles of LDL	A/C	2,742	0.31	0.14	$9x10^{-06}$	N	$1x10^{-769}$
CETP	rs3764261	HDL size	A/C	2,742	0.31	0.19	$6x10^{-10}$	N	$1x10^{-769}$
CETP	rs3764261	Large particles of HDL	A/C	2,742	0.31	0.22	$4x10^{-13}$	N	$1x10^{-769}$
CETP	rs3764261	HDL3 cholesterol subfraction	A/C	2,900	0.31	0.23	$1x10^{-14}$	N	$1x10^{-769}$
CETP	rs3764261	Apoliprotein AI concentration	A/C	2,885	0.31	0.19	$4x10^{-10}$	N	$1x10^{-769}$
LIPG	rs7241918	Apoliprotein AI concentration	G/T	2,885	0.17	-0.19	$2x10^{-07}$	N	$1x10^{-44}$
PLTP	rs6065906	Large particles of HDL	C/T	2,742	0.18	-0.18	$1x10^{-06}$	N	$5x10^{-40}$
PLTP	rs6065906	Medium particles of HDL	C/T	2,742	0.18	0.35	$1x10^{-21}$	N	$5x10^{-40}$
Overlap of Lipid Subfractions with LDL Loci									
SORT1	rs629301	Apolipoprotein B concentration	G/T	2,821	0.21	-0.19	$2x10^{-08}$	N	$5x10-^{241}$
ApoE	rs4420638	ApoE concentration	G/A	2,260	0.16	-0.62	$9x10^{-10}$	N	$2x10^{-178}$
Overlap of Lipid Subfractions with Triglyceride Loci									
GCKR	rs1260326	Apolipoprotein CIII concentration	T/C	2,484	0.45	0.18	$2x10^{-10}$	N	$2x10^{-239}$
LPL	rs12678919	Apoliprotein AI concentration	G/A	2,885	0.1	0.2	$1x10^{-05}$	N	$2x10^{-199}$
APOA1	rs964184	Medium particles of VLDL	G/C	2,742	0.14	0.26	$2x10^{-10}$	N	$7x10^{-224}$
APOA1	rs964184	Remnant like particles expressed as triglycerides	G/C	2,385	0.14	0.2	$5x10^{-06}$	N	$7x10^{-224}$
APOA1	rs964184	Remnant like particles expressed as cholesterol	G/C	2,468	0.14	0.19	$7x10^{-06}$	N	$7x10^{-224}$
APOA1	rs964184	Apolipoprotein B concentration	G/C	2,821	0.14	0.23	$4x10^{-09}$	N	$7x10^{-224}$

**LDL=low density lipoprotein, HDL=high density lipoprotein, VLDL=very low density lipoprotein*

The threshold used for significance is1.4x10-5. This corresponds to a Bonferroni correction for 23 subfractions and 151 SNPs found in the lipid subfraction dataset (0.05/(23*151)).

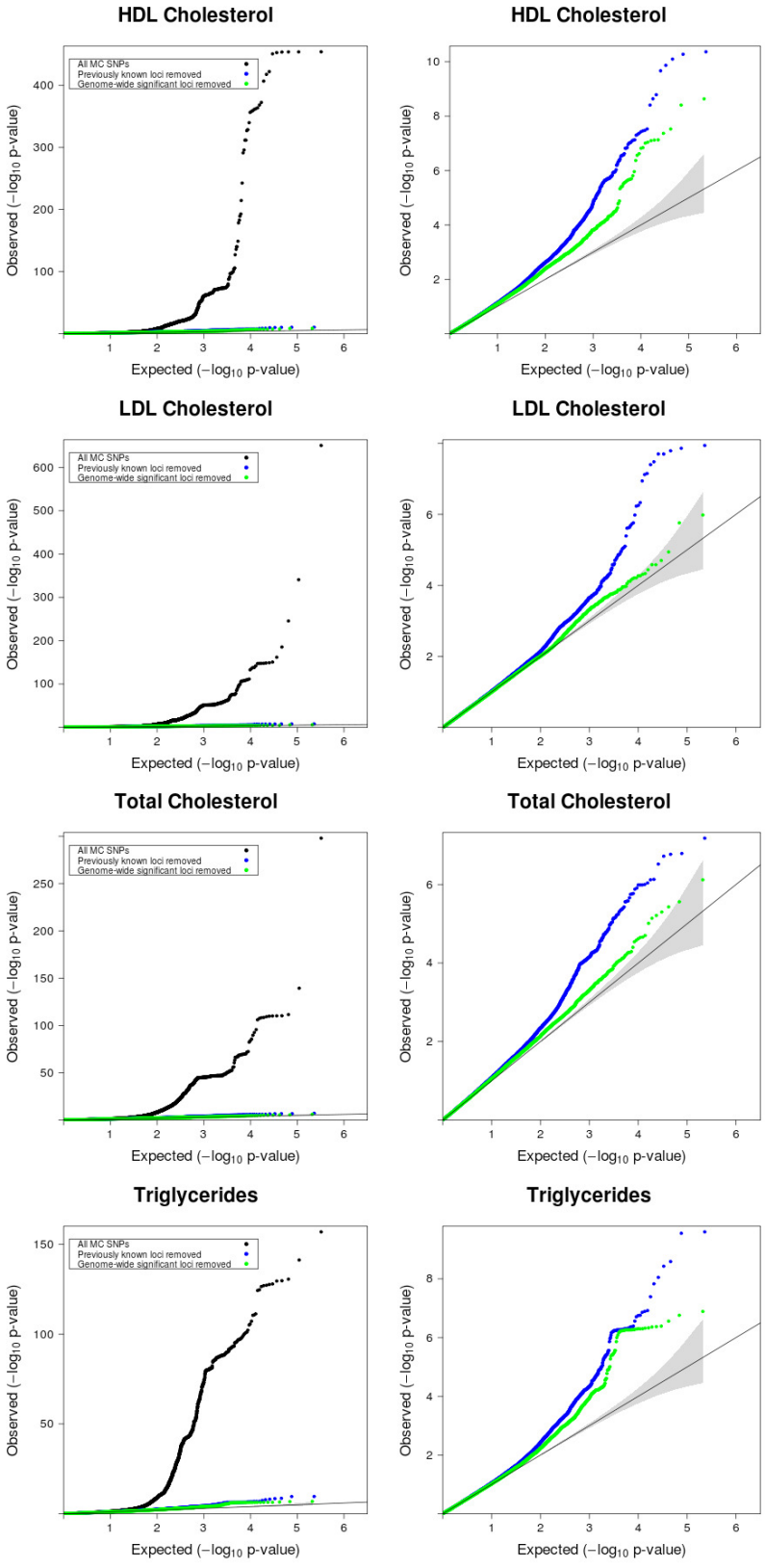
Supplementary Figures

Supplementary Figure S2.1: Study Design



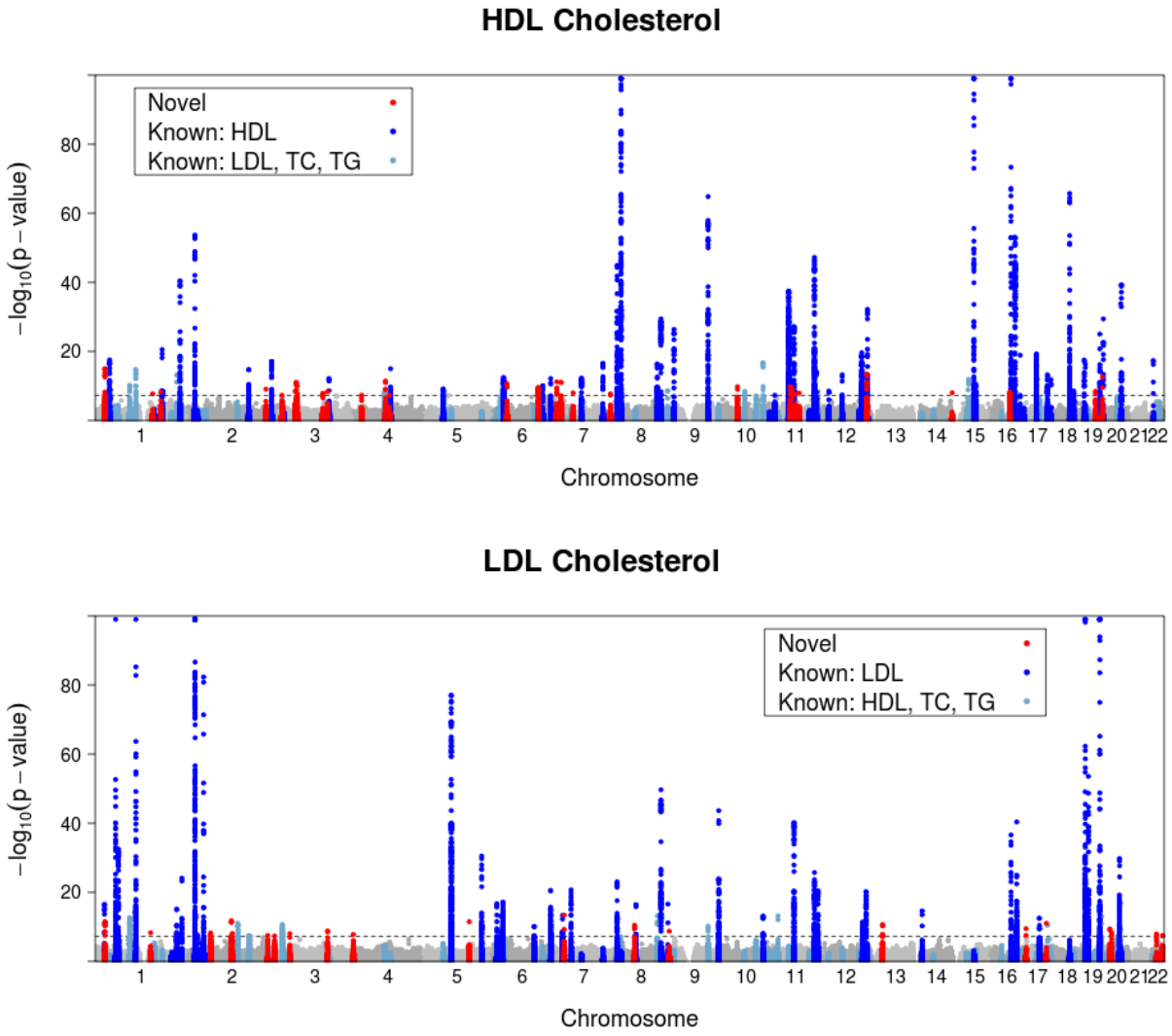
Supplementary Figure S2.2: QQ Plots of Metabochip Meta-Analysis P-value Distributions

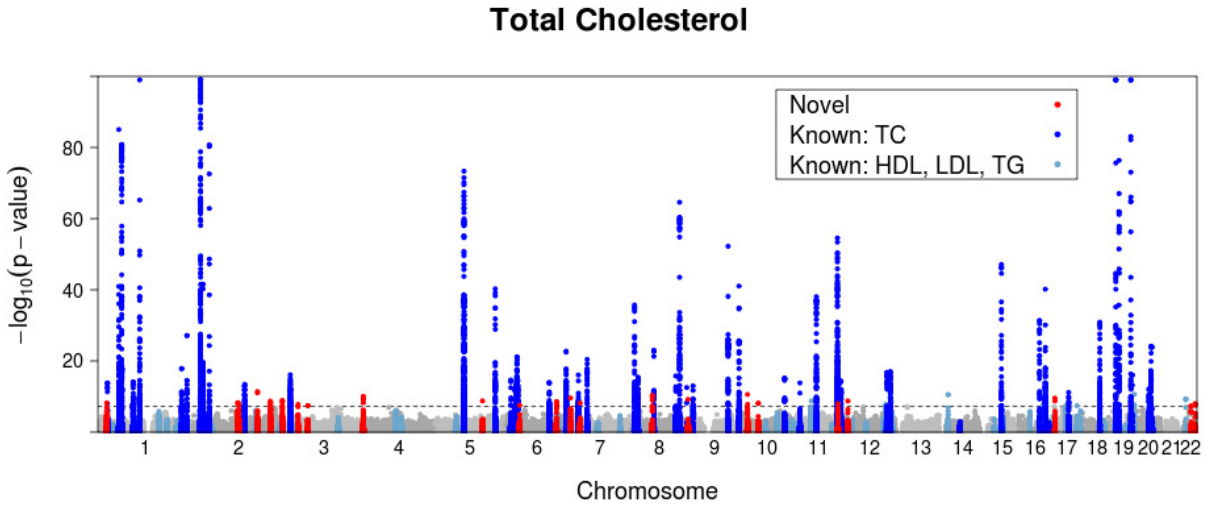
Quantile-quantile plots of metabochip meta-analysis observed association –log10 p-values plotted against expected –log10 p-values. Points in blue represent the p-value distribution after removing ± 1MB of previously known lipid loci. There is reduced inflation of p-values after removing ± 1MB of all genome-wide significant loci (shown in green). Lambda 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 Megabase 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 two traits (LDL cholesterol and triglyceride levels) but were at 1.05 for total cholesterol and 1.07 for HDL cholesterol. The interpretation of genomic control values from this experiment is complex because MetaboChip SNPs are heavily concentrated on regions associated with lipids and other cardiovascular traits. The initial genomic control values likely reflect this enrichment; the modestly high genomic control value after excluding confirmed regions of association could reflect a combination of polygenic effects, additional loci to discover, or population stratification.

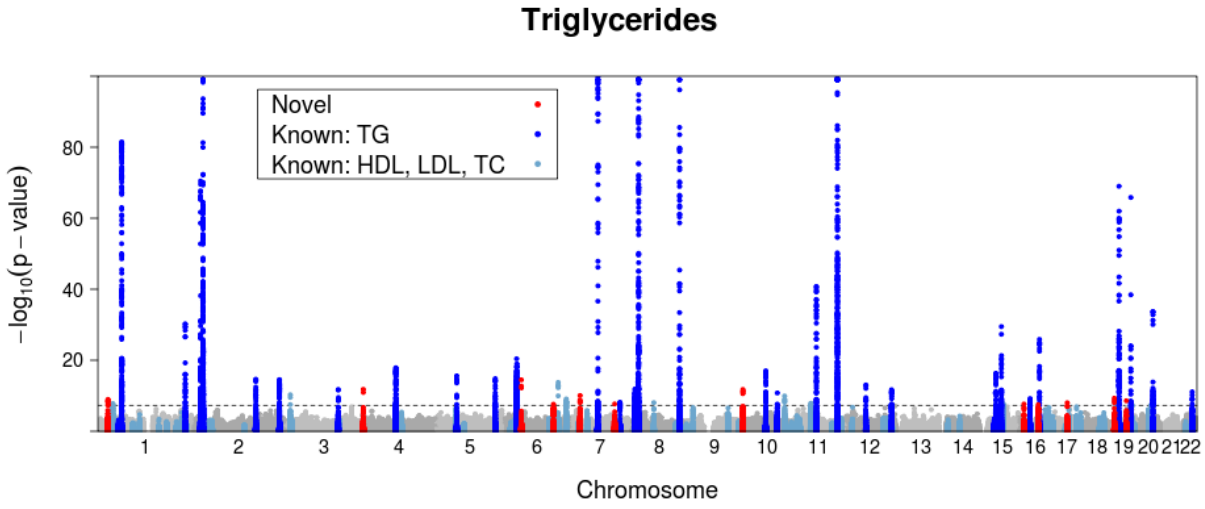

HDL Cholesterol
All MC SNPs
Previously known loci removed
Genome-wide significant loci removed
Observed (−log10 p-value)
Expected (−log10 p-value)
HDL Cholesterol
Observed (−log10 p-value)
Expected (−log10 p-value)
LDL Cholesterol
All MC SNPs
Previously known loci removed
Genome-wide significant loci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

Supplementary Figure S2.3: Manhattan Plots of Lipid-specific Association Results

Manhattan plots highlight significant SNP associations for each trait ($P< 5x10^{-8}$). Trait-specific novel loci are shown in red. Association results for known markers previously reported to be associated with lipid traits are shown in dark blue (when primary trait is the same trait) and light blue (when primary trait is a different lipid trait).

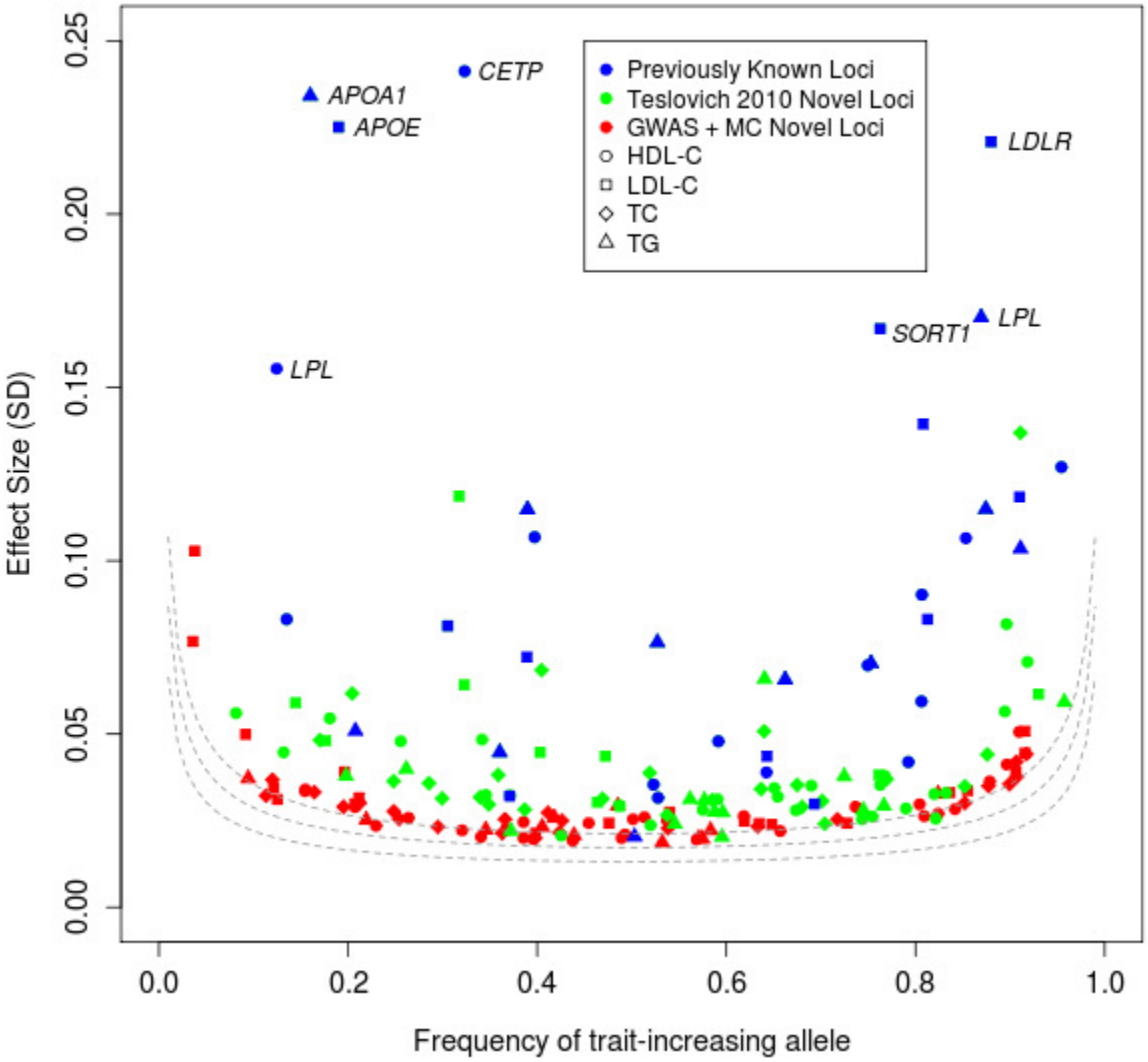



Total Cholesterol
Novel
Known: TC
Known: HDL, LDL, TG
$-\log_{10}(p-value)$
Chromosome


Triglycerides
Novel
Known: TG
Known: HDL, LDL, TC
$-\log_{10}(p-value)$
Chromosome

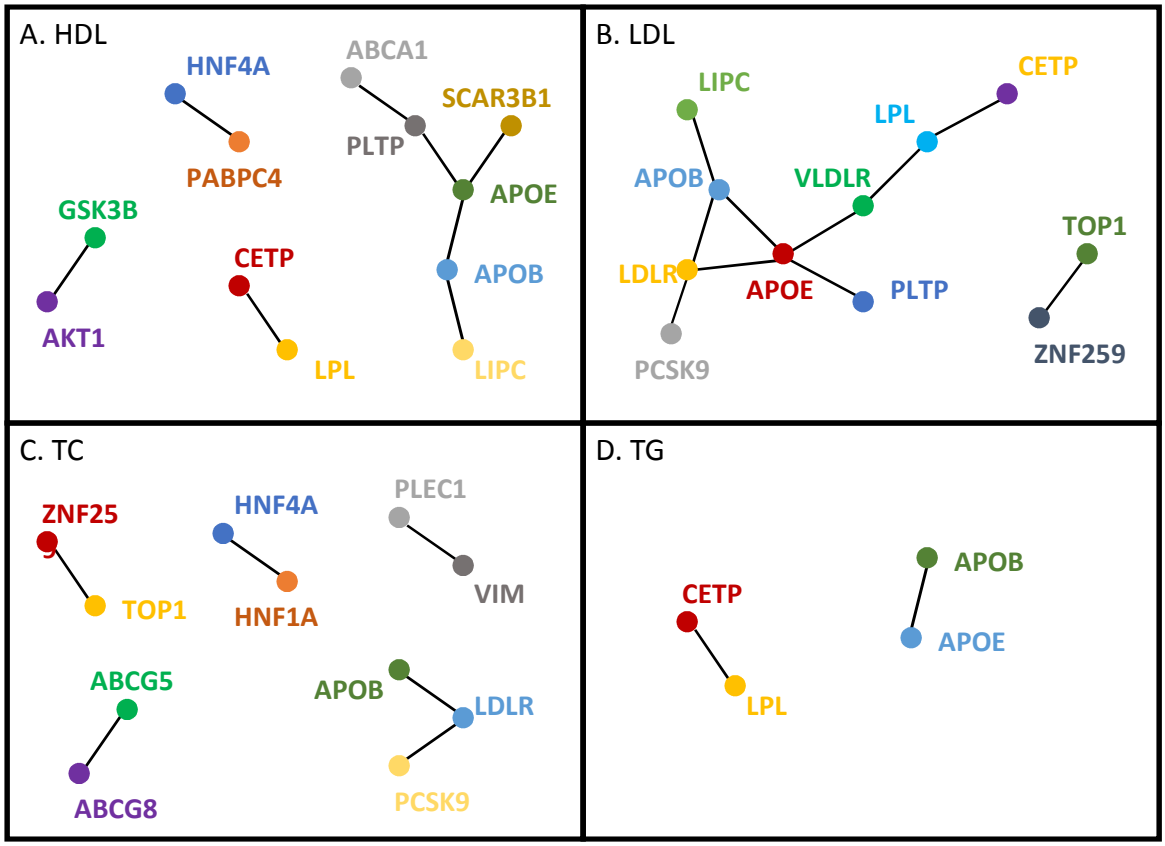
Supplementary Figure S2.4: Effect Size vs. Allele Frequency at Lipid Associated Loci

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 $5x10^{-8}$.



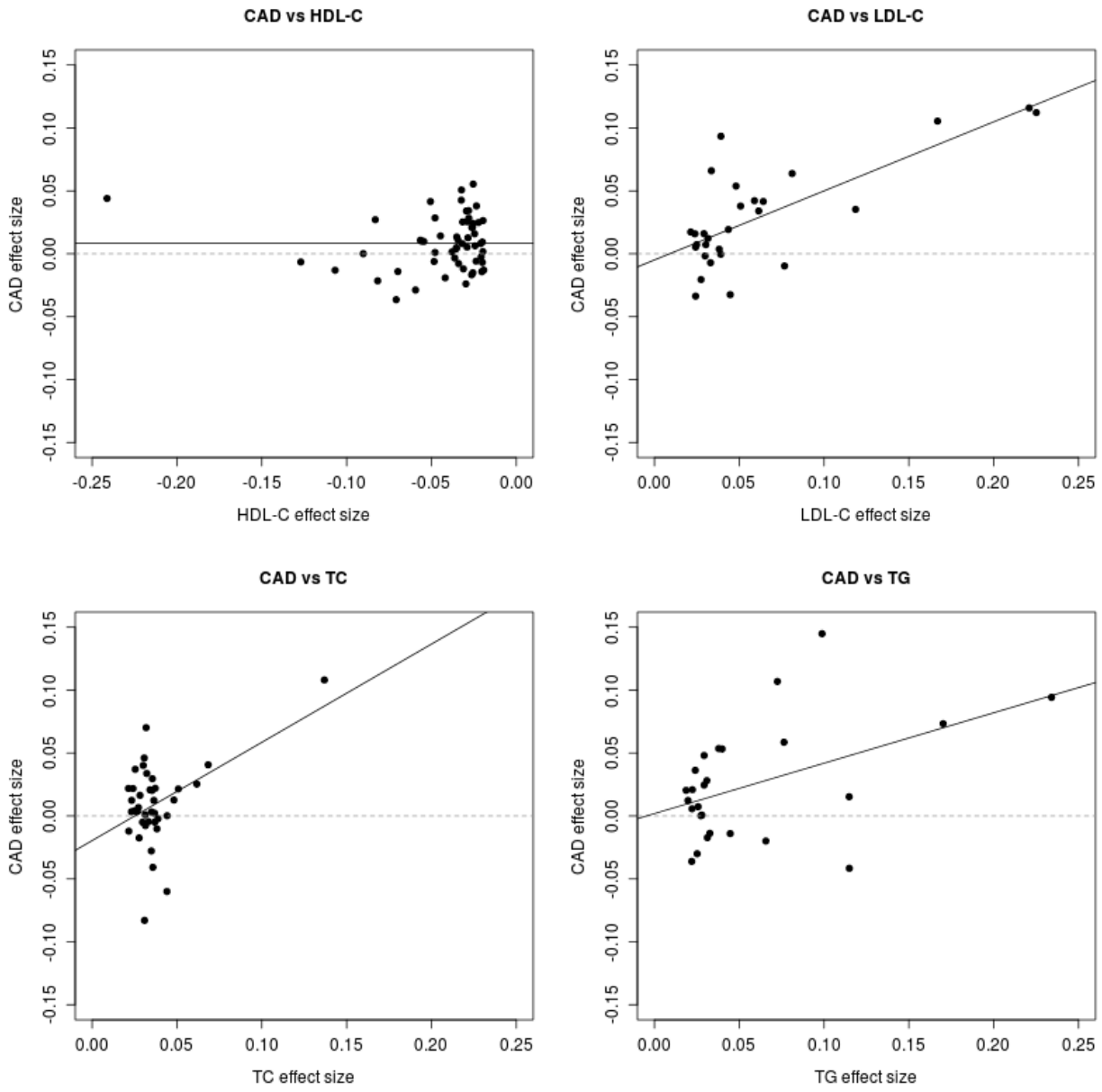
Supplementary Figure S2.5: Direct Protein-Protein Interactions from Dapple Analysis

Direct protein-protein interactions for A. HDL-C (8 interactions, *P* = .002), B. LDL-C (10 interactions, *P* = .0002), C. total cholesterol (6 interactions, *P* = .017), and D. triglycerides (2 interactions, *P* = .27) show connections between novel and known genes in the same pathways. We tested genes near previously known and new loci.



Supplementary Figure S2.6: Lipid vs. CAD Effect Sizes

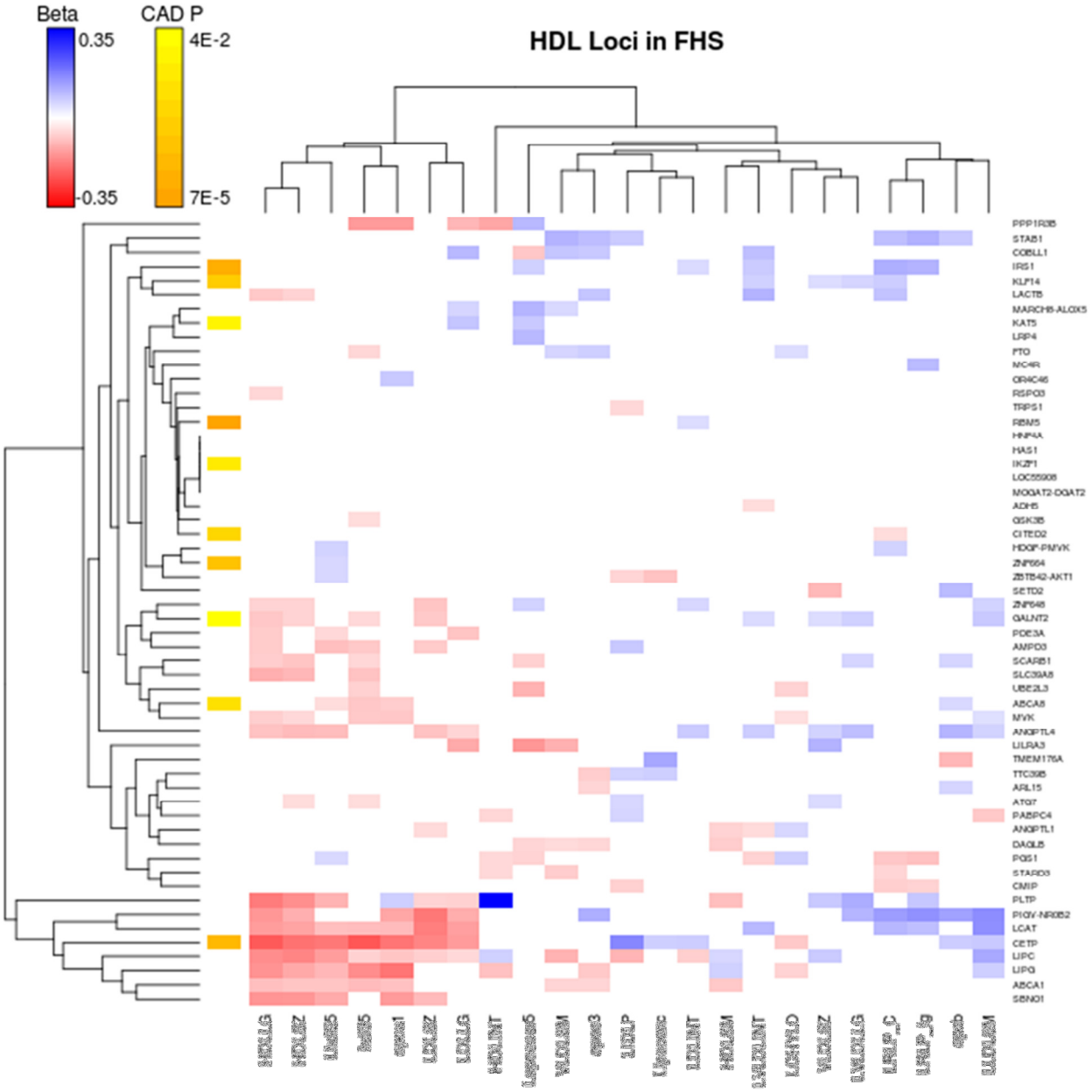
Plots show coronary artery disease (CAD) effect sizes against lipid effect sizes for SNPs showing primary association with each lipid trait. All effect sizes were oriented to the lipid trait-increasing (or trait-decreasing for HDL) allele. Diagonal lines represent regressions of predictor lipid effect 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 effect sizes were strongly associated with CAD effect sizes (Pearson r= 0.74, $P = 7 \times 10^{-6}$). The correlation between CAD effect size and triglyceride effect size (Pearson r= 0.46, $P = 0.02$) was higher than that observed for HDL (Pearson r=-9×10^{-4}, $P = 0.99$). Lipid effect sizes were transformed into SD units.



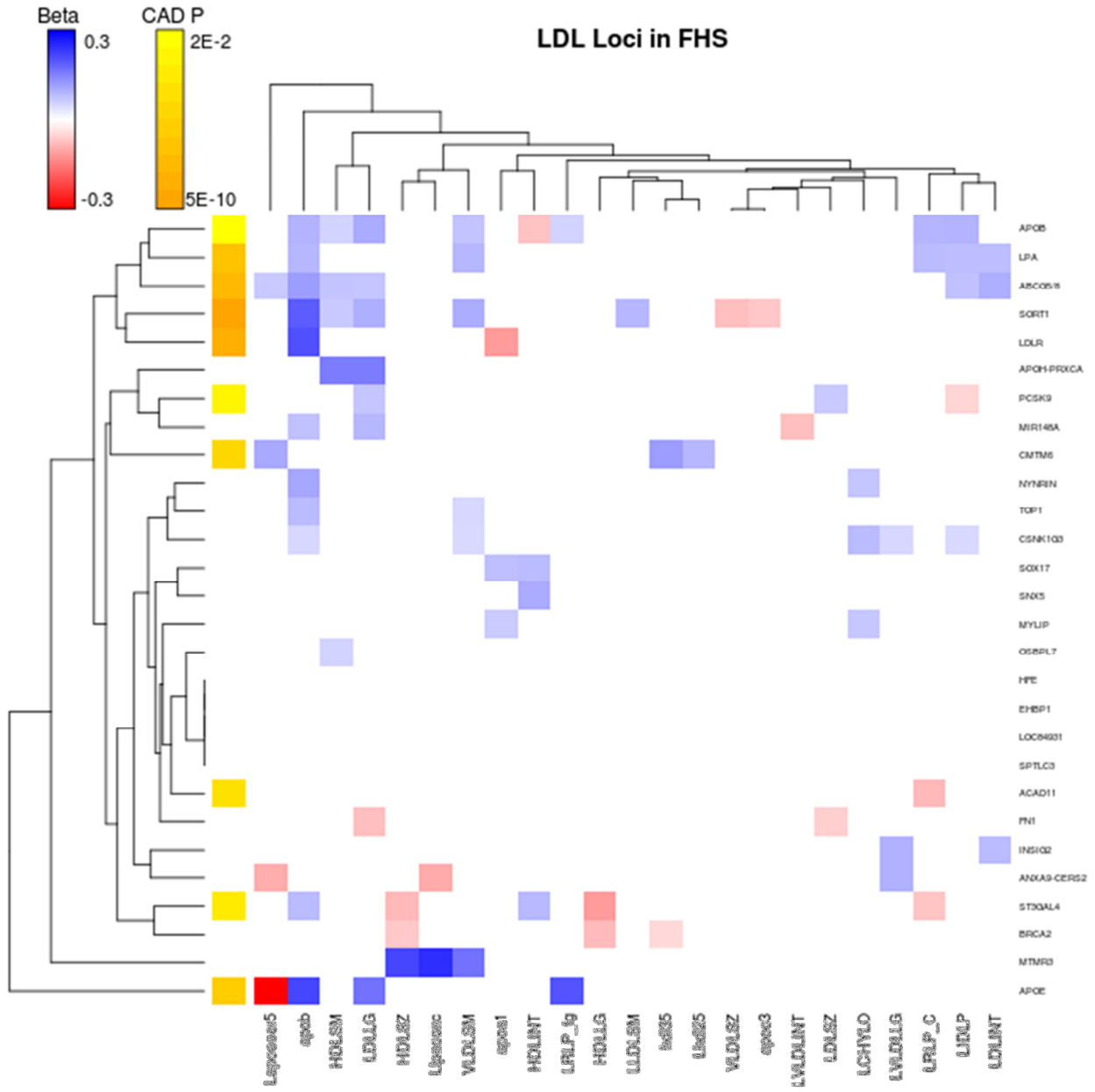
Supplementary Figure S2.7: Association with Lipid Subfractions

Heatmaps show effect sizes for association ($P < 0.10$) with 23 lipid subfractions in Framingham 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 association ($P < 0.05$) of lipid-associated SNPs with coronary artery disease (CAD) is 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. (E) shows a heatmap of correlations for the 23 lipid subfractions in Framingham. F-I show results from Women's Genome Health Study[1]. (J) shows a heatmap of lipid subfraction correlations in WGHS.

A.

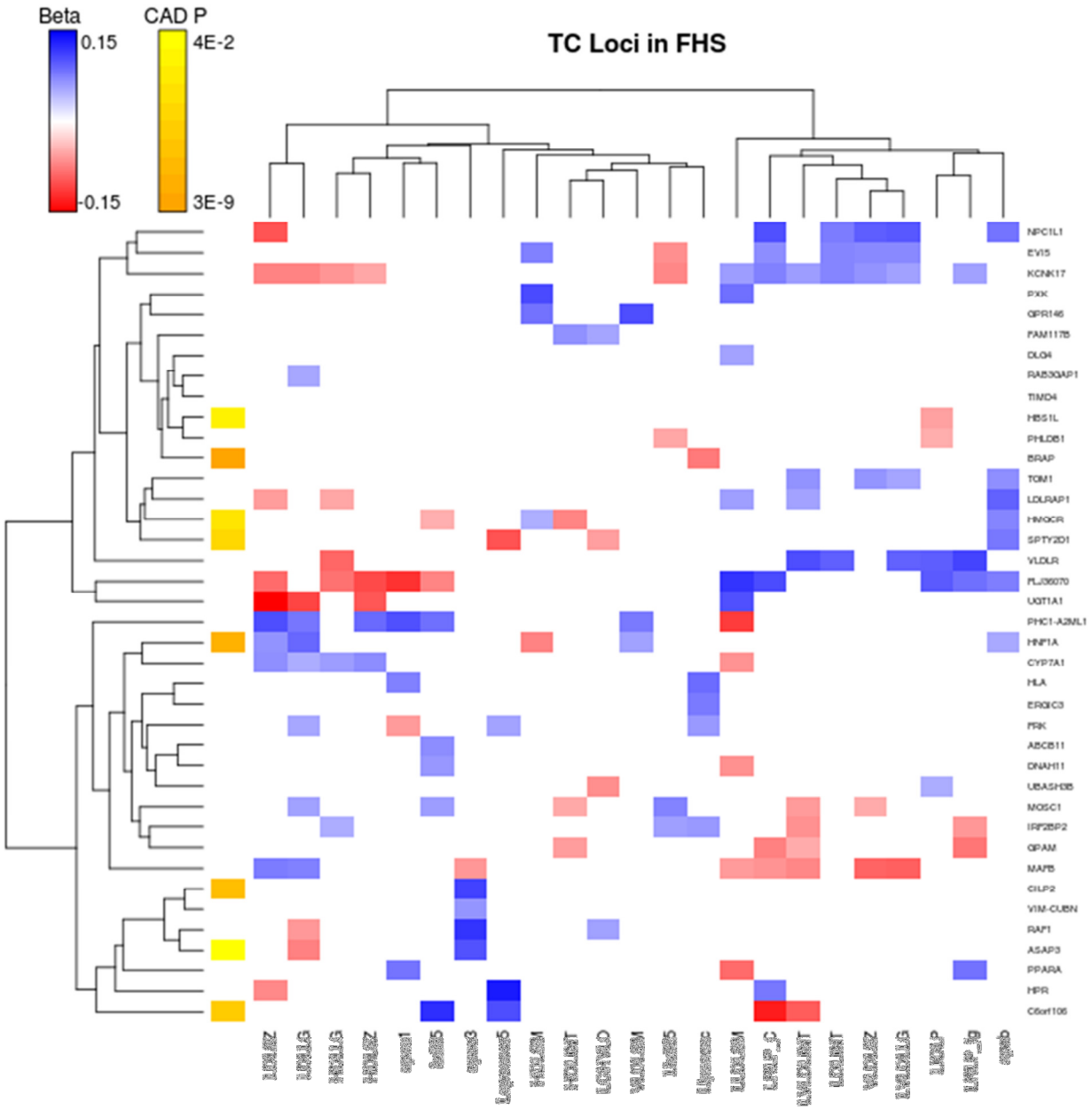


B.

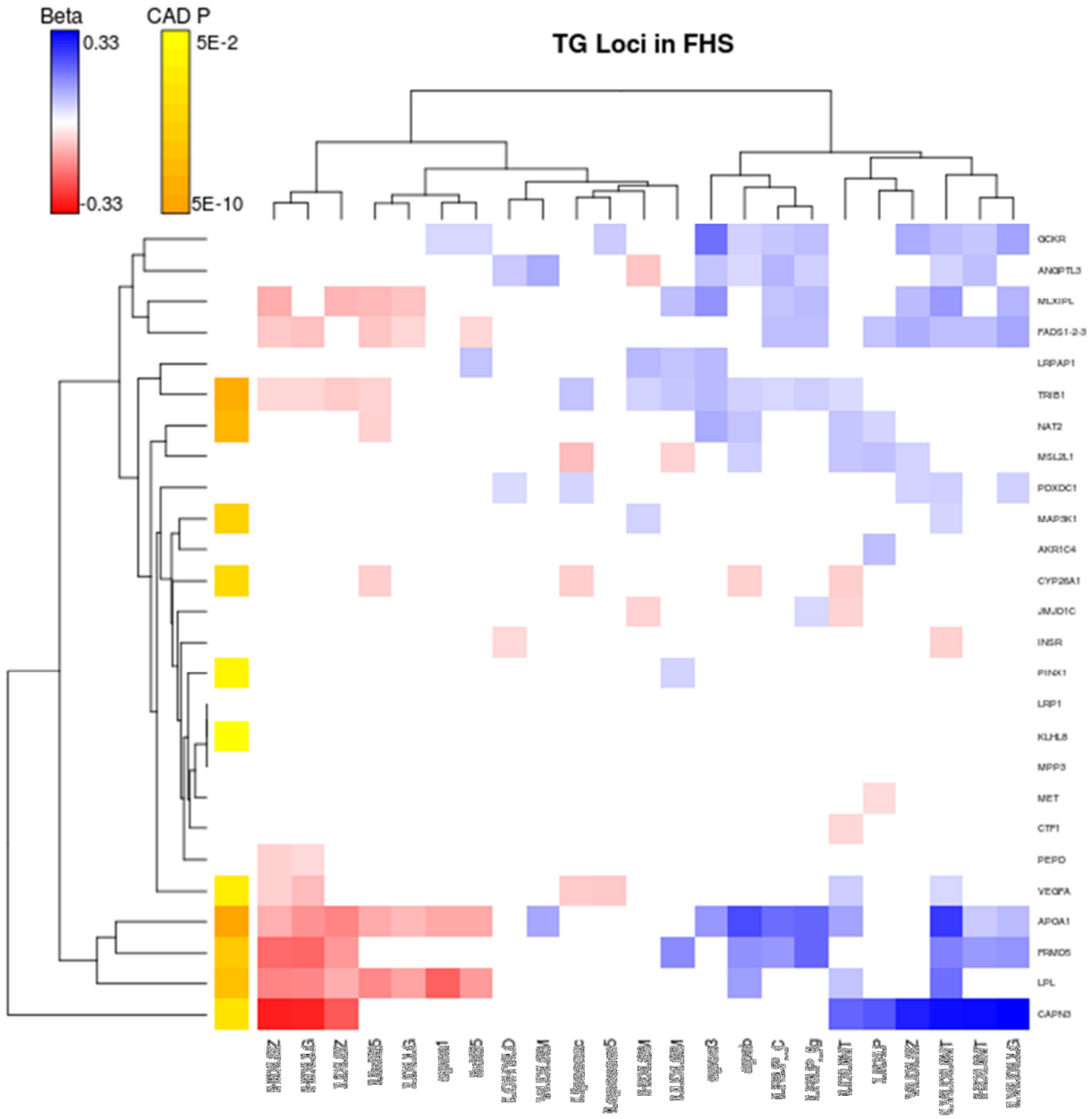


*The beta for the strongest association observed, rs4420638 at the APOE locus and Lapoeser5apc (beta = -0.62), is displayed as the minimum (-0.3) so that the color scale for the heatmap is more comparable to the heatmaps from the other 3 lipid traits.

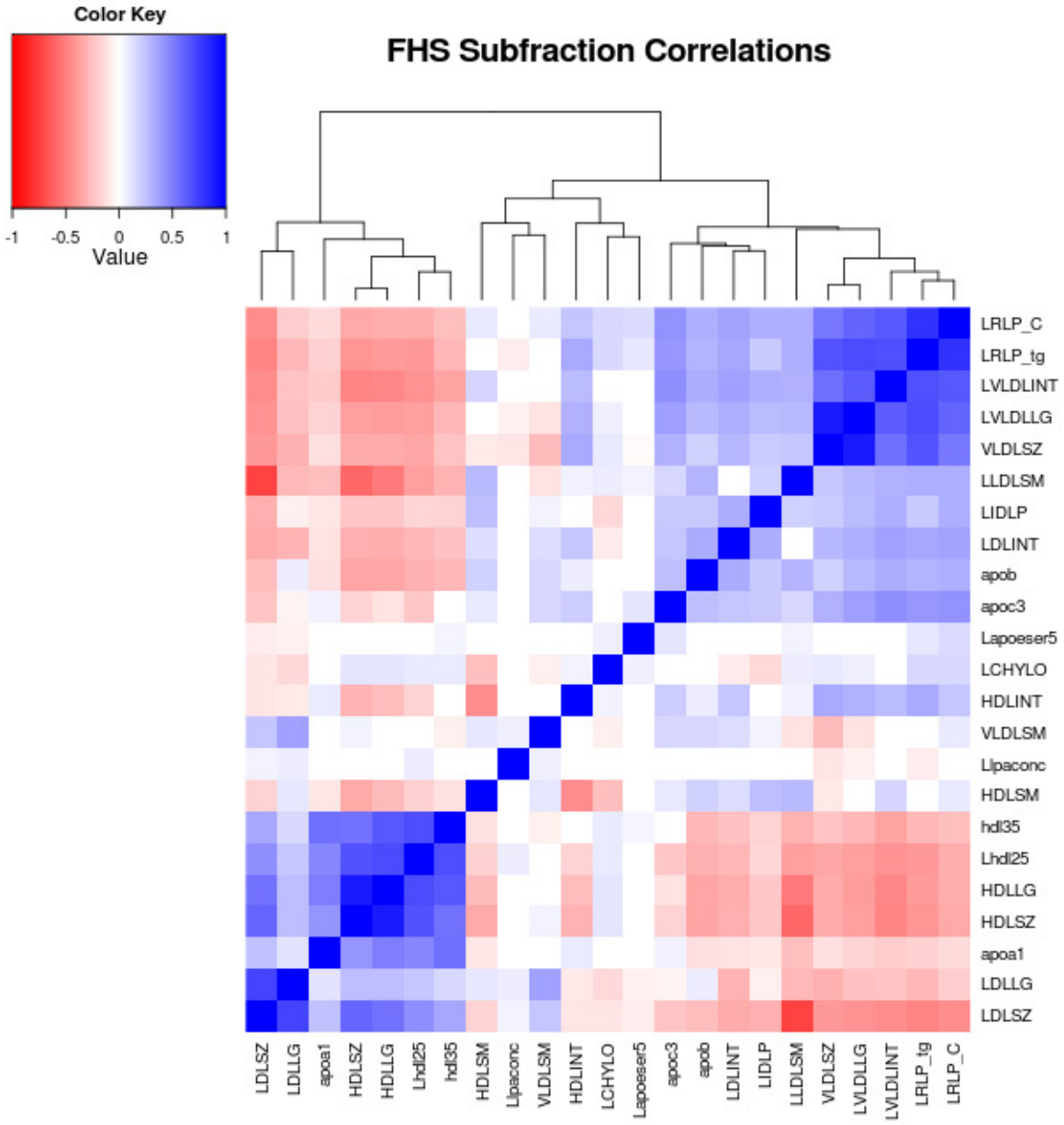
C.


Beta
0.15
-0.15
CAD P
4E-2
3E-9
TC Loci in FHS
NPC1L1
EVI5
KCNK17
PXK
QPRT46
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

D.


Beta
0.33
-0.33
CAD P
5E-2
5E-10
TG Loci in FHS
GCKR
ANGPTL3
NLX1PL
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

E.

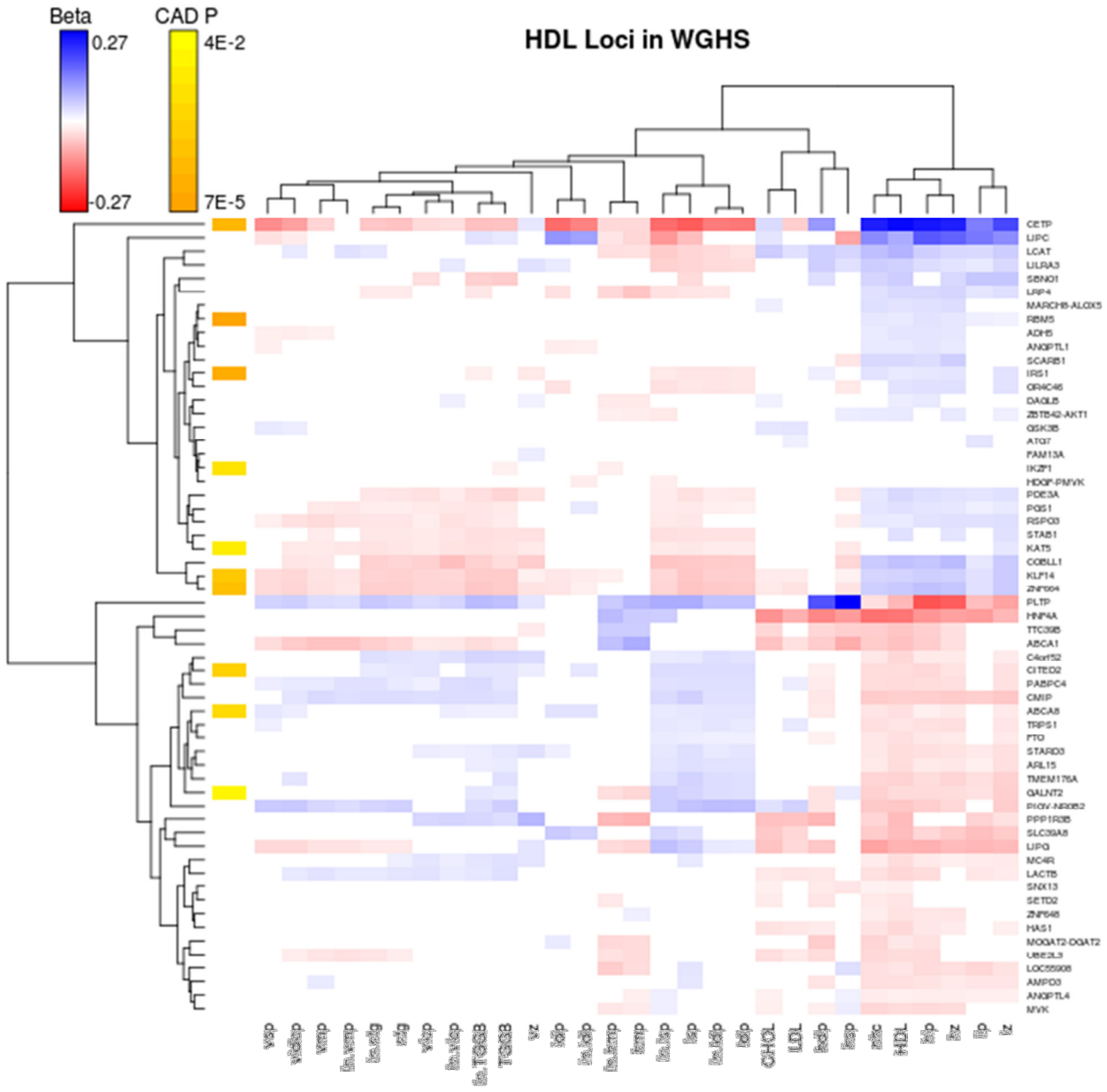


HDLLG: Large particles of high density lipoprotein concentrations determined by NMR, Exam 4
HDLSM: Small particles of high density lipoprotein concentrations determined by NMR, Exam 4
HDLSZ: Weighted average for HDL size based on measurements of HDLP1 through HDLP6, Exam 4
Lapoeser5*: ApoE concentrations 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

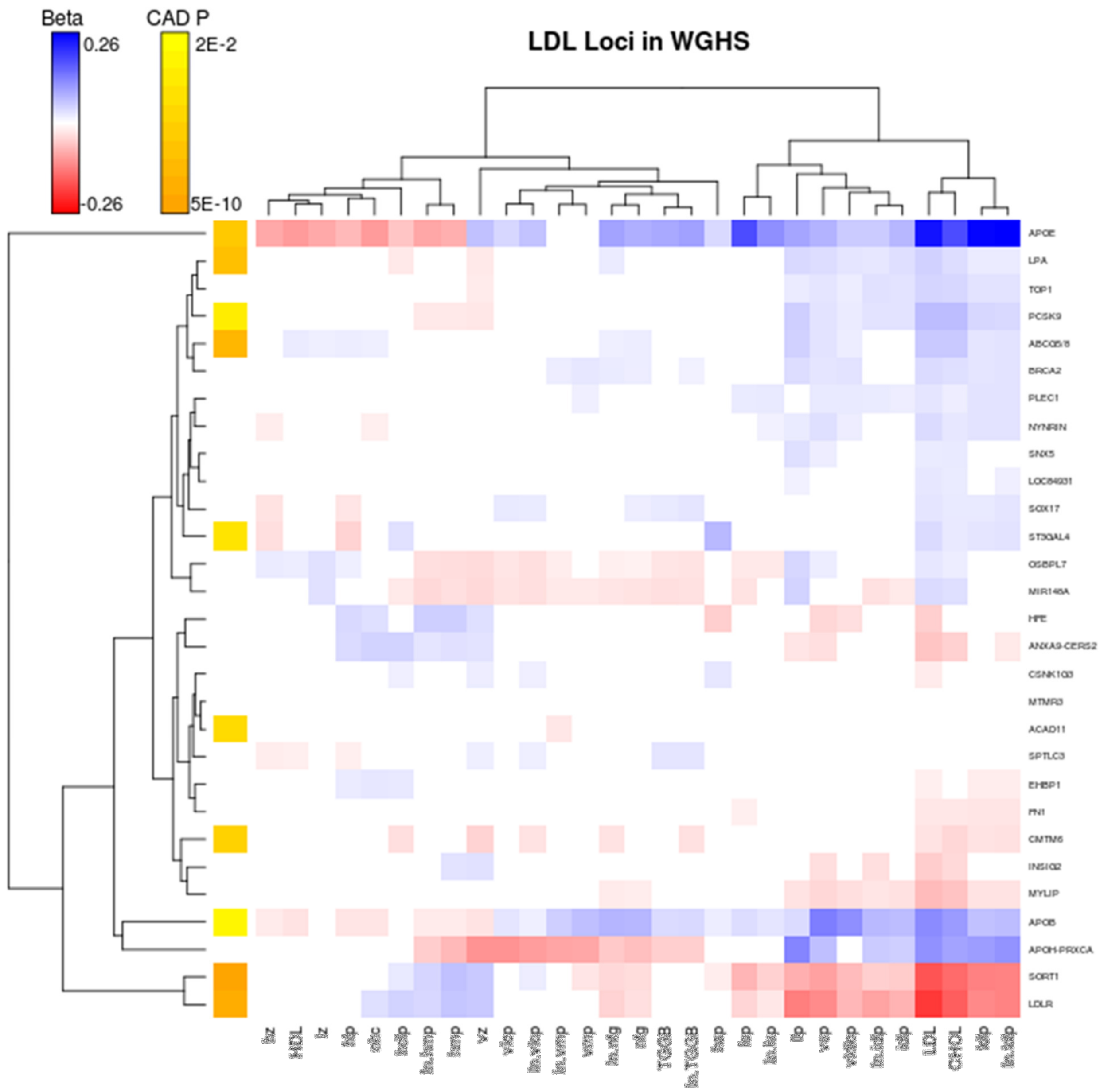
LRLP_C*: Remnant like particles measured using selective immunoseparation of lipoproteins using the Otsuka kit. Expressed as cholesterol in mg/dL, Exam 4
LRLP_tg*: Remnant like particles measured using selective immunoseparation of lipoproteins using the Otsuka kit. Expressed as triglycerides in mg/dL, 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.

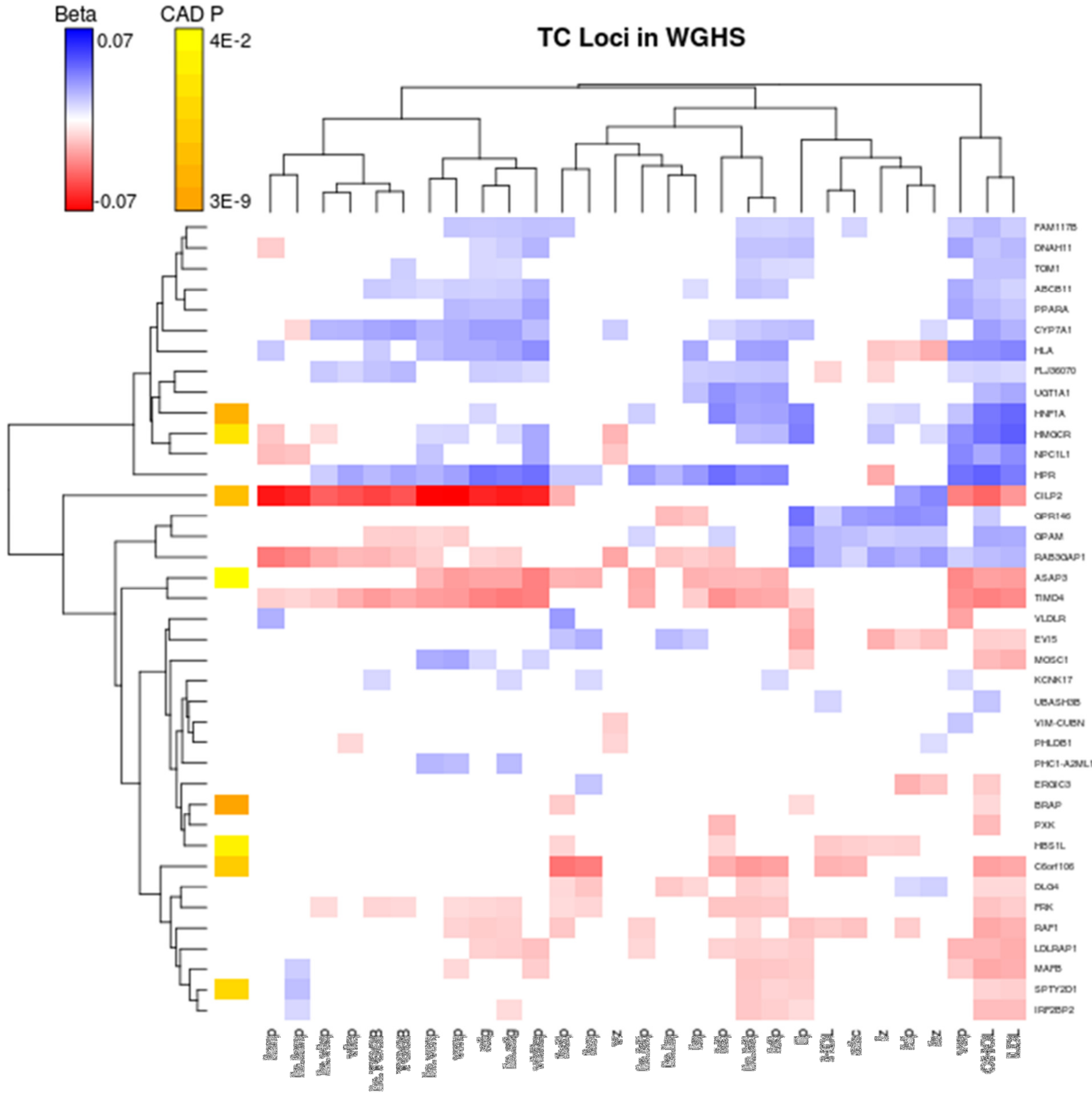
F.


Beta
0.27
-0.27
CAD P
4E-2
7E-5
HDL Loci in WGHS
CETP
LIPC
LCAT
LILRA3
SBNO1
LRP4
MARCH8-ALOX5
RBM5
ADH5
ANGPTL1
SCARB1
IRS1
GRHC46
DAGLB
ZBTB42-AKT1
GSK3B
ATG7
FAM13A
IKZF1
HDGF-PMVK
PDE3A
PGS1
RSPO3
STAB1
KAT5
COBLL1
KLF14
ZNF664
PLTP
HNF4A
TTC39B
ABCA1
C4orf52
CITED2
PABPC4
CMIP
ABCA8
TRPS1
FTO
STARD3
ARL15
TMEM176A
GALNT2
PIGV-NR0B2
PPP1R3B
SLC39A8
LIPG
MC4R
LACTB
SNX13
SETD2
ZNF648
HAS1
MOGAT2-DGAT2
UBE2L3
LOC55908
AMPD3
ANGPTL4
MVK
vsp
vldlp
vsmp
ln.xsmp
ln.sdg
sdg
vlvp
ln.vldp
ln.TGSB
TGSB
vz
ldp
ln.ldp
ln.lsmp
lsmp
ln.lsp
lsp
ln.ldp
ldp
cHCL
LDL
ln.idp
lisp
mic
HDL
hdp
hz
lip
zi

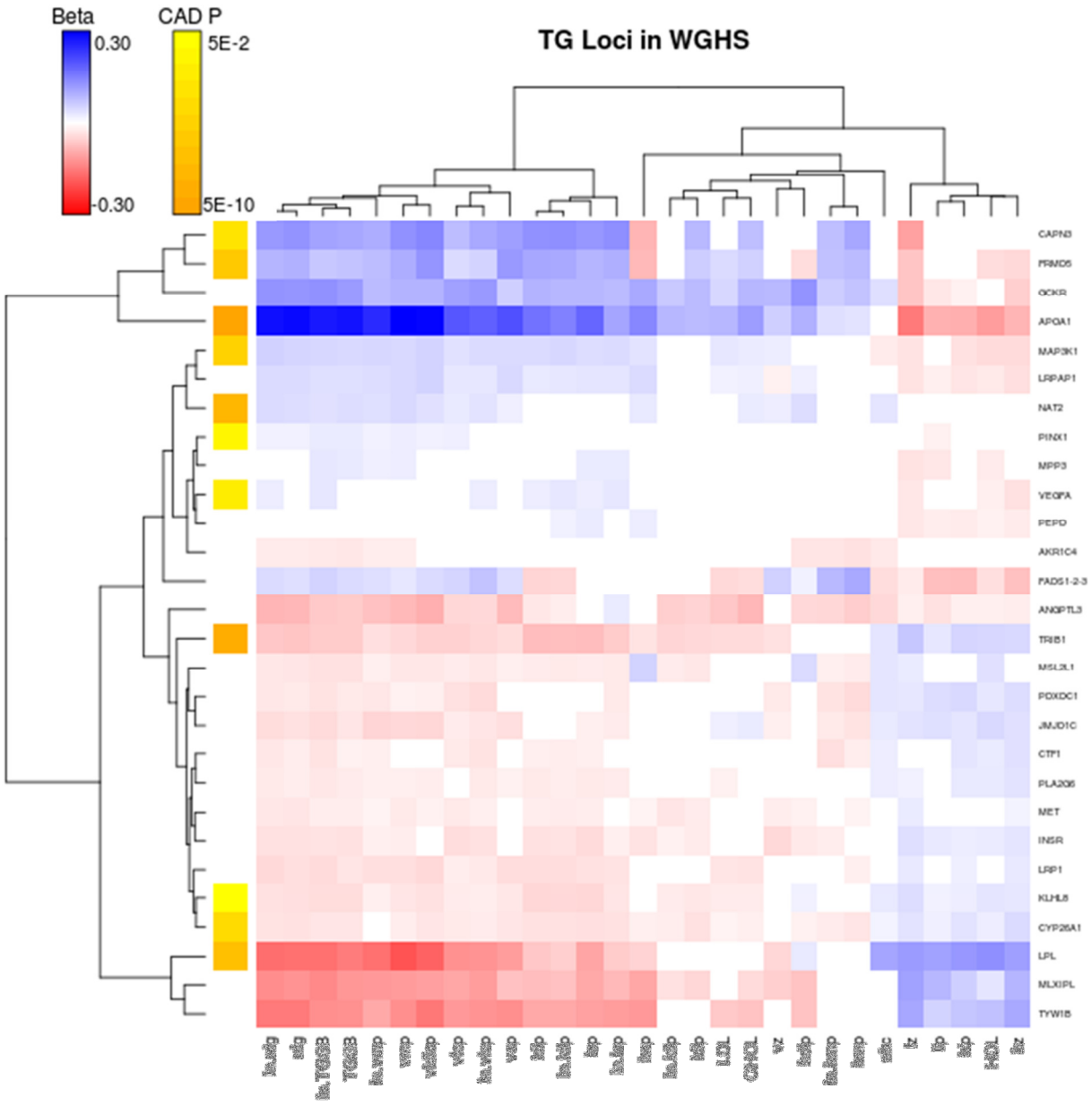
G.


Beta
0.26
-0.26
CAD P
2E-2
5E-10
LDL Loci in WGHS
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-PRKCA
SORT1
LDLR
HDL
LDL
CHOL

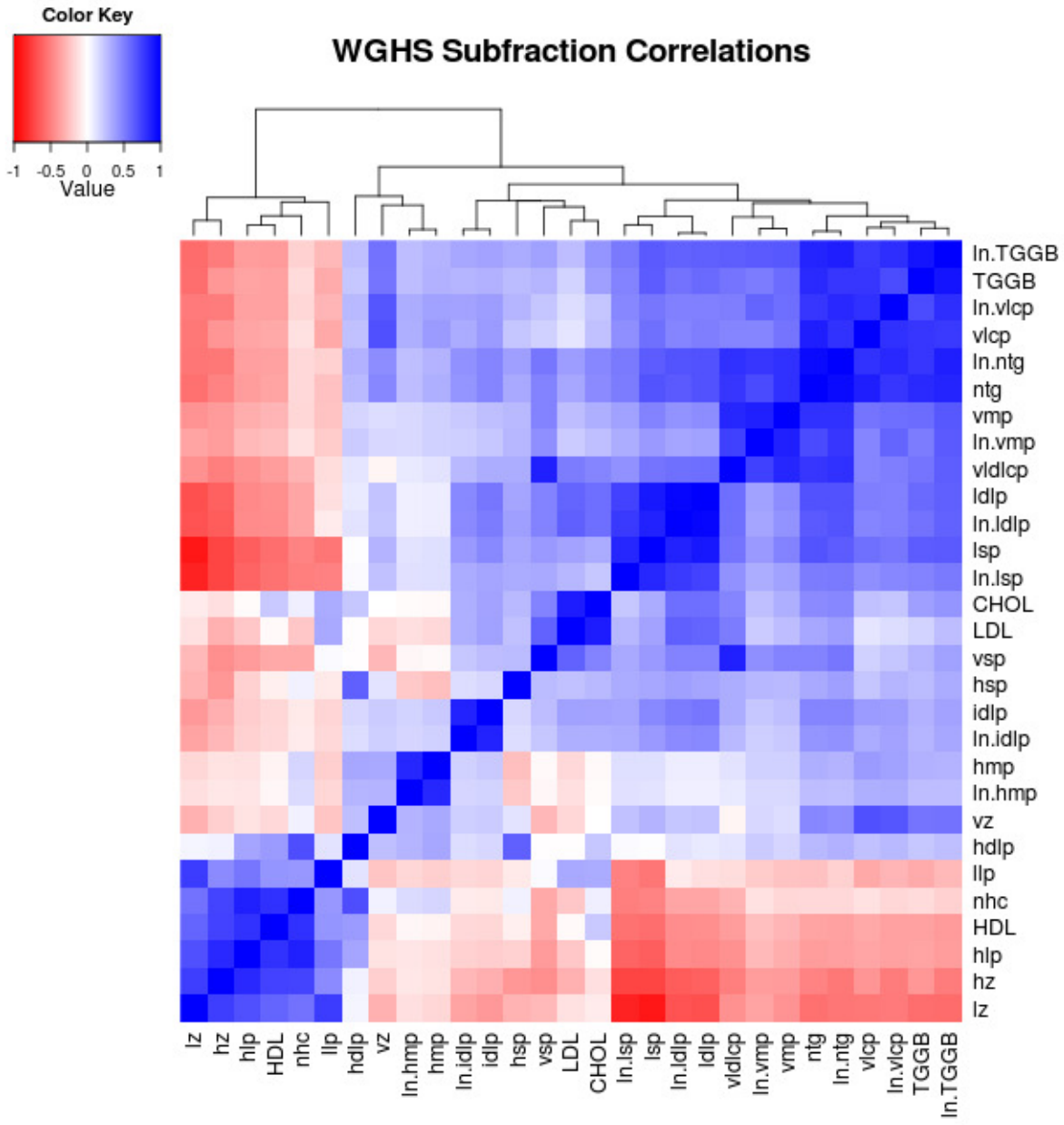
H.


Beta
0.07
-0.07
CAD P
4E-2
3E-9
TC Loci in WGHS
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

I.


Beta
0.30
-0.30
CAD P
5E-2
5E-10
TG Loci in WGHS
CAPN3
FRMD5
CCKR
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

J.



llp: LDL large
lsp: LDL small
ln.lsp: ln [LDL small]
lz: LDL mean size
idlp: IDL total
ln.idlp: ln [IDL total]
ldlp: LDL total
ln.ldlp: ln [LDL total]
LDL: LDL-C assay
hdlp: HDL total
hlp: HDL large

hmp: HDL medium
ln.hmp: ln [HDL medium]
hsp: HDL small
hz: HDL mean size
nhc: HDL-C by NMR
HDL: HDL-C assay
vldlcp: VLDL total
vlcp: VLDL large
ln.vlcp: ln [VLDL large]
vmp: VLDL medium
ln.vmp: ln [VLDL medium]

vsp: VLDL small
vz: VLDL mean size
ntg: TG by NMR
ln.ntg: ln [TG by NMR]
TGGB: TG assay
ln.TGGB: ln[TG assay]
CHOL: Total Cholesterol

Supplementary Note: Candidate Genes at Novel Loci

The list of notable genes in newly identified loci, below, is meant to provide an overview of the diverse set of loci associated with blood lipids in our study. Although the list can provide a starting point for exploration of these loci and help motivate follow-up studies and/or hypotheses, the list should not be considered exhaustive.

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 (Henkel et al. 2011). G6PC2 encodes a glucose-6-phosphatase catalytic subunit (MIM 608058). Variants at this locus have been implicated in liver enzyme and fasting glucose levels (Chambers et al. 2011; Chen et al. 2008).

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) (Ito et al. 2004).

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 (Ha et al. 2008).

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 andpromote 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 (Pare et al. 2009; Danik et al. 2009).

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 (Bisogno et al. 2003).

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) (Kim et al. 2007). SLC2A4 is an insulin-regulated glucose transporter (MIM 138190). The variant identified here was previously associated with alkaline phosphatase levels in plasma (Chambers et al. 2011).

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 (Denis et al. 2011).

***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 (Rohwedder et al. 2012).

***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 (Frayling et al. 2007; Fischer et al. 2009).

***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 growth (Liu et al. 2010). 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) (Wallrapp et al. 1998). Variants at this locus regulate persistence of fetal hemoglobinin adults and other haematological traits (Uda et al. 2008; Soranzo et al. 2009).

***HDGF* (hepatoma derived growth factor) and *PMVK* (phosphomevalonate kinase).**HDGFis 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) (Majumdar et al. 2011).

IKZF1 **(IKAROS family zinc finger 1)** is a transcription factor that regulates the low-density lipoprotein receptor in certain cell types (Loeper et al. 2008).

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 receptor-related 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 (can Beekum et al. 2008).

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 (Cho et al. 2012).

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 hepatocytecell proliferation, migration and survival (MIM 164860) (Yu et al. 2010; Zou et al. 2007).

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) (Zhou et al. 2012; Zheng et al. 2011).

***MOGAT2* (monoacylglycerol O-acyltransferase 2) and DGAT2 (diacylglycerol O-acyltransferase 2).** *MOGAT2* plays a central role in absorption of dietary fat in the small intestine (Cao et al. 2004). 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-1α (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 (Demirkan et al. 2012). About 300 kb away, *PLA2G10* encodes a protein that releases arachidonic acid from cell membrane phospholipids (MIM 603603).

***PEPD* (peptidase D)** encodesan 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 (Cho et al. 2012; Dastani et al. 2012).

***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) (Galliano et al. 2008).

***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) (Zhou et al. 2010).

***PIGV* (phosphatidylinositol glycan anchor biosynthesis, class V) and *NR0B2* (nuclear**

receptor subfamily 0, group B, member 2). PIGV is a mannosyl transferase 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) (Takeuchi et al. 2010).

RBM5 **(RNA binding motif protein 5)** is a 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 (Kazanskaya et al. 2008). Variants in this locus are associated with waist-hip ratio (Heid et al. 2010), bone mineral density (Duncan et al. 2011) and renal traits (Kim et al. 2011).

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) (Koharudin et al. 2009).

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 (Hicks et al. 2009).

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 (Li et al. 2011). Variants at this locus have been associated with waist-hip ratio (Heid et al. 2010).

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 (Ishikawa et al. 2012).

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 (Heid et al. 2010).

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) (Sarria et al. 1992). 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 (Wyne et al. 1996).

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 (Ding et al. 2012).

Chapter 3: Prioritizing Functional Variants in Genetic Association Studies

Introduction

Genome-wide association studies, which examine millions of genetic variants across thousands of individuals, have identified many complex trait associated loci. Most of these loci include many strongly associated variants in linkage disequilibrium with each other and exhibiting similar evidence for association. The large number of variants showing evidence for association in each locus makes it challenging to prioritize likely functional variants.

Information regarding biological plausibility can help prioritize SNPs for follow-up (Minelli et al. 2013). SNPs where annotation suggests a functional role are significantly enriched in loci associated with human diseases (Schaub et al. 2012). Importantly, the SNP most strongly supported by experimental evidence is often not the SNP where association peaks but another nearby SNP in linkage disequilibrium (Schaub et al. 2012). Several types of biological information have been shown to be useful, including impact on coding sequence (Hindorff et al 2009; Schork et al. 2013), impact on gene expression (Nicolae et al. 2010; Lappalainen et al. 2013), and impact on transcription factor binding motifs (Maurano et al. 2012; Trynka and Raychaudhuri 2013). Some of the earliest searches for overlap between association signals and functional annotation focused on eQTLs, which are extremely plentiful (Nicolae et al. 2010). Early enrichment analyses demonstrated strong enrichment of eQTLs near transcription factor binding sites, particularly near transcription start and end sites (Veyrieras et al. 2008). There is also strong enrichment of eQTLs in open chromatin regions and we now know that regulatory

annotation can help prioritize SNPs most likely to drive gene expression variation (Gaffney et al. 2012). Over time, these analyses have become increasingly sophisticated, extending from eQTLs to more complex traits and examining functional annotations specific to individual cell types or tissues (Global Lipids Genetics Consortium 2013; ENCODE Project Consortium 2012).

We define "causal" variants as the functional genetic variants that influence the risk for disease and explain the observed association. Overall, it is now well accepted that the search for causal variants for any trait may be aided by systematically modeling the features they share. For example, causal variants for lipid traits might preferentially overlap transcription factor binding sites active in liver, where important steps in lipid metabolism take place (Ernst et al. 2011). Intuitively, when choosing among two nearby lipid GWAS variants with similar association signals, we expect the one which overlaps a liver transcription factor binding site is more likely to be causal. Here, we set out to develop a method that quantifies the enrichment of particular annotations among the associated variants in a GWAS that is computationally efficient and reliably convergent so as to become a part of routine post GWAS analysis. In this way, we hope to prioritize variants for follow-up in a systematic and quantitative manner.

We propose two methods to study enrichment for GWAS: (i) a simpler approach that seeks causal variants in loci with genome-wide significant evidence for association and (ii) a Bayesian approach that allows for causal variants to reside in loci that do not reach genomewide significance. Our methods work with summary level data (effect sizes and standard errors, or p-values) and thus can be applied conveniently to large samples, including those derived through meta-analysis. Using the enrichment analysis results, our method computes a credible set of likely causal variants (Maller et al. 2012), narrowing the list of variants to be followed-up. Using simulations, we show that our method appropriately controls type I error rates and has

comparable or better power than the published method fGWAS (Pickrell 2014). We demonstrate real data applications of our method using publicly available datasets for lipids (Global Lipids Genetics Consortium 2013) and schizophrenia (Schizophrenia Working Group of the Psychiatric Genomics Consortium 2014). We explore the enrichment of different genomic features such as coding variation, overlap with complement genes, and CADD score (Kircher et al. 2014) in association data for age-related macular degeneration. We use the UK Biobank dataset (Sudlow et al. 2015, Bycroft et al. 2017) to explore the enrichment of eQTL and coding variation among GWAS associated variants and find eQTLs to be significantly enriched in 37 of 45 traits while nonsynonymous variants are significantly enriched in 19 of 45 traits.

Methods

We set out to quantify the relationship between the causal variants for a trait and a genomic feature of interest. When there is a set of similarly associated variants, our method aims to identify the features that would most effectively at separate out truly causal variants. More generally, our method aims to combine association summary statistics and biological feature annotation to prioritize variants for follow-up.

Consider the following contingency table:

Table 3.1: Example of a Contingency Table if Causal Variants are Known

	Annotated	**Not Annotated**
Causal	a	b
Non-causal	c	d

If we knew exactly which variants were causal, an odds ratio derived from this table would represent how likely a variant with annotation is to be causal relative to a variant without the annotation. Unfortunately, we do not know which variants are causal and, instead, expect that

in addition to each causal variant many nearby variants in linkage disequilibrium will show evidence for association in each locus. Thus, we use an iterative model to estimate expected cell counts and make inferences about the importance of a candidate biological feature.

We begin by dividing the variants into loci assuming each locus has at most one causal variant. We then use an initial estimate of the odds ratio to estimate the expected cell counts of the table. Next, we use these expected counts to update our odds ratio estimate and repeat the process until the estimates converge. This iterative algorithm is computationally efficient and can be implemented with only summary level data from single SNP association analysis, namely either effect sizes and standard errors or p-values.

To implement this method, we need to model the multi-SNP association with the trait, link annotation to the model and compute the conditional expected values of the cells given association summary statistics.

Modeling the Association

We begin with observed data from an association study. Let n be the sample size, $\boldsymbol{y}$ the trait vector denoting the trait values for each individual under study and $\boldsymbol{G}$ the $n \times p$ genotype matrix where p is the total number of SNPs in the study. Let there be L causal loci and let n_l be the number of variants in the l^{th} locus and $\boldsymbol{G_l}$ be the corresponding genotype matrix.

We model the trait as follows:

$$\boldsymbol{y} = \beta_0 \mathbf{1} + \sum_{l=1}^{L} (\boldsymbol{G_l Z_l}) \beta_l + \boldsymbol{\epsilon}$$

where $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \boldsymbol{I})$[10] and $\boldsymbol{Z_l}$ is the indicator vector for the l^{th} locus denoting which variant is causal.

To make estimating this model more tractable, we make two simplifying assumptions. First, we assume that there is exactly one causal variant in each causal locus. In practice, there may be loci with multiple signals in which case we use a conditional analysis approach (described later) to model more than one causal variant. Second, we assume that variants from different loci are not in linkage disequilibrium and, thus, that causal variants are not in linkage disequilibrium with each other. These assumptions allow us to conveniently process loci one at a time and derive an approximate solution to the variable selection problem. Since we expect most of the uncertainty about the identity of causal variants to be local, and caused by linkage disequilibrium, we expect this approximation captures and tackles the most interesting features of the data.

Note that our simplifying assumption ensures that each $\boldsymbol{Z_l}$ has exactly one entry as 1 and the rest as 0. The estimates from the summary level data are for single SNP analysis but we can use those to approximate the above multi-SNP model as described in the **Supplementary Methods**.

Linking the Annotations

Let $\boldsymbol{\delta_{li}}$ denote whether the i^{th} variant in the l^{th} locus has the feature or not, and let λ be an underlying parameter used to quantify the enrichment of variants with the feature among causal variants according to the following model:

$$P(\boldsymbol{Z_{li}} = 1 \,|\boldsymbol{\delta_{li}}, \lambda) = \frac{\exp(\mu + \lambda\delta_{li})}{1 + \exp(\mu + \lambda\delta_{li})}$$

which is a prior imposed on the missing data $\boldsymbol{Z}$ which denotes whether a variant is causal or not**.**

λ is our principal parameter of interest and represents how much more likely a variant with the feature is to be causal compared to the other variants in the locus. We estimate this λ defined as the log odds ratio from the contingency table (**Table 3.1**).

Conditional Expectation

Assuming exactly one causal variant per locus, for each SNP in a causal locus, the probability of being causal incorporating the prior based on annotations conditional on λ becomes:

$$P(\boldsymbol{Z_{li}} = 1 \mid \boldsymbol{y}, \boldsymbol{G}, \boldsymbol{\delta}, \lambda) = \frac{P(\boldsymbol{y} \mid \boldsymbol{G}, \boldsymbol{\delta}, \lambda, \boldsymbol{Z_{li}} = 1)\, P(\boldsymbol{Z_{li}} = \boldsymbol{1} \mid \boldsymbol{\delta}, \lambda)}{\sum_{j=1}^{n_l} P(\boldsymbol{y} \mid \boldsymbol{G}, \boldsymbol{\delta}, \lambda, \boldsymbol{Z_{lj}} = 1)\, P\big(\boldsymbol{Z_{lj}} = 1 \,\big|\, \boldsymbol{\delta}, \lambda\big)}$$

Note that since $\boldsymbol{Z}$ takes values 0 or 1, the above conditional probability is also the conditional expectation of $\boldsymbol{Z}$ which denotes the number of causal variants. We use these expected values to fill the cells of the contingency table (**Table 3.1**).

EM Algorithm

To implement this method, we begin by defining loci as described below. We start with an initial estimate of λ and calculate the expected cell counts of the table, which we use to estimate an updated value of λ. We repeat this process till our estimates converge (**Supplementary Methods**). Note that if any of the cell counts in the contingency table (**Table 3.1**) is small, the estimates may fail to converge. This generally occurs if there are too few annotated variants with significant p-values in the associated loci, so we report a failure to converge and expect that there is no enrichment in such cases.

Defining Loci and Modifying the Prior

Having described a simple strategy for calculating the probability that each variant is causal (within a locus) and for estimating our enrichment parameter λ, we now proceed to outline two approaches for dividing the genome into loci and interpreting the evidence for association.

Simpler Approach: Focusing on Loci that Reach Genome-wide Significance

In our first approach, we classify loci into two discrete groups at the beginning of our analysis. One group consists of trait associated loci, the other includes the remaining non-associated loci. We begin with the single-SNP analysis results and divide the genome into loci. Associated loci are regions near SNPs that are significantly associated with the trait (e.g., typically those with p-value $< 5 \times 10^{-8}$). We assume that each associated locus has exactly one causal variant and that background variants not in the associated loci are never causal. This simplification greatly improves computational efficiency, since it allows us to track only a simple count of variants in each category outside associated loci. We make further simplifying assumptions that the average chi-square for the background variants is 1. Suppose that there are b_1 and b_0 background variants with and without the annotation respectively, and L associated loci, we assign the background variants to the associated loci, so that the prior for a variant at each locus becomes

$$P(\boldsymbol{Z_{li}} = 1 \mid \boldsymbol{\delta_l}, \lambda) = \frac{\exp(\lambda \delta_{li})}{\frac{b_0}{L} + \frac{b_1}{L}\exp(\lambda) + \sum_{j=1}^{n_l} \exp(\lambda \delta_{lj})}$$

This ensures that the background variants are accounted for when estimating the enrichment parameter without having to iterate through the association results for the whole genome.

Proceeding one locus at a time, and using the assumption that there is exactly one causal variant per locus, we approximate (**Supplementary Methods**)

$$P(\boldsymbol{y} \mid \boldsymbol{G}, \boldsymbol{\delta}, \lambda, \boldsymbol{Z_l}) = P(\boldsymbol{y} \mid \boldsymbol{G}, \boldsymbol{\delta}, \lambda, \boldsymbol{Z_{li}} = 1) \approx \exp\left(\frac{1}{2}\frac{\hat{\beta}_{li}^2}{\mathrm{se}(\hat{\beta}_{li})^2}\right)$$

where $\boldsymbol{Z_l}: \boldsymbol{Z_{li}} = 1, \boldsymbol{Z_{lj}} = 0 \ \ \forall j \neq i$.

Hence,

$$P(\boldsymbol{Z_{li}} = 1 \mid \boldsymbol{y}, \boldsymbol{G}, \boldsymbol{\delta}, \lambda) \approx \frac{\exp\left(\frac{1}{2}\frac{\hat{\beta}_{li}{}^2}{\mathrm{se}(\hat{\beta}_{li})^2}\right)\exp(\lambda\delta_{li})}{\sum_{j=1}^{n_l}\exp\left(\frac{1}{2}\frac{\hat{\beta}_{lj}{}^2}{\mathrm{se}(\hat{\beta}_{lj})^2}\right)\exp(\lambda\delta_{lj}) + \left(\frac{b_0}{L} + \frac{b_1}{L}\exp(\lambda)\right)\exp\left(\frac{1}{2}\right)}$$

Here we assumed that the χ^2 test statistic T_{li} is

$$T_{li} = \left(\frac{\hat{\beta}_{li}}{se\ (\hat{\beta}_{li})}\right)^2$$

and thus, we can approximate

$$T_{li} \approx \ \mathrm{CHIDIST}^{-1}(P_{li}, \mathrm{df} = 1)$$

where P_{li} is the p-value for the i^{th} marker in the l^{th} locus and CHIDIST returns the one-tailed probability of the χ^2 distribution function. We can therefore invert the p-values to get the corresponding χ^2 test statistic, and implement our method even if the effect sizes and standard errors are not available.

The likelihood contribution of one locus l is

$$\boldsymbol{L_l}(\lambda) \approx \sum_i P(\boldsymbol{y} \mid \boldsymbol{G}, \boldsymbol{\delta}, \lambda, \boldsymbol{Z_{li}} = 1)\, P(\boldsymbol{Z_{li}} = 1 \mid \boldsymbol{\delta_l}, \lambda)$$

$$+ \exp\left(\frac{1}{2}\right)\left(\frac{b_0}{L} + \frac{b_1}{L}\exp(\lambda)\right)\frac{1}{\frac{b_0}{L} + \frac{b_1}{L}\exp(\lambda) + \sum_{j=1}^{n_l}\exp(\lambda\delta_{lj})}$$

$$= \left[\sum_{i=1}^{n_l} \exp\left(\frac{1}{2}\frac{\hat{\beta}_{li}{}^2}{\mathrm{se}(\hat{\beta}_{li})^2}\right) \exp(\lambda\delta_{li}) \right.$$

$$\left. + \exp\left(\frac{1}{2}\right)\left(\frac{b_0}{L} + \frac{b_1}{L}\exp(\lambda)\right)\right] \frac{1}{\frac{b_0}{L} + \frac{b_1}{L}\exp(\lambda) + \sum_{j=1}^{n_l} \exp(\lambda\delta_{lj})}$$

We assume that the loci are independent and approximate the joint likelihood by taking a product across loci as

$$\boldsymbol{L}(\lambda) \approx \prod_{\boldsymbol{l}} \boldsymbol{L_l}(\lambda)$$

An Alternative Bayesian Approach:

For the Bayesian approach, we divide the variants across the whole genome into loci based on LD patterns (Berisa and Pickrell 2016; Pickrell 2014), and assume that each locus l has n_l variants and a probability π_l of being causal. We assume that the loci are defined such that they contain at most one causal variant. Then,

$$\pi_l = 1 - \prod_{i=1}^{n_l} \frac{1}{1 + \exp(\mu + \lambda\delta_{li})}$$

which is obtained by calculating the probability – based on the parameters μ and λ – that none of the variants in a locus is causal, and then taking the complement of that. For this approach, we iterate through the association results for the whole genome to estimate both λ and μ.

The corresponding conditional probability is (**Supplementary Methods**):

$$P(\boldsymbol{Z_{li}} = 1 \,|\boldsymbol{y}, \boldsymbol{G}, \boldsymbol{\delta}, \lambda) = \left[\frac{\sum_k \exp(\mu + \lambda\delta_{lk})\ BF_{lj}}{1 + \sum_k \exp(\mu + \lambda\delta_{lk})\ BF_{lk}}\right]\left[\frac{\exp(\lambda\delta_{li})\, BF_{li}}{\sum_k \exp(\lambda\delta_{lk}) BF_{lk}}\right]$$

where BF_{ij} represents the Bayes Factor for the j^{th} SNP at the i^{th} locus and can be computed using summary level data based on single-SNP association analysis (Wen 2014):

$$BF_{li} = \left(1 + \frac{\phi^2}{\text{se}\left(\hat{\beta}_{li}\right)^2}\right)^{1/2} \exp\left(\frac{1}{2}\frac{\phi^2}{\text{se}\left(\hat{\beta}_{li}\right)^2\left(\text{se}\left(\hat{\beta}_{li}\right)^2 + \phi^2\right)}\right)\hat{\beta}_{li}^{\ 2}$$

where we assume a $N(0, \phi^2)$ prior on β_{li} and in practice, average over a range of values of ϕ^2.

The joint likelihood then becomes (**Supplementary Methods**):

$$P(\boldsymbol{y} \mid \boldsymbol{G}, \boldsymbol{\delta}, \mu, \lambda) \propto \prod_{l \in L_1} \pi_l \left[\sum_i BF_{li} \exp\left(\mu + \lambda\delta_{li}\right) + 1\right]$$

Prioritizing Variants for Follow-up

We use a Likelihood Ratio to test whether the enrichment parameter λ is significantly different from null (**Supplementary Methods**). For each method, we use the estimated λ and the corresponding log likelihoods to get the test statistic

$$\Lambda = 2(\boldsymbol{LLK}(\hat{\lambda}) - \boldsymbol{LLK}(0))$$

which follows a χ^2distribution with 1 degree of freedom.

If significant ($p < 0.05$), we use the estimated value of λ to calculate the posterior probability that each SNP in an associated locus is causal. The 95% credible set for a locus is the smallest set of variants in that locus whose posterior probabilities sum up to $\geq 95\%$ (Maller et al. 2012). Thus, we calculate the 95% Bayesian credible set for each locus to get a list of variants most likely to be causal. These credible sets can be used to prioritize variants for follow-up.

Loci with Multiple Causal Variants

The method described thus far assumes that each causal locus has exactly one true causal variant and that variants from different loci are not in linkage disequilibrium. In practice, conditional analyses of GWAS results often demonstrate that there are multiple causal variants in a locus (Hormozdiari et al. 2014). To enable us to prioritize causal variants when there are overlapping loci, we use conditional analysis. Specifically, when there are multiple nearby independent signals, we first define a super-locus including all variants near these signals. Then, we define a series of pseudo-loci by conditioning association statistics in turn on all but one of the top independent signals in the region. For example, if conditional analysis indicates 3 distinct association signals in one locus, we define 3 pseudo-loci for that locus. Each pseudo locus corresponds to one of the distinct association signals and uses association results obtained after conditioning on the top variants from the other two signals. This is, admittedly, a rather ad-hoc approach to approximate the multi-SNP association model that would be required. Observe that in such a scenario, summary level data based on single-SNP analysis can still be used when a method such as GCTA (Yang et al. 2011) is used to carry out approximate conditional analyses and define the pseudo-loci.

Application to Multiple Traits

Phenome-wide Association Studies (PheWAS) involve evaluating the association of a single genetic marker with multiple phenotypes (Ye et al. 2015). Decreases in genotyping costs have led to large biobanks genotyping all participants making it feasible to conduct PheWAS on a genome-wide scale (Bush et al. 2016). Electronic Health Records (EHR) and survey questionnaires are leveraged to construct thousands of phenotypes. For example, in the Michigan Genomics Initiative (Schmidt et al. 2017), International Classification of Diseases (ICD) codes

(https://www.cdc.gov/nchs/icd/icd9.htm) are used to track 8,940 phenotypes derived from distinct ICD-9 codes in the data. In an initial analysis of 21,241 individuals, Schmidt et al (2017) report 145 associated loci across 131 traits. We reason that some genomic features may be enriched in causal loci for all traits and thus applied our approach for across traits. To accommodate multiple traits, we defined 41 causal loci associated with at least one of the phenotypes under study. When causal loci for two or more phenotypes overlapped, we kept the phenotype with the more significant association. The rest of the algorithm works as described above.

Simulation

We use simulations based on real genotype data from a study on age-related macular degeneration (AMD) (Fritsche et al. 2016) to validate our method by ensuring that Type I error is controlled and investigate the power to detect different values of the enrichment parameter. The original study includes 17,832 European controls genotyped at 439,350 variants. To generate simulated GWAS datasets, we sampled 5,000 European ancestry individuals and a set of 3,000 loci, each with 100 variants. The 3,000 loci are selected so that they are at least10 Mb apart, so that a pair of variants from two different loci is not likely to be in linkage disequilibrium with each other.

For the baseline model, we assume that variants possess a feature of interest with probability 0.1 and that the value of λ used to determine the causal variant is 2.5 (odds ratio = 12.2). We use the parametric model

$$P(\boldsymbol{Z_{li}} = 1 \mid \boldsymbol{\delta_{li}}, \lambda) = \frac{\exp(\mu + \lambda \boldsymbol{\delta_{li}})}{1 + \exp(\mu + \lambda \boldsymbol{\delta_{li}})}$$

to determine which variants are causal, using $\mu = -8$ which generates approximately 100 causal loci under the null.

We vary the parameters to observe the behavior of our method under different scenarios. We also use the simulated data to compare our method with the published method fGWAS (Pickrell 2014).

Real Data Application

Age-related Macular Degeneration

We apply our method on sequencing data based on case-control study of age-related macular degeneration (AMD) with a sample size of 4,787 individuals with 2,394 cases and 2,393 controls. Studies have shown that up 70% of AMD risk can be attributed to genetic variation (Seddon et al. 2005). Previous association studies have found up to 52 signals in 34 different loci (Fritsche et al. 2016). The sequencing study samples were matched based on age and sex and restricted to European ancestry using LASER (Wang et al. 2015). The sequencing study involved 45.4 million variants most of which were intergenic and very rare, and led to the detection of 9 signals in 4 distinct loci *CFH*, *C2/CFB/SK1V2L*, *C3* and *ARMS2/HTRA1*. Most notably, the *CFH* locus has multiple signals as well as variants with very high odds ratios. We construct 9 pseudo-loci using conditional analysis results from Firth-adjusted logistic regression analysis. We use our method to do enrichment analysis on these data considering different genomic annotations – whether a variant is (i) non-synonymous, (ii) non-synonymous or frameshift, (iii) overlaps with a complement gene, or (iv) has a CADD score greater than 20 (Kircher et al. 2014).

Publicly Available GWAS Datasets: Lipids and Schizophrenia

We also implement our method on publicly available data-sets for a variety of traits including lipids (Global Lipids Genetics Consortium 2013) and schizophrenia (Schizophrenia Working Group of the Psychiatric Genomics Consortium 2014). We use the published lists of associated variants to define our associated loci to investigate the enrichment of non-synonymous variants.

Michigan Genomics Initiative

The Michigan Genomics Initiative (MGI) (https://www.michigangenomics.org/) is a large data repository where phenotype information is collected using both Electronic Health Records and questionnaires (Schmidt et al. 2017). Based on the February 2016 data freeze, 1,448 traits based on PheWAS codes (**Supplementary Figure S3.1**) are analyzed for a sample size of 21,241 participants.

We defined significant loci based on the GWAS signals for all PheWAS traits. We defined 1Mb loci (500kb on either side) from the top signals (p-value < 5x10 -9 to adjust for multiple testing as we considered multiple traits simultaneously), resulting in 41 significant loci. A handful of overlapping loci were merged as required, with the more significant trait's statistics being used.

The saddlepoint approximation (Dey et al. 2017) test calculates only the association p-values for each variant, which were inverted to approximate the χ^2 test statistics. We tested for enrichment of nonsynonymous variants in the trait associated loci.

We obtained eQTL data from 44 human tissues collected by GTEx (http://www.gtexportal.org/home/), and ranked the eGenes by first q-value and then effect size (GTEx Consortium 2015). Considering the top 5,000 eGenes for each of the 44 tissues, we

annotated whether variants were one of the eQTL SNPs on the GTEx list. We then used these to estimate the enrichment parameter and the posterior probabilities. Then we used these posterior probabilities to compute 95% credible sets at each locus (**Supplementary Figure S3.2**).

UK Biobank

We estimated enrichment of non-synonymous variants in GWAS results from UK Biobank data (Sudlow et al. 2015, Bycroft et al. 2017) of 408,961 white British European ancestry samples. Association results for 989 ICD-10 derived phenotypes were obtained using SAIGE (Zhou et al. 2017) and used to find genome-wide significant loci. 45 traits were found with at least 9 genome-wide significant loci. Non-synonymous variants were annotated using VEP (McLaren et al. 2016) and tested for enrichment in these traits with at least 9 genome-wide significant loci. Additionally, we investigated enrichment of eQTL SNPs for 44 tissues obtained from GTEx (GTEx Consortium 2015; http://www.gtexportal.org/home/) in the associated loci for the 45 traits.

Results

Simulation

Simulation results show that power to detect enrichment increases with sample size as well as with number of associated loci (**Figure 3.1**). As number of associated loci increases, we get a more accurate estimate of the enrichment parameter when combining information across many loci. With smaller sample sizes, small effect sizes are difficult to detect leading to fewer causal loci being detected and wider bounds on λ estimates.

Additionally, stronger enrichments are easier to detect and hence, the higher the true value of λ, the greater the power (**Figure 3.2**). For example, with sample size 5,000, as

simulating λ increases from 1 to 2 (odds ratio increases from 2.7 to 7.4), empirical power increases from 45.5% to 97.8% in the simpler approach and from 61.6% to 100% in the Bayesian approach. For simulation purposes, we do not consider the scenario where λ is less than zero as significant depletion of a rare annotation among associated variants is extremely difficult to detect (consider, for example, that if ~1% of the variants are coding, detecting a deficit of coding variants among disease associated loci might require 100s or 1000s of independent association signals).

We observe that the Bayesian approach has slightly greater power than the simpler binary classification approach. However, for large datasets we recommend using the binary approach as it is much faster. While the simpler approach controls the Type I error at 0.05, the Bayesian approach is over-conservative near the null (Type I error = 0.01) due to the approximations made when estimating μ.

Considering the estimation aspect of our method, observe that a theoretical lower bound of the standard error of the estimate can be calculated from the contingency table set-up. The standard error of the true log-odds from the contingency table is:

$$\text{se}(\log \text{OR}) = \sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}$$

which is dominated by the term $\frac{1}{a}$ as we expect the first cell of the contingency table (**Table 3.1**) to have the smallest cell count. Thus, if we assume that there are fewer than 100 causal loci, the standard error of the log-odds is greater than 0.1, which puts a minimum bound on the confidence interval length. Since we do not know the actual cell values of the table, the estimate we use is more uncertain and has a wider confidence interval than the true log-odds ratio. Thus, increasing sample size or varying other parameters does not allow us to improve the

standard error beyond this limit. Simulations show that our estimated confidence intervals are very similar to those obtained when the true causal variants are known, although they are slightly larger as expected (**Table 3.2**) due to the uncertainty involved in identifying associated loci.

As expected, the confidence interval length decreases as the value of the true underlying enrichment parameter increases (e.g. confidence interval length decreases from 1.40 to 0.61 as λ increases from 0 to 2.5 (**Table 3.2**)).

We calculate the coverage probability of our estimate, that is, the proportion of times the simulation λ lies in the confidence interval of the estimate and observe that in our method it lies in the 95% confidence interval at least 95% of the time (range 95% - 99%) (**Figure 3.3**). We observe that the confidence interval also always contains the true observed log-odds ratio, that is, the log-odds ratio calculated based on the causal variants used in simulation.

We compute credible sets of the causal variants. In the simpler approach, we calculate the posterior probability of being causal for each variant in the associated loci. Then, for each associated locus, the 95% credible set of variants is the minimal set whose posterior probabilities sum up to be equal to or more than 0.95. In the Bayesian approach, we do the same for every locus after removing loci where the total posterior probabilities are small .

We compute credible sets using both a flat prior and our method, and observe how the credible set size decreases when taking enrichment due to genomic features into account (**Table 3.3**). Note that in our simple simulation scenario, we get consistently smaller credible sets than under the null (that is, assuming no enrichment), and, as expected, the decrease in credible set size increases as enrichment increases (credible set size decrease of 12.7% for $\lambda = 2.5$).

We compare our method with the published method fGWAS and observe that our method tends to be more stable when λ values are very high or very close to 0. It is difficult to compare

estimates directly since the fGWAS approach gives estimates at two levels – a variant level enrichment parameter and a locus level enrichment parameter – but at the more extreme values of λ our method generates more reasonable confidence intervals than fGWAS. The power of the simpler approach is very similar to fGWAS, with the Bayesian approach having slightly more power (**Figure 3.4**).

Real Data Application

Age-Related Macular Degeneration

We use our method to analyze AMD sequencing data with a sample of 4,806 individuals. We construct 9 pseudo-loci based on conditional analysis as well as single SNP analysis results. For a locus with multiple signals, we obtain the pseudo loci by conditioning in turn by all top independent signals but one. For example, in a locus with 3 independent signals, we obtain 3 pseudo loci where all the variants in the locus are analyzed after conditioning on each pair of the 3 signals.

We expect annotating non-synonymous variants (estimated log-odds ratio = 3.05; odds ratio = 21.1) to lead to a high enrichment parameter, but observe that dichotomized CADD score (estimated log odds ratio = 5.54; odds ratio = 254.7), overlap with complement gene (estimated log odds ratio = 5.17; odds ratio = 175.9) or frameshift variations along with nonsynonymous variants (estimated log odds ratio = 5.14; odds ratio = 170.7) are estimated to be more enriched in the associated loci (**Table 3.4**).

We construct 95% credible sets at each locus based on the different enrichment parameters, and compare them to credible sets constructed assuming no enrichment (**Table 3.5**). Observe that while most of the credible sets are reduced in size to some extent compared to the no enrichment scenario, there are 5 loci where the credible set is reduced to exactly 1 variant for

at least one of the four annotations considered. We observe that the credible sets are reduced for at least seven of the nine loci for all four traits (**Table 3.5**). While there is considerable reduction in credible set size (for example, total credible set size across all 9 loci reduces from 13,669 to 640 when considering annotation with CADD score (**Table 3.5**)), it is mostly driven by one or two loci with hundreds or thousands of variants in the null credible set. We highlight some of the variants that are in these reduced sets (**Table 3.6**) which have at least 95% posterior probability for one of the four annotations considered, and note that all of them are present in credible sets for at least 2 kinds of annotations.

To illustrate our method, we focus on loci 8 and 9 in particular, which are pseudo-loci obtained using conditional analysis on chromosome 19 (**Figure 3.5**). There are several variants in high linkage disequilibrium with the top SNP, but considering the annotations, 3 variants stand out. Note that without using conditional analysis results, the variant chr19:6718146 would not be in any credible set as the credible sets would be dominated by the signal in locus 8 .

Publicly Available GWAS Datasets: Lipids and Schizophrenia

We do similar analyses for association results from lipids and schizophrenia to test for enrichment of non-synonymous variants (**Table 3.7**). While non-synonymous variants are found significantly enriched for HDL-cholesterol, LDL-cholesterol and triglycerides, the enrichment is not statistically significant for total cholesterol or schizophrenia. Using the estimated enrichment parameter to compute 95% credible sets leads to a reduction in the number of variants to potentially follow-up on. For example for HDL-cholesterol, there is a 16% reduction in credible set size as number of variants in 95% credible sets across all loci reduces from 547 to 452 when the estimated enrichment parameter is used (**Table 3.7**).

Michigan Genomics Initiative

We defined 41 loci based on association results from 38 traits. Nonsynonymous variants were found to be significantly enriched in trait-associated loci with an estimated enrichment parameter of 3.68 (odds ratio = 39.6; p-value = 2×10^{-7}). We computed the 95% credible set at each locus, and found that the credible sets at most loci were very similar to the 95% null credible sets computed assuming no enrichment of nonsynonymous variants. We observed that 21 of the loci had fewer than 5 variants in the credible sets in both cases.

We tested for enrichment in overlap with eQTL SNPs across all traits and obtained an estimated enrichment parameter of 4.41 (odds ratio = 82.3; p-value = 3×10^{-8}). We observed that 24 of the 41 loci had 5 or fewer variants in their 95% credible sets and 9 of those loci had exactly one variant in their 95% credible sets. We note that, for a pair of loci, using the estimated enrichment parameter decreases the credible sets from 46 variants to 1 variant and from 23 variants to 11 variants respectively.

Table 3.8 shows the annotated variants which have a posterior probability of $\geq 10\%$ and are present in the 95% credible sets for the 41 associated loci. We observed some biologically-plausible connections such as a variant in the credible set for 'disorders of lipid metabolism' annotated in liver, and a variant in the credible set for 'Skin cancer' annotated for 'skin - sun exposed lower leg'. However, not all the connections were obvious to us, so we considered estimating the enrichment for each eQTL tissue separately. However, since there are only 3,551 annotated variants for all 44 tissues in the 41 loci, considering only one tissue dropped this number to less than 150, and we often end up with fewer than 5 annotated variants in the credible sets. This leads to unstable estimates and inflated standard errors.

UK Biobank

We used association results from SAIGE (Zhou et al. 2017) to find trait-associated loci for 989 traits. We found 45 traits with at least 9 distinct genome-wide significant loci where loci were defined using both physical distance (100kb on either side) and p-value cut-off ($p < 5x10^{-8}$) so that loci extend 100kb on either side of the peaks. Nonsynonymous variants were found to be significantly enriched in 19 traits as detailed in **Table 3.9**. Additionally, we found nominally significant enrichment of eQTL SNPs in 37 of the 45 traits (**Table 3.10**).

Discussion

Recent work from various consortia (ENCODE Project Consortium 2012, Bernstein et al. 2010) has led to detailed mappings of the genomic regulatory regions and the functional properties therein. We attempt to integrate this knowledge with GWAS study results to provide a deeper insight into potentially causal variants. Our method works with summary level data, namely effect sizes and standard errors or p-values, and thus can be used with already published GWAS data.

There are some existing methods which have similar aims(Pickrell 2014; Kichaev et al. 2014). However, our method uses a multi-SNP model based on single-SNP analysis results and provides an easy to interpret enrichment parameter estimate. Simulations show that our method is more stable than existing methods when the true values are near zero. The algorithm is fast and efficient as it takes advantage of the fact that most SNPs in the genome are not associated with the trait of interest. We then use the estimated enrichment parameter to construct credible sets of SNPs prioritized for follow-up.

The summary statistics from any association study are sufficient to implement our method. This makes it easy to apply our method to previously published GWAS to compare how

different genomic features affect various traits. Another advantage is that we can easily use our method for meta-analysis data. Meta-analysis is a popular method to increase sample size by combining studies to get improved power. We can study enrichment easily in this combined sample as individual level data is not required.

Our method helps us interpret GWAS results in a systematic manner. We construct credible sets for each locus which narrows down the set of potential causal variants there. For example, in the AMD data, the total number of variants in credible sets for all the loci goes down from 13,669 to 640 when considering annotation with CADD score (**Table 3.5**). However, note that this decrease is mainly driven by one locus where the credible set sizes decreases from 13,425 to 591 and there is much smaller reduction in credible set sizes for the other loci.

Loci can be defined based on linkage disequilibrium or distance from top associated SNPs. Loci based on distance is simpler to implement but loci based on linkage disequilibrium may be better for refining the signals as they can allow for unequal sized loci. For loci with multiple signals, we recommend using conditional analysis results to define them as multiple pseudo-loci. This enables us to get credible sets for each independent signal which are not dominated by stronger signals nearby.

We recommend using the simpler binary classification approach in most cases for faster results as most GWAS have a large number of variants and comparatively fewer associated loci. However, for better power the Bayesian approach is preferred in situations where the sample size or the number of associated variants is relatively low.

We have shown that our method works well to estimate the enrichment parameter. However, our method does not work well in situations where the SNPs annotated with the feature of interest are likely to be depleted in the causal loci. This is because the features of

interest tend to be rare, and in case of depletion, we may end up with too few annotated variants in the associated loci to get stable estimates. Our method also requires either a sufficient number of causal loci (ideally >25), or loci with high effect sizes to achieve sufficient power.

While both methods described are easy to implement and quite effective, there are several directions in which it can be extended. We may wish to consider quantitative annotations, or categorical annotations with more than 2 levels. In our analysis, we dichotomized the CADD score, but using all the different levels available may lead to more informative results. Another option is to consider multiple genomic annotations simultaneously, which may or may not be correlated. We hope that our method can form the basis to develop tools which help us statistically quantify the relation between genomic features and phenotypes, and lead to a better understanding of the biological mechanisms behind complex genetic traits.

Table 3.2: Estimated Confidence Interval Lengths for the Enrichment Parameter

Simulating λ	*Mean estimated CI length*	*SD of estimated CI length*	*Actual observed CI length*	*SD of actual observed CI length*
0	1.40	0.17	1.36	0.25
0.5	1.18	0.10	1.07	0.12
1	1.00	0.08	0.87	0.07
1.5	0.84	0.05	0.72	0.04
2	0.71	0.03	0.62	0.02
2.5	0.61	0.02		

Confidence intervals reported are calculated using the Bayesian approach. Mean and SD calculated empirically based on 500 simulations for each parameter. Actual observed CI refers to CI based on using the known causal variants used in simulation.

Table 3.3: Decrease in Credible Set Size using Enrichment Parameter

Simulation λ	*Mean Credible Set Size with Estimate*	*Mean Credible Set Size under Null*	*Mean Decrease in Credible Set Size (SD)*
0	1099.59	1109.87	10.28 (15.42)
0.5	1128.34	1139.62	11.28 (13.06)
1	1132.18	1154.08	21.90 (18.73)
1.5	1129.53	1178.23	48.70 (30.11)
2	1113.07	1202.94	89.86 (44.31)
2.5	1012.76	1159.78	147.02 (61.98)

95% credible sets are calculated using the enrichment parameter estimated. Results are shown based on enrichment parameter for simpler approach, but Bayesian approach results similar.

Table 3.4: Estimated enrichment parameter in AMD data for different genomic features

Annotation	*Estimated λ*	*Odds Ratio*	*P-value*
Nonsynonymous	3.05	21.1	1×10^{-3}
Nonsyn + Frameshift	5.14	170.7	4×10^{-5}
CADD score	5.54	254.7	1×10^{-4}
Complement Genes	5.17	175.9	2×10^{-4}

Enrichment parameters estimated for different genomic features based on 9 associated 'pseudo-loci' in association results for age-related macular degeneration. Odds ratio $= \exp(\lambda)$.

Table 3.5: Credible Set Sizes Based on Different Genomic Features for AMD Data

Top SNP at Locus	*Null*	*Nonsyn*	*NFS*	*CADD*	*Complement*
chr1:196684392	20	19	20	19	19
chr1:196661505	90	53	62	9	88
chr1:196024122	13,425	4,537	7,483	591	1,402
chr1:196358288	6	5	30	1	67
chr6:31894355	2	2	2	1	1
chr6:32609038	111	27	111	11	27
chr10:124214600	11	8	2	1	11
chr19:6718387	2	1	1	1	2
chr19:6718146	2	1	1	6	1

Credible sets constructed based on estimated enrichment parameter at 9 'pseudo-loci' for AMD sequencing data. Null denotes credible sets constructed assuming no enrichment. Annotations are: Nonsyn = nonsynonymous; NFS = nonsynonymous or frameshift; CADD = CADD score > 20; Complement = nearest gene is a complement gene.

Table 3.6: Some Highlighted Variants in the Associated Loci for AMD Data

Locus	*Top SNP*	*Annotated*	*Credible Set*	*Posterior Probability*			
				Nonsyn	**NFS**	**CADD**	**Compl.**
4	chr1:196407807	CADD	All 4	0.16	0.15	0.97	0.04
5	chr6:31930462	CADD	Nonsyn, NFS, Compl.	0.79	1	0.99	0.02
5	chr6:31894355	Compl.	Nonsyn, NFS, Compl.	0.21	0.2	0.001	0.98
7	chr10:124214600	CADD	All 4	0.11	0.02	0.99	0.28
7	chr10:124214448	Nonsyn, NFS	Nonsyn, NFS, Compl.	0.65	0.94	0.001	0.08
8	chr19:6718387	Nonsyn, NFS, Compl.	Nonsyn, NFS, Compl.	0.99	0.99	0.04	0.90
8	chr19:6717655	CADD, Compl.	CADD, Compl.	0.004	0.0005	0.95	0.08
9	chr19:6718146	Nonsyn, NFS, Compl.	All 4	0.99	0.99	0.51	0.99

Highlighted variants selected such that posterior probability >95% for at least one annotation. Annotated denotes what features the top variant has; Credible Set denotes which credible sets the variant is present in; Posterior Probability is the estimated posterior probability each enrichment parameter; Nonsyn = nonsynonymous; NFS = nonsynonymous or frameshift; CADD = CADD score > 20; Compl. = nearest gene is a complement gene

Table 3.7: Estimated Enrichment Parameter for Nonsynonymous Variants in Publicly Available Datasets

Trait	*Estimated λ*	*Odds Ratio*	*P-value*	*Credible set size*	*Null credible set*
HDL	2.67	14.4	1.2×10^{-9}	452	547
LDL	2.85	17.3	2.0×10^{-5}	132	289
TC	1.50	4.5	0.53	-	788
TG	2.04	7.7	0.03	271	312
SCZ2	0.72	2.1	0.42	-	1,707

Lipid traits High Density Lipoprotein cholesterol (HDL), Low Density Lipoprotein Cholesterol (LDL), Total Cholesterol (TC) and Triglycerides (TG) data taken from published meta-analysis (Global Lipids Genetics Consortium 2013). Schizophrenia (SCZ2) data taken from Schizophrenia Working Group of the Psychiatric Genomics Consortium (2014). Nonsynonymous variants annotated using VEP (McLaren et al. 2016).

Table 3.8: Highlighted Variants in the Enrichment Analysis for eQTL SNPs in MGI Data

Trait	*SNP*	*Credible Set Size*	*Posterior Probability*	*Annotation*
Rhesus isoimmunization in pregnancy	1:25561667	2	0.33	23 tissues*
Rhesus isoimmunization in pregnancy	1:25583610	2	0.62	33 tissues*
Disorders of lipoid metabolism	1:109817192	5	0.25	Liver, Muscle Skeletal
Disorders of lipoid metabolism	1:109817590	5	0.26	Esophagus Mucosa, Pancreas
Disorders of lipoid metabolism	1:109818306	5	0.42	Brain Cortex, Skin-Not Sun Exposed Suprapubic, Whole Blood
Gout	4:89045331	6	0.86	Vagina
Skin Cancer	6:396321	1	1.00	Cells EBV-transformed lymphocytes, Small Intestine Terminal Ileum, Whole Blood
Fracture of pelvis	9:116113396	1	0.97	Whole Blood
Other venous embolism and thrombosis	9:136137065	11	0.33	Adrenal Gland
Other venous embolism and thrombosis	9:136149229	11	0.58	Colon Sigmoid, Pituitary, Uterus, Vagina
Skin Cancer	16:90024202	13	0.28	Spleen
Arrhythmia (cardiac) NOS	17:65200303	29	0.56	Brain Cortex
Skin Cancer	20:32665748	1	0.96	Skin-Sun Exposed Lower leg

Annotated variants in 95% credible sets that have at least 10% posterior probability of being causal are listed, along with the eQTL tissues they're annotated for. Credible Set Size denotes the number of variants present in the credible set for that locus.
**Some variants annotated for more than 20 eQTL tissues*

Table 3.9: Traits where Nonsynonymous Variants are found to be Significantly Enriched in UK Biobank Data

Trait	*No. of Loci*	*Enrichment Estimate*	*Odds Ratio*	*P-value*
Cholelithiasis and cholecystitis	30	4.38	79.85	5×10^{-08}
Disorders of lipoid metabolism	47	3.48	32.33	1×10^{-07}
Hypercholesterolemia	43	3.57	35.37	3×10^{-07}
Skin cancer	46	3.40	29.97	7×10^{-07}
Other non-epithelial cancer of skin	48	3.45	31.44	1×10^{-06}
Coronary atherosclerosis	51	3.54	34.31	4×10^{-06}
Hypothyroidism	49	3.46	31.81	2×10^{-05}
Diabetes mellitus	56	2.97	19.40	5×10^{-04}
Disorders of mineral metabolism	9	3.99	53.94	2×10^{-03}
Cataract	21	3.29	26.91	2×10^{-03}
Gout	11	4.24	69.14	0.02
Phlebitis and thrombophlebitis	11	3.43	30.86	0.03
Ischemic Heart Disease	29	2.57	13.04	0.03
Other chronic ischemic heart disease, unspecified	20	3.00	20.11	0.03
Other arthropathies	9	3.54	34.58	0.03
Phlebitis and thrombophlebitis of lower extremities	12	3.11	22.39	0.04
Inflammatory bowel disease and other gastroenteritis and colitis	20	3.29	26.80	0.04
Circulatory disease NEC	9	3.09	21.98	0.04
Other disorders of circulatory system	10	3.04	20.95	0.04

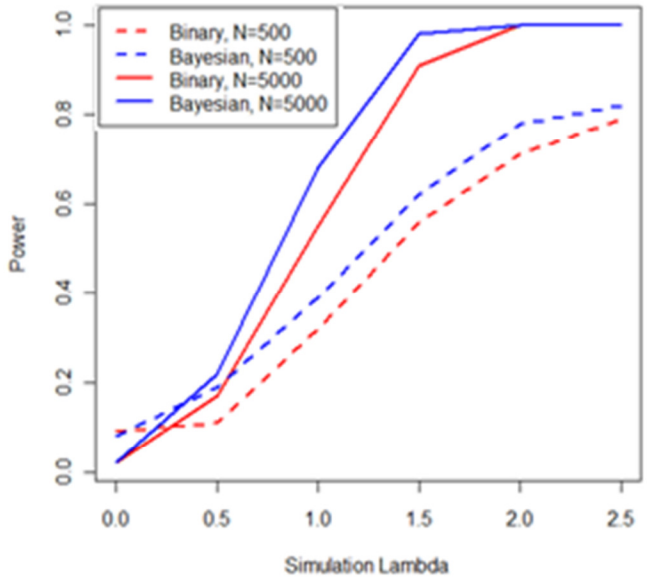
Traits where nonsynonymous variants were found to be significantly enriched in associated loci. Analysis done on 44 traits found to have at least 9 significantly associated loci. No. of loci denotes the number of distinct regions found to be genome-wide significant; Enrichment Estimate the estimated enrichment parameter and P-value the corresponding P-value.

Table 3.10: Traits where eQTL SNPs are found to be Significantly Enriched in UK Biobank Data

Trait	*No. of Loci*	*Enrichment Estimate*	*Odds Ratio*	*P-value*
Varicose veins	31	5.53	252.87	4×10^{-15}
Coronary atherosclerosis	51	4.93	138.93	2×10^{-14}
Disorders of lipoid metabolism	47	4.71	111.41	1×10^{-12}
Varicose veins of lower extremity	33	5.14	171.19	2×10^{-12}
Cardiac dysrhythmias	25	5.62	275.70	2×10^{-12}
Hypercholesterolemia	43	4.67	106.86	7×10^{-12}
Other chronic ischemic heart disease, unspecified	20	5.42	225.92	5×10^{-11}
Angina pectoris	22	5.41	223.47	7×10^{-11}
Ischemic Heart Disease	29	4.80	121.18	2×10^{-10}
Osteoarthrosis	22	5.55	256.69	3×10^{-10}
Diabetes mellitus	56	4.54	94.15	4×10^{-10}
Myocardial infarction	23	5.17	175.98	2×10^{-09}
Skin cancer	46	4.16	63.88	7×10^{-09}
Asthma	43	4.38	80.01	1×10^{-08}
Cholelithiasis and cholecystitis	30	4.64	103.85	2×10^{-08}
Overweight, obesity and other hyperalimentation	10	5.36	212.12	7×10^{-07}
Hypothyroidism	49	4.04	56.89	9×10^{-07}
Disorders of muscle, ligament, and fascia	28	4.39	80.59	1×10^{-06}
Abdominal hernia	16	5.29	199.16	2×10^{-06}
Other non-epithelial cancer of skin	48	3.75	42.36	2×10^{-06}
Inflammatory bowel disease and other gastroenteritis and colitis	20	4.81	123.29	3×10^{-06}
Diverticulosis and diverticulitis	38	4.33	76.31	4×10^{-06}
Nasal polyps	25	4.62	101.24	8×10^{-06}
Benign neoplasm of colon	28	4.38	79.84	2×10^{-05}
Ulcerative colitis	13	4.96	143.16	1×10^{-04}
Diffuse diseases of connective tissue	13	5.54	254.23	1×10^{-04}
Lupus (localized and systemic)	9	6.28	532.10	2×10^{-04}
Benign neoplasm of uterus	18	4.23	69.02	7×10^{-04}
Glaucoma	15	4.44	84.46	8×10^{-04}
Phlebitis and thrombophlebitis	11	4.85	127.34	9×10^{-04}
Pulmonary heart disease	16	3.98	53.75	0.01
Circulatory disease NEC	9	4.10	60.64	0.01
Other disorders of circulatory system	10	4.04	56.73	0.01
Urinary calculus	10	3.58	35.82	0.03
Gout	11	4.38	79.76	0.03
Psoriasis and related disorders	9	3.55	34.82	0.03
Cataract	21	3.72	41.27	0.05

Traits where eQTL SNPs obtained from GTEx (http://www.gtexportal.org/home/) are enriched in the associated loci for UK Biobank Data. Analysis done on 44 traits found to have at least 9 significantly associated loci.

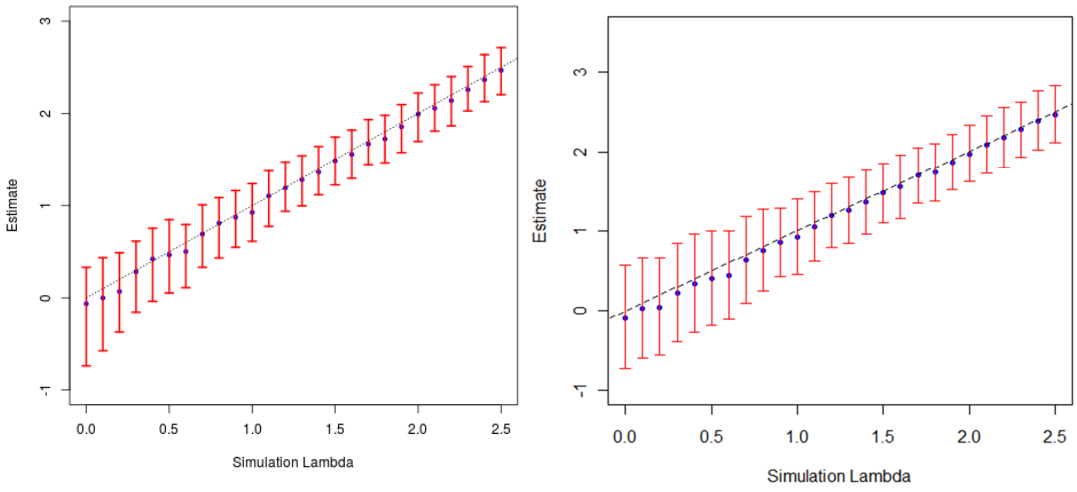
Figure 3.1: Power Curves for Different Sample Sizes



As expected the power improves with increasing sample size, and the Bayesian method has slightly better power than the binary classification method.

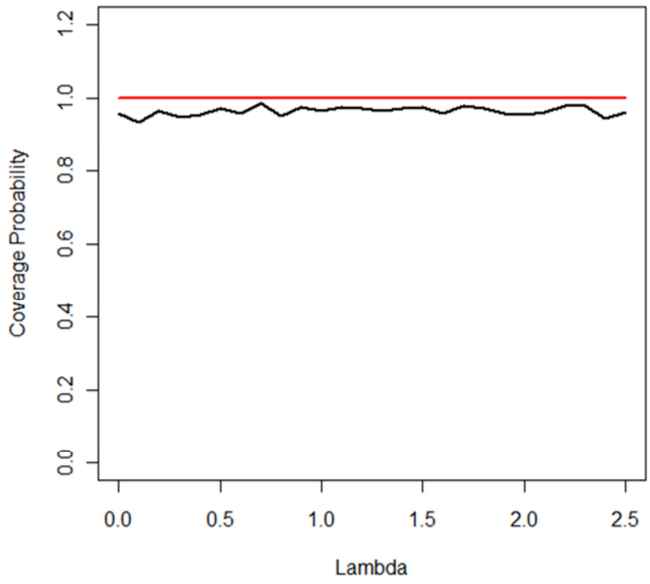
Figure 3.2: Empirical Mean and SD of Enrichment Parameter Estimate

(a) Simpler Approach (b) Bayesian Approach



Estimated λ with standard errors when we vary λ values for (a) Simpler approach and (b) Bayesian approach show that as the true value of λ increases, the standard errors decrease as we have better power for estimation.

Figure 3.3: Coverage Probability



The black line denotes the coverage probability of our estimate which is >95% for all the values of λ simulated.

Figure 3.4: Power Comparison with fGWAS

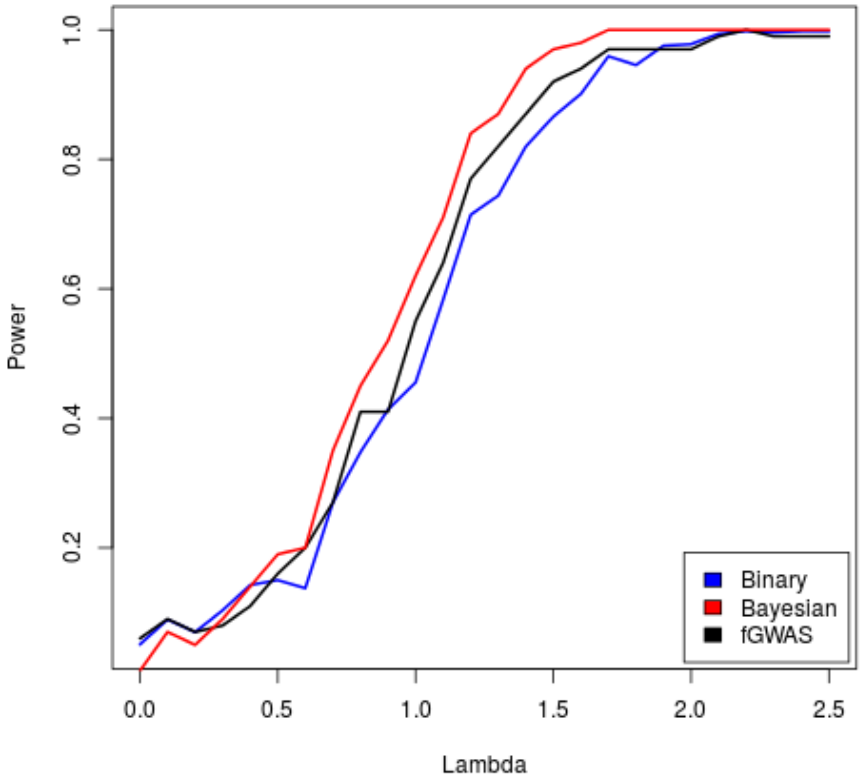
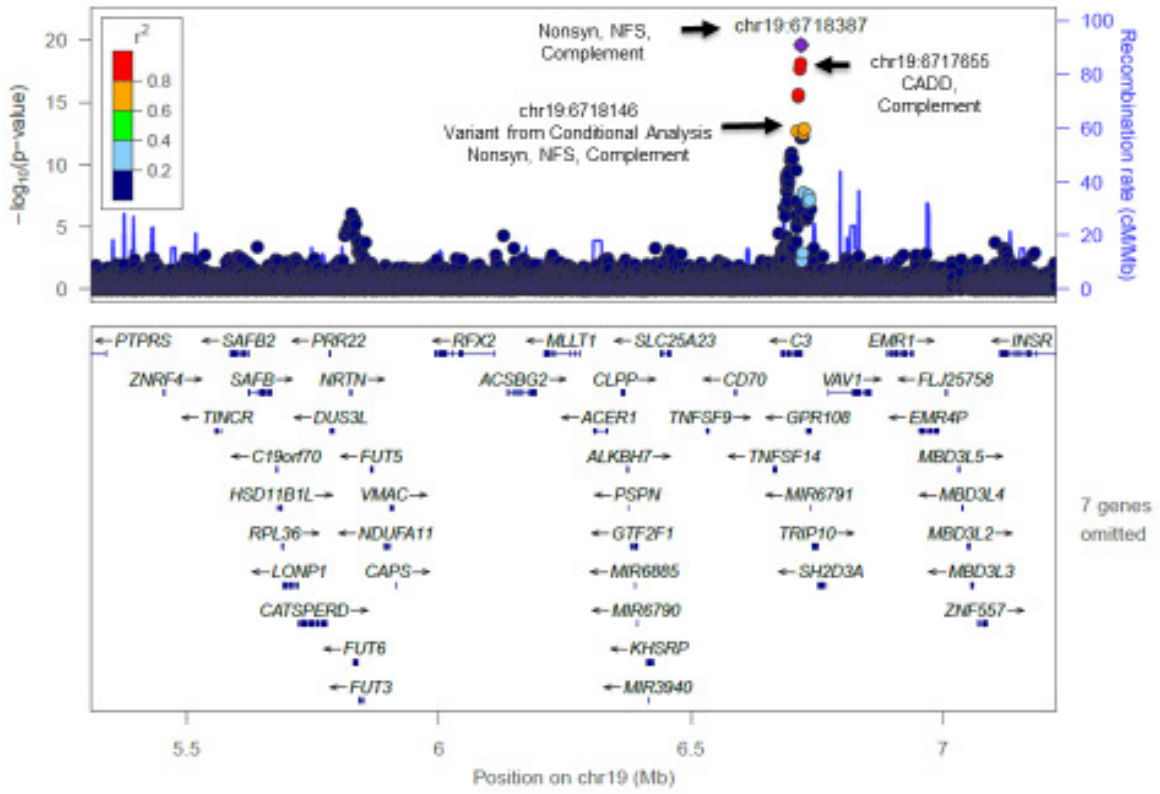


Figure 3.5: Locuszoom plot for region around chr19:6718387 for AMD



Locuszoom plot for region around chr19:6718387 highlighting annotated variants in credible sets from single-SNP analysis as well as conditional analysis results Credible sets for 'pseudo loci' 8 and 9 leads to 3 variants for follow-up.

Supplementary Methods

Algorithm

We wish to quantify the enrichment of variants annotated with a particular feature among causal variants. Now suppose the causal variants were known. Consider the following contingency table:

	Has Annotation	**No Annotation**
Causal	a	b
Non-causal	c	d

The log odds ratio of this table estimates the parameter λ from the logistic model

$$\log \frac{P(Z = 1 \mid \delta)}{P(Z = 0 \mid \delta)} = \mu + \lambda\delta$$

and thus, is an enrichment parameter as required. However, we do not observe the table in practice. We begin with some initial estimates of μ and λ, and compute the expected cell counts of the table and the corresponding log odds ratio under our model assumptions. We continue estimating in an iterative process till the estimates converge.

Modeling the Association

We begin with observed data from an association study. Let n be the sample size and $\boldsymbol{y}$ the trait vector, $\boldsymbol{G}$ the genotype matrix.

We model the phenotype-genotype associations using a standard multiple linear regression model:

$$\boldsymbol{y} = \beta_0 \mathbf{1} + \sum_{i=1}^{p} \beta_i \boldsymbol{g_i} + \boldsymbol{e_i}$$

where $\boldsymbol{g_i}$ indicates the i^{th} column of the matrix $\boldsymbol{G}$.

Simpler Approach: Focusing on Loci that Reach Genome-wide Significance

In this approach, we assume that each associated locus has exactly one causal variant and that the background variants not in the associated loci are not causal. We make further simplifying assumptions that the average chi-square for the background variants is 1.

Prior

Suppose that there are b_1 and b_0 background variants with and without the annotation respectively, and L associated loci, we assign the background variants to the associated loci, so that the prior for a variant at each locus becomes

$$P(\boldsymbol{Z_{li}} = 1 \mid \boldsymbol{\delta_l}, \lambda) = \frac{\exp(\lambda\delta_{li})}{\frac{b_0}{L} + \frac{b_1}{L}\exp(\lambda) + \sum_{j=1}^{n_l} \exp(\lambda\delta_{lj})}$$

Association Results

Note that the χ^2 test statistics from the single-SNP association results can be used to approximate the likelihood as follows:

$$T_{li} = 2\log\left(\frac{P(\boldsymbol{y} \mid \boldsymbol{G}, \boldsymbol{Z_{li}} = 1)}{P(\boldsymbol{y} \mid H_0)}\right)$$

where $T_{li} = (\hat{\beta}_{li}/\mathrm{se}(\hat{\beta}_{li}))^2$ is the χ^2 test statistic associated with the i^{th} variant of the l^{th} locus, and H_0 is the null case where the trait is not associated with any of the variants, and thus can be considered as a constant c independent of the genotypes.

Hence,

$$P(\boldsymbol{y} \mid \boldsymbol{G}, \boldsymbol{Z_{li}} = 1) \propto \exp\left(\frac{T_{li}}{2}\right)$$

Thus, for variants in the causal loci, we get

$$P(\boldsymbol{y} \mid \boldsymbol{G}, \boldsymbol{\delta}, \lambda, \boldsymbol{Z_{li}} = 1) \approx \exp\left(\frac{1}{2}\left(\frac{\hat{\beta}_{li}}{\mathrm{se}(\hat{\beta}_{li})}\right)^2\right)$$

Likelihood

The likelihood contribution of one locus l is

$$\boldsymbol{L_l}(\lambda) \approx \sum_i P(\boldsymbol{y} \mid \boldsymbol{G}, \boldsymbol{\delta}, \lambda, \boldsymbol{Z_{li}} = 1)\, P(\boldsymbol{Z_{li}} = 1 \mid \boldsymbol{\delta_l}, \lambda)$$

$$+ \exp\left(\frac{1}{2}\right)\left(\frac{b_0}{L} + \frac{b_1}{L}\exp(\lambda)\right)\frac{1}{\frac{b_0}{L} + \frac{b_1}{L}\exp(\lambda) + \sum_{j=1}^{n_l} \exp(\lambda\delta_{lj})}$$

$$= \left[\sum_{i=1}^{n_l} \exp\left(\frac{1}{2}\frac{\widehat{\beta_{li}}^2}{\mathrm{se}(\widehat{\beta_{li}})^2}\right)\exp(\lambda\delta_{li})\right.$$

$$\left. + \exp\left(\frac{1}{2}\right)\left(\frac{b_0}{L} + \frac{b_1}{L}\exp(\lambda)\right)\right]\frac{1}{\frac{b_0}{L} + \frac{b_1}{L}\exp(\lambda) + \sum_{j=1}^{n_l} \exp(\lambda\delta_{lj})}$$

We assume that the loci are independent so that the joint log-likelihood becomes

$$\boldsymbol{LLK}(\lambda) = \sum_{\boldsymbol{l}} \log\left(\boldsymbol{L_l}(\lambda)\right)$$

Testing

The Likelihood ratio test is

$$\Lambda = 2(\boldsymbol{LLK}(\hat{\lambda}) - \boldsymbol{LLK}(0))$$

which follows a χ^2distribution with 1 degree of freedom.

Posterior Probability

Although we assign the background variants to the associated loci, they are spread across the genome. To calculate the posterior probability for the variants actually present in the loci, we get

$$P(\boldsymbol{Z_{li}} = 1 \mid \boldsymbol{y}, \boldsymbol{G}, \boldsymbol{\delta}, \lambda) \approx \frac{\exp\left(\frac{1}{2}\frac{\hat{\beta}_{li}^{\,2}}{\mathrm{se}(\hat{\beta}_{li})^2}\right)\exp(\lambda\delta_{li})}{\left[\sum_{j=1}^{n_l} \exp\left(\frac{1}{2}\frac{\hat{\beta}_{lj}^{\,2}}{\mathrm{se}(\hat{\beta}_{lj})^2}\right)\exp(\lambda\delta_{lj}) + \exp\left(\frac{1}{2}\right)\left(\frac{b_0}{L} + \frac{b_1}{L}\exp(\lambda)\right)\right]}$$

However, this formula still takes the background variants into account so that the posterior probabilities of variants actually present in an associated locus sum to <1. To compute the credible sets, we use the modified posterior probabilities:

$$P(\boldsymbol{Z_{li}} = 1 \mid \boldsymbol{y}, \boldsymbol{G}, \boldsymbol{\delta}, \lambda) \approx \frac{\exp\left(\frac{1}{2}\frac{\hat{\beta}_{li}^{\,2}}{\mathrm{se}(\hat{\beta}_{li})^2}\right)\exp(\lambda\delta_{li})}{\left[\sum_{j=1}^{n_l}\exp\left(\frac{1}{2}\frac{\hat{\beta}_{lj}^{\,2}}{\mathrm{se}(\hat{\beta}_{lj})^2}\right)\exp(\lambda\delta_{lj})\right]}$$

An Alternative Bayesian Approach

Association Results

The phenotype-genotype association is modeled using a standard multiple linear regression model:

$$\boldsymbol{y} = \beta_0\mathbf{1} + \sum_{i=1}^{p}\beta_i\boldsymbol{g_i} + \boldsymbol{e_i}$$

where $\boldsymbol{g_i}$ indicates the i^{th} column of the matrix $\boldsymbol{G}$.

In the GWAS context, we expect most of the SNPs in the genome to not be associated with the trait and thus, only a small proportion of the β_is is non-zero. However, we do not know which SNPs are causal and a large number of SNPs are in linkage disequilibrium with one another. Let $\boldsymbol{Z}$ be a latent indicator variable denoting the causal SNPs. We assume that our genotype data is divided into *L* smaller segments or loci and that the vector $\boldsymbol{Z}$ can be split correspondingly into vectors as follows:

$$\boldsymbol{Z} = \boldsymbol{Z_1} \oplus \boldsymbol{Z_2} \oplus \ldots \oplus \boldsymbol{Z_L}$$

Let $\boldsymbol{Z_{ij}}$ be the indicator variable corresponding to the j^{th} SNP in the i^{th} locus and $\boldsymbol{\delta_{ij}}$ be an indicator variable denoting whether the SNP has the feature of interest.

Prior

We assume that the prior probability of a SNP being causal depends on whether it has the feature or not as follows:

$$\mathrm{P}(\mathbf{Z_{ij}} = 1) = \frac{\exp(\mu + \lambda \boldsymbol{\delta_{ij}})}{1 + \exp(\mu + \lambda \boldsymbol{\delta_{ij}})}$$

where λ is the parameter of interest which quantifies the enrichment of SNPs with feature among the causal variants.

Likelihood

Now, the likelihood of the observed data can be obtained by summing over all possible values of the latent variable $\boldsymbol{Z}$ as follows:

$$P(\boldsymbol{y} \mid \boldsymbol{G}, \mu, \lambda) = \sum_{\boldsymbol{Z}} P(\boldsymbol{y} \mid \boldsymbol{G}, \boldsymbol{Z})\, P(\boldsymbol{Z} \mid \mu, \lambda)$$

$$= P(\boldsymbol{y} \mid H_0) \sum_{\boldsymbol{Z}} \frac{P(\boldsymbol{y} \mid \boldsymbol{G}, \boldsymbol{Z})}{P(\boldsymbol{y} \mid H_0)} P(\boldsymbol{Z} \mid \mu, \lambda)$$

where H_0 is the case where the trait is not associated with any of the variants, and thus can be considered as a constant c independent of the genotypes. Thus,

$$P(\boldsymbol{y} \mid \boldsymbol{G}, \mu, \lambda) = c \sum_{\boldsymbol{z}} BF(\boldsymbol{z}) P(\boldsymbol{Z} = \boldsymbol{z} \mid \mu, \lambda)$$

where $BF(\boldsymbol{z})$ indicates the Bayes Factor for the selected $\boldsymbol{z}$.

Bayes Factor

Here,

$$BF_{ij} = \left(1 + \frac{\phi^2}{v_{ij}}\right)^{1/2} \exp\left(\frac{1}{2} \frac{\widehat{\beta_{ij}}^2}{v_{ij}} \frac{\phi^2}{v_{ij} + \phi^2}\right)$$

with $\widehat{\beta_{ij}}$ and v_{ij} being the estimated association effect size and variance for the j^{th} variant in the i^{th} locus respectively. The prior imposed on this effect size is $\beta \sim N(0, \phi^2)$ and we average over a range of possible values of ϕ (Wen 2014).

Now, note that the Bayes Factor can be approximated as

$$BF(\boldsymbol{z}) = \prod_{i \in L} BF(\boldsymbol{z_i})$$

where $BF(\boldsymbol{z_i})$ represents the Bayes Factor for the selected vector $\boldsymbol{z_i}$ at locus i.

Simplification of Likelihood

From the prior imposed on a variant being causal, we get

$$P(\boldsymbol{Z} \mid \mu, \lambda) = \prod_{i,j} P\left(\boldsymbol{Z_{ij}} \middle| \mu, \lambda\right)$$

$$= \prod_{i \in L} P(\boldsymbol{Z_i} \mid \mu, \lambda)$$

Thus,

$$P(\boldsymbol{y} \mid \boldsymbol{G}, \mu, \lambda) \propto \sum_{\boldsymbol{z}} \prod_{i \in L} BF(\boldsymbol{z_i}) P(\boldsymbol{Z} = \boldsymbol{z} \mid \mu, \lambda)$$

$$= \sum_{\boldsymbol{z_1}, \boldsymbol{z_2}, \ldots \boldsymbol{z_L}} \prod_{i \in L} BF(\boldsymbol{z_i}) P(\boldsymbol{Z_i} = \boldsymbol{z_i} \mid \mu, \lambda)$$

$$= \prod_{i \in L} \left(\sum_{\boldsymbol{z_i}} BF(\boldsymbol{z_i}) P(\boldsymbol{Z_i} = \boldsymbol{z_i} \mid \mu, \lambda) \right)$$

To iterate over all possible combinations of $\boldsymbol{z}$ is not computationally feasible. In order to make the inference procedure computationally tractable, we make some simplifying assumptions. Suppose that we can partition L into the set of loci without a causal variant (L_0) and the set of

loci with at least one causal variant (L_1) such that $L = L_0 \cup L_1$. In practice, we use association results based on single SNP analysis to approximate the partition. Observe that for a non-causal locus i, $BF(\boldsymbol{z_i}) = 1$, and hence,

$$\sum_{\boldsymbol{z_i}} BF(\boldsymbol{z_i}) P(\boldsymbol{Z_i} = \boldsymbol{z_i} \mid \mu, \lambda) = \sum_{\boldsymbol{z_i}} P(\boldsymbol{Z_i} = \boldsymbol{z_i} \mid \mu, \lambda) = 1$$

$$\Rightarrow P(\boldsymbol{y} \mid \boldsymbol{G}, \mu, \lambda) \propto \prod_{i \in L_1} \left(\sum_{\boldsymbol{z_i}} BF(\boldsymbol{z_i}) P(\boldsymbol{Z_i} = \boldsymbol{z_i} \mid \mu, \lambda) \right)$$

For the associated loci, we assume that the locus is defined to be small enough to contain a single causal variant. Then for such a locus i,

$$BF(\boldsymbol{z_i}) P(\boldsymbol{Z_i} = \boldsymbol{z_i} \mid \mu, \lambda) \to 0 \quad \text{for } \boldsymbol{z_i} : \sum_j \boldsymbol{z_{ij}} > 1$$

This means that there is at most one underlying causal variant in the locus, and terms involving indicator variables with more than one causal variant are negligible.

Let $S_i = \{ \boldsymbol{z_i} : \sum_j \boldsymbol{z_{ij}} = 1 \}$, that is, the set with exactly one causal variant in the locus.

$$\Rightarrow \sum_{\boldsymbol{z_i}} BF(\boldsymbol{z_i}) P(\boldsymbol{Z_i} = \boldsymbol{z_i} \mid \mu, \lambda) = \sum_{\boldsymbol{z_i} \in \boldsymbol{S_i}} BF(\boldsymbol{z_i}) P(\boldsymbol{Z_i} = \boldsymbol{z_i} \mid \mu, \lambda) + BF(\boldsymbol{0}) P(\boldsymbol{Z_i} = \boldsymbol{0} \mid \mu, \lambda)$$

Note that for $\boldsymbol{z_i^*} \in S_i : \boldsymbol{z_{ij}^*} = 1$, $BF(\boldsymbol{z_i^*}) = BF(\boldsymbol{z_{ij}})$ and $BF(\boldsymbol{0}) = 1$.

$$\Rightarrow \sum_{\boldsymbol{z_i}} BF(\boldsymbol{z_i}) P(\boldsymbol{Z_i} = \boldsymbol{z_i} \mid \mu, \lambda)$$

$$= \sum_j BF_{ij} P\left(\boldsymbol{Z_{ij}} = 1, \boldsymbol{Z_{ik}} = 0 \ \forall \, k \neq j \,\middle|\, \mu, \lambda\right) + P(\boldsymbol{Z_i} = 0 \mid \mu, \lambda)$$

$$= \sum_j \left(BF_{ij}\, P\left(\boldsymbol{Z_{ij}} = 1 \,\middle|\, \mu, \lambda\right) \prod_{k \neq j} P(\boldsymbol{Z_{ik}} = 0 \,|\mu, \lambda) \right) + P(\boldsymbol{Z_i} = \mathbf{0} \mid \mu, \lambda)$$

$$= \sum_j \left(BF_{ij} \frac{\exp(\mu + \lambda\delta_{ij})}{1 + \exp(\mu + \lambda\delta_{ij})} \prod_{k \neq j} \frac{1}{1 + \exp(\mu + \lambda\delta_{ik})} \right) + P(\boldsymbol{Z_i} = \mathbf{0} \mid \mu, \lambda)$$

$$= \pi_i \left(\sum_j BF_{ij} \exp(\mu + \lambda\delta_{ij}) + 1 \right)$$

where BF_{ij} is the Bayes Factor corresponding to the j^{th} variant in the i^{th} locus and π_i is the locus specific probability that there is no causal variant in the i^{th} locus.

Let the i^{th} locus have n_i variants. Then,

$$\pi_i = \prod_{j=1}^{n_i} \frac{1}{1 + \exp(\mu + \lambda\delta_{ij})}$$

$$\Rightarrow P(\boldsymbol{y} \mid \boldsymbol{G}, \mu, \lambda) \propto \prod_{i \in L_1} \pi_i \left[\sum_j BF_{ij} \exp\left(\mu + \lambda\delta_{ij}\right) + 1 \right]$$

Testing

Thus, the Likelihood Ratio Test is:

$$\Lambda = 2 \log\left(\frac{P(\boldsymbol{y} \mid \boldsymbol{G}, \hat{\mu}, \hat{\lambda})}{P(\boldsymbol{y} \mid \boldsymbol{G}, \widehat{\mu_0}, \lambda = 0)} \right)$$

where $\widehat{\mu_0}$ is the estimated value of μ under the model assumption that $\lambda = 0$ and Λ follows χ^2 with 1 degree of freedom.

Posterior Probabilities

Note that as $\boldsymbol{Z_{ij}}$ takes values 0 and 1, the posterior expectation of $\boldsymbol{Z_{ij}}$ is simply the posterior probability that $\boldsymbol{Z_{ij}} = 1$.

$$P(\boldsymbol{Z_{ij}} = 1 \mid \boldsymbol{y}, \boldsymbol{G}, \mu, \lambda) = \frac{P(\boldsymbol{y}, \boldsymbol{Z_{ij}} = 1 \mid \boldsymbol{G}, \mu, \lambda)}{P(\boldsymbol{y} \mid \boldsymbol{G}, \mu, \lambda)}$$

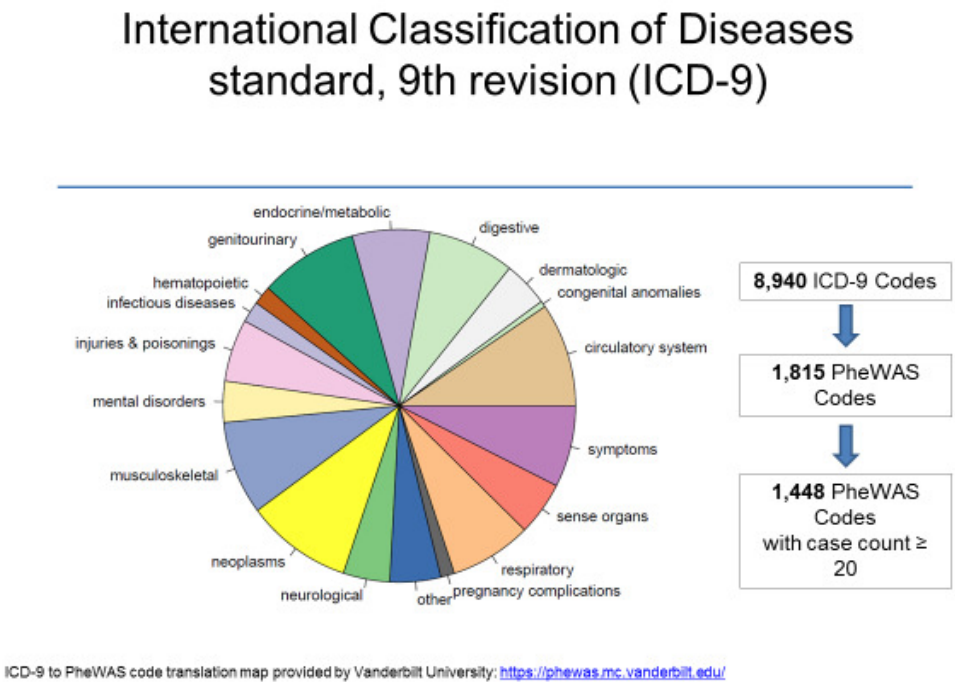
Hence,

$$P(\boldsymbol{Z_{ij}} = 1 \mid \boldsymbol{y}, \boldsymbol{G}, \mu, \lambda) = \frac{\left[\exp\left(\mu + \lambda\delta_{ij}\right)\right] \pi_i \, BF_{ij}}{\sum_k \left[\exp\left(\mu + \lambda\delta_{ij}\right)\right] \pi_i BF_{ik} + \pi_i}$$

$$= \frac{\sum_k \exp(\mu + \lambda\delta_{ik}) \; BF_{ik}}{1 + \sum_k \exp(\mu + \lambda\delta_{ik}) \; BF_{ik}} \left[\frac{\exp\left(\mu + \lambda\delta_{ij}\right) BF_{ij}}{\sum_k \exp(\mu + \lambda\delta_{ik}) \, BF_{ik}}\right]$$

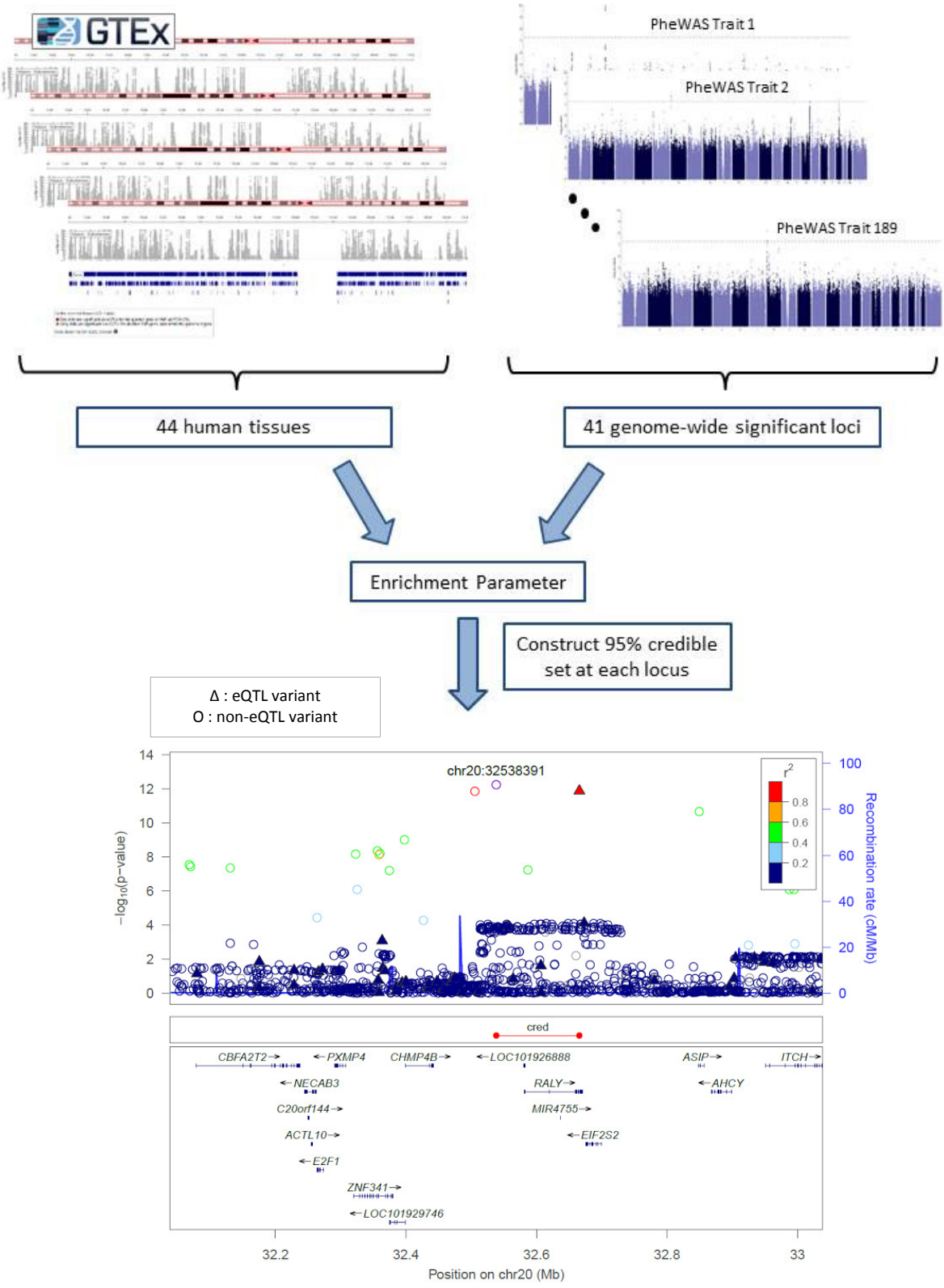
$$= \frac{\sum_k \exp(\mu + \lambda\delta_{ik}) \; BF_{ik}}{1 + \sum_k \exp(\mu + \lambda\delta_{ik}) \; BF_{ik}} \left[\frac{\exp\left(\lambda\delta_{ij}\right) BF_{ij}}{\sum_k \exp(\lambda\delta_{ik}) BF_{ik}}\right]$$

Supplementary Figure S3.1: Phenotypes for MGI Data



Phenotype data for MGI was based on ICD-9 codes. 8,940 ICD-9 codes were aggregated into 1,815 PheWAS Codes, out of which 1,448 had case count ≥ 20.

Supplementary Figure S3.2: Outline of Method to Estimate Enrichment of eQTL's in MGI Data



Chapter 4: Correcting for Sample Overlap in GWAS Meta-Analysis Using Summary Statistics

Introduction

Meta-analysis is a practical method to increase sample size and power of genome-wide association studies, enabling discovery of novel signals and refinement of discovered loci. Many consortia have been formed to investigate the genetic underpinnings of different traits (Sullivan 2010; Global Lipids Genetics Consortium 2013; The Coronary Artery Disease (C4D) Consortium 2010; International Consortium for Blood Pressure Genome-Wide Association Studies 2011), and often make their summary statistics publicly available. Newer studies may take advantage of these published statistics as a starting point for their own meta-analysis, further increasing sample sizes and increasing power to detect novel signals.

In these successive meta-analyses, an important issue to consider is the potential overlap in the set of participants among successive studies. Any overlap can lead to inflated type I error and false signals. The overlap can have many sources. For example, in some analyses publicly available controls are shared among different studies (Burton et al. 2007). Additionally, same cohorts may contribute to different meta-analysis efforts. For example, for type 2 diabetes, data from FUSION was used in both GoT2D GWAS (Fuchsberger et al. 2016) and 70KforT2D GWAS (Bonàs-Guarch et al. 2017).

Methods exist that account for overlap when individual-level data are available, or the number of participants contributing to both studies is known (Lin and Sullivan 2009). In this paper, we consider a different scenario in which only summary statistics (Z-score and sample

size at each marker) are available from potentially overlapping studies. Assuming that the samples belong to the same ancestry, we propose a method to identify overlap between pairs of studies using GWAS summary statistics, estimate the degree of overlap, and meta-analyze the studies appropriately accounting for the overlaps.

We test the accuracy of our method by constructing overlapping samples based on real GWAS datasets to construct artificial overlapping datasets to illustrate our method. The results indicate that our method works well to estimate and correct for the overlap and obtain well-calibrated summary statistics (Z-scores).

Material and Methods

Standard Meta-Analysis Method

A common approach in meta-analysis is to sum the Z-scores across studies, weighting them appropriately using the sample sizes (Stouffer et al. 1949). Suppose we have K studies, with $Z_k,\ k = 1, \ldots K$, being the Z-score from the k^{th} study and N_k the corresponding sample size. A standard meta-analysis uses weights $w_k,\ k = 1, \ldots K$, to combine the estimates as follows:

$$Z = \sum_{k=1}^{K} w_k \, Z_k \qquad \ldots \text{Equation (1)}$$

The Z_k's are assumed to be have standard normal distribution under the null hypothesis of no association between trait and genetic marker. Hence, the variance of the combined Z-score is:

$$\text{Var}(Z) = \sum_{k=1}^{K} w_k^2 \qquad \ldots \text{Equation (2)}$$

The weights are usually chosen based on per-study sample size so that larger studies have more weight (eg. $w_k = \frac{\sqrt{N_k}}{\sqrt{\sum_l N_l}}$). When the Z-scores are independent, these weights ensure that the combined Z-score is distributed as $N(0,1)$ under the null. However, when the studies have overlapping samples, the variance (2) becomes:

$$\mathrm{Var}(Z) = \sum_{k=1}^{K} w_k^2 + 2 \sum_{k=1}^{K} \sum_{l=k+1}^{K} w_k w_l \,\mathrm{Cov}\,(Z_k, Z_l) \qquad \text{... Equation (3)}$$

where the covariance terms $\mathrm{Cov}(Z_k, Z_l)$ depend on overlap between each pair of studies. Thus, using standard weights no longer leads to a $N(0,1)$ test statistic under the null. To account for this, we estimate this covariance and adjust the weights accordingly. The optimal weights can be shown to be (Lin and Sullivan 2009):

$$[w_1, \ldots w_K] = e^T \Omega^{-1} / e^T \Omega^{-1} e \qquad \text{... Equation (4)}$$

where e is a K x 1 vector of 1's and Ω is the estimated covariance matrix of $(Z_1, \ldots Z_K)$.

The covariance matrix Ω can be calculated easily if individual-level data are available, or if the exact number of overlapping samples between each pair of studies is known. We consider the more general case where the number of overlapping samples is not known and use the pair-wise correlation between Z-scores to estimate the overlap and adjust the weights as in (4).

Meta-Analysis Correcting for Sample Overlap

We develop a method to estimate the sample overlap and correct for it (**Figure 4.1**) using the correlation between Z-scores from each pair of studies. First, we stratify the Z-scores according to sample size at each marker because differences in the number of typed samples at each site could reflect success – or lack thereof – in genotyping across different studies. Second, we truncate the Z-scores using a cutoff value c ($|Z| < c$) to remove the effect of strongly

associated loci. Finally, we estimate the correlation from these stratified truncated observations, and used to estimate the covariance matrix in (4) and meta-analyze using the modified weights.

Correcting for Overlap in Meta-Analysis

Suppose there are K studies in a meta-analysis, and the Z-scores are combined in a weighted sum where w_k is the weight for the k^{th} study. If we can estimate the covariance between Z-scores of each pair of studies in the meta-analysis, we can meta-analyze using modified weights as in (4) as follows:

$$\hat{Z} = \frac{1}{\sqrt{\sum_k w_k^2 + \sum_k \sum_{l \neq k} w_k w_l \hat{r}_{kl}}} \sum_{k=1}^{K} w_k Z_k \qquad \text{... Equation (5)}$$

where $\hat{r}_{kl}$ is the estimated correlation between the Z-scores of the k^{th} and l^{th} studies under the null. Note that the Z-scores are assumed to have standard normal distribution under the null, and hence, covariance and correlation can be interchanged.

Using Truncated Z-scores to Estimate Covariance

We assume that (a) effect sizes at trait associated loci do not vary from study to study, a condition that should be approximately true given our assumption that all studies are of the same ancestry and (b) the degree of overlap is uniform across markers after accounting for sample size stratification. Furthermore, we assume that the Z-scores for a pair of studies have a bivariate normal distribution. Suppose that the trait under consideration is independent of genetic effects. Then the Z-scores are standard normally distributed for each study, and sample correlation of paired Z-scores can be used to estimate the correlation parameter of the bivariate normal distribution.

$$\begin{pmatrix} Z_i \\ Z_j \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_{ij} \\ \rho_{ij} & 1 \end{pmatrix} \right) \qquad \text{... Equation (6)}$$

However, Z-scores at trait associated loci are expected to show positive correlation even in independent samples as trait associated loci are expected to have same direction of effect, and using the sample correlation between observed Z-scores would lead to an over-estimation of the correlation. We also expect most traits for GWAS to be complex polygenic traits where there may be many variants with small effect sizes.

To exclude potentially causal loci, we use a cutoff c, and use markers with Z-scores in the interval $(-c, c)$ to estimate the correlation. For example, using $c = 1$ uses about 68% of the markers while excluding the more significant loci. We assume a truncated normal distribution on the Z-scores to estimate the maximum likelihood estimate of correlation, and use this to estimate the overlap. The likelihood of the observed Z-scores between studies i and j is:

$$L = \prod_{m} \frac{\phi(\boldsymbol{Z_{im}}, \boldsymbol{Z_{jm}} | \rho_{ij})}{P(|\boldsymbol{Z_{im}}| < c, |\boldsymbol{Z_{jm}}| < c \mid \rho_{ij})} \qquad \text{... Equation (7)}$$

where m ranges over all the markers present in both studies, and the Z-scores are assumed to follow a bivariate normal distribution with mean 0, variance 1 and correlation ρ_{ij}.

The estimated correlation obtained from (7) is then used in (5) to correctly meta-analyze the studies by modifying the weights to for overlap.

Stratification Based on Sample Size of Marker

For a pair of studies, if all markers are present in both studies, the overlap number is the same for each marker. However, it may happen that sample size varies across markers as some markers may be present only in a sub-cohort of a study. For example, **Figure 4.2** describes a simple scenario where two studies have a cohort overlapping (cohort 2). Markers absent in this overlapping cohort 2 would have an overlap of 0, and so they should be meta-analyzed without

correcting for overlap. However, markers present in the overlapping cohort should be meta-analyzed after correcting for an overlap the size of cohort 2.

Two problems arise if the overlapping number varies by marker. First, the estimated total covariance is biased downward by the markers where there is no overlap and we may apply an insufficient adjustment at many markers, leading to false signals. Secondly, when applying a constant correction for overlap, we may over-correct at markers with no overlap and lose power.

Ideally, clustering methods such as k-means clustering can be used to stratify the total sample size at each marker and works well when comparing a pair of studies. When many studies are included in a meta-analysis there may be a broad range of sample sizes(**Figure 4.3**) and using less refined clustering improves computational efficiency. Thus, we use markers that have at least 50% of total sample size, and bin them using relatively broad bin sizes. Then we estimate the correlation at each stratified level using (7) to estimate the overlap for that group of markers, and then correctly meta-analyze using (5).

If sample size per marker is not available, we can use the total sample size to estimate and correct for overlap. In this case, we omit the stratification step. However, if all markers are not present in all studies, we expect this to lead to errors in the final meta-analysis as some effect sizes may be over-corrected while others remain under-corrected.

Using Pair-wise Correlation of Z-scores to Estimate Effective Overlap Size

Consider a pair of studies with sample sizes n_1 and n_2, and suppose that the trait under investigation is independent of genetic effects. Then, we expect the Z-scores to be distributed as $N(0,1)$ for both studies. Let n_{12} be the number of samples overlapping between the two studies. Now, the Z-scores for each study can be considered as a weighted sum of the Z-scores for the

overlapping and non-overlapping parts. Assuming the weights are proportional to the sample size as follows:

$$Z_1 = \sqrt{(1-p_1)}Z_A + \sqrt{p_1}Z_C \qquad \text{... Equation (8)}$$

$$Z_2 = \sqrt{(1-p_2)}Z_A + \sqrt{p_2}Z_C \qquad \text{... Equation (9)}$$

where the weights used are $p_1 = n_{12}/n_1$ and $p_2 = n_{12}/n_2$, that is, the overlap proportions in each study and Z_A, Z_B, Z_C are standard normal variables. Then,

$$\text{Cov}(Z_1, Z_2) = E(Z_1 Z_2) = \sqrt{p_1 p_2} \qquad \text{... Equation (10)}$$

Thus, as the Z-scores have variance 1,

$$\text{Cor}(Z_1, Z_2) = \sqrt{p_1 p_2} = n_{12}/\sqrt{n_1 n_2} \qquad \text{... Equation (11)}$$

Hence, the effective overlapping number can be estimated using the sample correlation r_{12} between the Z-scores of the 2 studies as follows:

$$\hat{n}_{12} = \sqrt{n_1 n_2} r_{12} \qquad \text{... Equation (12)}$$

In case of GWAS where the trait is not independent of genetic effects, the estimated correlation from (7) can be used in (12) to get an estimate of the effective sample size.

Observe that (12) estimates the effective sample overlap which may be different from the actual sample overlap. For example, for two case-control studies k and l, the estimated correlation corresponds to:

$$\text{Cor}(Z_k, Z_l) \approx \frac{\left(n_{kl0}\sqrt{\frac{n_{k1}n_{l1}}{n_{k0}n_{l0}}} + n_{kl1}\sqrt{\frac{n_{k0}n_{l0}}{n_{k1}n_{l1}}}\right)}{\sqrt{n_k n_l}} \qquad \text{... Equation (13)}$$

where 1 refers to cases and 0 to controls (Lin and Sullivan 2009).

Hence, the estimated effective overlap sample size ($\hat{n}_{kl} = \sqrt{n_k n_l}\hat{r}_{kl}$) may correspond to a range of actual overlap numbers. We can readily derive two extreme possibilities. First, when the overlap is restricted to the cases, $\sqrt{\frac{n_{k1}n_{l1}}{n_{k0}n_{k1}}}\hat{n}_{kl}$ is a point estimate of the number of overlapping samples. Second, when the overlap is restricted to the controls, $\sqrt{\frac{n_{k0}n_{l0}}{n_{k1}n_{l1}}}\hat{n}_{kl}$ is an alternative point estimate of the overlap.

Similar issues may arise in GWAS for quantitative traits if overlap proportions vary by phenotype values. For example, if overlap is concentrated in participants with extremely high phenotype, the estimated effective overlap may be an over-estimate. Note that while the estimated correlation may correspond to a range of overlap proportions, the adjustments to the weights in (5) are still valid.

Meta-Analysis of Multiple Studies

Multiple studies can be meta-analyzed sequentially, that is, each new study can be meta-analyzed with the result from meta-analyzing the previous studies. For each marker for a pair of studies i and j, we meta-analyze them as described above and calculated the following quantities:

$$\text{Total Weight } W = \sqrt{w_i^2 + w_j^2 + 2 * w_i w_j r_{ij}}$$

$$\text{Effective Sample Size } N = n_i + n_j - n_{ij}$$

$$Z = \frac{1}{W}(w_i Z_i + w_j Z_j)$$

Observe that this ensures that the order the studies are analyzed in doesn't affect the results.

Simulation Set-up

We used actual genotypes from 5,000 GWAS individuals from a European population (Fritsche et al. 2016) to simulate a series of overlapping studies with phenotype independent of the genotypes. We simulated the phenotype as normal with mean zero and variance 1 and used 300,000 markers across the genome to run single marker tests for the overlapping studies. We then attempted to estimate sample overlap using GWAS summary statistics and to conduct a meta-analysis that accounted for this estimated overlap.

Artificially Creating Overlapping Datasets based on GWAS Data

We created another series of overlapping GWAS studies using actual lipids and type 2 diabetes) data. We first considered a quantitative trait and used real data from GWAS of HDL-cholesterol (Teslovich et al. 2010). We used 3 studies that contributed to the meta-analysis to artificially create a pair of overlapping studies. We used 2 studies with sample sizes 7,841 and 5,253 respectively, and meta-analyzed a study with sample size 2,485 with each of them to create 2 datasets with an overlap of 2,485. We then meta-analyzed these overlapping together using both the standard meta-analysis as well as our method correcting for overlap. We compared the results with the target results obtained when meta-analyzing the initial three studies directly without overlap.

We carried out a similar procedure for a case-control study using data for type 2 diabetes (Morris et al. 2012). We used 2 studies with sample sizes 6,528 and 16,503 respectively, and meta-analyzed a study with sample size 2,209 with each of them to create 2 datasets with an overlap of 2,209. We then meta-analyzed the overlapping datasets were meta-analyzed correcting for overlap and compared the Z-scores obtained to the target Z-scores obtained by meta-analyzing the studies without overlap.

Results

Simulation Results

Simulation results based on 300,000 markers from 5,000 individuals of European ancestry show that our method provides an accurate estimate of sample overlap (**Table 4.1**). For example, when GWAS studies overlapped by 17% of samples, we estimate overlapping proportions as 17.9% ± 0.8%. We then meta-analyzed GWAS summary statistics accounting for estimated overlap and observed that the mean genomic control was 1.01 as compared to 1.22 when meta-analyzed without considering overlap.

Artificially Created Overlapping Datasets: Quantitative Trait

We created overlapping datasets with sample sizes 10,326 and 7,738 where the overlap number is 2,485 using GWAS datasets for HDL-cholesterol (as per the scheme described in **Supplementary Figure S4.1**). We estimated the overlap proportion of 13.7% to be 14.1% when the Z-scores are truncated at the cutoff value $c = 1$ (**Table 4.2**), and meta-analyzed the data to obtain well-calibrated statistics (**Figure 4.4**).

We varied the cut-off value c and observed that as the cut-off value decreases, the length of our confidence interval for the effective overlap sample size generally increases (e.g. In the maximum sample size category where observed total sample size is 18,064, when cut-off is changed from 1 to 0.5, the confidence interval size increases from 134 to 581). This happens because the number of markers with Z-score within the cut-off limit decreases as the value of the cutoff becomes smaller. However, as the cut-off value increases, we do not observe any systematic pattern to the bias (**Table 4.2**). One exception is the category where there is no actual overlap (observed N = 13,094), where estimated effective overlap increases as the cut-off value increases. Thus, while a more stringent cut-off may be better in terms of truncating trait

associated variants with very small effect sizes, the reduction in the total number of markers decreases the accuracy of the estimate. Based on the datasets investigated, a cut-off value of c=1 appears to work well.

We compared the Z-scores of the corrected meta-analysis with the Z-scores obtained from meta-analyzing the studies without overlap (**Figures 4.4 and 4.5**) and found 100 markers (0.004%) with corrected p-values differing from the target p-values by >=2 on the $\log_{10}$ scale.

We examined these outliers and found that these occurred when the overlapping study had a different effect size compared to the non-overlapping studies. For example, a marker which had a Z-score of 8.9 in the overlapping sample and Z-scores of 5.0 and 2.3 in the non-overlapping samples leads to a corrected Z-score of 9.3 instead of the target value of 8.1 as the correction does not account for the additional deviation from null in the overlapping sample. Similarly, a marker with a Z-score of -1.2 in the overlapping sample and Z-scores of -6.7 and -5.0 in the non-overlapping samples leads to a corrected Z-score of -7.0 instead of the target value of -8.1 as there is overcorrection because of the overlapping Z-score being closer to the null than the others. Thus, population structure in the overlap affects the correction of the meta-analysis Z-scores. We observe that our method improves on the naïve meta-analysis for 11% of the outliers on the $\log_{10}$ scale and for 49% of the outliers on the $\log_e$ scale.

Artificially Created Overlapping Datasets: Case-Control Study

We created overlapping datasets with sample sizes 8,737 and 18,712 where the overlap number is 2,209 using GWAS datasets for type 2 diabetes (as per the scheme described in **Supplementary Figure S4.2**). We estimated the overlap proportion of 8.1% to be 8.5% (**Table 4.3**) when cutoff value for truncating Z-scores was 1, and meta-analyzed the data to obtain well-

calibrated statistics (**Figure 4.6**). A closer examination of the markers with p-value $>10^{-20}$ shows that most markers have corrected Z-scores close to the target Z-scores (**Figure 4.7).**

Varying the cut-off value allows for similar estimates with the confidence interval sizes varying accordingly (**Table 4.3**). We observe that for makers with no actual overlap (sample size = 23,031), increasing the cut-off value increases the estimated overlap as in the quantitative case. We expect this is because variants with small effect sizes may get included as we increase the cut-off value. Thus, for this example, a more stringent cut-off appears to work better.

Discussion

We describe a simple method to identify sample overlap based on GWAS summary statistics, to estimate the overlap, and to adjust for that overlap appropriately. Our method requires Z-scores and sample size at each marker for each study, which are usually available in published GWAS, and hence, can be used to meta-analyze publicly available GWAS data with newer datasets to increase sample size while accounting for any overlap. Not accounting for overlap generally leads to an inflation in type I error, potentially leading to false positive signals. Hence, our method helps to increase the power to detect weaker signals by aggregating sample size, while controlling type I error.

We recommend using our method only if all the samples are of the same ancestry. If the overlapping samples have significantly different effect sizes than the non-overlapping samples, the assumption of homogeneity of effect sizes is more likely to be violated and our method may actually perform worse than a naïve meta-analysis by over-correcting.

We assumed that the degree of overlap is uniform across markers after accounting for sample size stratification. Violation of this assumption leads to mis-calibration in correcting for overlap. For example, for a variant that is specific to a particular population (say, Finland) the

effect of overlap would be stronger than average if all the overlapping samples are drawn from that population (Finland).

While our method only requires the summary statistics, we do need the sample size for each marker to accurately estimate the overlap. Overlap can vary by marker, and hence, not stratifying by sample size can lead to mis-calibration of the summary statistics. For example, **Supplementary Figure S4.3** shows the comparison of corrected meta-analysis p-values when markers are not stratified by sample size with the target and we identify a significant variant not present in the overlapping sample. Not stratifying by sample size leads to an over-correction at this variant leading to a decrease in significance.

We recommend using a more stringent cut-off value for traits known to be highly polygenic to ensure that variants with small effect sizes are not included when estimating correlation. We note that this may lead to fewer variants based on which correlation is estimated, and so may lead to a loss in power.

We have implemented our method for a simple meta-analysis which uses every study available. It assumes there is no heterogeneity of effect sizes which is a rather stringent assumption. There exist newer meta-analysis approaches that modify the weighted Z-score to work under less stringent assumptions, e.g. meta-analysis using a subset based approach (Bhattacharjee et al 2012). Since our method works by estimating the covariance for Z-scores between a pair of studies, a direction for future research might be to extend our method to these approaches.

In conclusion, our proposed method is a simple yet effective way to adjust for sample overlap in GWAS in homogeneous populations while working with the constraints of summary level data.

Table 4.1: Overlap estimate and confidence interval when trait is independent of genotypes

N_1	N_2	*Overlap*	*Overlap Proportion*	*Estimated Overlap Proportion (SD)*
2,500	2,500	0	0.00	0 (-)
2,550	2,550	100	0.02	0.02 (0.0056)
2,750	2,750	500	0.09	0.09 (0.0066)
3,000	3,000	1,000	0.17	0.17 (0.0079)
5,000	5,000	5,000	1.00	1.00 (0.00)
3,000	2,100	100	0.02	0.02 (0.0051)
3,000	2,500	500	0.09	0.09 (0.0048)

Estimated overlap when phenotype is simulated independent of genotypes using real genotypes from 5,000 European samples across 300,000 markers. N_1 and N_2 denote the sample sizes of the observed overlapping samples, and Overlap the true overlap number. Overlap proportion is defined as Overlap / (N_1 + N_2).

Table 4.2: Estimated Sample Overlap for Artificially Created Overlapping Studies for HDL-cholesterol

Sample Size	*Cutoff = 0.5*		*Cutoff = 0.75*		*Cutoff = 1*		*Cutoff = 1.5*		*Cutoff = 2*	
	Estimate	**CI**	**Estimate**	**CI**	**Estimate**	**CI**	**Estimate**	**CI**	**Estimate**	**CI**
4,970	2,485	(2,473, 2,485)	2,485	(2,473, 2,485)	2,485	(2,478, 2,485)	2,485	(2,473, 2,485)	2,485	(2,473, 2,485)
10,223	2,691	(2,280, 2,982)	2,390	(2,214, 2,543)	2,409	(2,324, 2500)	2,463	(2,412, 2,521)	2,463	(2,412, 2,499)
12,811	2,091	(557, 2,887)	2,662	(2,381, 2,887)	2,448	(2,305, 2583)	2,476	(2,406, 2,558)	2,518	(2,457, 2,583)
13,094	-	(-2,792, -706)	-	(-674, -353)	-	(-546, -6)	-	(-128, 96)	40	(-32, 128)
18,064	2,941	(2,637, 3,218)	2,616	(2,503, 2,716)	2,546	(2,458, 2,592)	2,538	(2,503, 2,592)	2,581	(2,548, 2,637)

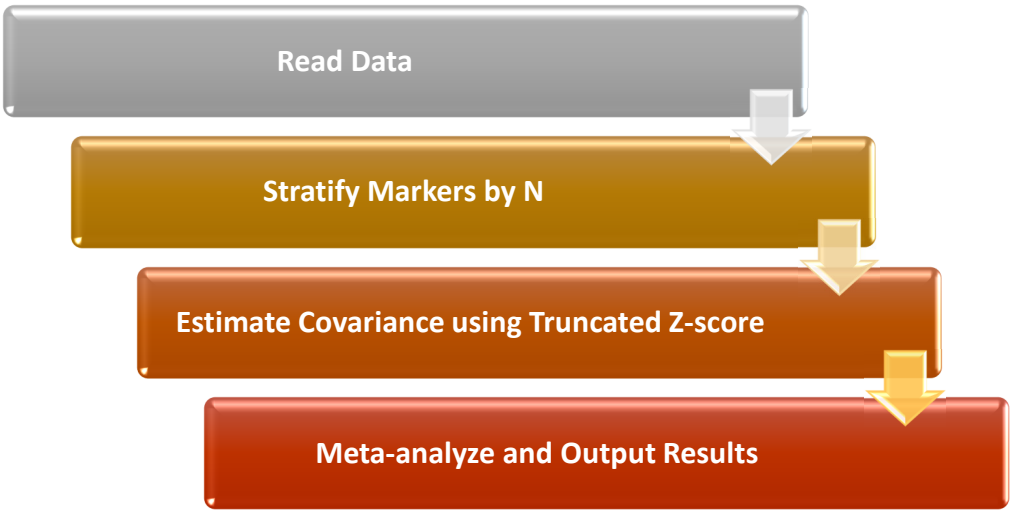
Three European GWAS datasets for HDL-cholesterol (Teslovich et al. 2010) are used to create two overlapping datasets. Datasets 1 and 2 are meta-analyzed together, and datasets 2 and 3 are meta-analyzed together to generate a pair of overlapping datasets whose overlap number equals the sample size of dataset 2 (2,485). The sample size column denotes the total observed sample size for the markers. For sample size = 13,094, the true overlap is 0, and for all other categories, the true overlap is 2,485. The cutoff value is used to truncate the Z-scores used to estimate the correlation and overlap (markers with abs(Z)<cutoff used in the estimation). Estimates <0 are not reported.

Table 4.3: Estimated Sample Overlap for Artificially Created Overlapping Studies for Type 2 Diabetes

Sample Size	*Cutoff = 0.5*		*Cutoff = 0.75*		*Cutoff =1*		*Cutoff = 1.5*		*Cutoff = 2*	
	Estimate	**CI**	**Estimate**	**CI**	**Estimate**	**CI**	**Estimate**	**CI**	**Estimate**	**CI**
4,418	2,209	(2,187, 2,209)	2,209	(2,187, 2,209)	2,209	(2,187, 2,209)	2,209	(2,187, 2,209)	2,209	(2,187, 2,209)
10,946	2,061	(1,779, 2,284)	2,118	(2,021, 2,197)	2,225	(2,175, 2,262)	2,303	(2,284, 2,328)	2,310	(2,284, 2,328)
20,921	1,696	(1,029, 2,282)	2,316	(2,122, 2,507)	2,656	(2,540, 2,765)	2,517	(2,443, 2,572)	2,372	(2,315, 2,411)
23,031	0	(-882, 207)	200	(0, 415)	176	(52, 311)	287	(208, 363)	397	(363, 467)
27,449	2,916	(2,685, 3,132)	2,391	(2,302, 2,493)	2,326	(2,238, 2,365)	2,335	(2,302, 2,365)	2,330	(2,302, 2,365)

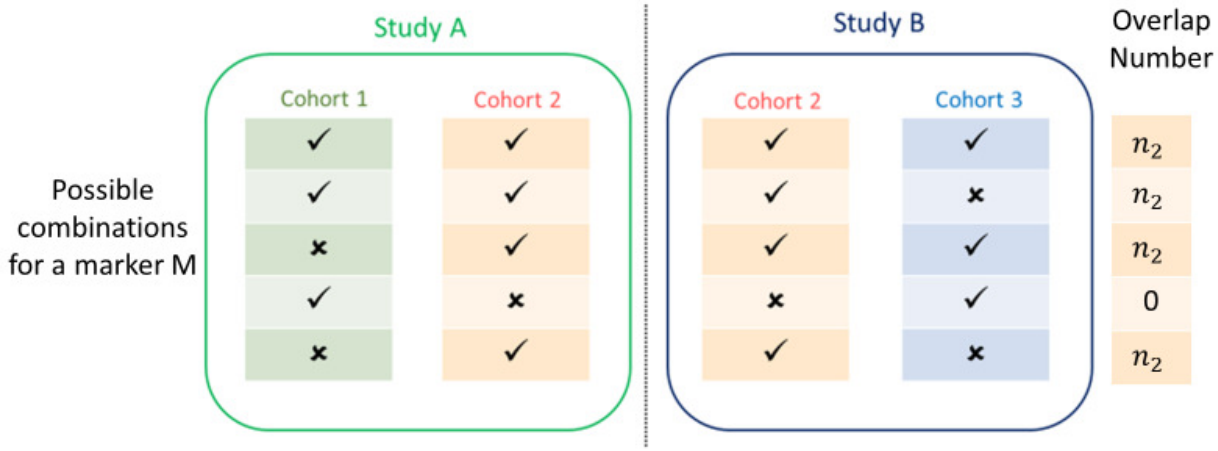
Three European GWAS datasets for type 2 diabetes (Morris et al. 2012) are used to create two overlapping datasets. Datasets 1 and 2 are meta-analyzed together, and datasets 2 and 3 are meta-analyzed together to generate a pair of overlapping datasets whose overlap number equals the sample size of dataset 2 (2,2095). The sample size column denotes the total observed sample size for the markers. For sample size = 23,031, the true overlap is 0, and for all other categories, the true overlap is 2,209. The cutoff value is used to truncate the Z-scores used to estimate the correlation and overlap (markers with abs(Z)<cutoff used in the estimation).

Figure 4.1: Outline of procedure to meta-analyze correcting for overlap



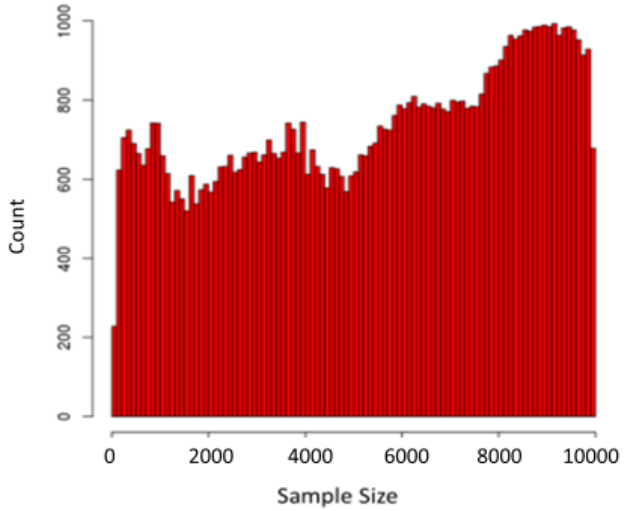
To correctly meta-analyze adjusting for potential overlap, the markers are first stratified by total observed sample size, and then the Z-scores truncated based on a pre-determined cutoff value and used to estimate the correlation between the paired Z-scores. Finally, the estimated correlation is used to adjust the weights in the meta-analysis so that covariance due to overlap is adjusted for.

Figure 4.2: Example Illustrating the Need for Stratification Based on Sample Size



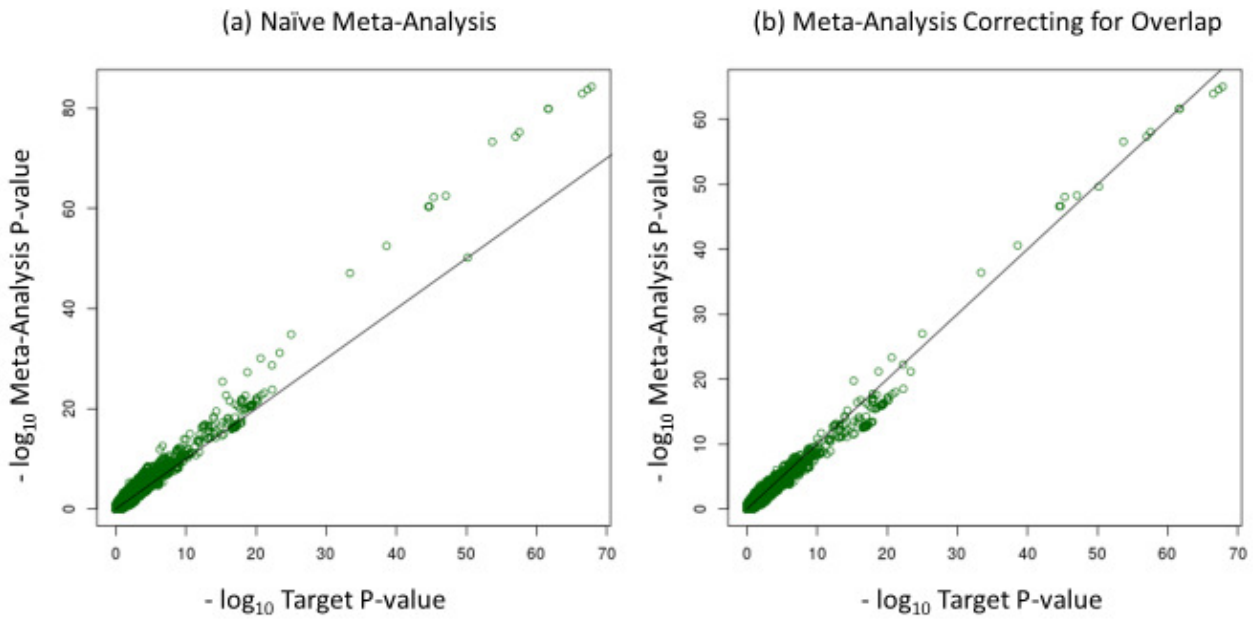
The above diagram shows a simple scenario where 2 overlapping studies Study A and Study B share one cohort Cohort 2. Thus the actual number of overlapping samples is the sample size of cohort 2 (n_2). However, if a marker is not present (that is, not genotyped or imputed) in cohort 2, the overlap number for it is 0. The diagram shows possible combinations for a marker, and why it is important to stratify based on observed sample size.

Figure 4.3 : Example of Sample Size Distribution in a Large Meta-Analysis



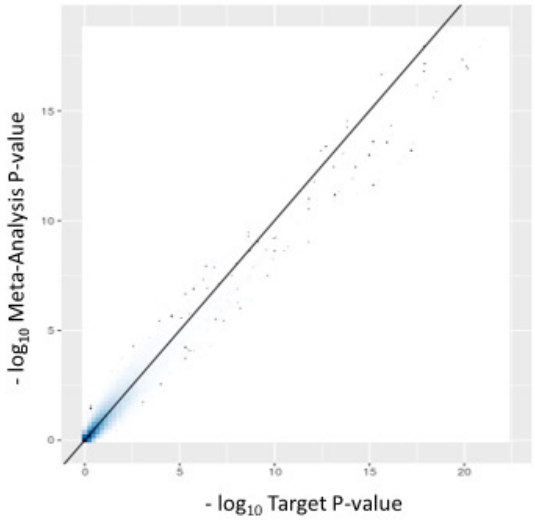
Sample size distribution of a meta-analysis for HDL cholesterol (Teslovich et al. 2010) for all markers with 73,588 unique values of sample size. This demonstrates that markers are not present in all samples as well as the fact that stratifying based on every possible sample size combination is not feasible in large-scale meta-analyses.

Figure 4.4: Performance of Meta-Analysis Correcting for Overlap in HDL-Cholesterol



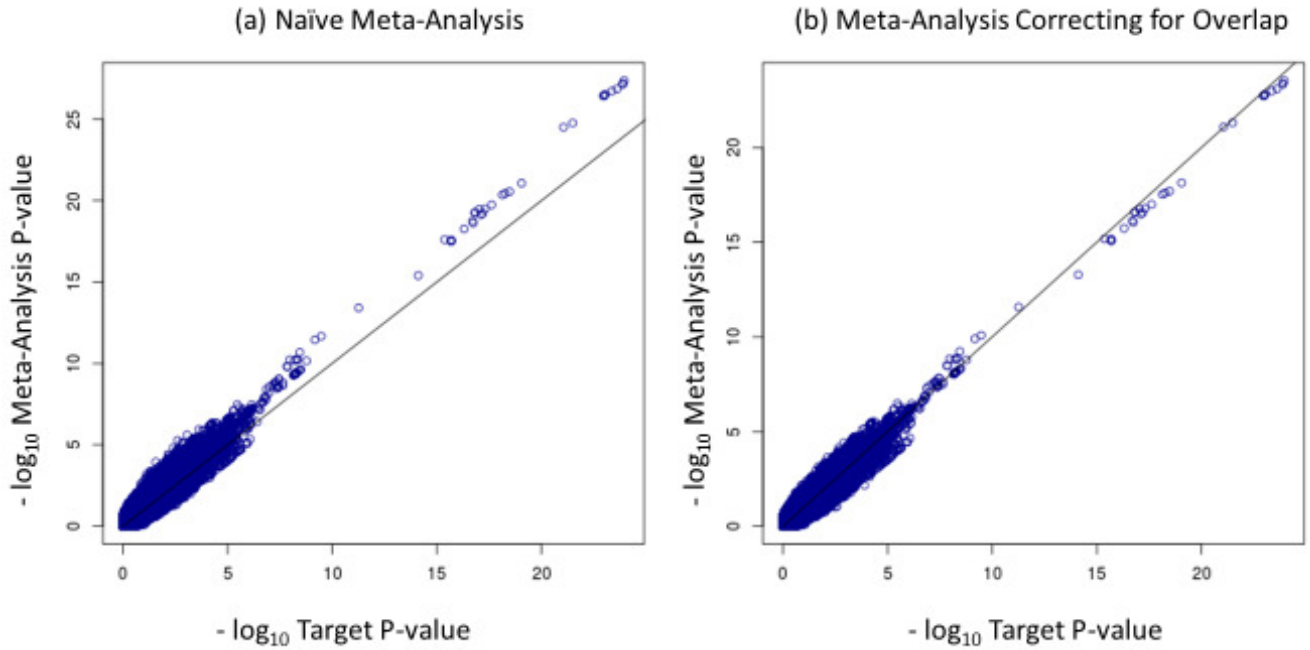
Comparison of target -log_{10} p-values obtained when meta-analyzing the 3 original non-overlapping studies for HDL-cholesterol together with (a) -log_{10} p-values obtained when naively meta-analyzing the overlapping studies; and (b) -log_{10} p-values obtained when meta-analyzing after adjusting for overlap.

Figure 4.5: Outliers in Meta-Analysis Correcting for Overlap in HDL-Cholesterol



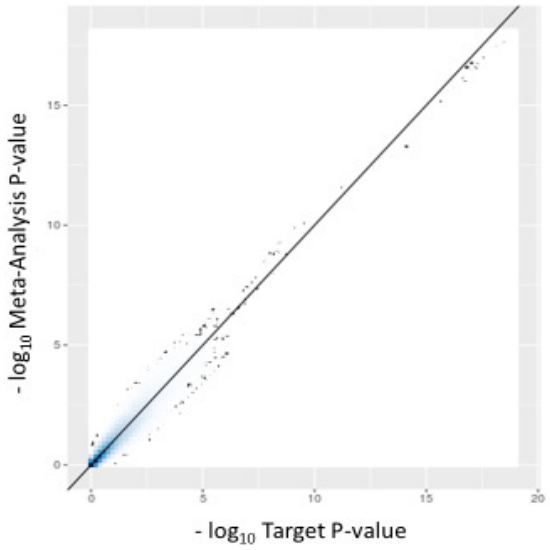
Investigating the markers that are outliers, we observe that zooming in on the markers with target –log_{10} p-value <20, the corrected meta-analysis p-values tend to be biased downward as expected since for most of the markers, the overlap is over-estimated slightly.

Figure 4.6: Performance of Meta-Analysis Correcting for Overlap in Type 2 Diabetes



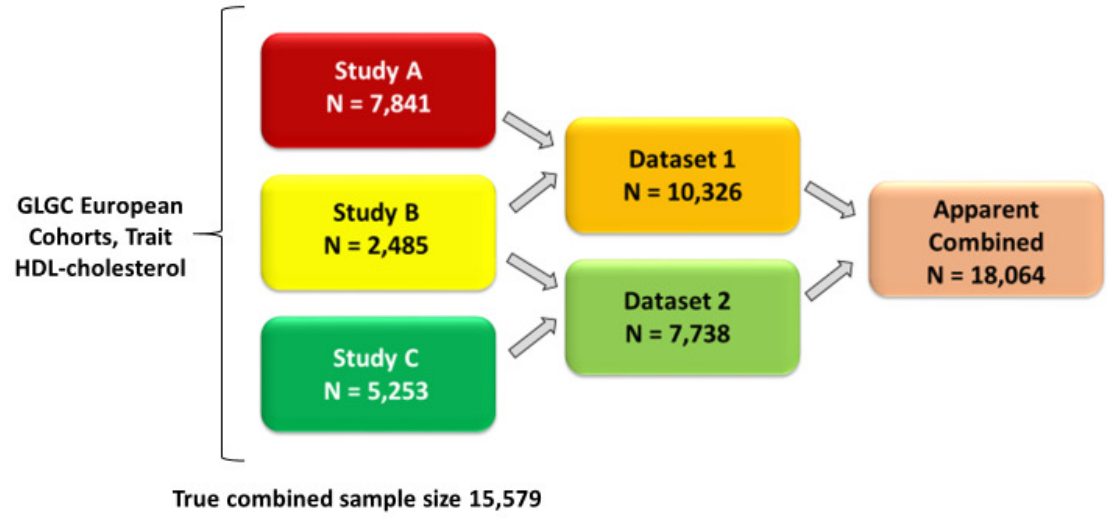
Comparison of target -log_{10} p-values obtained when meta-analyzing the 3 original non-overlapping studies for type 2 diabetes together with (a) -log_{10} p-values obtained when naively meta-analyzing the overlapping studies; and (b) -log_{10} p-values obtained when meta-analyzing after adjusting for overlap

Figure 4.7: Outliers in Meta-Analysis Correcting for Overlap in Type 2 Diabetes



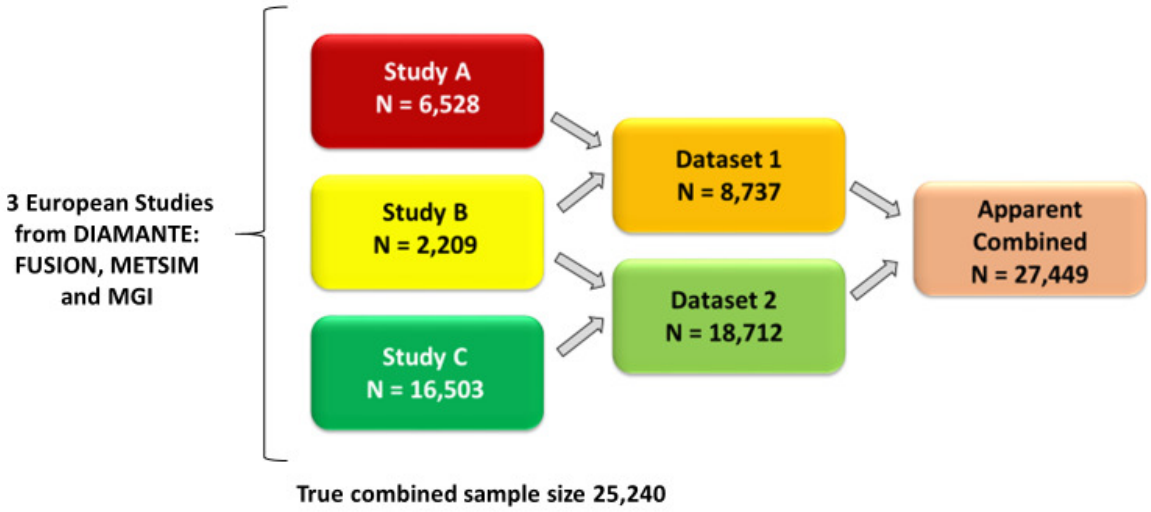
Markers with –log_{10}p-value< 20 for type 2 diabetes seem to be well calibrated: the color denotes the density at each point while the dots denote the outliers.

Supplementary Figure S4.1: Creating Overlapping Datasets for HDL-Cholesterol



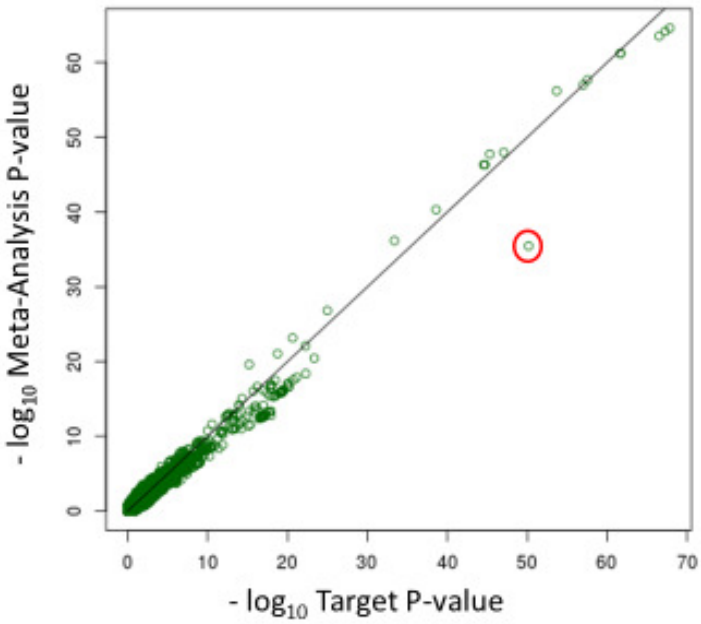
We artificially create an overlapping dataset for HDL-cholesterol using GWAS data (Teslovich et al. 2010). Study B is meta-analyzed with studies A and C respectively to create a pair of overlapping datasets. Studies A, B and C are meta-analyzed directly to get the target results without overlap.

Supplementary Figure S4.2: Creating Overlapping Datasets for Type 2 Diabetes



We artificially create an overlapping dataset for Type 2 Diabetes using GWAS data (Morris et al. 2012). Study B is meta-analyzed with studies A and C respectively to create a pair of overlapping datasets. Studies A, B and C are meta-analyzed directly to get the target results without overlap.

Supplementary Figure S4.3: Effect of Not Stratifying by Sample Size



The target p-value denotes the p-values obtained by meta-analyzing the samples without the overlap, and thus are the gold standard for our analysis. The meta-analysis p-value denotes the p-values obtained using our method to meta-analyze correcting for potential sample overlap. Overlapping samples were created using HDL-cholesterol GWAS data (Teslovich et al. 2010). When the markers are not stratified by sample size before estimating covariance and correcting for overlap, we may lose some signals. Markers such as the one circled in red in the figure are not present in the overlapping sample. Thus, correcting for overlap leads to a decrease in power for these markers.

Chapter 5: Conclusion

Meta-analysis is a powerful tool to jointly analyze genetic association results from multiple genome-wide association studies when individual level data is not available. Aggregating data across studies to increase sample size and power facilitates the discovery of trait-associated variants with modest effect sizes, and provides an opportunity to increase our understanding of genetic susceptibility. However, meta-analysis results may produce a large number of variants significantly associated with the trait of interest. The association findings from GWAS provide an initial guide for the development of medical treatments by pointing to a genomic region of interest and thus it is important to refine the lists of associated variants for further investigation. In this thesis, I have advanced our understanding of lipid genetics through a large scale meta-analysis, demonstrated the challenges in interpreting meta-analysis results, and provided a framework to integrate functional data to prioritize variants for follow-up. Additionally, I have developed a method to meta-analyze studies that may have overlapping samples.Chapter 2 described the largest genetic association study of blood lipid levels to date, where data on 94,595 individuals from the initial GWAS (Teslovich et al. 2010) were meta-analyzed with a follow-up study of 93,982 individuals genotyped on the Metabochip (Voight et al. 2012). In the manuscript Global Lipids Genetics Consortium et al. (2013), we discovered 62 novel genetic loci associated with lipids to contribute to the existing list of 95 known associated loci. Discovery efforts were followed by several downstream analyses to prioritize variants for follow-up such as literature review, pathway analysis, and investigation of regulation of mRNA

expression. However, the different sources of prioritization sometimes disagreed, establishing the difficulty in interpreting GWAS results and understanding causality. Fine mapping in 65 lipid-associated loci in different ancestries facilitated separation of the strongest signal from the prior GWAS signal in 12 regions. Based on our downstream analyses, we suggested a list of 70 genes from 44 of the novel loci that might be the focus of the first round of functional studies. However, the role of the remaining loci is unknown, leaving opportunities for future genetic studies to study their functional impact.

In Chapter 3, I have described a method to systematically incorporate functional information about the genome to prioritize variants for follow-up analyses. Summary statistics (effect sizes and standard errors or p-values) are weighted using genomic annotation to produce credible sets constituting a list of potentially causal variants. I have proposed two methods: one which uses association results across the whole genome and is more computationally intensive, and one which approximates the association results for variants not in the associated loci to increase computational efficiency. Simulation studies demonstrated the accuracy of our estimates and compared their power with fGWAS (Pickrell 2014). Real data applications to MGI (https://www.michigangenomics.org/) and UK Biobank data (Sudlow et al. 2015, Bycroft et al. 2017), which have hundreds of phenotypes, established the advantages of such a systematic approach. To sort through the large volumes of data, I aggregated data across related traits with few signals in the MGI data, which generated credible sets at associated loci. Applications to Age-related Macular Degeneration association data exhibited how using different genomic annotations can lead to different variants being prioritized. However, consolidating a unique list of potentially causal variants based on different genomic annotations remains a challenge.

As different studies and consortia make their summary statistics publicly available, it is challenging to consolidate their findings into one "bottom-line" p-value, which can be used as a guide to investigate and understand their genetic architecture. A major challenge in meta-analyzing these data together is the potential for sample overlap. In Chapter 4, I have described a meta-analysis approach that identifies and corrects for this overlap and accurately meta-analyzes potentially overlapping studies. This method works with summary statistics (Z-score and sample size) and thus, can be used on publicly available data. A caveat is that my method assumes that overlap does not vary by ancestry, which may not always be the case.

Prioritizing Variants for Follow-up Studies

Currently, integration of varying types of data is an emerging area of inquiry, as large volumes of data from both association studies as well as functional studies are being made publicly available. Investigators collaborate in large-scale consortia to generate association results for huge sample sizes, while large repositories of high-throughput genomics and epigenomic data are being built to enable functional annotation. For example, the systems genetics approach, or Genome Wide Network Study as coined by Björkegren et al. (2015), emphasizes combining data from intermediate phenotypes such as RNA, proteins, metabolites, and epigenetics in multiple disease-relevant tissues.

As demonstrated in the downstream analyses in Chapter 2 and the method developed in Chapter 3, coupling of GWAS findings with functional genomics data can potentially advance our understanding of disease etiology. The method described in Chapter 3 currently works only for binary genomic annotations; thus, in the examples described, I dichotomized non-binary annotations such as CADD scores. However, using the complete range of CADD score values

may lead to better estimates. Additionally, comparing multiple genomic annotations at a time may be of interest, which this model does not yet support.

A simplifying assumption I have used is that there is at most one causal variant per locus, which may not be realistic. An ad-hoc approach to adjust for violation of this assumption is described in the analysis of the AMD data, where conditional association results are used as "pseudo-loci." In such a scenario, summary statistics are no longer adequate and additional conditional analyses are required. Additionally, selection of the variants that the conditional analysis is based on may prove difficult. Multi-SNP models allowing for multiple causal variant per locus require individual level data (Kichaev et al. 2014) and hence, using summary statistics to model that remains difficult.

A key challenge is selection of functional data to use, and to consolidate results from diverse functional annotations in a systematic manner. Future work in this direction may involve extending the model to incorporate non-binary or multiple annotations. While the large volumes of genomics and epigenomics data from sources such as ENCODE (ENCODE Project Consortium 2012) and Roadmap Epigenomics Consortium (Bernstein et al. 2010) make it enticing to test all possible annotations for enrichment, multiple testing issues should be kept in mind since it is possible that some annotations would be found significantly enriched by chance. Thus, a rigorous framework to integrate functional data and association results is required to take advantage of the diverse functional annotations available.

Meta-Analysis of Studies with Sample Overlap

With decreasing genotyping and sequencing costs, there are more and more available sets of genotyped or sequenced controls that can be used in multiple studies. Other sources of overlap may include participants belonging to multiple studies, or the same study contributing to multiple

meta-analysis efforts. We aimed to meta-analyze data for the same trait while correcting for overlap to generate a consolidated list of associated variants. An ongoing project for the type 2 diabetes portal (www.type2diabetesgenetics.org/) is to generate a "bottom-line" p-value combining information across multiple potentially overlapping studies, that is, for each variant a single p-value combining information across all studies for type 2 diabetes should be reported. When overlap numbers are not known, and individual level data is not available, a method to meta-analyze correcting for overlap using summary statistics is required.

A limitation of the method described in Chapter 4 is that we essentially assume a fixed effects model and that allele frequencies and effect sizes do not vary with overlap. This assumption may work if we limit our attention to European populations, which have been studied extensively. However, currently available GWAS data from non-Europeans tend to have smaller sample sizes and thus it may be desirable to meta-analyze them with data from other ancestries to increase power. In such scenarios, as effect sizes and allele frequencies may vary by ancestry, correction for overlap becomes challenging. Future work in this direction may involve using publicly available allele frequencies (1000 Genomes Project Consortium 2010) to approximate the population structure in the corrected meta-analysis.

An additional question to consider is what happens when there is sample overlap between a pair of studies, but the overlap is between cases of one study and controls of another. Depending on how the phenotype is defined in each study, participants with borderline values of the trait may be categorized differently. For example, in the UK Biobank data (Sudlow et al. 2015; Bycroft et al. 2017), several phenotypes are recorded that are highly correlated such as "disorders of lipoid metabolism", "hyperlipidemia" and "hypercholesterolemia". However, cases for one trait may be controls for the other, depending on the attending physician's definition of

the trait in question, rather than inherent misclassification. Our method currently assumes that the studies have uniform definition of phenotypes, which is a realistic assumption in well-designed meta-analyses. However, this is an important issue to keep in mind when using previously published results.

In Summary

In this dissertation, I have discussed various methods to interpret results obtained from genome-wide association studies. I have focused my research on the use of summary statistics to take advantage of the growing repositories of publicly available data. I have described a large-scale meta-analysis for blood lipid levels, leading to new insights into lipid biology. Additionally, I have developed a method to integrate functional annotation of the genome with association results to prioritize lists of potential causal variants, as well as a method to meta-analyze studies with potentially overlapping variants by estimating and correcting for the overlap. It is my hope that the methods and tools developed can lead to advances in our understanding of the genetics of various complex traits and diseases.

Bibliography

1000 Genomes Project Consortium. "A map of human genome variation from population-scale sequencing." Nature467.7319 (2010): 1061-1073.

Ashburner, Michael, et al. "Gene Ontology: tool for the unification of biology." Nature genetics 25.1 (2000): 25-29.

Asselbergs, F.W. et al. "Large-scale gene-centric meta-analysis across 32 studies identifies multiple lipid loci." Am J Hum Genet 91 (2012): 823-38

Barter, Philip J., and Kerry-Anne Rye. "Cholesteryl ester transfer protein inhibition as a strategy to reduce cardiovascular risk." Journal of lipid research 53.9 (2012): 1755-1766.

Begum, Ferdouse, et al. "Comprehensive literature review and statistical considerations for GWAS meta-analysis." Nucleic acids research 40.9 (2012): 3777-3784.

Bernstein, Bradley E., et al. "The NIH roadmap epigenomics mapping consortium." Nature biotechnology 28.10 (2010): 1045-1048.

Bhattacharjee, Samsiddhi, et al. "A subset-based approach improves power and interpretation for the combined analysis of genetic association studies of heterogeneous traits." The American Journal of Human Genetics 90.5 (2012): 821-835.

Bisogno, Tiziana, et al. "Cloning of the first sn1-DAG lipases points to the spatial and temporal regulation of endocannabinoid signaling in the brain." J Cell Biol 163.3 (2003): 463-468.

Björkegren, Johan LM, et al. "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 (2015): 830-845.

Bonàs-Guarch, Sílvia, et al. "A comprehensive reanalysis of publicly available GWAS datasets reveals an X chromosome rare regulatory variant associated with high risk for type 2 diabetes." bioRxiv (2017): 112219.

Burton, Paul R., et al. "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls." Nature 447.7145 (2007): 661-678.

Burton, Paul R., et al. "Size matters: just how big is BIG? Quantifying realistic sample size requirements for human genome epidemiology." International journal of epidemiology38.1 (2008): 263-273.

Bush, William S., Matthew T. Oetjens, and Dana C. Crawford. "Unravelling the human genome-phenome relationship using phenome-wide association studies." Nature Reviews Genetics17.3 (2016): 129-145.

Buyske, Steven, et al. "Evaluation of the metabochip genotyping array in African Americans and implications for fine mapping of GWAS-identified loci: the PAGE study." PLoS One7.4 (2012): e35651.

Bycroft, Clare, et al. "Genome-wide genetic data on~ 500,000 UK Biobank participants." bioRxiv (2017): 166298.

Cao, J. et al. A predominant role of acyl-CoA:monoacylglycerol acyltransferase-2 in dietary fat absorption implicated by tissue distribution, subcellular localization, and up-regulation by high fat diet. Journal of Biological Chemistry279, 18878-86 (2004).

Castelli, W. P. "Cholesterol and lipids in the risk of coronary artery disease--the Framingham Heart Study." The Canadian journal of cardiology 4 (1988): 5A-10A.

Chambers, John C., et al. "Genome-wide association study identifies loci influencing concentrations of liver enzymes in plasma." Nature genetics 43.11 (2011): 1131-1138.

Chapman, Kay, et al. "Defining the power limits of genome□wide association scan meta□ analyses." Genetic epidemiology35.8 (2011): 781-789.

Chasman, Daniel I., et al. "Forty-three loci associated with plasma lipoprotein size, concentration, and cholesterol content in genome-wide analysis." PLoS genetics 5.11 (2009): e1000730.

Chen, Wei-Min, et al. "Variations in the G6PC2/ABCB11 genomic region are associated with fasting glucose levels." The Journal of clinical investigation 118.7 (2008): 2620.

Chen, Wenan, et al. "Fine mapping causal variants with an approximate Bayesian method using marginal test statistics." Genetics 200.3 (2015): 719-736.

Cho, Yoon Shin, et al. "Meta-analysis of genome-wide association studies identifies eight new loci for type 2 diabetes in east Asians." Nature genetics 44.1 (2012): 67-72.

Clarke, Robert, et al. "Cholesterol fractions and apolipoproteins as risk factors for heart disease mortality in older men." Archives of Internal Medicine 167.13 (2007): 1373-1378.

Cooper, Harris, Larry V. Hedges, and Jeffrey C. Valentine, eds. The handbook of research synthesis and meta-analysis. Russell Sage Foundation, 2009.

Coronary Artery Disease (C4D) Genetics Consortium. "A genome-wide association study in Europeans and South Asians identifies five new loci for coronary artery disease." Nature genetics 43.4 (2011): 339-344.

Cross-Disorder Group of the Psychiatric Genomics Consortium. "Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis." The Lancet 381.9875 (2013): 1371-1379.

Danik, J.S. et al. Novel loci, including those related to Crohn disease, psoriasis, and inflammation, identified in a genome-wide association study of fibrinogen in 17 686 women: the Women's Genome Health Study. Circ Cardiovasc Genet2, 134-41 (2009).

Dastani, Zari, et al. "Novel loci for adiponectin levels and their influence on type 2 diabetes and metabolic traits: a multi-ethnic meta-analysis of 45,891 individuals." PLoS genetics 8.3 (2012): e1002607.

Demirkan, Ayşe, et al. "Genome-wide association study identifies novel loci associated with circulating phospho-and sphingolipid concentrations." PLoS genetics 8.2 (2012): e1002490.

Demontis, Ditte, et al. "Discovery Of The First Genome-Wide Significant Risk Loci For ADHD." bioRxiv (2017): 145581.

Denis, Nicholas, et al. "Quantitative proteomic analysis of PCSK9 gain of function in human hepatic HuH7 cells." Journal of proteome research 10.4 (2011): 2011-2026.

DerSimonian, Rebecca, and Nan Laird. "Meta-analysis in clinical trials." Controlled clinical trials 7.3 (1986): 177-188.

Devlin, Bernie, and Kathryn Roeder. "Genomic control for association studies." Biometrics 55.4 (1999): 997-1004.

Dey, Rounak, et al. "A fast and accurate algorithm to test for binary phenotypes and its application to PheWAS." The American Journal of Human Genetics (2017).

Ding, Liang, et al. "Akt3 deficiency in macrophages promotes foam cell formation and atherosclerosis in mice." Cell metabolism 15.6 (2012): 861-872.

Duncan, E.L. et al. Genome-wide association study using extreme truncate selection identifies novel genes affecting bone mineral density and fracture risk. PLoS Genet7, e1001372 (2011).

Duncan, L. E., et al. "Largest GWAS of PTSD (N= 20 070) yields genetic overlap with schizophrenia and sex differences in heritability." Molecular Psychiatry (2017).

Dupuis, J. et al. New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. Nature Genetics 42 (2010): 105-16

ENCODE Project Consortium. "A user's guide to the encyclopedia of DNA elements (ENCODE)." PLoS biology 9.4 (2011): e1001046.

ENCODE Project Consortium. "An integrated encyclopedia of DNA elements in the human genome." Nature 489.7414 (2012): 57-74.

Ernst, Jason, et al. "Mapping and analysis of chromatin state dynamics in nine human cell types." Nature 473.7345 (2011): 43-49.

Evangelou, Evangelos, and John PA Ioannidis. "Meta-analysis methods for genome-wide association studies and beyond." Nature Reviews Genetics 14.6 (2013): 379-389.

Faye, Laura L., et al. "Re-ranking sequencing variants in the post-GWAS era for accurate causal variant identification." PLoS genetics 9.8 (2013): e1003609.

Fischer, Julia, et al. "Inactivation of the Fto gene protects from obesity." Nature 458.7240 (2009): 894-898.

Fitzgerald, Michael L., Kathryn J. Moore, and Mason W. Freeman. "Nuclear hormone receptors and cholesterol trafficking: the orphans find a new home." Journal of molecular medicine 80.5 (2002): 271-281.

Frayling, Timothy M., et al. "A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity." Science 316.5826 (2007): 889-894.

Freathy, Rachel M., et al. "Common variation in the FTO gene alters diabetes-related metabolic traits to the extent expected given its effect on BMI." Diabetes 57.5 (2008): 1419-1426.

Friedewald, William T., Robert I. Levy, and Donald S. Fredrickson. "Estimation of the concentration of low-density lipoprotein cholesterol in plasma, without use of the preparative ultracentrifuge." Clinical chemistry 18.6 (1972): 499-502.

Frikke-Schmidt, Ruth, et al. "Association of loss-of-function mutations in the ABCA1 gene with high-density lipoprotein cholesterol levels and risk of ischemic heart disease." Jama299.21 (2008): 2524-2532.

Fritsche, Lars G., et al. "A large genome-wide association study of age-related macular degeneration highlights contributions of rare and common variants." Nature genetics48.2 (2016): 134-143.

Fuchsberger, Christian, et al. "The genetic architecture of type 2 diabetes." Nature 536.7614 (2016): 41-47.

Gaffney, Daniel J., et al. "Dissecting the regulatory architecture of gene expression QTLs." Genome biology 13.1 (2012): R7.

Galliano, M.F. et al. Binding of alpha2ML1 to the low density lipoprotein receptor-related protein 1 (LRP1) reveals a new role for LRP1 in the human epidermis. PLoS ONE3, e2729 (2008).

Global Lipids Genetics Consortium. "Discovery and refinement of loci associated with lipid levels." Nature genetics 45.11 (2013): 1274-1283.

GTEx Consortium. "The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans." Science 348.6235 (2015): 648-660.

Ha, Vi Luan, et al. "ASAP3 is a focal adhesion-associated Arf GAP that functions in cell migration and invasion." Journal of Biological Chemistry 283.22 (2008): 14915-14926.

Hagberg, Carolina E., et al. "Vascular endothelial growth factor B controls endothelial fatty acid uptake." Nature 464.7290 (2010): 917-921.

Han, Buhm, and Eleazar Eskin. "Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies." The American Journal of Human Genetics 88.5 (2011): 586-598.

Heid, Iris M., et al. "Meta-analysis identifies 13 new loci associated with waist-hip ratio and reveals sexual dimorphism in the genetic basis of fat distribution." Nature genetics 42.11 (2010): 949-960.

Hicks, A.A. et al. Genetic determinants of circulating sphingolipid concentrations in European populations. PLoS Genet5, e1000672 (2009).

Hindorff, Lucia A., et al. "Potential etiologic and functional implications of genome-wide association loci for human diseases and traits." Proceedings of the National Academy of Sciences 106.23 (2009): 9362-9367.

Hormozdiari, Farhad, et al. "Identifying causal variants at loci with multiple signals of association." Genetics 198.2 (2014): 497-508.

International Consortium for Blood Pressure Genome-Wide Association Studies. "Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk." Nature 478.7367 (2011): 103-109.

Ioannidis, John PA, Thomas A. Trikalinos, and Muin J. Khoury. "Implications of small effect sizes of individual genetic variants on the design and interpretation of genetic association studies of complex diseases." American journal of epidemiology 164.7 (2006): 609-614.

Ishikawa, Kazunobu, Mohamad Navab, and Aldons J. Lusis. "Vasculitis, atherosclerosis, and altered HDL composition in heme-oxygenase-1-knockout mice." International journal of hypertension 2012 (2012).

Ito, Jin-ichi, et al. "Apolipoprotein AI induces translocation of protein kinase Cα to a cytosolic lipid-protein particle in astrocytes." Journal of lipid research 45.12 (2004): 2269-2276.

Kang, Hyun Min, et al. "Variance component model to account for sample structure in genome-wide association studies." Nature genetics 42.4 (2010): 348-354.

Kannel, William B., et al. "Factors of Risk in the Development of Coronary Heart Disease—Six-Year Follow-up ExperienceThe Framingham Study." Annals of internal medicine 55.1 (1961): 33-50.

Kaprio, Jaakko, et al. "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 (1991): 1330-1348.

Karczewski, Konrad J., et al. "Systematic functional regulatory assessment of disease-associated variants." Proceedings of the National Academy of Sciences 110.23 (2013): 9607-9612.

Kavvoura, Fotini K., and John PA Ioannidis. "Methods for meta-analysis in genetic association studies: a review of their potential and pitfalls." Human genetics 123.1 (2008): 1-14.

Kazanskaya, Olga, et al. "The Wnt signaling regulator R-spondin 3 promotes angioblast and vascular development." Development 135.22 (2008): 3655-3664.

Keating, Brendan J., et al. "Concept, design and implementation of a cardiovascular gene-centric 50 k SNP array for large-scale genomic association studies." PloS one3.10 (2008): e3583.

Kichaev, Gleb, et al. "Integrating functional data to prioritize causal variants in statistical fine-mapping studies." PLoS genetics 10.10 (2014): e1004722.

Kim, Young Jin, et al. "Large-scale genome-wide association studies in East Asians identify new genetic loci influencing metabolic traits." Nature genetics 43.10 (2011): 990-995.

Kim, Youngjun, et al. "A conserved phosphatase cascade that regulates nuclear membrane biogenesis." Proceedings of the National Academy of Sciences 104.16 (2007): 6596-6601.

Kircher, Martin, et al. "A general framework for estimating the relative pathogenicity of human genetic variants." Nature genetics 46.3 (2014): 310-315.

Koharudin, Leonardus MI, et al. "The phox domain of sorting nexin 5 lacks phosphatidylinositol 3-phosphate (PtdIns (3) P) specificity and preferentially binds to phosphatidylinositol 4, 5-bisphosphate (PtdIns (4, 5) P2)." Journal of Biological Chemistry 284.35 (2009): 23697-23707.

Li, Ruomei, et al. "Role of liver sinusoidal endothelial cells and stabilins in elimination of oxidized low-density lipoproteins." American Journal of Physiology-Gastrointestinal and Liver Physiology 300.1 (2011): G71-G81.

Lin, Dan-Yu, and Patrick F. Sullivan. "Meta-analysis of genome-wide association studies with overlapping subjects." The American Journal of Human Genetics 85.6 (2009): 862-872.

Liu, Y., et al. "Conditional ablation of Gsk-3β in islet beta cells results in expanded mass and resistance to fat feeding-induced diabetes in mice." Diabetologia 53.12 (2010): 2600-2610.

Lloyd-Jones, Donald, et al. "Heart disease and stroke statistics—2010 update." Circulation 121.7 (2010): e46-e215.

Loeper, Siobhan, Sylvia L. Asa, and Shereen Ezzat. "Ikaros modulates cholesterol uptake: a link between tumor suppression and differentiation." Cancer research 68.10 (2008): 3715-3723.

MacArthur, Jacqueline, et al. "The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog)." Nucleic acids research 45.D1 (2017): D896-D901.

Majumdar, Avijit, et al. "Nuclear translocation of cellular retinoic acid-binding protein II is regulated by retinoic acid-controlled SUMOylation." Journal of Biological Chemistry286.49 (2011): 42749-42757.

Maller, Julian B., et al. "Bayesian refinement of association signals for 14 loci in 3 common diseases." Nature genetics44.12 (2012): 1294-1301.

Malo, Nathalie, Ondrej Libiger, and Nicholas J. Schork. "Accommodating linkage disequilibrium in genetic-association analyses via ridge regression." The American Journal of Human Genetics 82.2 (2008): 375-385.

Manolio, Teri A., et al. "Finding the missing heritability of complex diseases." Nature 461.7265 (2009): 747-753.

Maurano, Matthew T., et al. "Systematic localization of common disease-associated variation in regulatory DNA." Science 337.6099 (2012): 1190-1195.

McCarthy, Mark I., et al. "Genome-wide association studies for complex traits: consensus, uncertainty and challenges." Nature reviews genetics 9.5 (2008): 356-369.

McLaren, William, et al. "The ensembl variant effect predictor." Genome biology 17.1 (2016): 122.

Minelli, Cosetta, et al. "Importance of Different Types of Prior Knowledge in Selecting Genome□Wide Findings for Follow□Up." Genetic epidemiology 37.2 (2013): 205-213.

Moonesinghe, Ramal, et al. "Required sample size and nonreplicability thresholds for heterogeneous genetic associations." Proceedings of the National Academy of Sciences 105.2 (2008): 617-622.

Morris, Andrew P., et al. "Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes." Nature genetics 44.9 (2012): 981.

Musunuru, Kiran, et al. "From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus." Nature 466.7307 (2010): 714-719.

Neale, Benjamin M., et al. "Meta-analysis of genome-wide association studies of attention-deficit/hyperactivity disorder." Journal of the American Academy of Child & Adolescent Psychiatry 49.9 (2010): 884-897.

Nicolae, Dan L., et al. "Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS." PLoS genetics 6.4 (2010): e1000888.

Palmen, Jutta, et al. "The functional interaction on in vitro gene expression of APOA5 SNPs, defining haplotype APOA5□ 2, and their paradoxical association with plasma triglyceride but

not plasma apoAV levels." Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease 1782.7 (2008): 447-452.

Panagiotou, Orestis A., et al. "The power of meta-analysis in genome-wide association studies." Annual review of genomics and human genetics 14 (2013): 441-465.

Paré, Guillaume, et al. "Novel associations of CPS1, MUT, NOX4, and DPEP1 with plasma homocysteine in a healthy population." Circulation: Cardiovascular Genetics 2.2 (2009): 142-150.

Pickrell, Joseph K. "Joint analysis of functional genomic data and genome-wide association studies of 18 human traits." The American Journal of Human Genetics 94.4 (2014): 559-573.

Pirastu, Nicola, et al. "A Genome-Wide Association Study in isolated populations reveals new genes associated to common food likings." Reviews in Endocrine and Metabolic Disorders 17.2 (2016): 209-219.

Plyte, Simon E., et al. "Glycogen synthase kinase-3: functions in oncogenesis and development." Biochimica et Biophysica Acta (BBA)-Reviews on Cancer 1114.2-3 (1992): 147-162.

Price, Alkes L., et al. "Principal components analysis corrects for stratification in genome-wide association studies." Nature genetics 38.8 (2006): 904-909.

Rahalkar, Amit R., and Robert A. Hegele. "Monogenic pediatric dyslipidemias: classification, genetics and clinical spectrum." Molecular genetics and metabolism 93.3 (2008): 282-294.

Ripke, Stephan, et al. "Genome-wide association analysis identifies 13 new risk loci for schizophrenia." Nature genetics45.10 (2013): 1150-1159.

Rohwedder, Ina, et al. "Plasma fibronectin deficiency impedes atherosclerosis progression and fibrous cap formation." EMBO molecular medicine 4.7 (2012): 564-576.

Rossin, Elizabeth J., et al. "Proteins encoded in genomic regions associated with immune-mediated disease physically interact and suggest underlying biology." PLoS genetics 7.1 (2011): e1001273.

Sanna, Serena, et al. "Fine mapping of five loci associated with low-density lipoprotein cholesterol detects variants that double the explained heritability." PLoS genetics 7.7 (2011): e1002198.

Sarria, A. J., S. R. Panini, and R. M. Evans. "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 (1992): 19455-19463.

Schadt, Eric E., et al. "Mapping the genetic architecture of gene expression in human liver." PLoS biology 6.5 (2008): e107.

Schaub, Marc A., et al. "Linking disease associations with regulatory information in the human genome." Genome research 22.9 (2012): 1748-1759.

Schizophrenia Working Group of the Psychiatric Genomics Consortium. "Biological insights from 108 schizophrenia-associated genetic loci." Nature 511.7510 (2014): 421-427.

Schmidt, Ellen M., et al. "The Michigan Genomics Initiative: A Model Framework for Genetic Discovery Using Patient Electronic Health Records." GENETIC EPIDEMIOLOGY. Vol. 41. No. 7. 111 RIVER ST, HOBOKEN 07030-5774, NJ USA: WILEY, 2017.

Schunkert, Heribert, et al. "Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease." Nature genetics 43.4 (2011): 333-338.

Seddon, Johanna M., et al. "The US twin study of age-related macular degeneration: relative roles of genetic and environmental influences." Archives of ophthalmology 123.3 (2005): 321-327.

Segrè, Ayellet V., et al. "Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits." PLoS genetics 6.8 (2010): e1001058.

Seminara, Daniela, et al. "The emergence of networks in human genome epidemiology: challenges and opportunities." Epidemiology 18.1 (2007): 1-8.

Servin, Bertrand, and Matthew Stephens. "Imputation-based analysis of association studies: candidate regions and quantitative traits." PLoS genetics 3.7 (2007): e114.

Shi, Jianxin, et al. "Common variants on chromosome 6p22. 1 are associated with schizophrenia." Nature 460.7256 (2009): 753-757.

Solovieff, Nadia, et al. "Pleiotropy in complex traits: challenges and strategies." Nature Reviews Genetics 14.7 (2013): 483-495.

Soranzo, Nicole, et al. "A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the HaemGen consortium." Nature genetics41.11 (2009): 1182-1190.

Speliotes, Elizabeth K., et al. "Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index." Nature genetics 42.11 (2010): 937-948.

Stouffer, Samuel A., et al. "The American soldier: Adjustment during army life.(Studies in social psychology in World War II), Vol. 1." (1949).

Sudlow, Cathie, et al. "UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age." PLoS medicine 12.3 (2015): e1001779.

Sullivan, Patrick F. "The psychiatric GWAS consortium: big science comes to psychiatry." Neuron 68.2 (2010): 182-186.

Takeuchi, Hiroshi, et al. "Characterization of PXK as a protein involved in epidermal growth factor receptor trafficking." Molecular and cellular biology 30.7 (2010): 1689-1702.

Teslovich, Tanya M., et al. "Biological, clinical and population relevance of 95 loci for blood lipids." Nature 466.7307 (2010): 707-713.

Toker, Alex, and Lewis C. Cantley. "Signalling through the lipid products of phosphoinositide-3-OH kinase." Nature 387.6634 (1997): 673-676.

Trynka, Gosia, and Soumya Raychaudhuri. "Using chromatin marks to interpret and localize genetic associations to complex human traits and diseases." Current opinion in genetics & development 23.6 (2013): 635-641.

Uda, Manuela, et al. "Genome-wide association study shows BCL11A associated with persistent fetal hemoglobin and amelioration of the phenotype of β-thalassemia." Proceedings of the National Academy of Sciences 105.5 (2008): 1620-1625.

van Beekum, Olivier, et al. "The adipogenic acetyltransferase Tip60 targets activation function 1 of peroxisome proliferator-activated receptor γ." Endocrinology 149.4 (2007): 1840-1849.

Veyrieras, Jean-Baptiste, et al. "High-resolution mapping of expression-QTLs yields insight into human gene regulation." PLoS genetics 4.10 (2008): e1000214.

Visscher, Peter M., et al. "Five years of GWAS discovery." The American Journal of Human Genetics 90.1 (2012): 7-24.

Voight, Benjamin F., et al. "Plasma HDL cholesterol and risk of myocardial infarction: a mendelian randomisation study." The Lancet 380.9841 (2012): 572-580.

Voight, Benjamin F., et al. "The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits." PLoS genetics 8.8 (2012): e1002793.

Voight, Benjamin F., et al. "Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis." Nature genetics 42.7 (2010): 579-589.

Wallrapp, Christine, et al. "The product of the mammalian orthologue of the Saccharomyces cerevisiae HBS1 gene is phylogenetically related to eukaryotic release factor 3 (eRF3) but does not carry eRF3□like activity." FEBS letters 440.3 (1998): 387-392.

Wang, Chaolong, et al. "Improved ancestry estimation for both genotyping and sequencing data using projection procrustes analysis and genotype imputation." The American Journal of Human Genetics 96.6 (2015): 926-937.

Welch, Carrie L., et al. "Genetic regulation of cholesterol homeostasis: chromosomal organization of candidate genes." Journal of lipid research 37.7 (1996): 1406-1421.

Wen, Wanqing, et al. "Meta-analysis identifies common variants associated with body mass index in east Asians." Nature genetics 44.3 (2012): 307-311.

Wen, Xiaoquan. "Bayesian model selection in complex linear systems, as illustrated in genetic association studies." Biometrics 70.1 (2014): 73-83.

Willer, Cristen J., et al. "Newly identified loci that influence lipid concentrations and risk of coronary artery disease." Nature genetics 40.2 (2008): 161-169.

Willer, Cristen J., Yun Li, and Gonçalo R. Abecasis. "METAL: fast and efficient meta-analysis of genomewide association scans." Bioinformatics 26.17 (2010): 2190-2191.

Wood, Andrew R., et al. "Defining the role of common variation in the genomic and biological architecture of adult human height." Nature genetics 46.11 (2014): 1173-1186.

Wyne, Kathleen L., et al. "Expression of the VLDL receptor in endothelial cells." Arteriosclerosis, thrombosis, and vascular biology 16.3 (1996): 407-415.

Yang, Jian, et al. "Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits." Nature genetics 44.4 (2012): 369-375.

Yang, Jian, et al. "GCTA: a tool for genome-wide complex trait analysis." The American Journal of Human Genetics 88.1 (2011): 76-82.

Ye, Zhan, et al. "Phenome-wide association studies (PheWASs) for functional variants." European Journal of Human Genetics 23.4 (2015): 523-529.

Yu, XueJun, et al. "Hepatocyte growth factor protects endothelial progenitor cell from damage of low-density lipoprotein cholesterol via the PI3K/Akt signaling pathway." Molecular biology reports 37.5 (2010): 2423-2429.

Zeggini, E., & Ioannidis, J. P. (2009). Meta-analysis in genome-wide association studies. Pharmacogenomics, 10(2), 191-201.

Zheng, Biqiang, et al. "MicroRNA-148a suppresses tumor cell invasion and metastasis by downregulating ROCK1 in gastric cancer." Clinical Cancer Research 17.24 (2011): 7574-7583.

Zhou, Qiong L., et al. "A novel pleckstrin homology domain-containing protein enhances insulin-stimulated Akt phosphorylation and GLUT4 translocation in adipocytes." Journal of Biological Chemistry 285.36 (2010): 27581-27589.

Zhou, Wei, et al. "Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies." bioRxiv (2017): 212357.

Zhou, Xin, et al. "Altered expression of miR-152 and miR-148a in ovarian cancer is related to cell proliferation." Oncology reports 27.2 (2012): 447-454.

Zou, Chunbin, et al. "Lack of Fas antagonism by Met in human fatty liver disease." Nature medicine 13.9 (2007): 1078-1085.