Modern Statistical Methods for Genetics and Genomic Studies

HARVARD UNIVERSITY
Graduate School of Arts and Sciences

DISSERTATION ACCEPTANCE CERTIFICATE

The undersigned, appointed by the

Department of Biostatistics

have examined a dissertation entitled

"Modern Statistical Methods for Genetics and Genomic Studies"

presented by Xihao Li

candidate for the degree of Doctor of Philosophy and hereby certify that it is worthy of acceptance.

Signature Xihong Lin
Xihong Lin (Jan 11, 2021 14:27 EST)

Typed name: Prof. Xihong Lin

Signature JWMiller
Jeffrey Miller (Jan 12, 2021 15:24 EST)

Typed name: Prof. Jeffrey Miller

Signature Liming Liang
Liming Liang (Jan 12, 2021 21:35 EST)

Typed name: Prof. Liming Liang

Signature .....

Typed name:

Date: January 11, 2021

Modern Statistical Methods for Genetics and Genomic Studies

A Dissertation Presented

by

Xihao Li

to

The Department of Biostatistics

The Graduate School of Arts and Sciences

In Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

in the Subject of

Biostatistics

Harvard University

Cambridge, Massachusetts

January 2021

Dissertation advisor: Professor Xihong Lin Xihao Li

Modern Statistical Methods for Genetics and Genomic Studies

ABSTRACT

Recent scientific advances in genetics and genomic studies have enabled the characterization and prediction of functional genomic elements across the human genome, including biological evidence which assesses different aspects of functional consequences of genetic variants through a diverse set of in silico functional annotations; and genetic evidence which assesses how genetic variants are associated with complex phenotypes or traits from large-scale sequencing studies. In this dissertation, we present novel statistical methods that performs integrative analysis of data arising from these complementary lines of evidence to better understand the functional annotation landscape of coding and noncoding genetic variants and uncover the genetic architecture of human disease or traits.

In Chapter 1, we propose Multi-dimensional Annotation Class Integrative Estimation (MACIE), an unsupervised multivariate mixed model framework capable of integrating annotations of diverse origin to assess multi-dimensional functional roles for both coding and noncoding variants. MACIE effectively summarizes these diverse and complementary functional annotations into measures that can predict the multi-faceted biological functions of any given genetic variant, and thus provides richer and more interpretable information than existing one-dimensional scores in the presence of multiple aspects of functionality. Applied to a variety of

independent coding and non-coding datasets, MACIE demonstrates powerful and robust performance in discriminating between functional and non-functional variants. We also show an application of MACIE to fine-mapping using lipids GWAS summary statistics data from the European Network for Genetic and Genomic Epidemiology Consortium.

Large-scale whole genome sequencing (WGS) studies have enabled the analysis of rare variants (RVs) associated with complex phenotypes. Commonly used RV association tests (RVATs) have limited scope to leverage variant functions. In Chapter 2, we propose STAAR (variant-Set Test for Association using Annotation infoRmation), a scalable and powerful RVAT method that effectively incorporates both variant categories and multiple complementary annotations using a dynamic weighting scheme. STAAR accounts for population structure and relatedness, and is scalable for analyzing very large cohort and biobank WGS studies of continuous and dichotomous traits. We apply STAAR to identify RVs associated with four lipid traits using data from the Trans-Omics for Precision Medicine (TOPMed) program. We discover and replicate novel RV associations, including disruptive missense RVs of *NPC1L1* and an intergenic region near *APOC1P1* associated with low-density lipoprotein cholesterol.

Meta-analysis of WGS studies has provided an exciting solution to leverage large sample sizes for the discovery of coding and noncoding RVs associated with complex human traits. Existing RV meta-analysis approaches are not scalable when applied to WGS data due to the very large number of RVs whose summary-level information needs to be stored and shared. In Chapter 3, we extend the method in Chapter 2 and propose MetaSTAAR as a powerful and resource-efficient RV meta-analysis framework scalable to large cohort and biobank WGS studies with

hundreds of millions of RVs across the genome, while accounting for relatedness and population structure for both quantitative and dichotomous traits. Through meta-analysis of four lipid traits from 14 studies of the TOPMed program, we demonstrate that MetaSTAAR performed resource-efficient RV meta-analysis at scale and identified several conditionally significant RV associations with lipids.

TABLE OF CONTENTS

ABSTRACT iii
TABLE OF CONTENTS vi
LIST OF TABLES viii
LIST OF FIGURES ix
ACKNOWLEDGEMENTS xi
CHAPTER I 1
Abstract 1
Introduction 2
Results 5
Discussion 19
Methods 22
CHAPTER II 27
Abstract 27
Introduction 28
Results 30
Discussion 46
Methods 49
CHAPTER III 66
Abstract 66
Introduction 67
Results 69
Discussion 81
Methods 84
REFERENCES 93

APPENDIX A **106**
APPENDIX B **112**
APPENDIX C **139**

LIST OF TABLES

Table 2.1 Gene-centric analysis results of both unconditional analysis and analysis conditional on known common and low-frequency variants. 42

Table 2.2 Genetic region (2-kb sliding window) analysis results of both unconditional analysis and analysis conditional on known common and low-frequency variants. 43

Table 3.1 Comparison of runtimes and storage of MetaSTAARWorker and RareMetalWorker. 74

Table 3.2 Gene-centric meta-analysis results of both unconditional analysis and analysis conditional on known common and low-frequency variants. 76

LIST OF FIGURES

Figure 1.1 Heatmap demonstrating the correlation between individual and integrative functional scores for ClinVar pathogenic and benign noncoding variants. 4

Figure 1.2 ROC curves comparing the performances of MACIE and other functional scores in discriminating between ClinVar pathogenic and benign nonsynonymous coding variants. 10

Figure 1.3 ROC curves comparing the performances of MACIE and other functional scores in discriminating between loss-of-function (LOF) and functional (FUNC) nonsynonymous coding variants within 13 exons that encode functionally critical domains of *BRCA1* based on saturation genome editing (SGE) data. Here the LOF class is our putative functional class and the FUNC class is our putative non-functional class..... 12

Figure 1.4 ROC curves comparing the performances of MACIE and other functional scores in discriminating between a, CAGE identified promoters and non-promoters and b, CAGE identified enhancers and non-enhancers among noncoding variants from 1000 Genomes Project Phase 3 data. For CAGE Enhancer predictions, LINSIGHT was excluded as it uses the FANTOM5 enhancer label as one of the genomic features in building the LINSIGHT score. 14

Figure 1.5 ROC curves comparing the performances of MACIE and other functional scores for prediction of a, validated regulatory variants in lymphoblastoid cell lines (LCLs) from massively parallel reporter assays (MPRAs) and b, dsQTLs identified using DNase I sequencing data in LCLs against control variants..... 16

Figure 1.6 LocusZoom plot (41) for GWAS associations of LDL-C at the *APOE* locus. The lipids GWAS summary statistics were from the European Network for Genetic and Genomic Epidemiology (ENGAGE) Consortium (*n* = 58,381) (40). 19

Figure 2.1 STAAR workflow. 31

Figure 2.2 Correlation heatmap of functional annotation scores. 32

Figure 2.3 Genetic region (2-kb sliding window) unconditional analysis results of LDL-C in the discovery phase using the TOPMed cohort. 41

Figure 3.1 MetaSTAAR workflow. 70

Figure 3.2 Genetic region (2-kb sliding window) unconditional meta-analysis results of LDL-C using the TOPMed data. 79

ACKNOWLEDGEMENTS

First and foremost, I would like to express my deepest gratitude to my advisor, Professor Xihong Lin, for your enduring patience and unwavering support of my doctoral journey in every aspect. You have brought me into the field of biostatistics since I started as a graduate student in the Department of Biostatistics. During the past years, your passion and insights have not only paved the way for me to become a keen researcher in biostatistics and statistical genetics, but also shaped me into a scholar with methodological rigor, scientific knowledge, and critical thinking. Your understanding, kindness, and encouragement have always enlightened me whenever I was facing difficulties in my research and beyond. You will always be my role model and I would never be able to arrive where I am without your guidance.

I would like to sincerely thank my dissertation committee, Professor Liming Liang and Professor Jeffrey Miller, for your helpful insights and research ideas, support and encouragement during my committee meetings and much more. I would like to truly thank Zilin Li, Godwin Yung, Hufeng Zhou, Sheila Gaynor, Ryan Sun, Corbin Quick, Rounak Dey, Han Chen, Yaowu Liu, Zhonghua Liu, Haoyu Zhang, and all members of Lin Lab for your helpful suggestions and discussions. I would also like to thank Drs. Pradeep Natarajan, Gina Peloso, Jerome Rotter, Cristen Willer, and other investigators from the Trans-Omics of Precision Medicine Program Lipids Working Group for providing technical, material, and administrative support. I am so fortunate to have the opportunity learning, working, and collaborating with you all.

My sincere gratitude to Professors Marcello Pagano, Lee-Jen Wei, Garrett Fitzmaurice, David Wypij, Paige Williams, and Rajarshi Mukherjee for giving me the opportunities to serve as a

teaching assistant and summer course instructor; to all of the faculty members in the Department of Biostatistics for your professional and personal help; and to Jelena Follweiler, Trevor Bierig, and the wonderful staff for always being so friendly and supportive. I would like to gratefully thank Derek Shyr, Tom Chen, Xiao Wu, Larry Han, Molei Liu, Siyuan Ma, Boyu Ren, and my fellow graduate students for our friendship.

Finally, and most importantly, I would like to extend my heartfelt gratitude to my family - my parents, Hong Zhang and Jinzhong Li, my grandmother Yaping Yang, my wife Shijia Bian, and my parents-in-law Ping Li and Wenbing Bian for your endless love, company, blessings, and encouragement every single day. Thank you, mom and dad, for raising me and teaching me to be a better person. Thank you, Shijia and Arya, for making my life so beautiful and bright. Whenever I was having a hard time, you would always be by my side. Without your unconditional love and support, I would never reach this far and fulfill my Ph.D. journey.

To my beloved family.

CHAPTER I

A Multi-dimensional integrative scoring framework for predicting functional variants in the human genome

Xihao Li, Godwin Yung, Hufeng Zhou, Ryan Sun, Zilin Li, Yaowu Liu, Iuliana Ionita-Laza and Xihong Lin

Abstract

Attempts to identify and prioritize functional DNA elements in coding and noncoding regions, particularly through use of in silico functional annotation data, continue to increase in popularity. However, specific functional roles may vary widely from one variant to another, making it challenging to summarize different aspects of variant function. Here we propose Multi-dimensional Annotation Class Integrative Estimation (MACIE), an unsupervised multivariate mixed model framework capable of integrating annotations of diverse origin to assess multi-dimensional functional roles for both coding and noncoding variants. Unlike existing one-dimensional scoring methods, MACIE views variant functionality as a composite attribute encompassing multiple different characteristics, and estimates the joint posterior functional probability vector of each genomic position, a quantity that offers richer and more interpretable information in the presence of multiple aspects of functionality. Applied to a variety of independent coding and non-coding datasets, MACIE demonstrates powerful and robust performance in discriminating between functional and non-functional variants. We also show an application of MACIE to fine-mapping using lipids GWAS summary statistics data from the European Network for Genetic and Genomic Epidemiology Consortium.

Introduction

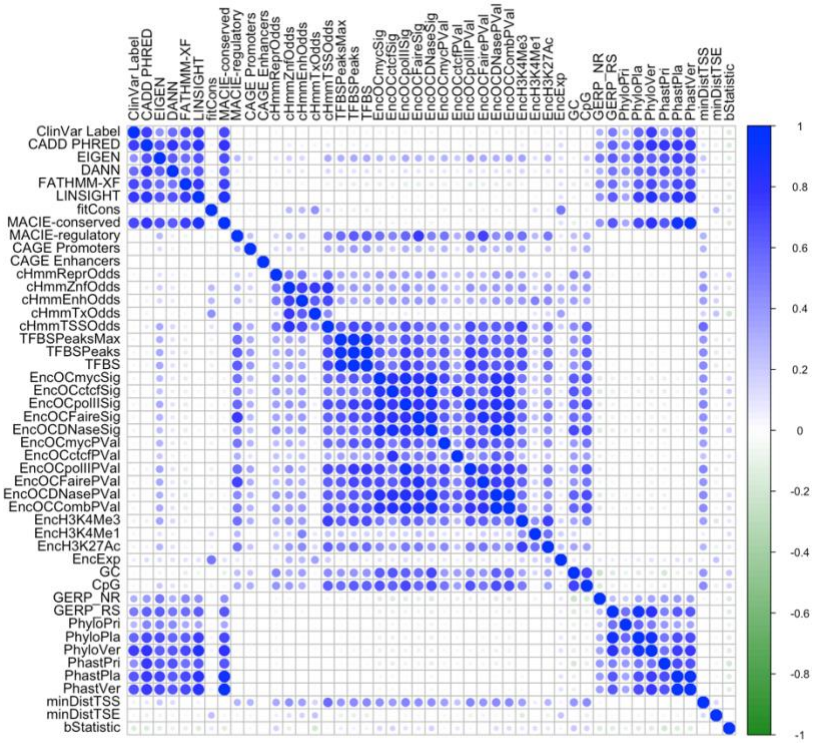
Recent scientific advances have enabled the identification of functional genomic elements through a diverse set of functional annotations, including proteins functional scores (1, 2), evolutionary conservation scores (3-5), and epigenetics scores from the Encyclopedia of DNA Elements (ENCODE) (6). Other initiatives such as the Roadmap Epigenomics project (7) and FANTOM5 project (8, 9) also provide evidence for potential regulatory variants in the human genome. Although different functional annotations capture different aspect of variant function, yet they provide complementary information on each other (10). Thus, to achieve a comprehensive understanding of the biological function of genomic variants, multi-faceted information from different functional annotations should be integrated simultaneously. However, it remains unclear how to summarize these diverse functional annotations in an insightful and interpretable manner.

Current algorithmic scoring frameworks utilize a variety of statistical and machine-learning methods to aggregate information from multiple sources of individual annotations into one-dimensional scores to measure functional impact of genetic variants. Supervised tools such as CADD (11), DANN (12), GWAVA (13), FATHMM-MKL (14), and FATHMM-XF (15) build machine learning classifiers on training sets with pre-labeled functional statuses, e.g., fine-mapped pathogenic or disease-associated variants labeled against benign or neutral variants. Such supervised approaches rely strongly on the quality of labels in the training set. Therefore, they may demonstrate suboptimal performance when inaccurate or biased labels are used. Unsupervised methods such as EIGEN (16), GenoCanyon (17), PINES (18), and FUN-LDA (19)

do not depend on labeled training data. They possess advantages in studying non-coding regions, where our current lack of knowledge often precludes gold-standard training data labels. A third group of methods including fitCons (20) and LINSIGHT (21) use evolution-based approaches that characterize the potential effect of natural selection at each genomic location using polymorphism and divergence data. Recent reviews provide a more detailed discussion of available functional annotation tools (22, 23).

Although existing methods attempt to integrate functional annotations through various approaches, to the best of our knowledge, these methods all summarize the annotation information with a single rating. In doing so, they implicitly assume that variant function can be described along a single axis, with variants being more functional on one end of the axis and less functional on the other end. This assumption may be reasonable if interest lies in predicting a specific aspect of variant function (e.g. regulatory behavior) and all annotations used as input are intended to predict that same aspect. However, if multiple aspects of variant function are simultaneously of interest, then it is unclear how to interpret the one-dimensional consolidation of annotations measuring different aspects of function, especially when these annotations appear to provide orthogonal information, e.g., weak correlation between evolutionary conservation scores and regulatory scores (Figure 1.1). Therefore, it is of interest to construct multi-dimensional integrative scores capable of capturing multiple facets of variant function simultaneously.

Figure 1.1 Heatmap demonstrating the correlation between individual and integrative functional scores for ClinVar pathogenic and benign noncoding variants.



In this chapter we propose Multi-dimensional Annotation Class Integrative Estimation (MACIE), an unsupervised multivariate mixed model framework capable of synthesizing multiple categories of annotations and producing interpretable multi-dimensional integrative scores. Instead of a single rating, MACIE explicitly defines variant function as a vector of latent binary outcomes, each outcome capturing functionality corresponding to a specific class of annotations. Correlations within and between the different classes of annotations are explicitly modeled, another advancement over existing methods. Using the Expectation-Maximization algorithm, MACIE calculates the joint posterior probability vector of a genomic position being functional (Methods).

Because of its multivariate formulation, MACIE is able to provide detailed and nuanced assessments of variant functionality. Output from MACIE is highly interpretable due to the specificity allowed by multiple functional classes. Additionally, the MACIE framework allows for considerable versatility to incorporate data in a manner that is most biologically relevant to the scientific question of interest. We apply MACIE to multiple independent coding and noncoding testing sets and show that, compared to current state-of-the-art integrative scores, MACIE consistently provides robust and best or near best performance in discriminating between functional and non-functional variants.

Results

Construction of MACIE training sets

MACIE scores were computed for a, nonsynonymous coding and b, noncoding and synonymous coding variants separately because the two types of variants are expected to have highly different functional profiles (16). All nonsynonymous coding annotations and some noncoding and synonymous coding annotations were downloaded from EIGEN. The remaining noncoding and synonymous coding annotations were downloaded from CADD full database (11) v1.3.

Nonsynonymous coding variants

For the nonsynonymous coding training set, we randomly extracted 10% of the variants with a match in the dbNSFP database (24). This database excludes synonymous variants that fall in coding regions but do not alter protein function. Only one unique variant per position was selected, and variants residing in sex chromosomes X and Y were removed to mitigate potential sources of bias. The final set included approximately 2.2 million variants. For each variant in the

training set, four protein substitution damage scores (SIFT (1), PolyPhenDiv, PolyPhenVar (2), Mutation Assessor (25)) and eight evolutionary conservation scores (GERP_NR and GERP_RS(5); PhyloP primate (PhyloPri), placental mammal (PhyloPla), and vertebrate (PhyloVer)(4); PhastCons primate (PhastPri), placental mammal (PhastPla), and vertebrate (PhastVer)(3)) were extracted from the EIGEN database (16). Thus we defined the two-class MACIE model ($M = 2$) for nonsynonymous coding variants to assess damaging protein coding function and evolutionarily conserved function. Full information on the MACIE model for nonsynonymous coding variants and the list of individual functional scores are given in Methods and Supplementary Table 1.1.

Noncoding and synonymous coding variants

For the noncoding and synonymous coding training set, we extracted a random sample comprising 10% of the variants in the 1000 Genomes Project dataset that were located within 500 base pairs (bp) upstream of a gene start site and did not possess a match in dbNSFP. Duplicated variants with multiple alternative alleles and variants in sex chromosomes X and Y were again removed to mitigate potential bias. The final training set included 36,431 variants. For each variant in the training set, the same eight evolutionary conservation scores used for coding variants were extracted from the EIGEN full database (16). A total of twenty-eight transformed epigenetic scores were additionally extracted from the CADD database (11) v1.3, including a collection of regulatory annotations from the ENCODE Project (6), three transcription factor binding site scores, GC content, CpG content, five chromatin state probabilities derived from the 15 state ChromHMM model (26), a background selection score (27), and physical distance metrics (11). We then defined the two-class MACIE model ($M = 2$)

for noncoding and synonymous coding variants to assess evolutionarily conserved function and epigenetic regulatory function. Full information on the MACIE model for noncoding and synonymous coding variants and the list of individual functional scores are given in Methods and Supplementary Table 1.1. Detailed information on pre-processing steps for the epigenetic scores are given in Supplementary Table 1.2.

Benchmarking the performance of MACIE with other integrative scoring methods

We compared the predictive performance of MACIE against existing state-of-the-art variant classifiers including CADD (11), FATHMM-XF (15), EIGEN (16), fitCons (20), LINSIGHT (21), and DANN (12) over a range of realistic variant assessment scenarios. Specifically, we assessed the ability of each score to identify clinically significant variants from ClinVar (28, 29); loss-of-function variants in the *BRCA1* gene uncovered through saturation genome editing (SGE) (30); promoters and enhancers from the FANTOM5 project defined by cap analysis of gene expression (CAGE) (8, 9); and experimentally verified functional variants from massive parallel reporter assays (MPRA) (31, 32). Some alternative scoring methods were excluded due to difficulties related to providing a proper comparison of results. For example, LINSIGHT is designed to predict the deleteriousness of noncoding variants, so we did not include it in the comparison for nonsynonymous coding variants.

Distribution of posterior probabilities for noncoding and synonymous coding variants in the training set

In Supplementary Table 1.3 we provide the posterior probabilities of each functional class averaged across all the noncoding and synonymous coding variants in the training set. The

predicted MACIE score for a given variant can be interpreted as the posterior probability of that variant belonging to (0,0), neither conserved nor regulatory classes; (1,0), the conserved but not the regulatory class; (0,1), the regulatory but not the conserved class; and (1,1), both conserved and regulatory classes. The four MACIE scores necessarily sum up to 1. A chi-squared test comparing observed and expected percentages under independence of evolutionary conservation and regulatory classes gives a significant *P* value of less than 2.2×10^{-16}, suggesting that the two classes are correlated. Since the observed percentage of functional variants that belong to (1,1) is statistically significantly greater than the expected percentage under independence (3.15% > 1.96%), we find strong evidence of enrichment of regulatory activity in conserved regions. Additionally, the MACIE model for noncoding and synonymous coding variants estimates that 8.05% and 24.34% of the variants show evolutionarily conserved and regulatory functionality, respectively. This is consistent with the prediction from LINSIGHT and other previous studies that approximately 7% - 9% of noncoding sites are under evolutionary constraint (21, 33), as well as an estimated upper bound of 25% of the functional fraction within the human genome (34).

ClinVar pathogenic and benign variants

We first validated our methods on a testing set consisting of all variants recorded in the ClinVar database (28, 29). Variant effect predictor (VEP) information was extracted from GENCODE (35) and used to separate nonsynonymous coding variants from noncoding and synonymous coding variants in ClinVar. The two MACIE models described above were then applied to the respective partitions. We combined the ClinVar categories “pathogenic” and “likely pathogenic” into a single pathogenic class and treated these variants as the putatively functional class.

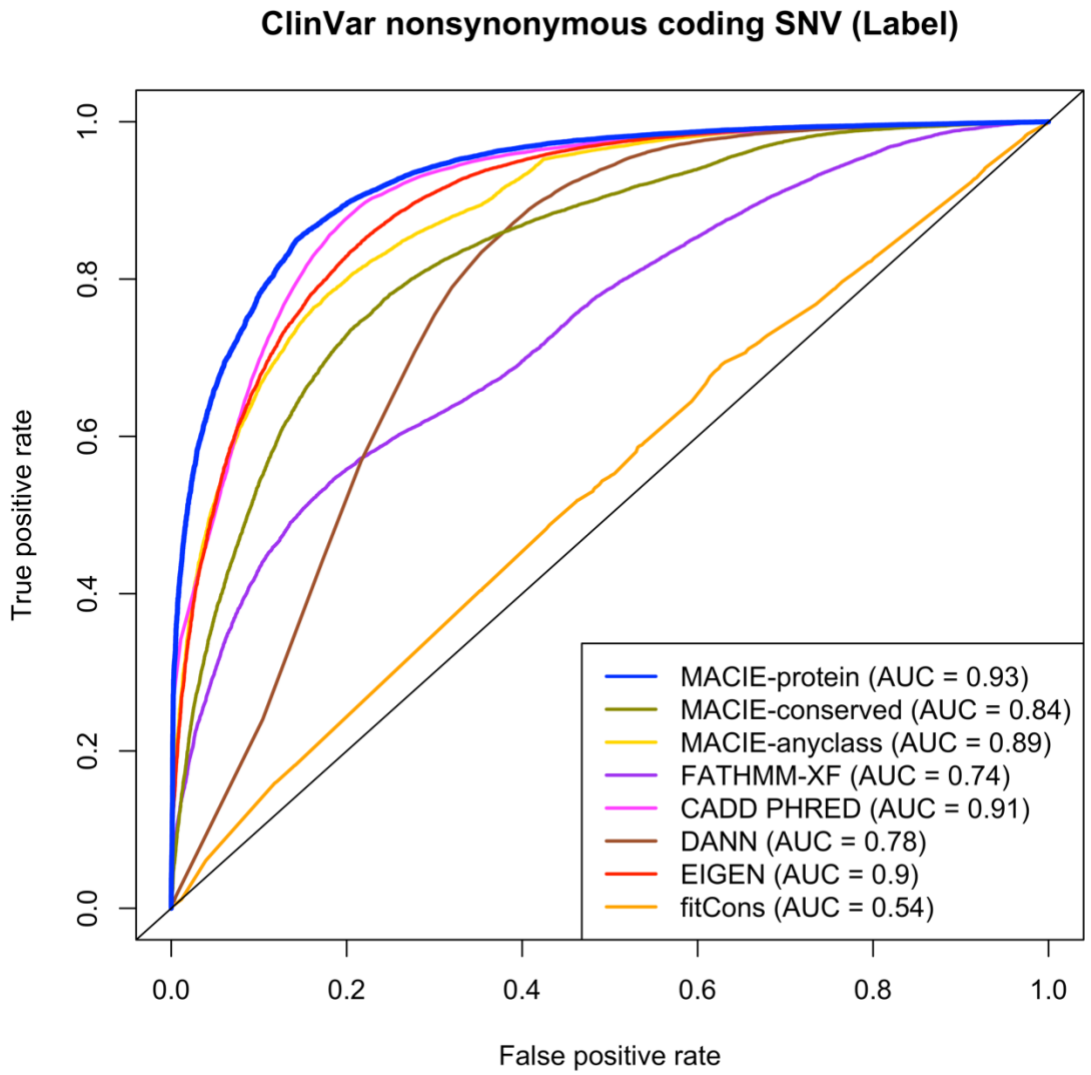
Similarly, we combined the ClinVar categories “benign” and “likely benign” into a single benign class and treated these variants as the putatively non-functional class. The remaining variants were categorized as having uncertain significance.

We first tested MACIE’s ability to distinguish pathogenic variants ($n = 33{,}714$) from their benign counterparts ($n = 14{,}410$) among ClinVar nonsynonymous variants through two approaches. First, we calculated two marginal MACIE scores: a, MACIE-damaging protein function score (denoted by MACIE-protein) as the sum of the posterior probabilities of “damaging protein functional/not conserved” and “damaging protein functional/conserved”; b, MACIE-conserved score as the sum of the posterior probabilities of “damaging protein functional/conserved” and “not damaging protein functional/conserved”. We also considered the posterior probability of either damaging protein functional or conserved (denoted by MACIE-anyclass) by summing the posterior probabilities corresponding to at least one functional class. This example illustrates the versatility of MACIE’s posterior probability outputs, which can be summed to form new probability measures with various informative interpretations depending on the specific needs of each analysis.

Figure 1.2 provides the receiver operating characteristic (ROC) curves and area under the curves (AUC) for the three MACIE approaches and seven one-dimensional scores for ClinVar nonsynonymous variants. Of the methods considered, MACIE-damaging protein function score delivered the highest discrimination power (AUC = 0.93), followed by CADD (AUC = 0.91), EIGEN (AUC = 0.90), and MACIE-anyclass (AUC = 0.89). These four methods substantially outperformed the supervised DANN (AUC = 0.78), the supervised FATHMM-XF (AUC =

0.74), and the evolution-based fitCons (AUC = 0.54). Similar results were observed when distinguishing between pathogenic missense (as opposed to all nonsynonymous) variants ($n = 21{,}409$) from their benign counterparts ($n = 14{,}035$) in ClinVar (Supplementary Figure 1.1).

Figure 1.2 ROC curves comparing the performances of MACIE and other functional scores in discriminating between ClinVar pathogenic and benign nonsynonymous coding variants.



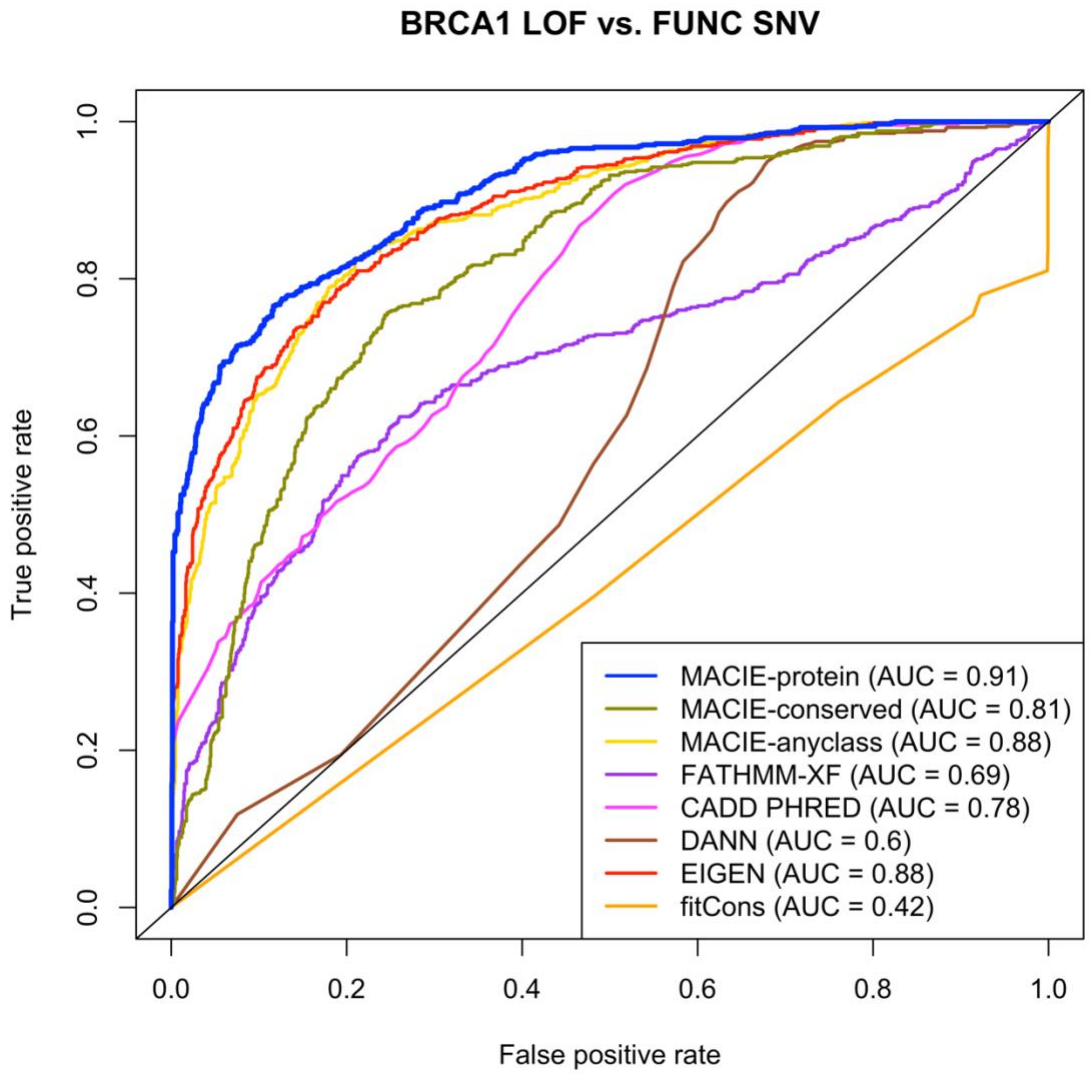
Next, we identified 40,109 noncoding variants from ClinVar database in total, including 6,551 pathogenic variants, and 33,558 benign variants. For these noncoding variants, we chose to

calculate a marginal MACIE-conserved score, as ClinVar pathogenic noncoding variant labels track closely with evolutionary conservation scores (Figure 1.1). ROC curves and AUCs for discriminating between the pathogenic and benign variants are provided in Supplementary Figure 1.2. MACIE-conserved score showed comparable performance (AUC = 0.95) to FATHMM-XF score, which showed the highest discrimination power (AUC = 0.97). The outperformance of FATHMM-XF in this specific example should be expected because FATHMM-XF is a supervised machine-learning method trained on labels that bear many similarities to the labels defined in ClinVar, while MACIE is an unsupervised method. We performed Wilcoxon rank-sum tests to compare the distribution of integrative scores between ClinVar pathogenic and benign noncoding variants for each method. The Wilcoxon test *P* values for both FATHMM-XF and MACIE-conserved scores were less than 2.2×10^{-308}, representing high discriminative abilities for each score. MACIE-conserved score substantially outperformed the unsupervised method EIGEN (AUC = 0.84) and the evolution-based method fitCons (AUC = 0.55).

Loss-of-function nonsynonymous coding variants in *BRCA1*

We evaluated MACIE's performance in predicting the deleteriousness of nonsynonymous coding variants located within 13 exons that encode functionally critical domains of *BRCA1*. A two-component Gaussian mixture model was fit based on the saturation genome editing function scores to classify all *BRCA1* variants as loss-of-function (LOF), intermediate (INT), or functional (FUNC), in a decreasing order of severity (30). Thus, FUNC corresponds to benign variants in this experiment. We selected reported LOF nonsynonymous coding variants ($n = 674$) as the putative functional set and designated FUNC nonsynonymous coding variants ($n = 1{,}443$) as the putative non-functional set. Among all the methods compared (Figure 1.3), MACIE-

Figure 1.3 ROC curves comparing the performances of MACIE and other functional scores in discriminating between loss-of-function (LOF) and functional (FUNC) nonsynonymous coding variants within 13 exons that encode functionally critical domains of *BRCA1* based on saturation genome editing (SGE) data. Here the LOF class is our putative functional class and the FUNC class is our putative non-functional class.



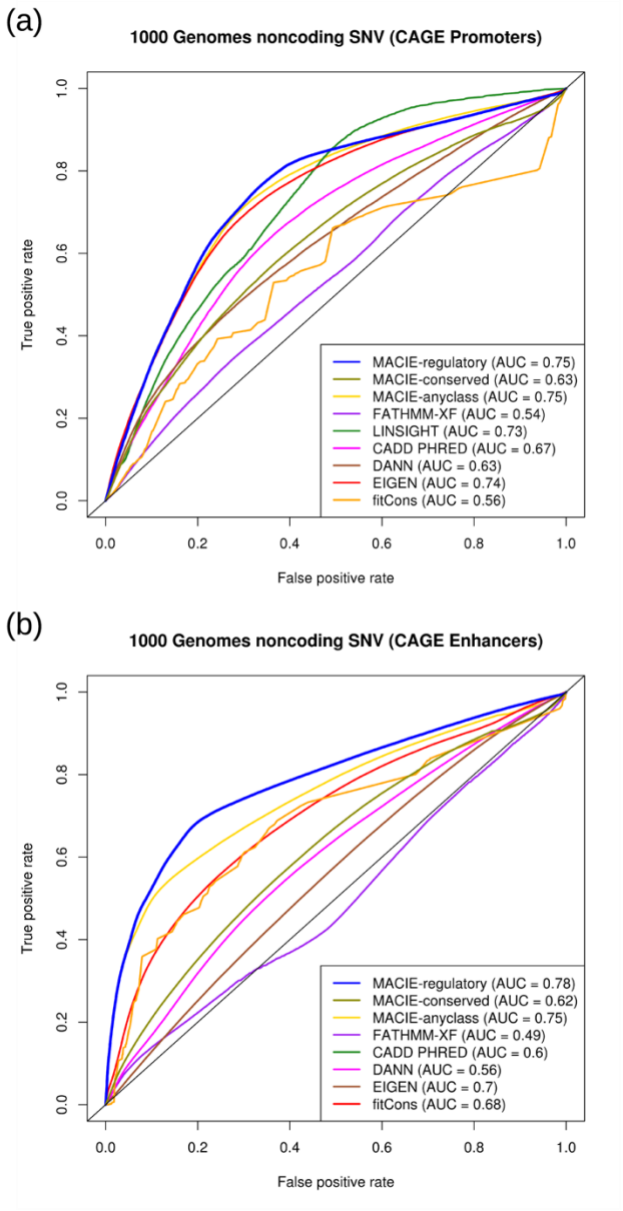
damaging protein function score showed the highest predictive power (AUC = 0.91), followed by EIGEN (AUC = 0.88) and MACIE-anyclass (AUC = 0.88). The top three scores were much more powerful than CADD (AUC = 0.78), FATHMM-XF (AUC = 0.69), DANN (AUC = 0.60) and fitCons (AUC = 0.42). The Wilcoxon test *P* value for MACIE-damaging protein function

score was the lowest ($P = 7.60 \times 10^{-203}$), and was orders of magnitude smaller than EIGEN ($P = 7.22 \times 10^{-179}$), CADD ($P = 1.81 \times 10^{-95}$) and other integrative scores. We observed similar results when distinguishing between *BRCA1* LOF nonsynonymous coding variants ($n = 674$) and ClinVar benign nonsynonymous coding variants ($n = 14{,}410$) (Supplementary Figure 1.3).

FANTOM5 CAGE-defined promoters and enhancers among 1000 Genomes noncoding variants

We tested the ability of MACIE to identify promoter regions defined by the cap analysis of gene expression conducted during the FANTOM5 project (8, 9). A total of 110,895 out of approximately 80 million noncoding variants from the 1000 Genomes Project Phase 3 data (36) were mapped to such regions and therefore labeled as CAGE promoters. For each identified CAGE promoter variant, we used the 1000 Genomes Project database to randomly select a matched control variant (non-promoter) that possessed the same minor allele frequency (MAF) and same minimum distance to any gene transcription start site that was located at least 500 kilobase (kb) away from the promoter variant, yielding a total number of 97,298 variants in the control set (it was not possible to find a matched control for each CAGE variant). Similar to the previous analysis, we calculated a marginal MACIE-regulatory score by summing the two probabilities corresponding to the regulatory class (denoted by MACIE-regulatory). ROC curves and AUCs for discriminating between CAGE promoters and non-promoters are provided in Figure 1.4a. MACIE-regulatory and MACIE-anyclass scores showed the highest discrimination power (AUC = 0.75), followed by EIGEN with AUC = 0.74. The Wilcoxon test *P* value for MACIE-regulatory score was less than 2.2×10^{-308}, indicating high discrimination ability.

Figure 1.4 ROC curves comparing the performances of MACIE and other functional scores in discriminating between a, CAGE identified promoters and non-promoters and b, CAGE identified enhancers and non-enhancers among noncoding variants from 1000 Genomes Project Phase 3 data. For CAGE Enhancer predictions, LINSIGHT was excluded as it uses the FANTOM5 enhancer label as one of the genomic features in building the LINSIGHT score.



FATHMM-XF (AUC = 0.54) and fitCons (AUC = 0.56) scores performed poorly due to the inability of these one-dimensional scores to capture epigenetic functionality. We also performed a similar analysis by contrasting CAGE-identified enhancers ($n = 520{,}987$) versus non-

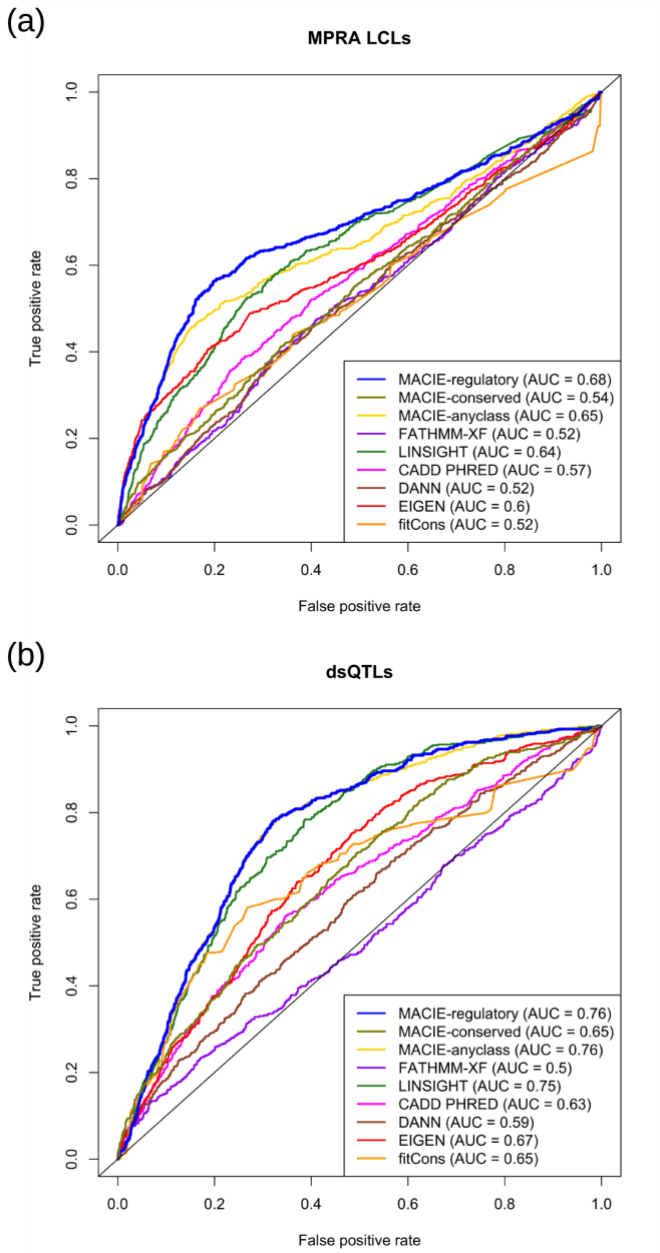
enhancers ($n = 448{,}253$) using noncoding variants from the 1000 Genomes Project. The results were similar, with MACIE-regulatory score displaying the highest predictive power and significantly outperforming all other state-of-the-art methods (Figure 1.4b).

MPRA validated variants and dsQTLs in lymphoblastoid cell lines

We examined the performance of MACIE for predicting cell type/tissue-specific regulatory variants using test sets from the massively parallel reporter assay. The MPRA dataset included validated regulatory variants in lymphoblastoid cell lines (LCLs) (31). We paired each positive variant ($n = 693$) with four control variants from MPRA where neither allele showed significant differential expression at a Bonferroni corrected *P* value threshold of 0.1 ($n = 2{,}772$) (37). Figure 1.5a shows that MACIE-regulatory score produced the highest discrimination power (AUC = 0.68), outperforming the second-best performing method (LINSIGHT, AUC = 0.64).

Finally, we evaluated the performance of our proposed method on a collection of dsQTLs that were identified using DNase I sequencing data from human lymphoblastoid cell lines (38). Variants possessing association *P* values less than 1×10^{-5} and residing within 100 bp of their corresponding DNase I-hypersensitive sites were chosen as the putatively functional set ($n = 560$) (39). The control set of variants was randomly selected from a larger set of common variants (MAF > 5%) falling in the top 5% of DNase I sensitivity sites used to identify dsQTLs in the original study ($n = 2{,}240$). We observed that MACIE-regulatory score exhibited a larger AUC (AUC = 0.76) than all other methods (Figure 1.5b). MACIE-anyclass score also delivered robust performance on MPRA validated and dsQTLs datasets.

Figure 1.5 ROC curves comparing the performances of MACIE and other functional scores for prediction of a, validated regulatory variants in lymphoblastoid cell lines (LCLs) from massively parallel reporter assays (MPRAs) and b, dsQTLs identified using DNase I sequencing data in LCLs against control variants.



In summary, MACIE consistently ranked as one of the most powerful, robust and interpretable methods across a variety of settings and scientific questions. Our results show that while one-dimensional scores have gaps in coverage, a multi-dimensional scoring method offers robust and

interpretable predictive performance. The ability of MACIE to interrogate variant functionality from multiple perspectives, at a level that is highly competitive with or better than state of the art methods, is unmatched by existing integrative functional scoring methods.

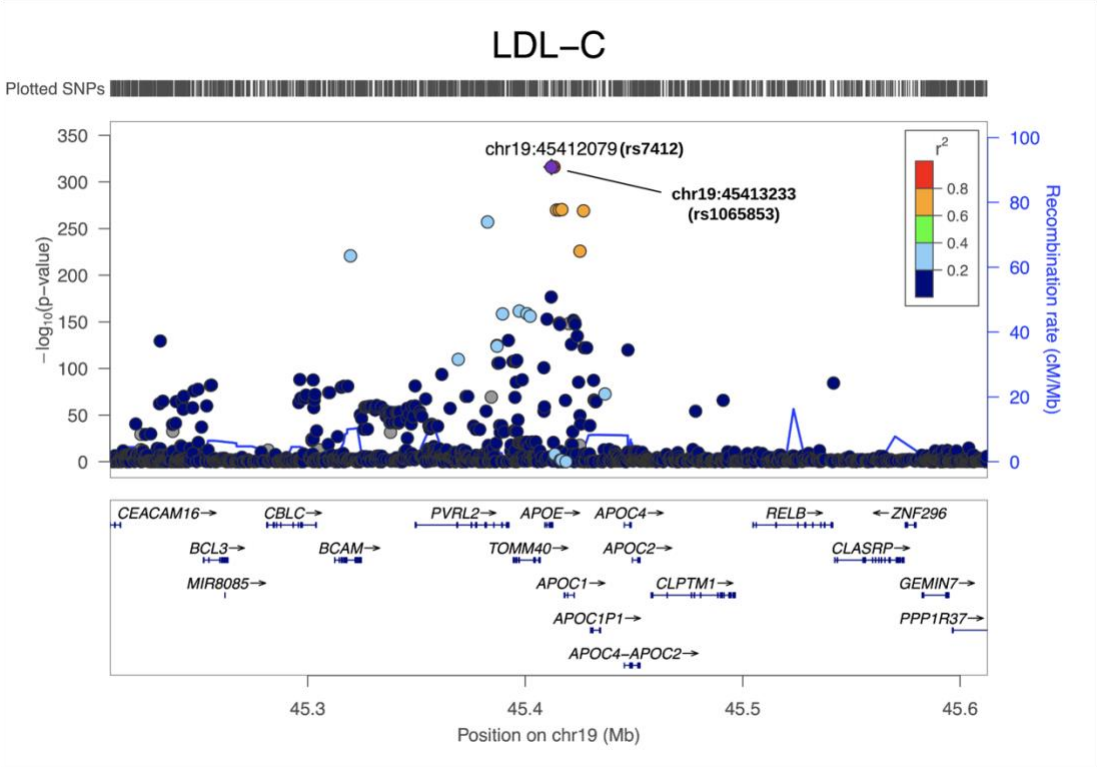
MACIE prioritizes functional variants using lipids GWAS data

To illustrate the utility of MACIE scores in identifying plausible functional causal variants in genetic association studies, we applied MACIE to the publicly available lipids GWAS data from the European Network for Genetic and Genomic Epidemiology (ENGAGE) Consortium (40). This dataset consists of lipids GWAS summary statistics for 9.6 million single nucleotide variants (SNVs) across 62,166 samples (Supplementary Table 1.4). We focused on genome-wide significant ($P < 5 \times 10^{-8}$) SNVs associated with low-density lipoprotein cholesterol (LDL-C), high-density lipoprotein cholesterol (HDL-C), triglycerides (TG), and total cholesterol (TC). In total, we found 8, 9, 6, and 11 nonsynonymous coding SNVs that were predicted to belong to the protein damaging class with probability greater than 0.9 for LDL-C, HDL-C, TG, TC, respectively; 640, 377, 322, and 846 synonymous or noncoding SNVs that were predicted to belong to the regulatory class with probability greater than 0.9; 50, 64, 39, and 61 SNVs that were predicted to belong to the evolutionarily conserved class with probability greater than 0.9; and 9, 8, 10, 12 SNVs that were predicted to belong to both evolutionarily conserved and regulatory class with probability greater than 0.9 (Supplementary Tables 1.5-1.12). Compared to the total number of marginally significant SNVs for each trait (Supplementary Table 1.4), the MACIE scores reduce the number of SNVs prioritized for follow-up by an order of magnitude, saving much cost and effort in effectively pinpointing SNVs with relevant biological function.

For example, for LDL-C, the single most significant SNV was rs7412 (chr19:45412079 C/T; $P < 1 \times 10^{-316}$). We predicted this known common missense SNV to be functional, as both MACIE-protein and MACIE-conserved scores provided a prediction greater than 0.95. These predictions highlight the multiple functional roles of this SNV. It is also worth noticing that the second most significant SNV rs1065853 (chr19:45413233 G/T; $P < 1 \times 10^{-316}$) is in extremely high linkage disequilibrium (LD) with the leading SNV rs7412 (Figure 1.6). MACIE scores indicate that rs1065853 (upstream variant of *APOC1*) may possess a regulatory role since its MACIE-regulatory score is greater than 0.99, possibly suggesting that both the missense and regulatory variants can be putatively causal in affecting LDL-C levels. Similar results were observed for TC (Supplementary Figure 1.4). For HDL-C, although the single most significant SNV was rs17231506 (chr16:56994528 C/T; $P = 6.88 \times 10^{-316}$), the MACIE prediction was less than 0.01 for both classes. By scanning across the *CETP* locus and nearby noncoding regions associated with HDL-C, we found that two SNVs, rs72786786 (chr16:56985514 G/A; $P = 2.52 \times 10^{-253}$) and rs12720926 (chr16:56998918 A/G; $P = 1.89 \times 10^{-260}$), both under moderate to high LD with the leading SNV (Supplementary Figure 1.5), possess a MACIE-regulatory score greater than 0.99. These two SNVs may be more functionally important than rs17231506 and may provide more information regarding risk-perturbing biological mechanisms associated with this locus and can be prioritized for functional follow-up. For TG, there is also a lack of functional evidence for the leading SNV rs964184 (chr11:116648917 G/C; $P = 1.74 \times 10^{-157}$) in the *APOA1/C3/A4/A5* gene cluster region. However, a SNV rs2075290 (chr11:116653296 C/T; $P = 2.13 \times 10^{-103}$) in moderate LD with rs964184 at this locus has a MACIE-regulatory score of 0.88 (Supplementary Figure 1.6). These examples illustrate how

MACIE scores can be used to supplement previous literature and provide additional information to aid prioritization of putatively functional causal variants for functional follow-up.

Figure 1.6 LocusZoom plot (41) for GWAS associations of LDL-C at the *APOE* locus. The lipids GWAS summary statistics were from the European Network for Genetic and Genomic Epidemiology (ENGAGE) Consortium (*n* = 58,381) (40).



The MACIE-protein and MACIE-conserved scores for rs7412 are 0.96 and 0.97, respectively. The MACIE-conserved and MACIE-regulatory scores for rs1065853 are < 0.01 and > 0.99, respectively.

Discussion

As the amount of publicly available annotation data increases and our understanding of variant functional effects continues to grow, describing variant functionality with a flexible yet practically interpretable and intuitive vocabulary will only become more important. Existing one-dimensional integrative scores cannot capture the multi-faceted functional profile of a variant

because such ratings necessarily combine diverse, and possibly unrelated, sets of annotations into a single outcome. Oftentimes, they also ignore or do not fully take into account the correlation between individual annotations. Current supervised methods further demonstrate performance profiles that are linked strongly to the quality of training set labels. These supervised scores may lack robustness in the absence of gold-standard training sets.

In this chapter we have proposed MACIE, an unsupervised multivariate mixed model framework that allows for multiple, possibly correlated, binary functional statuses. This framework offers several fundamental advancements over existing methods. First, MACIE provides multi-dimensional scores that measure functionality across multiple different functional classes. As posterior predictive probabilities, these scores are interpretable and scientifically relevant. They can be further summarized into marginal measures such as "probability of function according to at least one class of annotations" or "probability of function according to all classes of annotations". Classes of annotation can be defined separately for different types of variants, for example, coding and noncoding variants.

Second, the MACIE model accommodates correlations both within- and between- classes. It has been reported that, while some of the available annotations measure similar notions of functionality, others provide distinct and complementary information (10, 23). By flexibly modeling potential, complex correlations across all the annotations, MACIE reflects this underlying biology. In doing so, it is better able to assign each annotation and group of annotations the appropriate amount of influence.

In multiple independent testing datasets, we showed that MACIE delivers powerful and robust performance in discriminating between functional and non-functional variants. Using lipids GWAS summary statistics data from the ENGAGE consortium, we also illustrated that MACIE offers an effective tool for fine-mapping studies to prioritize top hit in silico functional variants for experimental follow-up. MACIE scores have already been used, for example, to identify and characterize inflammation and immune-related risk variants in squamous cell lung cancer (42). Finally, the proposed MACIE scores can be used as a weighting scheme to further empower variant-set analyses of rare variants (43).

Our proposed MACIE framework provides a multi-dimensional functional class extension of several existing unsupervised single scoring frameworks, such as EIGEN (16). MACIE fits a mixed model to the set of annotations for several latent functional classes and outputs the corresponding posterior component probabilities, which are highly interpretable. If we assume that there exists a single latent dichotomous variable summarizing functional status and that all annotations are independent conditional on the univariate functional status, then MACIE reduces to the GenoCanyon framework (17).

The versatility of the MACIE approach does introduce additional decisions that investigators need to make. For example, one needs to decide which set of annotations to include and how to group the annotations. The exponential family assumption in the model may also require a proper transformation for each individual annotation score before fitting the model. Operationally, users need to consider the trade-off between a more complex model (e.g., by increasing the number of classes or the number of functional scores in each class) and

computation time. Such issues will become more relevant when extending the MACIE framework to integrate cell type-specific, tissue-specific, species-specific, or phenotype-related annotations (18, 19, 37). Nevertheless, these choices again highlight the flexibility of the MACIE approach. Unlike other one-dimensional algorithms that rely on assumptions more likely to be satisfied when the number of annotations is small, the MACIE statistical model scales well with increasing annotation data. Thus, MACIE can be expected to provide more meaningful predictions as the availability of annotation scores continues to expand and the quality of these data improves.

A final important consideration in practical analysis concerns the differences between supervised and unsupervised methods. The performance of unsupervised scores may lag behind supervised methods when training datasets with relevant, high-quality labels are available. We observed this behavior when comparing MACIE to FATHMM-XF in ClinVar noncoding variants. Future extensions of interest include development of tools capable of integrating both supervised and unsupervised methods to further improve prediction accuracy (37).

Methods

The MACIE generalized linear mixed model (GLMM)

Suppose there are N genetic variants in total and we are interested in M latent annotation classes, each containing L_j annotation scores. For example, the first class may consist of $L_1 = 4$ protein functional scores and the second class may consist of $L_2 = 8$ evolutionary conservation scores. For genetic variant i and annotation class j, we denote the set of L_j annotations as $\boldsymbol{y}_{ij} =$

$\left(y_{ij1}, \dots, y_{ijL_j}\right)^T$, such that each variant is described by $L = \sum_{j=1}^{M} L_j$ annotations in total. We want to estimate for each variant i the vector of binary functional statuses $\boldsymbol{c}_i = (c_{i1}, \dots c_{iM})$, where c_{ij} is the unobserved latent functional status for class j. Continuing our example, c_{i1} would denote membership in the evolutionarily conserved function class while c_{i2} would denote membership in the regulatory function class. Conditional on c_{ij} and a random effect term b_{ijk}, we assume that $\boldsymbol{y}_{ij}$ follows a GLMM,

$$g_{jk}\left(E\left(y_{ijk}|c_{ij}, b_{ijk}\right)\right) = \beta_{0jk} + \beta_{1jk}c_{ij} + b_{ijk},$$

where $\boldsymbol{b}_{ij} = \boldsymbol{\Lambda}_j \boldsymbol{f}_{ij} \sim \mathrm{MVN}\left(\mathbf{0}, \boldsymbol{\Lambda}_j \boldsymbol{\Lambda}_j^T\right)$ is modeled using a factor analysis model with $\boldsymbol{f}_{ij} \sim MVN\left(\mathbf{0}, \mathbf{I}_{P_j \times P_j}\right)$ and $P_j < L_j$. Note that $\boldsymbol{\Lambda}_j \boldsymbol{\Lambda}_j^T$ is a flexible model for capturing the correlation between annotations of group j, conditional on c_{ij}, while reducing the number of covariance parameters that need to be estimated (44).

The Expectation-Maximization (EM) algorithm

The MACIE score for a given genetic variant i is defined by $p(\boldsymbol{c}_i|\boldsymbol{y}_i)$, that is, the posterior probability of the unobserved class label $\boldsymbol{c}_i$, conditional on the observed annotations $\boldsymbol{y}_i$. Given this is a missing data problem (from an unsupervised perspective), the EM algorithm provides a natural solution (45). We first write out the complete-data log-likelihood

$$\log f(\boldsymbol{y}, \boldsymbol{c}, \boldsymbol{b}) = \sum_{i=1}^{N}\left(\sum_{j=1}^{M}\sum_{k=1}^{L_j} \log f_{jk}\left(y_{ijk}|c_{ij}, b_{ijk}; \boldsymbol{\beta}_{jk}, \phi_{jk}\right) + \sum_{j=1}^{M} \log f\left(\boldsymbol{b}_{ij}; \boldsymbol{\theta}\right) + \log p(\boldsymbol{c}_i; \boldsymbol{\gamma})\right)$$

where $\boldsymbol{\beta}, \boldsymbol{\phi}, \boldsymbol{\gamma}, \boldsymbol{\theta}$ are (unknown) model parameters. Given that both $\boldsymbol{c}$ and $\boldsymbol{b}$ are unobserved, we proceed with the following EM algorithm.

i) Initiate reasonable parameter values. At iteration r with parameter estimates $\left(\hat{\beta}_{jk}^{(r)}, \hat{\phi}_{jk}^{(r)}, \widehat{\boldsymbol{\Lambda}}_j^{(r)}, \widehat{\boldsymbol{\gamma}}^{(r)}\right)$

ii) (E-step 1) Compute $\hat{f}^{(r)}(\boldsymbol{c}_i, \boldsymbol{b}_i | \boldsymbol{y}_i) = \hat{f}^{(r)}(\boldsymbol{b}_i | \boldsymbol{y}_i, \boldsymbol{c}_i)\hat{p}^{(r)}(\boldsymbol{c}_i | \boldsymbol{y}_i)$ via

$$f(\boldsymbol{b}_i | \boldsymbol{y}_i, \boldsymbol{c}_i) = \prod_{j=1}^{M} \frac{f(\boldsymbol{y}_{ij} | c_{ij}, \boldsymbol{b}_{ij}) f(\boldsymbol{b}_{ij})}{\int f(\boldsymbol{y}_{ij} | c_{ij}, \boldsymbol{b}_{ij}) f(\boldsymbol{b}_{ij}) \mathrm{d}\boldsymbol{b}_{ij}}$$

$$p(\boldsymbol{c}_i | \boldsymbol{y}_i) = \frac{p(\boldsymbol{c}_i, \boldsymbol{y}_i)}{p(\boldsymbol{y}_i)} = \frac{\prod_{j=1}^{M}\left[\int f(\boldsymbol{y}_{ij} | c_{ij}, \boldsymbol{b}_{ij}) f(\boldsymbol{b}_{ij}) \mathrm{d}\boldsymbol{b}_{ij}\right] \cdot p(\boldsymbol{c}_i)}{\sum_{\boldsymbol{c} \in \{0,1\}^M} \prod_{j=1}^{M}\left[\int f(\boldsymbol{y}_{ij} | c_{ij}, \boldsymbol{b}_{ij}) f(\boldsymbol{b}_{ij}) \mathrm{d}\boldsymbol{b}_{ij}\right] \cdot p(\boldsymbol{c})}$$

iii) (E-step 2) Compute expected score functions with respect to the posterior distribution of $f(\boldsymbol{c}_i, \boldsymbol{b}_i | \boldsymbol{y}_i)$, i.e. $E_{\boldsymbol{c},\boldsymbol{b}} S(\boldsymbol{\beta}_{jk}), E_{\boldsymbol{c},\boldsymbol{b}} S(\boldsymbol{\Lambda}_j), E_{\boldsymbol{c},\boldsymbol{b}} S(\phi_{jk}), E_{\boldsymbol{c},\boldsymbol{b}} S(\boldsymbol{\gamma})$, where $S(\{\boldsymbol{\beta}_{jk}, \boldsymbol{\Lambda}_j, \phi_{jk}, \boldsymbol{\gamma}\}) = \partial \log f(\boldsymbol{y}, \boldsymbol{c}, \boldsymbol{b}) / \partial\{\boldsymbol{\beta}_{jk}, \boldsymbol{\Lambda}_j, \phi_{jk}, \boldsymbol{\gamma}\}$ are the complete data score functions of $\boldsymbol{\beta}_{jk}, \boldsymbol{\Lambda}_j, \phi_{jk}, \boldsymbol{\gamma}$, respectively.

iv) (M-step) Update $\left(\hat{\beta}_{jk}^{(r+1)}, \hat{\phi}_{jk}^{(r+1)}, \widehat{\boldsymbol{\Lambda}}_j^{(r+1)}, \widehat{\boldsymbol{\gamma}}^{(r+1)}\right)$ by solving the expected score equations from (iii).

v) Iterate between (ii) - (iv) until convergence of parameters.

The algorithm proceeds until the relative change in the estimated parameters is sufficiently small ($< 10^{-4}$) with a maximum of 200 iterations. The final converged value of $\hat{p}(\boldsymbol{c}_i | \boldsymbol{y}_i)$ corresponds to the MACIE score for genetic variant i. Further details are available online (46).

Data analysis using the MACIE GLMM

We used the proposed framework to fit the MACIE GLMM models for a, nonsynonymous coding variants and b, noncoding and synonymous variants separately. For nonsynonymous coding variants, we considered fitting a two-class MACIE model $(M = 2)$ where the damaging protein function class included four protein substitution scores: SIFT, PolyPhenDiv, PolyPhenVar (dichotomous) and Mutation Assessor (continuous), with two latent factors of $\mathbf{\Sigma}_1$; and the evolutionary conserved class included eight conservation scores: GERP_NR, GERP_RS, PhyloPri, PhyloPla, PhyloVer (continuous), and PhastPri, PhastPla, PhastVer (dichotomous), with two latent factors of $\mathbf{\Sigma}_2$ (Supplementary Table 1.1). As such, the MACIE score predicted for each nonsynonymous coding variant is a vector of length 4, representing the estimated joint posterior probabilities of belonging to (0,1) - "not damaging protein functional and conserved"; (1,0) - "damaging protein functional and not conserved"; (0,0) - "not damaging protein functional and not conserved"; (1,1) - "both damaging protein functional and conserved". The MACIE GLMM regression paramete estimates from the training set of nonsynonymous coding variants are presented in Supplementary Table 1.13.

For noncoding and synonymous coding variants, we considered fitting a two-class MACIE model $(M = 2)$, where the evolutionary conserved class included the same eight conservation scorers as the nonsynonymous coding model, with two latent factors of $\mathbf{\Sigma}_1$, and the regulatory class included a total of twenty-eight transformed (continuous) epigenetic scores scores, consisting of three histone marks and 12 open chromatin marks from the ENCODE Project, three transcription factor binding site scores, GC content, CpG content, five chromatin state probabilities derived from the 15 state ChromHMM model, a background selection score, and physical distance metrics, with three latent factors of $\mathbf{\Sigma}_2$ (Supplementary Table 1.1). As such, the

MACIE score predicted for each noncoding or synonymous coding variant is also a vector of length 4, representing the estimated joint posterior probabilities of belonging to (0,1) - “not conserved and regulartory functional”; (1,0) - “conserved and not regulatory functional”; (0,0) - “not conserved and not regulatory functional”; (1,1) - “both conserved and regulatory functional”. The MACIE GLMM regression parameter estimates from the training set of noncoding and synonymous coding variants are presented in Supplementary Table 1.14.

CHAPTER II

Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole genome sequencing studies at scale

Xihao Li, Zilin Li, Hufeng Zhou, Sheila M. Gaynor, Yaowu Liu, Han Chen, Ryan Sun, Rounak Dey, et al., NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium, TOPMed Lipids Working Group, Benjamin M. Neale, Shamil R. Sunyaev, Gonçalo R. Abecasis, Jerome I Rotter, Cristen J. Willer, Gina M. Peloso, Pradeep Natarajan and Xihong Lin

Abstract

Large-scale whole genome sequencing (WGS) studies have enabled the analysis of rare variants (RVs) associated with complex phenotypes. Commonly used RV association tests (RVATs) have limited scope to leverage variant functions. We propose STAAR (variant-Set Test for Association using Annotation infoRmation), a scalable and powerful RVAT method that effectively incorporates both variant categories and multiple complementary annotations using a dynamic weighting scheme. For the latter, we introduce "annotation Principal Components", multi-dimensional summaries of *in silico* variant annotations. STAAR accounts for population structure and relatedness, and is scalable for analyzing very large cohort and biobank WGS studies of continuous and dichotomous traits. We applied STAAR to identify RVs associated with four lipid traits in 12,316 discovery samples and 17,822 replication samples from the Trans-Omics for Precision Medicine program. We discovered and replicated novel RV associations,

including disruptive missense RVs of *NPC1L1* and an intergenic region near *APOC1P1* associated with low-density lipoprotein cholesterol.

Introduction

An increasing number of whole genome/exome sequencing (WGS/WES) studies are being conducted to investigate the genetic bases of human diseases and traits, including the Trans-Omics for Precision Medicine Program (TOPMed) of the National Heart, Lung and Blood Institute (NHLBI) and the Genome Sequencing Program (GSP) of the National Human Genome Research Institute (NHGRI). Such studies enable assessment of associations between complex traits and both coding and non-coding rare variants (RVs; minor allele frequency (MAF) < 1%) across the genome. However, single-variant analyses typically have low power to identify associations with rare variants (47-49). To improve power, variant-set tests have been proposed to jointly test the effects of given sets of multiple rare variants. These methods include the burden test (50-53), Sequence Kernel Association Test (SKAT) (54), and their various combinations (55-58). In parallel, external biological information provided by functional annotations, such as conservation scores and predicted enhancer status, has been successfully used for prioritizing plausibly causal common variants in fine-mapping studies, partitioning heritability in GWAS, and predicting genetic risk (59-63). It is of substantial interest to incorporate variant functional annotations effectively, to boost the power of RV analysis of WGS association studies (64, 65).

Variant functional annotations take two forms: (i) qualitative functional groupings into genomic elements, such as Variant Effect Predictor (VEP) categories (35, 66), and (ii) quantitative functional scores available for variants across the genome, including protein functional scores (1, 2), evolutionary conservation scores (3, 4), epigenetic measures (6), and integrative functional scores (11). Different annotation scores capture diverse aspects of variant function (22, 23). Given the diversity of available annotations, efforts have been made to aggregate the evidence they provide on genomic function (10). Simultaneous use of multiple, varied functional annotation scores in variant-set tests could improve rare variant association study (RVAS) power, for example, by optimally selecting and weighting plausibly-causal rare variants (67).

To boost power for variant-set tests in WGS RVAS, we propose the variant-Set Test for Association using Annotation infoRmation (STAAR), a general framework that dynamically incorporates both qualitative functional categories and quantitative complementary annotation scores using a unified omnibus multi-dimensional weighting scheme. For the latter, to effectively capture the multi-faceted biological impact of a variant, we introduce annotation Principal Components (aPCs), multi-dimensional summaries of annotation scores that can be leveraged in the STAAR framework.

Recent methods (68-70) have incorporated functional annotations in genetic association studies. However, these methods are not scalable to analyze large-scale WGS studies while accounting for relatedness and population structure. Large scale WGS and WES studies, such as TOPMed and GSP, include a considerable fraction of related and ancestrally diverse samples. STAAR

accounts for both relatedness and population structure, as well as longitudinal follow-up designs, for both quantitative and dichotomous traits, using a Generalized Linear Mixed Models (GLMM) framework (71) that includes linear and logistic mixed models (72, 73). Using sparse Genetic Relatedness Matrices (GRMs) (74), STAAR is computationally scalable for very large WGS studies and biobanks of hundreds of thousands of samples.

We perform herein extensive simulation studies to demonstrate that STAAR can achieve substantially greater power compared to conventional variant-set tests, while maintaining accurate type I error rates for both quantitative and dichotomous phenotypes. We then apply STAAR to perform WGS gene-centric and sliding window-based genetic region analysis of 12,316 discovery samples and 17,822 replication samples with four quantitative lipid traits: low-density lipoprotein cholesterol (LDL-C), high-density lipoprotein cholesterol (HDL-C), triglycerides (TG), and total cholesterol (TC) from the NHLBI TOPMed program. We show that STAAR outperforms existing methods and identifies novel and replicated associations, including with LDL-C in disruptive missense RVs of *NPC1L1*, and in an intergenic region near *APOC1P1*.

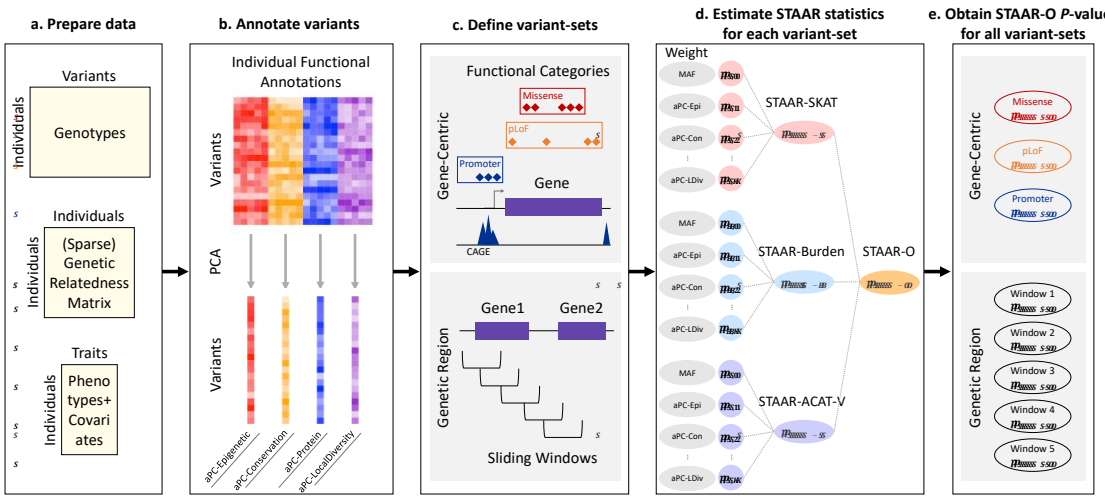
Results

Overview of methods

STAAR is a general framework for analyzing WGS RVAS at scale by using both qualitative functional categories as well as multiple in silico variant annotation scores within a variant-set, while accounting for population structure and relatedness by fitting linear and logistic mixed models for quantitative and dichotomous traits using fast and scalable algorithms. For each

variant-set, there are two main components of the STAAR framework: (i) using annotation PCs to capture and prioritize multi-dimensional variant biological functions, and (ii) testing the association between each variant-set and phenotypes by incorporating these annotation PCs as well as other integrative functional scores and MAFs in the STAAR test statistics using an omnibus weighting scheme (Figure 2.1).

Figure 2.1 STAAR workflow.

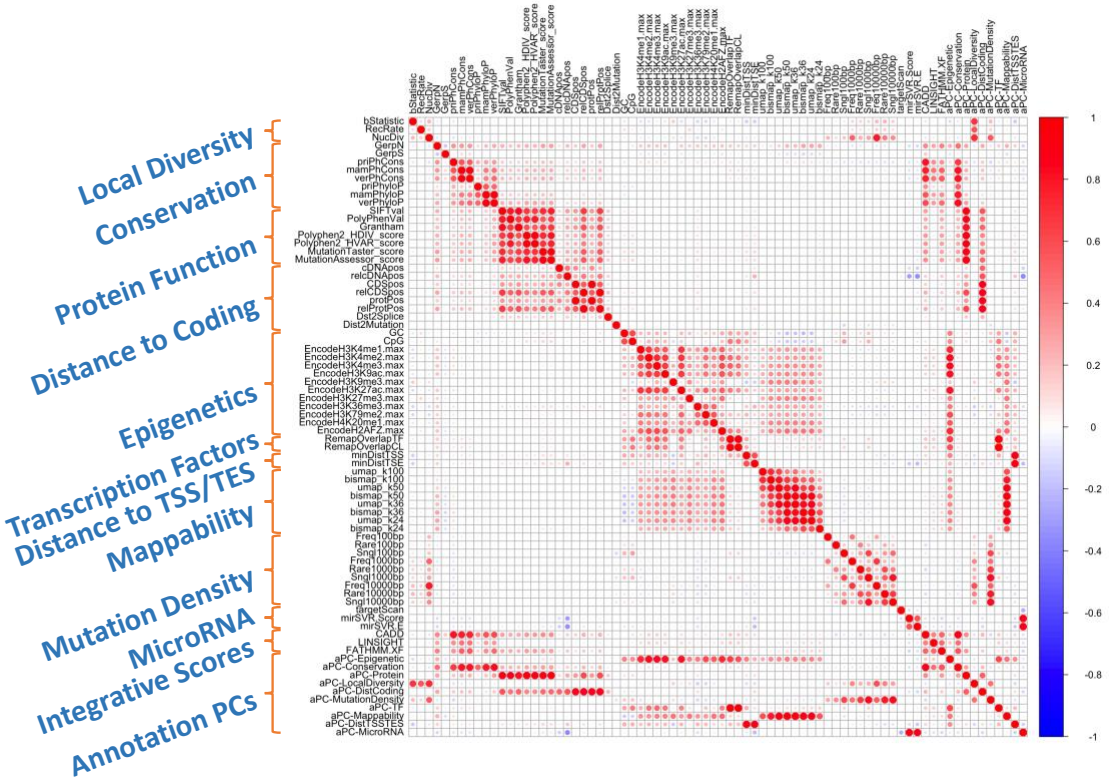


a, Prepare the input data of STAAR, including genotypes, phenotypes, covariates, and (sparse) genetic relatedness matrix. b, Annotate all variants in the genome and calculate the annotation principal components for different classes of variant function. c, Define two types of variant-sets: gene-centric analysis by grouping variants into functional genomic elements for each protein-coding gene; genetic region analysis using agnostic sliding windows. d, Estimate STAAR statistics for each variant-set. e, Obtain STAAR-O *P* values for all variants sets that are defined in c and report significant findings.

Variants often influence genes and gene products through multiple mechanisms. We extract a broad set of variant functional annotations (Supplementary Table 2.1), including both individual and ensemble functional scores, from various databases, such as ENCODE (6), Roadmap Epigenomics (7), and other evolutionary and protein annotation databases (11, 24, 75). A correlation heatmap across variants in the genome (Figure 2.2) shows that the correlation

structure among all individual annotations is approximately block-diagonal, with highly correlated blocks representing different classes of variant function, e.g., epigenetic function, evolutionary conservation, protein function, local nucleotide diversity. We introduce annotation Principal Components defined as the first PCs calculated from the set of individual functional annotation scores in each functional block (Supplementary Table 2.1 and Methods). Annotation PCs effectively reduce the dimensionality of the large number of individual annotations and summarize multiple aspects of variant function.

Figure 2.2 Correlation heatmap of functional annotation scores.



Pairwise correlations between 76 individual and integrative functional annotations using variants from the pooled samples of lipid traits in the TOPMed data. The cells in the visualization are colored by Pearson's correlation coefficient values with deeper colors indicating higher positive (red) or negative (blue) correlations. Each annotation principal component (aPC) is the first PC calculated from the set of individual functional annotations that measure similar biological function. These aPCs are then transformed into the PHRED-scaled scores for each variant across the genome (Methods).

The STAAR framework first calculates a set of multiple candidate test statistics using different annotation weights under a particular testing approach (Figure 2.1d). For each type of RV test, STAAR then uses ACAT (aggregated Cauchy association test) method to combine the resulting *P* values calculated using different weights in order to effectively and powerfully aggregate the association strength from all annotations in a data-adaptive manner (Fig. 2.1d and Methods). The ACAT method for combining *P* values is accurate and computationally efficient, while accounting for arbitrary correlation structure between tests (55, 76). To leverage the advantages of different types of tests, we propose an omnibus test in the STAAR framework (STAAR-O) by combining *P* values across different types of multiple-annotation-weighted variant-set tests using the ACAT method (Figure 2.1d and Methods).

Simulation studies

To evaluate the type I error and power of STAAR compared to conventional variant-set tests, we performed simulation studies under a variety of configurations. We followed the steps described in Data simulation (Methods) to generate both continuous and dichotomous phenotypes. We generated genotypes by simulating 20,000 sequences for 100 different regions with each spanning 1 megabase (Mb). The data were generated to mimic the linkage disequilibrium (LD) structure of an African American population by using the calibration coalescent model (COSI) (77). We randomly selected 5-kilobase (kb) regions from these 1-Mb regions and considered

sample sizes of 2,500, 5,000, and 10,000 for each replicate. The simulation studies focused on aggregating uncommon variants with MAF < 5%.

Type I error simulations

The empirical type I error rates for STAAR-O were evaluated based on 10^9 simulations at $\alpha = 10^{-5}, 10^{-6}, 10^{-7}$ for continuous and dichotomous traits (Supplementary Table 2.2). The results show that the type I error rate for STAAR-O appeared to be well controlled for both continuous and dichotomous traits at all α levels. For continuous traits, STAAR-O delivered accurate empirical type I error rates. For dichotomous traits and the smallest α level considered of 10^{-7}, STAAR-O was slightly conservative for moderate sample sizes (2,500 individuals); however, its type I error rate came close to the nominal level with larger sample sizes.

Empirical power simulations

Next, we evaluated the power of STAAR empirically by incorporating MAF and 10 annotations into its analysis and comparing results with conventional variant-set tests in a variety of configurations. Power was estimated as the proportion of P values less than $\alpha = 10^{-7}$ based on 10^4 replicates. Causality of variants was allowed to be dependent on different sets of annotations through a logistic model (Methods). We considered different proportions of causal variants (5%, 15%, 35% on average) in the signal region. For both continuous and dichotomous traits, STAAR-O incorporating all 10 annotations had higher power than the conventional variant-set tests in terms of signal region detection (Supplementary Figures 2.1-2.4). Power simulation results of STAAR-O for different magnitudes of effect sizes and different proportions of effect

size directions yielded the same conclusion (Supplementary Figures 2.1, 2.5 and 2.6). Overall, our simulation studies showed that STAAR-O could provide considerably higher power than conventional variant-set tests.

Association analysis of lipid traits in the TOPMed WGS data

We applied STAAR to identify RV-sets associated with four quantitative lipid traits (LDL-C, HDL-C, TG and TC) using TOPMed WGS data (78, 79). LDL-C and TC were adjusted for the presence of medications as before (78). DNA samples were sequenced at >30× target coverage. The discovery phase consists of four study cohorts of TOPMed Freeze 3. The replication phase consists of ten different study cohorts in TOPMed Freeze 5 that were not in Freeze 3 (Supplementary Note and Supplementary Table 2.3).

Sample-level and variant-level quality control (QC) were performed (78, 79). There were 12,316 discovery samples, which had 155 million single nucleotide variants (SNVs), and 17,822 replication samples, which had 188 million SNVs. The TOPMed data consist of ancestrally diverse and multi-ethnic related samples. Race/ethnicity was defined using a combination of self-reported race/ethnicity and study recruitment information. The discovery cohorts consist of 4,580 (37.2%) Black or African American, 6,266 (50.9%) White, 543 (4.4%) Asian American, and 927 (7.5%) Hispanic/Latino American. Among all samples in discovery phase, 3,577 (29.0%) had first-degree relatedness, 491 (4.0%) had second-degree relatedness, and 273 (2.2%) had third-degree relatedness (Supplementary Figure 2.7). Among all SNVs observed in the discovery samples, there were 6.5 million (4.2%) common variants (MAF > 5%), 5.3 million (3.4%) low

frequency variants (1% ≤ MAF ≤ 5%), and 143.2 million (92.4%) rare variants (MAF < 1%). The race/ethnicity distribution, related sample distribution, and variant number distribution for replication phase and pooled samples (samples from both discovery phase and replication phase) are given in Supplementary Table 2.4.

Our study used the proposed STAAR-O method to perform (i) gene-centric analysis using RV-sets based on functional categories, and (ii) genetic region analysis using variant-sets defined by 2-kb sliding windows with 1-kb skip length across the genome. We adjusted for age, age^2, sex, race/ethnicity, study, and the first 10 ancestral PCs, while controlling for relatedness using linear mixed models, with inverse-rank normal transformation applied to phenotypes (Methods). Race/ethnicity was included as a covariate to adjust for sociocultural and environmental factors, while genetic ancestry differences were captured by the inclusion of the ancestral PCs. In addition to the two MAF weights (49), we incorporated 13 aggregated functional annotation scores in STAAR-O: 3 integrative scores (CADD (11), LINSIGHT (21), and FATHMM-XF (15)) and 10 aPCs. Figure 2.2 summarizes the correlation among all functional annotations, including 60 individual scores, 3 integrative scores, and 10 aPCs.

Gene-centric association analysis of coding and non-coding rare variants

We performed gene-centric analysis to identify whether rare variants in coding, promoter, and enhancer regions of genes are associated with lipid traits using STAAR-O. For each of the four lipid traits, we analyzed five functional categories (masks) of coding and non-coding variants: (i) pLoF (stop gain, stop loss and splice) RVs, (ii) missense RVs, (iii) synonymous RVs, (iv)

promoter RVs, and (v) enhancer RVs. The pLoF, missense, and synonymous RVs were defined by GENCODE VEP categories (35, 66). The promoter RVs were defined as RVs in the +/- 3-kb window of transcription starting site (TSS) with overlap of Cap Analysis of Gene Expression (CAGE) sites. The enhancer RVs were defined as RVs in GeneHancer predicted regions with overlap of CAGE sites (8, 9, 80). Within each gene functional category, we tested for an association between rare variants (MAF < 1%) in the functional category and lipid traits using STAAR-O with the 13 aggregated functional annotations described above. For missense RVs, we incorporated an additional annotation functional category predicting functionally "disruptive" variants determined by MetaSVM (81), which measures the deleteriousness of missense mutations. The overall distributions of STAAR-O *P* values were well calibrated for all four lipid phenotypes (Supplementary Figure 2.8). We considered in unconditional analysis a Bonferroni-corrected genome-wide significance threshold of $\alpha = 0.05/(20{,}000 \times 5) = 5.00 \times 10^{-7}$ accounting for five different masks across protein-coding genes.

STAAR-O identified 21 genome-wide significant associations with four lipid phenotypes using unconditional analysis of the discovery samples (Supplementary Table 2.5 and Supplementary Figure 2.9). After conditioning on known lipids-associated variants (40, 78, 82-96), 11 out of the 21 associations remained significant at the Bonferroni correction level $0.05/21 = 2.38 \times 10^{-3}$ using the discovery samples. These included associations with LDL-C (pLoF RVs in *PCSK9* and *APOB*, missense RVs in *PCSK9*, *NPC1L1*, and *APOE*), association with HDL-C (pLoF RVs in *APOC3*), association with TG (pLoF RVs in *APOC3*), and associations with TC (pLoF RVs in *PCSK9* and *APOB*, missense RVs in *PCSK9* and *LIPG*) (Table 2.1). Of these 11 associations, 10 were replicated at the Bonferroni-corrected level $0.05/11 = 4.55 \times 10^{-3}$ after adjusting for

known lipid-associated variants. The association between *APOC3* pLoF RVs and HDL-C was unreported in a previous study using the same TOPMed Freeze 3 data (78).

The association between missense RVs in *NPC1L1* and LDL-C was not detected by the conventional variant-set tests and has not been observed in previous studies (78, 85, 97, 98). In the discovery phase, its unconditional STAAR-O *P* value was 1.29×10^{-7}, while the most significant conventional variant-set test was the burden test with $P = 7.04 \times 10^{-6}$. This association was not driven by any single RV (minimum single RV *P* value $> 10^{-3}$) but was due to the aggregated effects of multiple missense RVs. The *P* value of the burden test additionally weighted by MetaSVM was the smallest of all annotations ($P = 3.15 \times 10^{-9}$), highlighting the significant association between disruptive missense RVs in *NPC1L1* and LDL-C (Supplementary Figure 2.10). Among all 174 missense RVs in *NPC1L1* from the discovery samples, the disruptive missense RVs as predicted by MetaSVM were enriched among variants with higher aPC-Conservation scores (Supplementary Table 2.6). This contributed to the test weighted by aPC-Conservation being the most significant across all quantitative annotation-weighted tests included in STAAR-O (burden $P = 3.12 \times 10^{-7}$). As aPC-Conservation summarizes variants' evolutionary conservation scores, it is informative in predicting whether or not variants are deleterious and thus functional (99, 100). Conditioning on the ten known common variants in *NPC1L1* associated with LDL-C (Supplementary Table 2.7) (40, 87-90, 94-96), the association between disruptive missense RVs in *NPC1L1* and LDL-C remained significant after Bonferroni correction with the conditional analysis $P = 9.27 \times 10^{-9}$ in discovery phase. This association was validated in replication phase with $P = 2.59 \times 10^{-4}$ and with $P = 4.02 \times 10^{-11}$ in pooled samples in conditional analysis. This significant association was also validated using

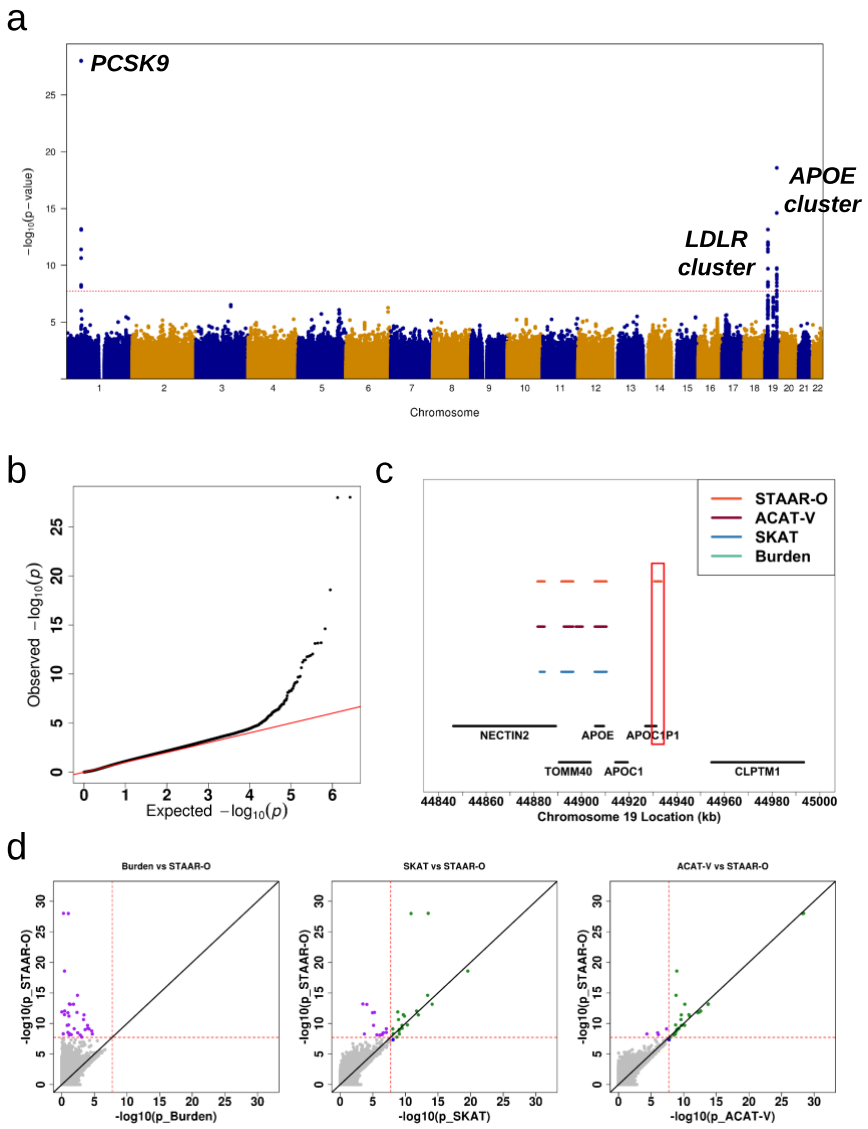
whole exome sequencing data from the UK Biobank (101) ($n = 40{,}519$) with $P = 2.49 \times 10^{-4}$ in the conditional analysis.

Genetic region analysis of rare variants

We performed genetic region analysis to determine whether RVs within sliding windows are associated with lipid traits. The sliding windows were defined to be 2 kb in length, start at position 0 base pairs (bp) for each chromosome, and have a skip length of 1 kb. Windows with a total minor allele count less than 10 were excluded from the analysis, resulting in a total of 2.66 million 2-kb overlapping windows, with a median of 104 RVs in each sliding window among discovery samples. For each 2-kb window, we tested for an association between the RVs in the window and each lipid trait using STAAR-O by incorporating 13 aggregated quantitative annotations. The overall distributions of STAAR-O *P* values were well calibrated for all four lipid phenotypes (Figure 2.3b and Supplementary Figures 2.11b, 2.12b and 2.13b). Using the Bonferroni correction, we set the genome-wide significance threshold at $\alpha = 0.05/(2.66 \times 10^{6}) = 1.88 \times 10^{-8}$ across sliding windows (Figure 2.3a and Supplementary Figures 2.11a, 2.12a, and 2.13a). Supplementary Table 2.8 summarizes the significant 2-kb sliding windows identified using STAAR-O. Overall, by dynamically incorporating multiple functional annotations capturing different aspects of variant function, STAAR-O was able to detect more significant sliding windows, and showed consistently smaller *P* values for top sliding windows compared with conventional variant-set tests weighted using MAFs (Figure 2.3c,d and Supplementary Figures 2.11c-f, 2.12c and 2.14). Burden tests were not able to detect any window that reached significance.

Among the 59 genome-wide significant sliding windows detected by STAAR-O in unconditional analysis, 17 remained significant at the Bonferroni correction level $0.05/59 = 8.47 \times 10^{-4}$ after conditioning on known lipids-associated variants using the discovery samples (Table 2.2). For LDL-C, the significant sliding windows were located in gene *PCSK9* or in a 50-kb region on chromosome 19 including the *APOE* cluster. For TC, all of the significant sliding windows were located in the same areas as for LDL-C. For TG, STAAR-O detected two consecutive significant sliding windows within *APOC3*, whereas no significant sliding windows were detected for HDL-C. Of these 17 associations, six were replicated at level $0.05/17 = 2.94 \times 10^{-3}$ after Bonferroni correction and another four were replicated at level $0.05/9 = 5.56 \times 10^{-3}$ after Bonferroni correction for nine non-overlapping sliding windows in conditional analysis of replication samples (63), including a sliding window located downstream of *APOC1P1* (chromosome 19: 44,931,528 bp - 44,933,527 bp), which was significantly associated with LDL-C but undetected by the burden test, SKAT, and ACAT-V (Table 2.2 and Figure 2.3c).

Figure 2.3 Genetic region (2-kb sliding window) unconditional analysis results of LDL-C in the discovery phase using the TOPMed cohort.



a, Manhattan plot showing the associations of 2.66 million 2-kb sliding windows for LDL-C versus $-\log_{10}(P)$ of STAAR-O. The horizontal line indicates a genome-wide *P* value threshold of 1.88×10^{-8} (*n* = 12,316). b, Quantile-quantile plot of 2-kb sliding window STAAR-O *P* values for LDL-C (*n* = 12,316). c, Genetic landscape of the windows significantly associated with LDL-C that are located in the 150-kb region on chromosome 19. Four statistical tests were compared: Burden, SKAT, ACAT-V and STAAR-O. A dot indicates that the sliding window at this location is significant using the statistical test that the color of the dot represents (*n* = 12,316). d, Scatterplot of *P* values for the 2-kb sliding windows comparing STAAR-O with Burden, SKAT and ACAT-V tests. Each dot represents a sliding window with x-axis label being the $-\log_{10}(P)$ of the conventional test and y-axis label being the $-\log_{10}(P)$ of STAAR-O (*n* = 12,316).

Table 2.1 Gene-centric analysis results of both unconditional analysis and analysis conditional on known common and low-frequency variants.

Trait	Gene	Chr. no.	Category	Discovery			Replication			Pooled			Variants (adjusted)
				No. of SNVs	STAAR-O (Unconditional)	STAAR-O (Conditional)	No. of SNVs	STAAR-O (Unconditional)	STAAR-O (Conditional)	No. of SNVs	STAAR-O (Unconditional)	STAAR-O (Conditional)	
LDL-C	*PCSK9*	1	Putative loss of function	5	3.09E-38	1.94E-07	8	6.97E-27	5.29E-10	9	4.59E-65	7.52E-17	rs28362286, rs28362263, rs11591147, rs12117661
	APOB	2	Putative loss of function	11	1.91E-14	2.38E-14	5	1.97E-09	1.76E-09	16	3.91E-21	4.08E-21	rs934197
	PCSK9	1	Missense	92	1.09E-16	2.65E-08	129	1.90E-06	1.15E-06	167	2.11E-15	1.14E-14	rs28362286, rs28362263, rs11591147, rs12117661
	NPC1L1	7	Missense	174	1.29E-07	3.83E-07	219	2.19E-03	3.28E-03	293	3.25E-10	1.58E-09	rs10234070, rs73107473, rs2072183, rs41279633, rs17725246, rs2073547, rs10260606, rs217386, rs7791240, rs2300414
	NPC1L1	7	Disruptive missense	94	3.15E-09*	9.27E-09*	129	1.46E-04*	2.59E-04*	173	8.05E-12*	4.02E-11*	rs10234070, rs73107473, rs2072183, rs41279633, rs17725246, rs2073547, rs10260606, rs217386, rs7791240, rs2300414
	APOE	19	Missense	54	3.11E-10	9.88E-11	58	6.61E-05	3.47E-04	88	1.07E-13	2.02E-12	rs7412, rs429358
HDL-C	*APOC3*	11	Putative loss of function	5	2.20E-07	6.82E-07	6	5.73E-18	2.89E-17	7	3.18E-23	4.51E-22	rs66505542
TG	*APOC3*	11	Putative loss of function	5	1.10E-14	5.53E-14	6	2.67E-49	2.73E-46	7	3.98E-56	1.04E-52	rs66505542, rs964184, rs7350481
TC	*PCSK9*	1	Putative loss of function	5	4.60E-33	2.04E-10	8	1.83E-25	9.74E-11	9	9.83E-58	4.23E-20	rs28362286, rs11591147, rs191448952
	APOB	2	Putative loss of function	11	7.29E-13	8.78E-13	5	2.62E-09	2.30E-09	16	9.76E-20	1.01E-19	rs934197
	PCSK9	1	Missense	92	6.00E-15	1.11E-06	131	2.14E-05	1.13E-05	169	5.18E-12	3.16E-12	rs28362286, rs11591147, rs191448952
	LIPG	18	Missense	62	9.61E-08	4.34E-06	68	3.45E-04	1.47E-01	101	2.04E-09	5.62E-04	rs4939883, rs7241918, rs149615216

*Burden test *P* value. A total of 12,316 discovery samples, 17,822 replication samples and 30,138 pooled samples from the TOPMed program were considered in the analysis. Results for the conditionally significant genes (unconditional STAAR-O $P < 5.00 \times 10^{-7}$; conditional STAAR-O $P < 2.38 \times 10^{-3}$) using discovery samples are presented in the table. Chr. no., chromosome number; category, functional category; no. of SNVs, number of RVs with a MAF < 1% of the particular functional category in the gene; STAAR-O, STAAR-O *P* value; variants (adjusted), adjusted variants in the conditional analysis.

Table 2.2 Genetic region (2-kb sliding window) analysis results of both unconditional analysis and analysis conditional on known common and low-frequency variants.

Trait	Chr. no.	Start location	End location	Gene	Discovery			Replication			Pooled			Variants (adjusted)
					No. of SNVs	STAAR-O (Unconditional)	STAAR-O (Conditional)	No. of SNVs	STAAR-O (Unconditional)	STAAR-O (Conditional)	No. of SNVs	STAAR-O (Unconditional)	STAAR-O (Conditional)	
LDL-C	1	55045498	55047497	*PCSK9*	114	7.83E-09	1.06E-04	124	3.33E-06	4.10E-04	186	1.89E-15	2.90E-09	rs28362286, rs28362263, rs11591147, rs12117661
	1	55046498	55048497	*PCSK9*	124	5.32E-09	2.13E-05	130	1.79E-06	8.79E-05	191	1.33E-15	1.15E-09	rs28362286, rs28362263, rs11591147, rs12117661
	19	44881528	44883527	*NECTIN2*	118	7.31E-10	1.81E-08	155	5.16E-04	2.42E-01	202	8.15E-08	5.26E-06	rs7412, rs429358
	19	44882528	44884527	*NECTIN2*	104	2.08E-10	3.90E-09	133	1.23E-01	3.59E-01	176	1.38E-08	7.47E-07	rs7412, rs429358
	19	44893528	44895527	*TOMM40*	110	2.64E-19	2.33E-11	136	4.54E-09	2.60E-02	187	7.29E-29	7.62E-13	rs7412, rs429358
	19	44894528	44896527	*TOMM40*	120	2.44E-15	4.31E-11	153	7.62E-05	1.74E-02	205	6.73E-20	5.28E-13	rs7412, rs429358
	19	44905528	44907527	*APOE*	91	1.73E-10	1.64E-10	115	1.22E-02	4.91E-03	169	7.68E-12	9.00E-12	rs7412, rs429358
	19	44906528	44908527	*APOE*	84	1.67E-09	1.90E-10	115	8.65E-03	3.24E-03	165	8.34E-11	6.25E-12	rs7412, rs429358
	19	44907528	44909527	*APOE*	113	1.01E-09	1.97E-10	143	5.92E-03	3.58E-03	205	4.88E-11	8.71E-12	rs7412, rs429358
	19	44908528	44910527	*APOE*	140	6.30E-10	1.32E-10	152	4.14E-03	6.10E-03	228	2.40E-11	5.21E-12	rs7412, rs429358
	19	44931528	44933527	*APOC1P1*	114	6.63E-09	7.60E-04	123	5.78E-11	5.40E-03	181	1.34E-19	4.15E-06	rs7412, rs429358
TG	11	116828930	116830929	*APOC3*	125	4.63E-10	2.80E-09	155	1.35E-36	3.94E-34	207	7.32E-45	2.73E-41	rs66505542, rs964184, rs7350481
	11	116829930	116831929	*APOC3*	109	3.61E-10	5.99E-10	140	2.85E-36	4.25E-34	187	5.75E-45	2.17E-41	rs66505542, rs964184, rs7350481
TC	1	55045498	55047497	*PCSK9*	114	3.05E-09	2.86E-07	130	3.12E-06	1.92E-06	189	2.22E-15	9.21E-14	rs28362286, rs11591147, rs191448952
	1	55046498	55048497	*PCSK9*	124	2.24E-09	2.06E-07	138	2.19E-06	1.34E-06	195	1.78E-15	7.04E-14	rs28362286, rs11591147, rs191448952
	19	44893528	44895527	*TOMM40*	111	9.35E-13	4.37E-07	146	1.12E-07	4.02E-01	196	7.57E-21	7.91E-08	rs7412, rs429358
	19	44894528	44896527	*TOMM40*	120	1.80E-09	1.99E-06	164	1.08E-04	8.31E-01	213	8.40E-14	2.19E-07	rs7412, rs429358

A total of 12,316 discovery samples, 17,822 replication samples and 30,138 pooled samples from the TOPMed program were considered in the analysis. Results for the conditionally significant sliding windows (unconditional STAAR-O $P < 1.88 \times 10^{-8}$; conditional STAAR-O $P < 8.47 \times 10^{-4}$) using discovery samples are presented in the table. Chr. no., chromosome number; start location, start location of the 2-kb sliding window; end location, end location of the 2-kb sliding window; no. of SNVs, number of RVs (MAF < 1%) in the 2-kb sliding window; STAAR-O, STAAR-O *P* value; variants (adjusted), adjusted variants in the conditional analysis. The physical positions of each window are on build hg38.

The top variant of the significant sliding window located downstream of *APOC1P1* was rs370625306 ($\mathrm{MAF} = 0.005, P = 8.71 \times 10^{-8}$), which was not significant at a Bonferroni-corrected threshold ($\alpha = 0.05/(1.51 \times 10^{7}) = 3.31 \times 10^{-9}$) in individual variant analysis. This rare variant and the second top variant in these windows (rs9749443, $\mathrm{MAF} = 0.009, P = 2.46 \times 10^{-5}$) were upweighted by aPC-Epigenetic in STAAR-O (Supplementary Figure 2.15). Specifically, the aPC-Epigenetic scores of rs370625306 and rs9749443 ranked in the top 10% and top 30% among all RVs, respectively, in each sliding window. Conditioning on the two known common variants rs7412 and rs429358 in *APOE* associated with LDL-C (85), the strength of association of both sliding windows was reduced but remained significant (Table 2.2). Similar results were found after further conditioning on *APOE* haplotypes using these two SNPs (Supplementary Table 2.8). This suggests that the effects of RVs in this sliding window are not fully captured by the two known common LDL-associated variants. STAAR-O also identified and replicated two highly significant windows in *APOC3* associated with TG in conditional analysis that were undetected by SKAT and burden test (102).

STAAR identifies more associations using relevant tissue functional annotations

To evaluate the effect of tissue specificity, we compared the performance of STAAR-O in both gene-centric and genetic region analysis by incorporating liver (a central hub for lipid metabolism), heart, and brain annotations. For each tissue, we calculated a tissue-specific aPC from tissue-specific DNase, H3K4me3, H3K27ac and H3K27me3 from ENCODE (Supplementary Table 2.9) (6, 103). We used tissue-specific CAGE sites with overlap of RVs in the +/- 3-kb window of TSS and GeneHancers to define promoter and enhancer RV masks in gene-centric analysis. To make a fair comparison between tissues, we calculated STAAR-O *P*

values based solely on the tissue-specific aPC and without incorporating the MAF and other annotations.

Overall, the use of liver annotation resulted in more significant levels of association than heart and brain annotations, as would be expected for lipid traits, although no additional replicated conditionally significant association was detected by using tissue-specific annotations. STAAR-O identified 9 and 8 replicated conditionally significant associations by using liver annotation in gene-centric and genetic region analysis, respectively (Supplementary Tables 2.10 and 2.11). Among these 17 significant associations, two were not seen when heart annotation was used and two were not seen when brain annotation was used, and no additional associations were detected by using heart and brain annotations (Supplementary Tables 2.10 and 2.11). Furthermore, more suggestive significant associations were detected when using liver annotation than the other two tissues at various levels of unconditional *P* value thresholds in the discovery phase (Supplementary Figures 2.16 and 2.17).

Computation cost

We developed an R package, STAAR, to perform scalable variant-set association tests incorporating multiple variant annotations for WGS RVAS. Using sparse GRMs (74), STAAR scales well both in terms of computation time and memory for very large-scale WGS association studies, such as sample sizes in TOPMed, GSP, and UK Biobank. The computation time for STAAR-O to perform WGS gene-centric and genetic region analysis on 30,000 related samples using the TOPMed data requires 15 hours for 100 2.10 GHz computing cores with 6 GB memory

for each lipid trait. Analyzing 500,000 simulated related samples mimicking the UK Biobank sample size requires 26 hours for WGS analysis using the same approach and computational resources (Methods).

Discussion

We propose STAAR as a general, computationally scalable framework that effectively incorporates multiple qualitative and quantitative variant functional annotations to boost power for variant-set tests for continuous and binary traits in WGS RVAS, while accounting for both population structure and relatedness using GLMMs.

We highlighted STAAR-O, the omnibus test that aggregates multiple annotation-weighted tests in the STAAR framework. We focused on two types of WGS RV association analyses using STAAR-O: gene-centric analyses by grouping coding and noncoding variants into functional categories for each protein-coding gene, and agnostic genetic region analyses using sliding windows. In extensive simulation studies, we demonstrated that STAAR-O achieves substantial power gain compared with conventional variant-set tests weighted by MAF, while maintaining accurate type I error rates for both quantitative and dichotomous phenotypes.

In a WGS RV analysis of lipid traits using the TOPMed data, STAAR-O identified several conditionally significant functional categories associated with lipid traits in gene-centric analysis (including *NPC1L1* missense RVs and LDL-C; *APOC3* pLoF RVs and HDL-C; and *LIPG* missense RVs and TC) that were missed by the previous study using the same TOPMed data

(78). Earlier studies reported marginal association between inactivating mutations (pLoF RVs and frameshift indels) in *NPC1L1* and LDL-C with $P = 0.04$ (98), which was replicated using the pooled TOPMed samples ($P = 0.02$), although no significant association between pLoF RVs and LDL-C was found ($P = 0.15$). STAAR-O identified much more significant novel association, which replicated, between missense RVs in *NPC1L1* and LDL-C, which was driven by disruptive missense RVs (conditional $P = 4.02 \times 10^{-11}$ in pooled samples). None of these disruptive missense RVs was reported in ClinVar (104), suggesting that the findings from emerging WGS studies can help guide the expansion of the ClinVar database. *NPC1L1* is the direct molecular target of the lipid-lowering drug ezetimibe, which reduces the absorption of cholesterol by binding to *NPC1L1* (105). STAAR-O also suggested several conditional associations in the discovery phase that were validated in our replication phase and achieved significance in pooled samples (Supplementary Table 2.12).

In agnostic sliding-window based genetic region analysis, STAAR-O detected and replicated 10 sliding windows after conditioning on known variants, including association between an intergenic region located downstream of *APOC1P1* and LDL-C, that were not detected using conventional tests. This detected *APOC1P1* region is located in the hepatic control region 2 (HCR-2) that regulates hepatic expression of apolipoproteins. By further conditioning on the APOE haplotypes and rs35136575, a common variant previously found in the downstream HCR-2 associated with LDL-C (106), the strength of association was reduced but remained significant (Supplementary Table 2.8). This discovery is due to upweighting several plausibly causal rare variants that have regulatory functions using aPC-Epigenetic scores in STAAR-O (Supplementary Figure 2.15 and Supplementary Table 2.13). These results highlight that

incorporating multiple functional annotations using STAAR can effectively boost power for WGS RVAS.

To capture multiple aspects of variant functionality, we introduced annotation PCs by performing dimension reduction of a large number of diverse individual annotations from various external databases. See Methods for an example demonstrating that aPCs explain diverse and complementary functionality of known LDL-associated functional rare variants, and STAAR provides greater power for RV association tests by upweighting these variants using aPCs.

In practice, STAAR is very flexible and users can determine the set of individual annotations to calculate aPCs and the number of aPCs and integrative functional scores and other qualitative scores to be used, as well as tissue, cell-type and phenotype-specific variant annotations (18, 19, 107). In this chapter, we group the individual annotations based on biological knowledge; users can also apply data-driven approaches, such as clustering, to group annotations for aPC calculation. We also demonstrate that STAAR detects more associations using relevant tissue functional annotations. It will be of interest, in future research, to incorporate improved rare variant effect size models in the weights to further improve power for RVAS (108, 109).

The STAAR procedure is fast and scalable for very large WGS studies and biobanks of hundreds of thousands to millions of samples for both quantitative and dichotomous phenotypes as it uses estimated sparse GRMs (74) to fit the null GLMM and to scan the genome. Besides using sliding windows of a pre-specified fixed window length, STAAR could be extended to flexibly detect

the sizes and locations of coding and non-coding rare variant association regions using the dynamic window analysis method SCANG (110). In addition, STAAR could be extended to settings with survival, unbalanced case-control, and multiple phenotypes, and hence could provide a comprehensive framework for WGS RVAS. Thus, STAAR provides a powerful and flexible tool for variant association discovery in many settings to explore the molecular basis of common diseases.

Methods

Notations and model

Suppose there are n subjects with M total variants sequenced across the whole genome. Given a genetic set of p variants, for subject i, let Y_i denote a continuous or dichotomous trait with mean μ_i; $\boldsymbol{X}_i = \left(X_{i1}, \dots, X_{iq}\right)^T$ denote q covariates, such as age, gender, ancestral principal components; and $\boldsymbol{G}_i = \left(G_{i1}, \dots, G_{ip}\right)^T$ denote the genotype information of the p genetic variants in a variant-set.

When the data consist of unrelated samples, we consider the following Generalized Linear Model (GLM)

$$g(\mu_i) = \alpha_0 + \boldsymbol{X}_i^T \boldsymbol{\alpha} + \boldsymbol{G}_i^T \boldsymbol{\beta}, \qquad (2.1)$$

where $g(\mu) = \mu$ for a continuous normally distributed trait, $g(\mu) = \text{logit}(\mu)$ for a dichotomous trait, α_0 is an intercept, $\boldsymbol{\alpha} = \left(\alpha_1, \dots, \alpha_q\right)^T$ is a vector of regression coefficients for $\boldsymbol{X}_i$, and $\boldsymbol{\beta} = \left(\beta_1, \dots, \beta_p\right)^T$ is a vector of regression coefficients for $\boldsymbol{G}_i$.

When the data consist of related samples, we consider the following Generalized Linear Mixed Model (GLMM) (71-73)

$$g(\mu_i) = \alpha_0 + \boldsymbol{X}_i^T \boldsymbol{\alpha} + \boldsymbol{G}_i^T \boldsymbol{\beta} + b_i, \qquad (2.2)$$

where the random effects b_i account for remaining population structure unaccounted by ancestral PCs, relatedness, and other between-observation correlation. We assume that $\boldsymbol{b} = (b_1, \dots, b_n)^T \sim N(\mathbf{0}, \sum_{l=1}^{L} \theta_l \boldsymbol{\Phi}_l)$ with variance components θ_l and known covariance matrices $\boldsymbol{\Phi}_l$. The random effects $\boldsymbol{b}$ can be decomposed into a sum of multiple random effects to account for different sources of relatedness and correlation as $\boldsymbol{b} = \sum_{l=1}^{L} \boldsymbol{b}_l$ with $\boldsymbol{b}_l \sim N(\mathbf{0}, \theta_l \boldsymbol{\Phi}_l)$. For example, $\boldsymbol{b}_1$ accounts for population structure and family relatedness by using the Genetic Relatedness Matrices (GRMs) as its covariance matrix $\boldsymbol{\Phi}_1$ (111, 112). A sparse GRM can be used to scale up computation (74). Additional random effects $\boldsymbol{b}_2, \cdots, \boldsymbol{b}_L$ can be used to account for complex sampling designs, such as correlation between repeated measures from longitudinal studies using subject-specific random intercepts and slopes and hierarchical designs. The remaining variables are defined in the same way as those in the GLM (2.1). Under both the GLM and the GLMM, we are interested in testing the null hypothesis of whether the variant-set is associated with the phenotype, adjusting for covariates and relatedness, which corresponds to $H_0: \boldsymbol{\beta} = \mathbf{0}$, that is, $\beta_1 = \beta_2 = \cdots = \beta_p = 0$.

Conventional variant-set tests

Conventional score-based aggregation methods allow for jointly testing the association between variants in the genetic set and phenotype. In particular, burden tests (50-53) assume that $\beta_j = w_j\beta$, where β is a constant for all variants, such that the corresponding burden test statistic to test $H_0: \boldsymbol{\beta} = \mathbf{0} \Leftrightarrow H_0: \beta = 0$ is given by

$$Q_{Burden} = \left(\sum_{j=1}^{p} w_j S_j\right)^2,$$

where $S_j = \sum_{i=1}^{n} G_{ij}(Y_i - \hat{\mu}_i)$ is the score statistic of the marginal model for variant j and $\hat{\mu}_i$ is the estimated mean of Y_i under the null GLM $g(\mu_i) = \alpha_0 + \boldsymbol{X}_i^T\boldsymbol{\alpha}$ or the null GLMM $g(\mu_i) = \alpha_0 + \boldsymbol{X}_i^T\boldsymbol{\alpha} + b_i$. Q_{Burden} asymptotically follows a chi-square distribution with 1 degree of freedom under the null hypothesis, and its *P* value can be obtained analytically while accounting for linkage disequilibrium (LD) between variants (49, 73).

For SKAT (54), the β_j's are assumed to be independent and identically distributed (i.i.d.) following an arbitrary distribution, with $E(\beta_j) = 0$ and $Var(\beta_j) = w_j^2\tau$. The null hypothesis of no variant-set effect $H_0: \boldsymbol{\beta} = \mathbf{0}$ is equivalent to $H_0: \tau = 0$, and the corresponding SKAT test statistic is given by

$$Q_{SKAT} = \sum_{j=1}^{p} w_j^2 S_j^2.$$

Q_{SKAT} asymptotically follows a mixture of chi-square distributions under the null hypothesis, and its *P* value can be obtained analytically while accounting for LD between variants (49, 73).

Further, the recently proposed ACAT-V test uses a combination of transformed variant *P* values rather than operating on the test statistics directly(55). The ACAT-V test statistic is given by

$$Q_{ACAT-V} = \overline{w^2\mathrm{MAF}(1-\mathrm{MAF})}\tan\big((0.5-p_0)\pi\big) + \sum_{j=1}^{p'} w_j^2\mathrm{MAF}_j\big(1-\mathrm{MAF}_j\big)\tan\Big(\big(0.5-p_j\big)\pi\Big),$$

where p' is the number of variants with minor allele count (MAC) greater than 10 and p_j is the association *P value* of individual variant j corresponding the individual variant score statistics S_j for those variants with MAC > 10. p_0 is the burden test *P* value of extremely rare variants with MAC ≤ 10 and $\overline{w^2\mathrm{MAF}(1-\mathrm{MAF})}$ is the average of the weights $w_j^2\mathrm{MAF}_j\big(1-\mathrm{MAF}_j\big)$ among the extremely rare variants with MAC ≤ 10. Q_{ACAT-V} can be well approximated by a Cauchy distribution under the null hypothesis, and its *P* value can be obtained analytically while accounting for LD between variants (55). For binary traits in highly unbalanced designs, one can improve individual *P* value calculations using Saddlepoint approximation (113, 114).

These conventional approaches consider a weight w_j defined as a threshold indicator or a function of minor allele frequency (MAF) for variant j, i.e. $w_j = Beta\big(\mathrm{MAF}_j; a_1, a_2\big)$ (49). Common choices of the parameters are $a_1 = 1$ and $a_2 = 25$ which upweights rarer variants, or $a_1 = 1$ and $a_2 = 1$, which corresponds to equal weights for all variants. In WGS studies, the vast majority of rare variants across the genome are not causal. Thus, choosing their weights

according to MAF will incorrectly upweight many such “noise” variants in a variant-set and result in a loss of statistical power. Weighting using multiple variant functional annotations will help overcome this deficiency.

Calculation of annotation principal components using individual functional annotations

To effectively capture the multi-faceted biological impact of a variant while reducing dimensionality, we propose variant annotation Principal Components (aPCs) as the PC summary of the functional annotation data by incorporating individual scores extracted from various functional databases (6, 7, 11, 24, 75, 115). We first group the individual scores into 10 major functional categories based on a priori knowledge, each capturing a specific aspect of variant biological function, including epigenetics, conservation, protein function, local nucleotide diversity, distance to coding, mutation density, transcription factors, mappability, distance to TSS/TES, and micro RNA (Figure 2.2). For each category, we then center and standardize all individual scores within the category, such that higher value of each individual score indicates increased functionality of that annotation, and calculate aPC as the first PC from the standardized individual scores (Supplementary Table 2.1). To facilitate better interpretation, these aPCs are then transformed into the PHRED-scaled scores for each variant across the genome, defined as $-10 \times \log_{10}(rank(-score)/M)$, where M is total number of variants sequenced across the whole genome.

Unlike ancestral PCs that are subject-specific and are calculated using genotypes across the genome to control for population structure, annotation PCs are variant-specific and are calculated

using functional annotations for individual variants and are used to summarize multi-facet functions of individual variants. Complementary to other existing single-dimension integrative functional scores, annotation PCs summarize multiple aspects of variant function, with different blocks captured by different annotation PCs in the heatmap (Figure 2.2).

STAAR incorporating multiple functional annotations

STAAR constructs the weights by modeling the probability of a variant being causal using its functional annotation information via qualitative annotations (e.g. functional categories) and quantitative annotations (e.g. annotation PCs and integrative annotations), as well as modeling the effect sizes of causal variants. Specifically, we consider the effect of variant j on a phenotype can be written as

$$\beta_j = c_j \gamma_j,$$

where c_j is the latent binary indicator of whether variant j is causal, and γ_j is the effect size of variant j if it is causal. The burden test, SKAT, and ACAT-V make direct assumptions on the variance of β_j using MAF information. This newly proposed variant effect model is expected to increase association power since a variant's causal status can be prioritized using its functional annotations (59, 60). Let $\pi_j = E(c_j)$ denote the probability of variant j being causal, then the effect of variant j given above is equivalent to

$$\beta_j = (1 - \pi_j)\delta_0 + \pi_j \gamma_j,$$

where δ_0 is the Dirac delta function indicating that with probability $1 - \pi_j$, variant j has no association with the phenotype.

Define $\hat{\pi}_{jk}$ as the estimated probability of jth variant being causal using the kth annotation ($k = 0, \cdots, K$), e.g., $\hat{\pi}_{j1}$ measures the estimated probability that the jth variant is causal using epigenetic annotation, aPC-Epigenetic. We estimate $\hat{\pi}_{jk}$ using the empirical CDF of the kth annotation for variant j using its rank among all variants as

$$\hat{\pi}_{jk} = ECDF_k(A_{jk}) = \frac{rank(A_{jk})}{M},$$

where A_{jk} is the kth annotation for the jth variant. For $k = 0$, we set $A_{j0} = 1$ as the intercept, which gives $\hat{\pi}_{j0} = 1$. For a quantitative annotation, A_{jk} represents its numeric value, e.g., the kth annotation PC. The quantitative A_{jk} we consider in this chapter include 10 aPCs (Supplementary Table 2.1) and existing integrative scores, including CADD (11), LINSIGHT (21), and FATHMM-XF (116). For a qualitative annotation, we define $A_{jk} = 1$ for variants in the functional group (yes) and $A_{jk} = 0$ for variants otherwise (no). For example, A_{jk} denotes whether a variant is a disruptive missense variant using MetaSVM (81). Hence, $\hat{\pi}_{jk} = 1$ for variants in the functional group and $\hat{\pi}_{jk} = 0$ otherwise, e.g., disruptive missense variants (yes/no). This corresponds to the RV tests using variants of this functional group.

In the STAAR framework, we model the effect sizes of causal variants γ_j in the same way as that used in conventional variant-set tests. Specifically, we assume $|\gamma_j| \propto w_j$, where w_j is assumed as a function of MAFs. For simplicity, we model w_j using $Beta(\mathrm{MAF}_j; a_1, a_2)$ and set (a_1, a_2) to be $(1,1)$ or $(1,25)$. Then, the burden test statistic using kth variant functional annotation as the

weight, e.g., aPC-Epigenetic, is given by $Q_{Burden,k} = \left(\sum_{j=1}^{p} \hat{\pi}_{jk} w_j S_j\right)^2$, whose P value is denoted by $p_{Burden,k}$ $(k = 0, \cdots, K)$. Under the assumption of SKAT, by estimating the probability of jth variant being causal using the kth annotation $(k = 0, \cdots, K)$, we have $E(\beta_j) = 0$ and $Var(\beta_j) = Var(c_j \gamma_j) = \pi_{jk} w_j^2 \tau_k$. Hence, the SKAT test statistic using kth variant functional annotation as the weight is given by

$$Q_{SKAT,k} = \sum_{j=1}^{p} \hat{\pi}_{jk} w_j^2 S_j^2,$$

whose P value is denoted by $p_{SKAT,k}$ $(k = 0, \cdots, K)$. In the ACAT-V test, the test statistic using kth variant functional annotation as the weight is given by

$$Q_{ACAT-V,k} = \overline{\hat{\pi}_{\cdot k} w^2 \mathrm{MAF}(1 - \mathrm{MAF})} \tan\left(\left(0.5 - p_{0,k}\right)\pi\right) + \sum_{j=1}^{p'} \hat{\pi}_{jk} w_j^2 \mathrm{MAF}_j\left(1 - \mathrm{MAF}_j\right) \tan\left(\left(0.5 - p_j\right)\pi\right),$$

where $\overline{\hat{\pi}_{\cdot k} w^2 \mathrm{MAF}(1 - \mathrm{MAF})}$ is the average of the weights $\hat{\pi}_{jk} w_j^2 \mathrm{MAF}_j\left(1 - \mathrm{MAF}_j\right)$ among the extremely rare variants with MAC ≤ 10. The P value of $Q_{ACAT-V,k}$ is denoted by $p_{ACAT-V,k}$ $(k = 0, \cdots, K)$.

We denote by $p_{Burden,k}$, $p_{SKAT,k}$, $p_{ACAT-V,k}$ the P values of burden, SKAT, and ACAT-V tests, respectively calculated using the kth annotation as the weight. For each type of RV tests, to robustly aggregate information from multiple annotations to boost power RV association tests in a data-adaptive manner, we propose to use the STAAR framework to combine individual

annotation weighted tests using the ACAT *P* value combination method (55, 117). Specifically, we define STAAR-Burden (STAAR-B), STAAR-SKAT (STAAR-S), and STAAR-ACAT-V (STAAR-A) as

$$T_{STAAR-B} = \sum_{k=0}^{K} \frac{tan\{(0.5 - p_{Burden,k})\pi\}}{K+1},$$

$$T_{STAAR-S} = \sum_{k=0}^{K} \frac{tan\{(0.5 - p_{SKAT,k})\pi\}}{K+1},$$

$$T_{STAAR-A} = \sum_{k=0}^{K} \frac{tan\{(0.5 - p_{ACAT-V,k})\pi\}}{K+1}.$$

The *P value* of $T_{STAAR-S}$, $T_{STAAR-B}$, and $T_{STAAR-A}$ can be approximated by

$$p_{STAAR-B} \approx \frac{1}{2} - \frac{\{arctan(T_{STAAR-B})\}}{\pi},$$

$$p_{STAAR-S} \approx \frac{1}{2} - \frac{\{arctan(T_{STAAR-S})\}}{\pi},$$

$$p_{STAAR-A} \approx \frac{1}{2} - \frac{\{arctan(T_{STAAR-A})\}}{\pi}.$$

To further aggregate information from different types tests and different weights, we propose an omnibus test in the STAAR framework (STAAR-O) by combining STAAR-B, STAAR-S and STAAR-A using the ACAT method (55, 117). We define the STAAR-O test statistic as

$$T_{STAAR-O} = \frac{1}{3|\mathcal{A}|} \sum_{(a_1,a_2)\in\mathcal{A}} \left[tan\left\{\left(0.5 - p_{STAAR-B(a_1,a_2)}\right)\pi\right\} + tan\left\{\left(0.5 - p_{STAAR-S(a_1,a_2)}\right)\pi\right\} + tan\left\{\left(0.5 - p_{STAAR-A(a_1,a_2)}\right)\pi\right\}\right],$$

where $p_{STAAR-B(a_1,a_2)}$, $p_{STAAR-S(a_1,a_2)}$, and $p_{STAAR-A(a_1,a_2)}$ denote the *P value*s of STAAR-B, STAAR-S, and STAAR-A using $w_j = Beta(\mathrm{MAF}_j; a_1, a_2)$, $\mathcal{A}$ is the set of specified values of (a_1, a_2), and $|\mathcal{A}|$ is the size of set $\mathcal{A}$. In practice, we set $\mathcal{A} = \{(1,25), (1,1)\}$. The *P value* of $T_{STAAR-O}$ could then be accurately approximated by

$$p_{STAAR-O} \approx \frac{1}{2} - \frac{\{arctan(T_{STAAR-O})\}}{\pi}.$$

By combining different types of tests into an omnibus test, STAAR-O has a robust power with respect to the sparsity of causal variants and the directionality of effects of causal variants in a variant-set, as well as variant multi-facet functions and MAFs. Specifically, by including the burden test, STAAR-O is powerful when majority of variants in a variant-set are causal and have effects in the same direction; by including SKAT, STAAR-O is powerful when not a small number of variants in a variant-set are causal with effects in different directions, or when variants in a variant-set are in high LD; by including ACAT-V, STAAR-O is powerful when a small number of variants in a variant-set are causal or a good number of extremely rare variants are causal; by weighting each type of tests using multiple annotation PCs and other integrative functional scores and qualitative annotations, STAAR-O is powerful when any of these variant functional annotations can pinpoint causal variants and help boost power.

Data simulation

Type I error simulations

We performed extensive simulation studies to evaluate whether the proposed STAAR framework preserves the desired type I error rate. We generated continuous traits from a linear model defined as

$$Y_i = 0.5X_{1i} + 0.5X_{2i} + \epsilon_i,$$

where $X_{1i} \sim N(0,1), X_{2i} \sim \text{Bernoulli}(0.5)$, and $\epsilon_i \sim N(0,1)$. Dichotomous traits were generated from a logistic model defined as

$$\text{logit } P(Y_i = 1) = \alpha_0 + 0.5X_{1i} + 0.5X_{2i},$$

where X_{1i} and X_{2i} were defined the same as continuous traits and α_0 was determined to set the prevalence to 1%. In this setting, we used a balanced case-control design. We generated genotypes by simulating 20,000 sequences for 100 different regions each spanning 1 Mb. The data were generated to mimic the LD structure of an African American population by using the calibration coalescent model (COSI) (77). In each simulation replicate, 10 annotations were generated as $A_1, \ldots, A_{10}$ i.i.d. $N(0,1)$ for each variant, and we randomly selected 5-kb regions from these 1-Mb regions for type I error simulations. We applied STAAR-B, STAAR-S, STAAR-A, and STAAR-O by incorporating MAFs and the 10 annotations and repeated the procedure with 10^9 replicates to examine the type I error rate at $\alpha = 10^{-5}, 10^{-6}, 10^{-7}$ levels. Total sample sizes considered were 2,500, 5,000, and 10,000.

Empirical power simulations.

Next, we carried out simulation study under a variety of configurations to assess the power gain by incorporating multiple functional annotations using STAAR compared to conventional variant-set tests that use MAFs as weights. In each simulation replicate, we randomly selected 5-kb regions from these 1-Mb regions for power simulations. For each selected 5-kb region, we generated causal variants according to a logistic model defined as

$$\text{logit}\, P\big(c_j = 1\big) = \delta_0 + \delta_{k_1} A_{j,k_1} + \delta_{k_2} A_{j,k_2} + \delta_{k_3} A_{j,k_3} + \delta_{k_4} A_{j,k_4} + \delta_{k_5} A_{j,k_5},$$

where $\{k_1, \cdots, k_5\} \subset \{1, \cdots, 10\}$ were randomly sampled for each region. For different regions, causality of variants was allowed to be dependent on different sets of annotations. We set $\delta_{k_l} = \log(5)$ for all annotations and varied the proportions of causal variants in the signal region by setting $\delta_0 = \text{logit}(0.0015)$, $\text{logit}(0.015)$, and $\text{logit}(0.18)$ for averaging 5%, 15% and 35% causal variants in the signal region, respectively.

We generated continuous traits from a linear model given by

$$Y_i = 0.5X_{1i} + 0.5X_{2i} + \beta_1 G_{1j} + \cdots + \beta_s G_{sj} + \epsilon_i,$$

where $X_{1i}, X_{2i}, \epsilon_i$ were defined the same as the type I error simulations, $G_{1j}, \ldots, G_{sj}$ were the genotypes of the *s* causal variants in the signal region, and $\beta_1, \ldots, \beta_s$ were the corresponding effect sizes of causal variants. Dichotomous traits were generated from a logistic model given by

$$\text{logit}\, P(Y_i = 1) = 0.5X_{1i} + 0.5X_{2i} + \beta_1 G_{1j} + \cdots + \beta_s G_{sj},$$

where α_0, X_{1i}, X_{2i} were defined the same as the type I error simulations, $G_{1j}, \ldots, G_{sj}$ were the genotypes of the s causal variants in the signal region, and $\beta_1, \ldots, \beta_s$ were the corresponding log ORs of the s causal variants.

Under both settings, we model the effect sizes of causal variants using $\beta_j = \gamma_j = c_0 |\log_{10} MAF_j|$. The effect size of causal variant was therefore a decreasing function of MAF. For continuous traits, c_0 was set to be 0.13. For dichotomous traits, c_0 was set to be 0.255, which gives an odds ratio of 3 for a variant with MAF of 5×10^{-5}. For each setting, we additionally varied the proportions of causal variant effect size directions by setting 100%, 80%, and 50% variants to have positive effects. Finally, we performed simulations using different magnitudes of effect sizes by varying the values of c_0 across a wide range. We applied STAAR-B, STAAR-S, STAAR-A, and STAAR-O using MAFs and all 10 annotations in the weighting scheme, and repeated the procedure with 10^4 replicates to examine the powers at $\alpha = 10^{-7}$ level. Total sample sizes considered were 10,000 across all settings.

Computation cost

To test the computation time of 500,000 related samples, we simulated 1,000 genomic regions, each with 100 variants, for 1 million haplotypes of 125,000 families with 2 parents and 2 children per family. The computation time for WGS RVAS was estimated by analyzing 2.5 million variant-sets with on average 100 variants in each set using STAAR.

Statistical analysis of lipid traits in the TOPMed data

The TOPMed WGS data consist of ancestrally diverse and multi-ethnic related samples (79). Race/ethnicity was defined using a combination of self-reported race/ethnicity and study

recruitment information. The discovery cohorts consist of 4,580 (37.2%) Black or African American, 6,266 (50.9%) White, 543 (4.4%) Asian American, and 927 (7.5%) Hispanic/Latino American. The replication cohorts consist of 3,534 (19.8%) Black or African American, 11,662 (65.4%) White, 132 (0.7%) Asian American, and 2,494 (14.0%) others. The "others" category in the replication cohort includes many Hispanic/Latino American as well as a cohort of Samoans.

We applied STAAR-O to identify RV-sets associated with four quantitative lipid traits (LDL, HDL, TG and TC) using the TOPMed WGS data. LDL-C and TC were adjusted for the presence of medications as before (78). Linear regression model adjusting for age, age^2, sex was first fit for each study-race/ethnicity-specific group. In addition, for Old Order Amish (OOA), we also adjusted for *APOB* p.R3527Q in LDL-C and TC analyses and adjusted for *APOC3* p.R19Ter in TG and HDL-C analyses (78). The residuals were rank-based inverse normal transformed and rescaled by the standard deviation of the original phenotype within each group. We then fit a heteroscedastic linear mixed model (HLMM) for the rank normalized residuals, adjusting for 10 ancestral PCs, study-ethnicity group indicators, and a variance component for empirically derived kinship matrix plus separate group-specific residual variance components to account for population structure and relatedness. The output of HLMM was then used to perform following variant set analyses for rare variants (MAF < 1%) by scanning the genome, including gene-centric analysis using five variant categories (pLoF RVs, missense RVs, synonymous RVs, promoter RVs, and enhancer RVs) for each protein coded gene, and agnostic genetic region analysis using 2-kb sliding windows across the genome with a 1-kb skip length. The WGS RVAS analysis was performed using the R package STAAR (version 0.9.5).

The aPCs provide diverse and complementary information on variant functionality, and are incorporated in rare variant association tests using an omnibus weighting scheme via the proposed STAAR method. We demonstrate using the following example that STAAR boosts the rare variant association test power by properly upweighting known LDL-associated functional rare variants. For example, the association between a 2-kb sliding window located at 55,038,498 bp - 55,040,497 bp on chromosome 1 and LDL-C using STAAR-O is more significant than conventional tests in unconditional analysis (Supplementary Table 2.14). This power gain of STAAR-O is due to upweighting functional variants, e.g., the known tolerated missense variant rs11591147 within the sliding window through incorporating multiple aPCs (118). Specifically, the aPC-Epigenetic, aPC-Protein, and aPC-Mappability PHRED scores are greater than 20 (top 1% across the genome), and the aPC-MutationDensity, aPC-TF, and CADD PHRED scores are greater than 10 (top 10% across the genome) for this variant, highlighting the multi-dimensional functionality of this variant. The aPC-Protein and aPC-Mappability weighted SKAT *P value*s are 6.69×10^{-13} and 3.78×10^{-12}, which are more significant than SKAT ($P = 1.12 \times 10^{-9}$) and burden test ($P = 4.68 \times 10^{-4}$).

Statistical analysis of LDL-C in the UK Biobank data

We used UK Biobank whole exome sequences (WES) from the functionally equivalent (FE) pipeline. Sample and variant quality control measures were previously described (101, 119). In brief, samples with mismatch between genetically inferred and reported sex, high rates of heterozygosity or contamination (D-stat > 0.4), low sequence coverage (less than 85% of targeted bases achieving 20× coverage), duplicates, and WES variants discordant with genotyping chip were removed. A total of 43,243 individuals with genetically inferred European

ancestry were included; 40,519 of those had data on LDL cholesterol. Total cholesterol was adjusted by dividing the value by 0.8 among individuals reporting lipid lowering medication use after 1994 or statin use at any time point. LDL cholesterol was calculated from adjusted total cholesterol levels by the Friedewald equation for individuals with triglyceride levels < 400 mg/dl. If LDL cholesterol levels were directly measured, then their values were divided by 0.7 among reporting lipid lowering medication use after 1994 or statin use at any time point. Residuals were created after adjustment for age, age^2, sex, and the first 10 ancestral principal components. Residuals were then rank-based inverse-normal transformed and multiplied by the standard deviation. Analyses were restricted to missense variants in the *NPC1L1* gene predicted to be damaging according to the MetaSVM prediction algorithm and conditioned on ten known common variants in *NPC1L1* associated with LDL-C (rs10234070, rs73107473, rs2072183, rs41279633, rs17725246, rs2073547, rs10260606, rs217386, rs7791240, rs2300414) obtained from the UK Biobank imputed genotype data. We performed a burden test for the association between disruptive missense RVs in *NPC1L1* and LDL-C.

Code availability

STAAR is implemented as an open source R package available at https://github.com/xihaoli/STAAR and https://content.sph.harvard.edu/xlin/software.html.

Data availability

This chapter used the TOPMed Freeze 5 Whole Genome Sequencing data and lipids phenotype data. The genotype and phenotype data are both available in dbGAP. The discovery phase used

the data from the following four study cohorts, where the accession numbers are provided in parenthesis: Framingham Heart Study (phs000974.v1.p1), Old Order Amish (phs000956.v1.p1), Jackson Heart Study (phs000964.v1.p1), and Multi-Ethnic Study of Atherosclerosis (phs001416.v1.p1). The replication phase used the data from the following ten study cohorts: Atherosclerosis Risk in Communities Study (phs001211), Cleveland Family Study (phs000954), Cardiovascular Health Study (phs001368), Diabetes Heart Study (phs001412), Genetic Study of Atherosclerosis Risk (phs001218), Genetic Epidemiology Network of Arteriopathy (phs001345), Genetics of Lipid Lowering Drugs and Diet Network (phs001359), San Antonio Family Heart Study (phs001215), Genome-wide Association Study of Adiposity in Samoans (phs000972) and Women's Health Initiative (phs001237). The sample sizes, ethnicity and phenotype summary statistics of these cohorts are given in Supplementary Table 2.3.

The functional annotation data are publicly available and were downloaded from the following links: GRCh38 CADD v1.4 (https://cadd.gs.washington.edu/download), ANNOVAR dbNSFP v3.3a (https://annovar.openbioinformatics.org/en/latest/user-guide/download), LINSIGHT (https://github.com/CshlSiepelLab/LINSIGHT), FATHMM-XF (http://fathmm.biocompute.org.uk/fathmm-xf), CAGE (https://fantom.gsc.riken.jp/5/data), GeneHancer (https://www.genecards.org), and Umap/Bismap (https://bismap.hoffmanlab.org). In addition, recombination rate and nucleotide diversity were obtained from Gazal et al(120). The tissue-specific functional annotations were downloaded from ENCODE (https://www.encodeproject.org/report/?type=Experiment).

CHAPTER III

Powerful and resource-efficient meta-analysis of rare variant associations in large whole-genome sequencing studies at scale

Xihao Li, Zilin Li, Corbin Quick, et. al, NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium, TOPMed Lipids Working Group, Jerome I. Rotter, Cristen J. Willer, Pradeep Natarajan, Gina M. Peloso and Xihong Lin

Abstract

Meta-analysis of whole-genome/exome sequencing (WGS/WES) studies has provided an exciting solution to leverage large sample sizes for the discovery of coding and noncoding rare variants (RVs) associated with complex human traits. Existing RV meta-analysis approaches are not scalable when applied to WGS/WES data due to the very large number of RVs whose summary-level information needs to be stored and shared. We propose MetaSTAAR, a powerful and resource-efficient RV meta-analysis framework scalable to large cohort and biobank WGS/WES studies with hundreds of millions of RVs across the genome, while accounting for relatedness and population structure for both quantitative and dichotomous traits. Through meta-analysis of four lipid traits in 30,138 ancestrally diverse samples from 14 studies of the Trans-Omics for Precision Medicine program, we demonstrated that MetaSTAAR performed resource-efficient RV meta-analysis at scale and identified several conditionally significant RV associations with lipids.

Introduction

Ongoing large-scale whole-genome/exome sequencing studies, such as the Trans-Omics for Precision Medicine (TOPMed) Program of the National Heart, Lung and Blood Institute (NHLBI) (79), the Genome Sequencing Program (GSP) of the National Human Genome Research Institute, and UK Biobank WES Program (101), have provided invaluable insights into uncovering the genetic contributions of both coding and noncoding RVs (minor allele frequency (MAF) < 1%) to many complex diseases and traits. Because single-variant analyses are typically underpowered to identify RV associations (49), variant set tests have been proposed by jointly analyzing the effects of multiple RVs to improve power (51-55). In addition, it is well known that single studies are underpowered to detect small to moderate genetic effects (121). As such, meta-analysis of data from comparable WGS/WES studies provides a natural and cost-effective solution to augment sample sizes and increase power for genetic discovery (122). Compared to the joint analysis of pooled individual-level data, meta-analysis only requires summary-level data to be shared from each study, which protects the data privacy of study participants, bypasses the cumbersome genotype and phenotype data harmonization, and results in smaller shareable data sizes. More importantly, the statistical power of meta-analysis is asymptotically equivalent to that of pooled analysis (123), making meta-analysis an essential tool for analyzing RV associations in large-scale WGS/WES studies, especially when individual-level data across studies cannot be shared.

Existing methods have been proposed to perform meta-analysis of RVs in genetic association studies (124-127). However, these methods require $\mathcal{O}(n^2)$ computation time and summary-level data storage for a participating study, where n is the sample size, which are not scalable to large

WGS studies. Here we propose the Meta-analysis of variant-Set Test for Association using Annotation infoRmation (MetaSTAAR), a general framework to perform RV meta-analysis for large-scale WGS studies with hundreds of millions of RVs across the genome. MetaSTAAR accounts for relatedness and population structure for both quantitative and dichotomous traits by fitting the null GLMMs using sparse genetic relatedness matrices (GRMs) (43, 74, 128). By calculating and storing a new form of summary-level data shared across studies, MetaSTAAR is computationally scalable and highly resource-efficient for RV meta-analysis of large-scale WGS data, which only requires $\mathcal{O}(n)$ computation time and summary-level data storage without information loss (Methods). Furthermore, MetaSTAAR dynamically incorporates multiple functional annotations to empower RV meta-analysis and could be applied to any analysis units, including gene-centric analysis by grouping variants into functional categories for each gene and genetic region analysis using sliding windows (43). MetaSTAAR also enables conditional analysis to identify RV association signals independent of known variants.

In the present study, we performed extensive simulation studies to demonstrate that MetaSTAAR maintains accurate type I error rates and achieves greater power by incorporating relevant functional annotations for both quantitative and dichotomous phenotypes. By applying MetaSTAAR to perform WGS RV meta-analysis of 30,138 related and ancestrally diverse samples from 14 participating studies with four quantitative lipid traits: low-density lipoprotein cholesterol (LDL-C); high-density lipoprotein cholesterol (HDL-C); triglycerides (TG) and total cholesterol (TC) from the NHLBI TOPMed program, we show that MetaSTAAR is computationally scalable and resource-efficient for large-scale WGS RV meta-analysis, requiring at least 100 times smaller storage and computation time than existing methods. MetaSTAAR also

identifies several conditionally significant RV associations with lipids, after adjusting for known lipid-associated variants.

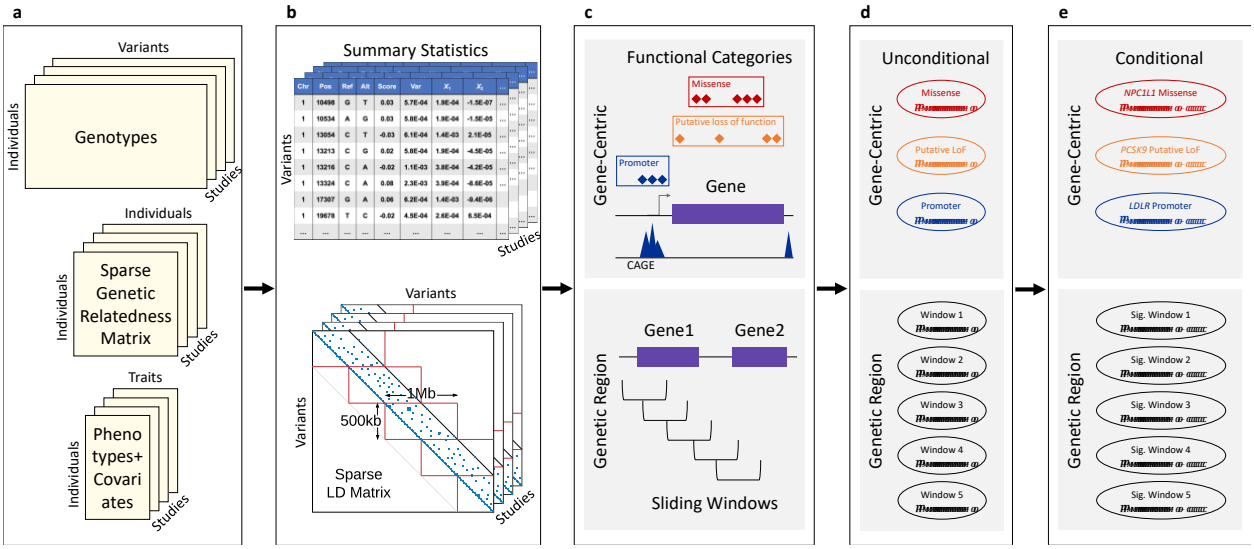
Results

Overview of methods

MetaSTAAR is a general framework to perform powerful and resource-efficient meta-analysis of RV associations in WGS studies at scale, while accounting for relatedness and population structure for both quantitative and dichotomous traits using fast and scalable algorithms. There are two main steps of the MetaSTAAR framework: (i) generating summary-level data for each participating study, referred to as MetaSTAARWorker, and (ii) testing for association between each variant set and phenotypes via meta-analysis by combining these summary-level data across studies and incorporating multiple functional annotations, including annotation principal components (aPCs) (43) (Figure 3.1).

For each participating study, MetaSTAARWorker first fits the null GLMM, including linear and logistic mixed model for quantitative and dichotomous trait, to account for relatedness and population structure (72, 73). It uses sparse GRM and allows for study-specific covariates (for example, ancestral principal components) in fitting the null mixed model to ensure computational efficiency while preserving accuracy (74, 128). MetaSTAARWorker then calculates single-variant score statistics and their variances (summary statistics) for all polymorphic variants in the study, which can be used to perform single-variant meta-analysis(129). For meta-analysis of RVs, one of the most time-consuming and resource-

Figure 3.1 MetaSTAAR workflow.



a, Input data of MetaSTAAR for each study, including genotypes, phenotypes, covariates, and sparse genetic relatedness matrix is prepared. b, Summary statistics and sparse LD matrices for each study are generated using MetaSTAARWorker. c, All RVs in the merged variant list are annotated (including annotation principal components) and two types of variant sets are defined: gene-centric analysis by grouping variants into functional genomic elements for each protein-coding gene; and genetic region analysis using agnostic sliding windows. d, The MetaSTAAR-O *P* values for all variant sets defined in c are obtained. e, The conditional MetaSTAAR-O *P* values for all significant variant sets from d after adjusting for known variants are obtained and reported.

demanding components is generating the variance-covariance matrices to represent the linkage disequilibrium (LD) structure among RVs. To address this issue, MetaSTAARWorker decomposes the variance-covariance matrix of RVs as the difference between the sparse LD matrix and the cross product of a low-rank dense matrix which captures the covariate effects (**Methods**). It stores the low-rank dense matrix along with the single-variant summary statistics, and stores the LD matrix in sparse matrix format. By storing these two matrices separately, MetaSTAARWorker only requires $\mathcal{O}(n)$ computation and storage without information loss. Compared with existing methods that require $\mathcal{O}(n^2)$ computation and storage (124, 125),

MetaSTAARWorker can efficiently reduce the summary-level data storage, while being able to reconstruct the variance-covariance matrix of RVs.

After collecting the summary-level data from each participating study, MetaSTAAR combines the summary statistics into a merged variant list for any user-specified variant set. MetaSTAAR then calculates the aggregated score statistics and their variance-covariance matrix that corresponds to all RVs in the merged variant list, by using the summary statistics and sparse LD matrices from each study. Since the vast majority of RVs sequenced across the genome are extremely rare variants, a considerable number of RVs are study-specific for WGS/WES meta-analysis (Supplementary Table 3.1a). As such, if a genetic variant is monomorphic in a study, MetaSTAAR will set its single-variant score statistic and the corresponding row and column in the variance-covariance matrix to 0 for that study (124, 125). With the aggregated score statistics and their variance-covariance matrix of a given variant set, MetaSTAAR performs powerful RV meta-analysis by incorporating multiple functional annotations in the weighting scheme using the STAAR framework and outputs the meta-analysis STAAR-O (MetaSTAAR-O) *P* value for the variant set (43). In addition, MetaSTAAR allows dissecting RV association signals independent of a given set of known variants via conditional analysis (Methods).

Application to the TOPMed Lipids WGS data

We applied MetaSTAAR to identify RV associations with four quantitative lipid traits (LDL-C, HDL-C, TG and TC) by meta-analysis of 14 study cohorts in TOPMed Freeze 5 WGS data consisting of 30,138 individuals (Supplementary Note of Appendix B): the Framingham Heart

Study (FHS), the Old Order Amish (OOA), the Jackson Heart Study (JHS), the Multi-Ethnic Study of Atherosclerosis (MESA), the Atherosclerosis Risk in Communities Study (ARIC), the Cleveland Family Study (CFS), the Cardiovascular Health Study (CHS), the Diabetes Health Study (DHS), the Genetic Study of Atherosclerosis Risk (GeneSTAR), the Genetic Epidemiology Network of Arteriopathy (GENOA), the Genetics of Lipid Lowering Drugs and Diet Network (GOLDN), the San Antonio Family Heart Study (SAFS), the Genome-wide Association Study of Adiposity in Samoans (SAS), and the Women's Health Initiative (WHI). LDL-C and TC were adjusted for the presence of medications(78), and DNA samples were sequenced at > 30x target coverage. We performed sample- and variant-level quality control for each participating study(78, 79). Race/ethnicity was defined using a combination of self-reported race/ethnicity and study recruitment information. There were 30,138 ancestrally diverse and multi-ancestry-related samples from these 14 studies in total, consisting of 8,114 (26.9%) Black or African-American individuals, 17,928 (59.5%) White, 675 (2.2%) Asian American and 3,421 (11.4%) Hispanic/Latino American and Samoans. Among these samples, 6,690 (22.2%) had first-degree relatedness, 938 (3.1%) had second-degree relatedness and 769 (2.6%) had third-degree relatedness. There were 255 million of single-nucleotide variants (SNVs) observed overall, consisting of 6.3 million (2.5%) common variants (MAF > 5%), 4.9 million (1.9%) low-frequency variants (1% ≤ MAF ≤ 5%) and 244 million RVs (MAF < 1%). The study-specific demographics, summaries of lipid levels and variant number distributions are given in Supplementary Tables 3.1a and 3.1b.

Runtime and resource requirements of MetaSTAARWorker

We evaluated the computational performance of MetaSTAARWorker, including runtime and resource requirements. For each study, we first applied inverse rank normal transformation to phenotypes, adjusted for age, age^2, sex, race/ethnicity, and the first ten ancestral principal components, and controlled for relatedness using heteroscedastic linear mixed models (HLMM) with sparse GRMs plus ancestry-specific residual variance components (Methods). We then used MetaSTAARWorker to generate and store the summary statistics of all variants and sparse LD matrices of variants whose MAFs are below a study-specific threshold (Supplementary Table 3.2). The MAF threshold is dependent on the relative sample size between studies to ensure all RVs in the pooled samples are included in the meta-analysis. It requires 3 hours for 100 2.10 GHz computing cores with 12 GB memory to generate the summary-level data for each study and each trait. Each trait requires 590 GB on average to store these summary-level data of all 14 cohorts (Supplementary Table 3.2).

We then considered multiple subsets of individuals from the TOPMed lipids Freeze 5 WGS data and compared the computational performance of MetaSTAARWorker and the existing method RareMetalWorker (RMW). Note that RMW does not allow for HLMM of a given study, linear models were performed using both methods for a fair comparison. In summary, MetaSTAARWorker requires at least 100 times smaller storage and computation time than RMW (Table 3.1). In addition, the ratio between RMW and MetaSTAARWorker of both storage and computation time increases as the sample size increases, which is anticipated due to the different order of computation complexity and storage for the two methods.

Table 3.1 Comparison of runtimes and storage of MetaSTAARWorker and RareMetalWorker.

Region	Sample size	MetaSTAARWorker		RareMetalWorker (RMW)		MetaSTAARWorker/RMW	
		CPU hours (h)	Storage (GB)	CPU hours (h)	Storage (GB)	CPU hours	Storage
chromosome 6: 160 Mb – 161 Mb	4,791	0.10	0.01	2.05	1.77	4.69%	0.46%
	12,316	0.14	0.02	10.47	3.77	1.34%	0.41%
	30,138	0.21	0.03	69.94	10.14	0.31%	0.26%
chromosome 16: 0 Mb -12 Mb	4,791	1.58	0.24	80.28	65.33	1.97%	0.37%
	12,316	2.65	0.46	358.04*	123.94*	0.74%	0.37%
	30,138	3.59	0.80	2303.78*	328.79*	0.16%	0.24%

*****Predicted numbers based on partial results. Runtimes and storage of MetaSTAARWorker v0.9.6 (linear model) and RareMetalWorker v4.15.1 (linear model) to generate sparse LD matrices and covariance matrices, respectively. Three datasets from TOPMed Freeze 5 total cholesterol WGS data were used in this benchmarking test: MESA cohort (n = 4,791); TOPMed Freeze 3 data (n = 12,316, including 4 study cohorts FHS, JHS, MESA and OOA described in the Supplementary Note) and TOPMed Freeze 5 data (n = 30,138, including 14 study cohorts described in the Supplementary Note). Two variant sets were considered in this test: all uncommon variants (MAF $\leq$ 5%) from 160 Mb to 161 Mb on chromosome 6 and all uncommon variants from 0 Mb to 12 Mb on chromosome 16. MetaSTAARWorker was performed at a 2.10 GHz computing core with 12 GB memory and RareMetalWorker was performed at the same core with 30 GB memory.

Gene-centric meta-analysis of coding and noncoding RVs

We applied MetaSTAAR-O to perform gene-centric meta-analysis of coding, promoter, and enhancer RVs of genes associated with lipid traits. RVs (pooled MAF < 1%) from five functional categories (masks) of each gene were aggregated and analyzed for each of the four lipid traits, including (i) putative loss-of-function (stop gain, stop loss and splice) RVs, (ii) missense RVs, (iii) synonymous RVs, (iv) promoter RVs with overlap of cap analysis of gene expression (CAGE) sites (8), and (v) enhancer RVs with overlap of CAGE sites (9, 80), where each mask was defined in Methods. We incorporated 10 aPCs (including 1 liver-specific aPC) (43), CADD

(11), LINSIGHT (21), FATHMM-XF (116) and MetaSVM (81) (for missense RVs only) along with the two MAF weights (49) in MetaSTAAR-O. Overall, the distribution of MetaSTAAR-O *P*-values was well calibrated for all four lipid phenotypes (Supplementary Figure 3.1). At a Bonferroni-corrected significance threshold of $\alpha = 0.05/(20{,}000 \times 5) = 5.00 \times 10^{-7}$ accounting for five different masks across protein-coding genes, MetaSTAAR-O identified 53 genome-wide significant associations with four lipid phenotypes using unconditional meta-analysis (Supplementary Table 3.3 and Supplementary Figure 3.2). After conditioning on known lipids-associated variants, 45 out of the 53 associations remained significant at the Bonferroni-corrected threshold of $0.05/53 = 9.43 \times 10^{-4}$, including associations with LDL-C (putative loss-of-function RVs in *PCSK9* and *APOB*, missense RVs in *PCSK9*, *ABCG5*, *NPC1L1*, *LDLR*, and *APOE*, synonymous RVs in *RNF20*, promoter RVs in *LDLR* and *APOE*, enhancer RVs in *LDLR*), associations with HDL-C (putative loss-of-function RVs in *APOC3*, missense RVs in *CD36*, *ABCA1*, *APOC3*, *PCSK7*, *SCARB1*, *CETP*, *LCAT*, and *LIPG*), associations with TG (putative loss-of-function RVs in *APOC3*, missense RVs in *APOA5*, *APOA4*, *APOC3*, *PAFAH1B2*, *APOE*, and *COL18A1*, promoter RVs in *APOA5*, *APOA4*, *APOC3*, and *APOE*, enhancer RVs in *APOA5*, *APOA1*, and *COL18A1*), and associations with TC (putative loss-of-function RVs in *PCSK9* and *APOB*, missense RVs in *PCSK9*, *ABCG5*, *NPC1L1*, *ABCA1*, *LIPG*, *LDLR*, and *APOE*, promoter RVs in *APOE*, enhancer RVs in *LDLR*) (Table 3.2). We then compared the results obtained from MetaSTAAR-O with the results from the joint analysis of pooled data using STAAR-O. All significant and conditionally significant findings using STAAR-O could be detected by MetaSTAAR-O (Table 3.2). Furthermore, the *P* values from MetaSTAAR-O and STAAR-O were highly concordant, with $r^2 > 0.99$ of significant and suggestive significant masks defined by various levels of unconditional *P* value thresholds for

Table 3.2 Gene-centric meta-analysis results of both unconditional analysis and analysis conditional on known common and low-frequency variants.

Trait	Gene	Chr. no.	Category	No. of SNVs	MetaSTAAR-O (Unconditional)	MetaSTAAR-O (Conditional)	Variants (adjusted)
LDL-C	*PCSK9*	1	Putative loss of function	9	3.07E-63	6.42E-63	rs11591147, rs28362263, rs505151, rs12117661, rs472495
	APOB	2	Putative loss of function	16	1.14E-20	2.42E-20	rs1367117, rs563290, rs533617
	PCSK9	1	Missense	167	1.55E-15	3.11E-14	rs11591147, rs28362263, rs505151, rs12117661, rs472495
	ABCG5	2	Missense	148	2.66E-08	6.28E-08	rs4245791
	NPC1L1	7	Missense	293	4.88E-10	2.40E-09	rs217381
	LDLR	19	Missense	192	5.33E-27	5.16E-27	rs12151108, rs688, rs6511720
	APOE	19	Missense	88	2.25E-13	1.91E-12	rs7412, rs429358, rs35136575
	RNF20	9	Synonymous	58	4.25E-08	4.25E-08	n/a
	LDLR	19	Promoter	150	1.46E-17	1.98E-05	rs12151108, rs688, rs6511720
	APOE	19	Promoter	102	7.52E-12	9.98E-12	rs7412, rs429358, rs35136575
	LDLR	19	Enhancer	170	2.17E-17	2.95E-05	rs12151108, rs688, rs6511720
HDL-C	*APOC3*	11	Putative loss of function	7	2.26E-22	7.49E-21	rs964184, rs12269901
	CD36	7	Missense	237	3.18E-07	4.03E-08	rs3211938
	ABCA1	9	Missense	346	6.72E-11	2.00E-11	rs4149310, rs1883025, rs11789603
	APOC3	11	Missense	19	4.93E-07	9.42E-07	rs964184, rs12269901
	PCSK7	11	Missense	116	1.17E-09	2.43E-09	rs964184, rs12269901
	SCARB1	12	Missense	120	7.76E-11	7.41E-11	rs10773112, rs4765127
	CETP	16	Missense	101	1.00E-13	1.73E-08	rs247616, rs5883, rs7499892, rs17231520, rs5880
	LCAT	16	Missense	63	1.56E-10	1.17E-10	rs1109166
	LIPG	18	Missense	101	1.62E-07	1.18E-07	rs8086351, rs9958734
TG	*APOC3*	11	Putative loss of function	7	3.57E-54	1.88E-51	rs964184, rs9804646, rs3135506, rs2266788
	APOA5	11	Missense	64	2.14E-07	2.37E-09	rs964184, rs9804646, rs3135506, rs2266788
	APOA4	11	Missense	118	1.15E-08	8.05E-10	rs964184, rs9804646, rs3135506, rs2266788
	APOC3	11	Missense	18	2.99E-12	1.50E-12	rs964184, rs9804646, rs3135506, rs2266788
	PAFAH1B2	11	Missense	31	2.71E-09	3.76E-10	rs964184, rs9804646, rs3135506, rs2266788
	APOE	19	Missense	89	1.43E-11	1.30E-10	rs12721054, rs5112, rs429358
	COL18A1	21	Missense	588	1.07E-08	1.07E-08	n/a
	APOA5	11	Promoter	15	1.32E-10	4.74E-12	rs964184, rs9804646, rs3135506, rs2266788
	APOA4	11	Promoter	198	8.12E-11	1.45E-09	rs964184, rs9804646, rs3135506, rs2266788
	APOC3	11	Promoter	62	4.72E-11	1.80E-11	rs964184, rs9804646, rs3135506, rs2266788
	APOE	19	Promoter	104	9.50E-18	3.83E-10	rs12721054, rs5112, rs429358
	APOA5	11	Enhancer	13	2.38E-10	8.90E-12	rs964184, rs9804646, rs3135506, rs2266788
	APOA1	11	Enhancer	357	5.04E-10	2.87E-10	rs964184, rs9804646, rs3135506, rs2266788
	COL18A1	21	Enhancer	312	3.97E-09	3.97E-09	n/a
TC	*PCSK9*	1	Putative loss of function	9	4.46E-57	1.23E-56	rs11591147, rs28362263, rs505151, rs12117661, rs2495477
	APOB	2	Putative loss of function	16	3.52E-19	7.85E-19	rs1367117, rs10692845, rs533617
	PCSK9	1	Missense	169	1.94E-11	1.15E-11	rs11591147, rs28362263, rs505151, rs12117661, rs2495477
	ABCG5	2	Missense	157	4.74E-09	1.21E-08	rs4245791
	NPC1L1	7	Missense	301	3.92E-08	1.57E-07	rs217381

Table 3.2 (Continued)

ABCA1	9	Missense	346	6.90E-08	3.72E-08	rs1800978, rs4149310, rs3847302
LIPG	18	Missense	101	2.69E-08	1.39E-08	rs9958734
LDLR	19	Missense	200	1.41E-22	8.33E-23	rs73015024, rs688, rs2278426, rs6511720
APOE	19	Missense	90	1.18E-08	1.49E-08	rs7412, rs429358, rs12721054
APOE	19	Promoter	105	1.92E-07	7.20E-08	rs7412, rs429358, rs12721054
LDLR	19	Enhancer	176	1.22E-15	7.15E-04	rs73015024, rs688, rs2278426, rs6511720

A total of 30,138 samples from 14 study cohorts in TOPMed program were considered in the meta-analysis. Results for the conditionally significant genes (unconditional MetaSTAAR-O $P < 5.00 \times 10^{-7}$; conditional MetaSTAAR-O $P < 9.43 \times 10^{-4}$) are presented in the table. Chr. no., chromosome number; category, functional category; no. of SNVs, number of RVs (pooled MAF < 1%) of the particular functional category in the gene; MetaSTAAR-O, MetaSTAAR-O *P* value; variants (adjusted), adjusted variants in the conditional analysis; n/a, no variant adjusted in the conditional analysis.

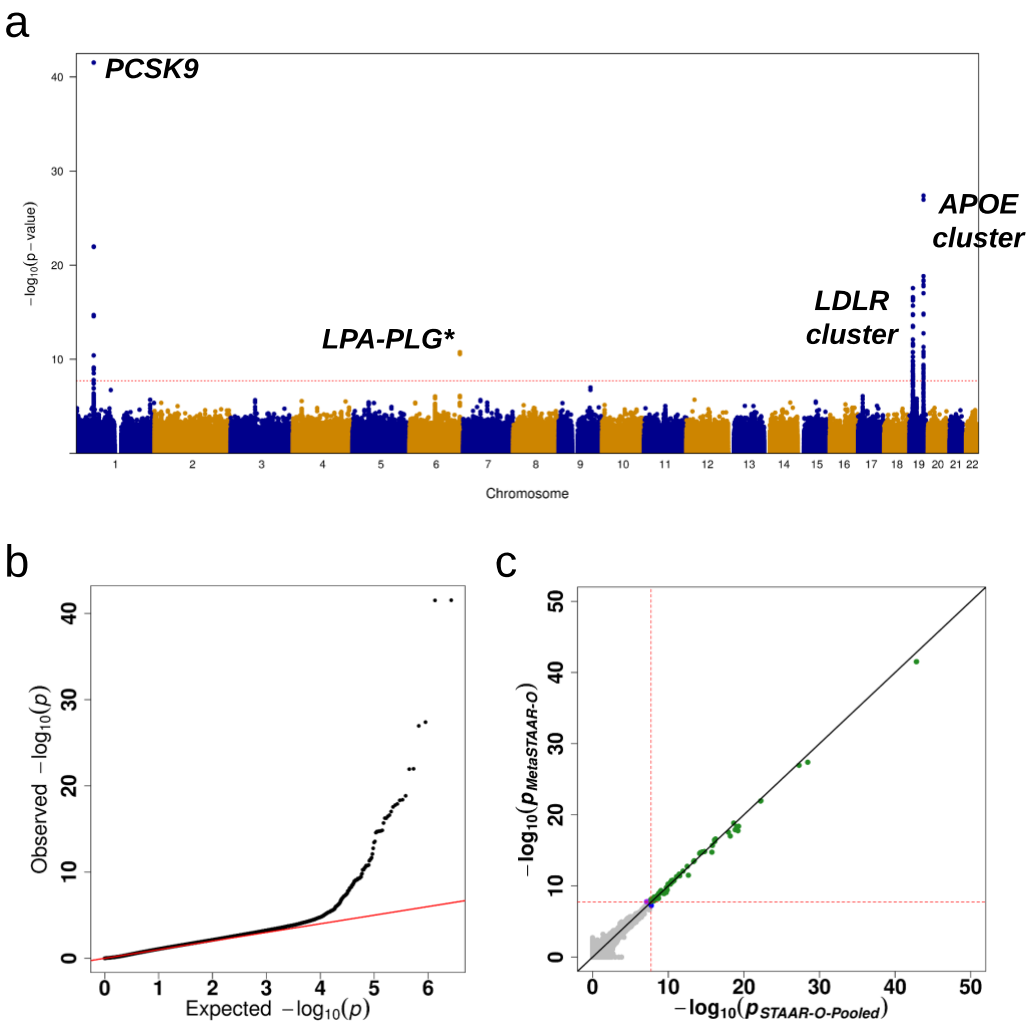
each lipid phenotype (Supplementary Table 3.4a and Supplementary Figure 3.3). In addition, MetaSTAAR-O and STAAR-O delivered highly concordant *P* values in conditional analysis of significant masks ($r^2 > 0.99$) (Supplementary Table 3.4b and Supplementary Figure 3.4).

Genetic region meta-analysis of RVs

We next applied MetaSTAAR-O to perform genetic region meta-analysis of RVs within sliding windows associated with lipid traits. We considered sliding windows to be 2 kb in length, started at position 0 bp for each chromosome and had a skip length of 1 kb. Windows with at least two RVs were included in the meta-analysis, resulting in a total of 2.68 million 2-kb overlapping windows. Same annotations were incorporated as the gene-centric analysis. Overall, the distribution of MetaSTAAR-O *P*-values was well calibrated for all four lipid phenotypes (Figure 3.2b and Supplementary Figures 3.5b, 3.6b and 3.7b). At a Bonferroni-corrected significance threshold of $\alpha = 0.05/(2.68 \times 10^6) = 1.86 \times 10^{-8}$ across sliding windows (Figure 3.2a and Supplementary Figures 3.5a, 3.6a and 3.7a), MetaSTAAR-O identified 268 genome-wide significant associations with four lipid phenotypes using unconditional meta-analysis. After conditioning on known lipids-associated variants, 143 out of the 268 associations remained significant at the Bonferroni-corrected threshold of $0.05/268 = 1.87 \times 10^{-4}$. (Supplementary Tables 3.5-3.8). We also compared the results of MetaSTAAR-O with that of pooled analysis using STAAR-O. Reassuringly, the *P* values from MetaSTAAR-O and STAAR-O were highly concordant, with $r^2 > 0.99$ of significant and suggestive significant sliding windows defined by various levels of unconditional *P* value thresholds for each lipid phenotype (Supplementary Table 3.9a and Supplementary Figures 3.2c, 3.5c, 3.6c and 3.7c). MetaSTAAR-O and STAAR-O

also delivered highly concordant *P* values in conditional analysis of significant sliding windows ($r^2 > 0.99$) (Supplementary Table 3.9b and Supplementary Figure 3.8).

Figure 3.2 Genetic region (2-kb sliding window) unconditional meta-analysis results of LDL-C using the TOPMed data.



a, Manhattan plot showing the associations of 2.68 million 2-kb sliding windows for LDL-C (low-density lipoprotein cholesterol) versus $-\log_{10}(P)$ of MetaSTAAR-O. The horizontal line indicates a genome-wide *P* value threshold of 1.86×10^{-8} (*n* = 30,138). b, Quantile-quantile plot of 2-kb sliding window MetaSTAAR-O *P* values for LDL-C (*n* = 30,138). c, Scatterplot of *P* values for 2-kb sliding windows comparing MetaSTAAR-O with STAAR-O from the joint analysis of pooled individual-level data (STAAR-O-Pooled). Each dot represents a sliding window with x-axis label being the $-\log_{10}(P)$ of STAAR-O-Pooled and y-axis label being the $-\log_{10}(P)$ of MetaSTAAR-O (*n* = 30,138). *Intergenic sliding window.

Simulation studies

We performed simulation studies to evaluate the type I error and power of MetaSTAAR under a variety of configurations. We considered five participating studies in the meta-analysis, each with a sample size of 10,000. Quantitative and dichotomous phenotypes were generated by following the steps described in Data simulation (Appendix C). For each study, genotypes were generated by simulating 20,000 sequences for 20-Mb to mimic the LD structure of an African American population using the calibration coalescent model (COSI) (77). We randomly selected 2-kb regions from the 20-Mb region in simulation studies.

Type I error simulations

For RV meta-analysis of both quantitative and dichotomous traits, we performed 10^9 simulations using MetaSTAAR and evaluated the empirical type I error rates for the burden(51-53), SKAT(54), ACAT-V(55) and STAAR-O tests at $\alpha = 10^{-5}, 10^{-6}, 10^{-7}$ (Supplementary Table 3.10). The results show that all of these four tests based on MetaSTAAR well controlled the type I error rate for both continuous and dichotomous traits at all α levels.

Empirical power simulations

We then examined the empirical power of MetaSTAAR-O under a variety of configurations. MAF and ten annotations were incorporated in the meta-analysis, and power was evaluated as the proportions of P values less than $\alpha = 10^{-7}$ based on 10^4 simulations. We considered different proportions of causal variants (5, 15 and 35% on average) in the signal region, and

allowed the causality of variants to be dependent on different sets of annotations through a logistic model (Methods). The results show that MetaSTAAR-O incorporating all ten annotations had higher power of detecting signal regions than the burden, SKAT, and ACAT-V tests implemented in MetaSTAAR for both quantitative and dichotomous traits across different proportions of effect size directions (Supplementary Figures 3.9-3.12). Our simulation studies indicated that MetaSTAAR-O could achieve considerable power gain through the incorporation of multiple relevant annotations.

Discussion

In this study, we proposed MetaSTAAR as a computationally scalable and resource-efficient framework to perform RV association meta-analysis in large WGS/WES studies, while accounting for population structure and relatedness for both quantitative and dichotomous traits.

We highlighted MetaSTAARWorker, the preliminary step of MetaSTAAR that generates and stores summary-level data, including summary statistics and sparse LD matrices, for each participating study. Existing methods stores the full variance-covariance matrix of RVs and requires $\mathcal{O}(n^2)$ computation and storage, which is not scalable of large-scale WGS studies. Instead, MetaSTAARWorker stores the sparse LD matrix and low-rank matrix that captures the covariate effects separately, and hence only requires $\mathcal{O}(n)$ computation and storage without information loss. MetaSTAARWorker was benchmarked to be at least 100 times smaller in computation time and storage than existing methods using TOPMed WGS data. This notable

gain in computation and storage efficiency of MetaSTAARWorker has made it possible for large-scale WGS RV association meta-analysis.

MetaSTAAR framework enables the dynamic incorporation of multiple functional annotations to boost RV meta-analysis power and allows for any analysis units. MetaSTAAR also provides conditional analysis to distinguish novel RV association signals independent of known variants. In the present study, we focused on gene-centric and genetic region meta-analysis of RVs using MetaSTAAR-O. In a WGS RV meta-analysis using the TOPMed Freeze 5 data consisting of 14 study cohorts, MetaSTAAR-O identified 45 conditionally significant functional categories lipid traits in gene-centric meta-analysis, including *NPC1L1* missense RVs and LDL-C; *CD36*, *APOC3*, *SCARB1* missense RVs and HDL-C; and *NPC1L1* missense RVs and TC that were missed by the burden, SKAT, and ACAT-V tests (Supplementary Table 3.3). In genetic region analysis, MetaSTAAR-O identified 143 conditionally significant sliding windows after conditioning on known variants (Supplementary Tables 3.5-3.8), including the association between a 2-kb sliding window (chromosome 1: 55,051,498 - 55,053,497 bp) located within *PCSK9* and LDL-C, that were not detected using tests without incorporating annotations (Supplementary Table 3.5). These results demonstrate that incorporating multiple functional annotations using MetaSTAAR-O can effectively boost power for WGS RV meta-analysis.

MetaSTAAR framework delivers comparable statistical power in detecting RV association signals compared to the joint analysis of pooled individual-level data. In both gene-centric and genetic region analysis, we showed that the *P* values from MetaSTAAR-O and STAAR-O of

pooled analysis were highly concordant in unconditional analysis of various levels of *P* value thresholds (Figure 3.2c, Supplementary Figures 3.3, 3.5c, 3.6c, 3.7c and Supplementary Tables 3.4a, 3.9a) and conditional analysis of significant variant sets (Supplementary Figures 3.4, 3.8 and Supplementary Tables 3.4b, 3.9b) for each lipid phenotype. The computation time for MetaSTAAR-O to perform WGS RV meta-analysis of 30,000 related samples from 14 study cohorts using the TOPMed data requires 10 hours for 100 2.10 GHz computing cores with 12 GB memory for each lipid trait, which is also comparable to the pooled analysis. These results guarantee that our proposed MetaSTAAR framework provides comparable performance in RV association analysis compared to pooled analysis, while bypassing the cumbersome data harmonization across studies and protecting the data privacy of study participants.

In practice, MetaSTAAR is very flexible and users can determine the RV analysis units and the annotations to be used (23, 43). In this study, we grouped the RVs by functional categories for each protein-coding gene and agnostic sliding windows with fixed length; users can also apply dynamic window analysis with flexible locations and sizes (110). In addition, MetaSTAAR generates phenotype-independent sparse LD matrix for quantitative traits in unrelated samples, hence further saving computation resources in phenome-wide association study (PheWAS) (Methods). It is of interest to extend MetaSTAAR to related samples in PheWAS settings.

Overall, the proposed MetaSTAAR framework is fast, scalable and highly resource-efficient for large WGS/WES studies of hundreds of thousands of samples and hundreds of millions of

variants. Our method is currently the only available solution to perform RV meta-analysis at the scale of large WGS studies.

Methods

Notations and model

Suppose there are K participating studies in the meta-analysis. For the kth study, suppose there are n_k subjects with M_k total variants sequenced in a given variant set. Let $\boldsymbol{Y}_k = \left(Y_{1,k}, \dots, X_{n_k,k}\right)^T$ denote a continuous or dichotomous trait vector with mean $\widehat{\boldsymbol{\mu}}_k = \left(\hat{\mu}_{1,k}, \dots, \hat{\mu}_{n_k,k}\right)^T$; $\boldsymbol{X}_k$ denote the $n_k \times q_k$ design matrix of covariates, such as age, gender, (study-specific) ancestral principal components; and $\boldsymbol{G}_k$ denote the $n_k \times M_k$ genotype matrix of the M_k genetic variants in the variant set. We let $\hat{\boldsymbol{e}}_k = (\hat{e}_{1,k}, \dots, \hat{e}_{n_k,k})^T$ denote the trait residuals adjusting for covariates, population stratification and relatedness, which is generated as follows.

When the data consist of unrelated samples, we consider the following null Generalized Linear Model (GLM)

$$g(\boldsymbol{\mu}_k) = \alpha_{0,k}\mathbf{1}_{n_k} + \boldsymbol{X}_k\boldsymbol{\alpha}_k, \qquad (3.1)$$

where $g(\mu) = \mu$ for a continuous normally distributed trait, $g(\mu) = \text{logit}(\mu)$ for a dichotomous trait, $\alpha_{0,k}$ is an intercept, $\mathbf{1}_{n_k}$ is a column vector of 1's with length n_k, $\boldsymbol{\alpha}_k = \left(\alpha_{1,k}, \dots, \alpha_{q_k,k}\right)^T$ is a vector of regression coefficients for $\boldsymbol{X}_k$. We calculate $\widehat{\boldsymbol{\Sigma}}_k = \hat{\phi}_k\mathbf{I}_{n_k}$ for linear model, where $\hat{\phi}_k$ is an estimate of the residual variance ϕ_k, $\mathbf{I}_{n_k}$ is the identity matrix of dimension $n_k \times n_k$; and

$\widehat{\boldsymbol{\Sigma}}_k = \text{diag}\left(1/\left(\hat{\mu}_{i,k}\left(1-\hat{\mu}_{i,k}\right)\right)\right)$ for logistic model, where $\hat{\mu}_{i,k}$ is the fitted value for individual i under the null model (3.1), and obtain $\hat{\boldsymbol{e}}_k = (\boldsymbol{Y}_k - \widehat{\boldsymbol{\mu}}_k)/\hat{\phi}_k$.

When the data consist of related samples, we consider the following null Generalized Linear Mixed Model (GLMM)(71-73)

$$g(\boldsymbol{\mu}_k) = \alpha_{0,k}\mathbf{1}_{n_k} + \boldsymbol{X}_k\boldsymbol{\alpha}_k + \boldsymbol{b}_k, \tag{3.2}$$

where the random effects $\boldsymbol{b}_k$ account for remaining population structure unaccounted by ancestral PCs and relatedness. We assume that $\boldsymbol{b}_k = \left(b_{1,k}, \dots, b_{n_k,k}\right)^T \sim N(\mathbf{0}, \theta_k\boldsymbol{\Phi}_k)$ with variance component θ_k and known sparse genetic relatedness matrix $\boldsymbol{\Phi}_k$. The remaining variables are defined in the same way as those in the GLM (1). We calculate $\widehat{\boldsymbol{\Sigma}}_k = \widehat{\boldsymbol{R}}_k + \hat{\theta}_k\boldsymbol{\Phi}_k$ with $\widehat{\boldsymbol{R}}_k = \hat{\phi}_k\mathbf{I}_{n_k}$ for linear mixed model; and $\widehat{\boldsymbol{R}}_k = \text{diag}\left(1/\left(\hat{\mu}_{i,k}\left(1-\hat{\mu}_{i,k}\right)\right)\right)$ for logistic mixed model, where $\hat{\mu}_{i,k}$ is the fitted value for individual i under the null model (3.2), and obtain $\hat{\boldsymbol{e}}_k = (\boldsymbol{Y}_k - \widehat{\boldsymbol{\mu}}_k)/\hat{\phi}_k$. Note that we allow for heteroscedastic models with group-specific residual variance components in both linear model and linear mixed model for quantitative traits.

Summary-level data shared by MetaSTAARWorker

We describe the summary-level data to be shared by MetaSTAARWorker, including summary statistics and sparse LD matrices. For the kth study, we first computed and shared a vector of score statistics $\boldsymbol{U}_k = \boldsymbol{G}_k^T\hat{\boldsymbol{e}}_k$ and a vector of corresponding variances $\boldsymbol{V}_k = \left(V_{1,k}, \dots, V_{M_k,k}\right)^T$, where $V_{j,k} = \boldsymbol{G}_{\cdot,j,k}^T\boldsymbol{P}_k\boldsymbol{G}_{\cdot,j,k}$, $\boldsymbol{G}_{\cdot,j,k} = \left(G_{1,j,k}, \dots, G_{n_k,j,k}\right)^T$ and $\boldsymbol{P}_k = \widehat{\boldsymbol{\Sigma}}_k^{-1} -$

$\widehat{\boldsymbol{\Sigma}}_k^{-1}\boldsymbol{X}_k\left(\boldsymbol{X}_k^T\widehat{\boldsymbol{\Sigma}}_k^{-1}\boldsymbol{X}_k\right)^{-1}\boldsymbol{X}_k^T\widehat{\boldsymbol{\Sigma}}_k^{-1}$. We also computed and shared a matrix $\boldsymbol{\Lambda}_k = \boldsymbol{G}_k^T\widehat{\boldsymbol{\Sigma}}_k^{-1}\boldsymbol{X}_k\left(\boldsymbol{X}_k^T\widehat{\boldsymbol{\Sigma}}_k^{-1}\boldsymbol{X}_k\right)^{-1/2}$ which captures the covariate effects. Note that $\boldsymbol{\Lambda}_k$ has the same number of rows as $\boldsymbol{U}_k$ and $\boldsymbol{V}_k$, and was shared in the summary statistics.

We next computed and shared the sparse LD matrix $\widetilde{\boldsymbol{G}}_k^T\widehat{\boldsymbol{\Sigma}}_k^{-1}\widetilde{\boldsymbol{G}}_k$, where $\widetilde{\boldsymbol{G}}_k$ denotes the genotype matrix of variants below a study-specific MAF threshold. Let $\widetilde{\boldsymbol{U}}_k = \widetilde{\boldsymbol{G}}_k^T\hat{\boldsymbol{e}}_k$ and $\widetilde{\boldsymbol{\Lambda}}_k = \widetilde{\boldsymbol{G}}_k^T\widehat{\boldsymbol{\Sigma}}_k^{-1}\boldsymbol{X}_k\left(\boldsymbol{X}_k^T\widehat{\boldsymbol{\Sigma}}_k^{-1}\boldsymbol{X}_k\right)^{-1/2}$ denote the corresponding partition of $\boldsymbol{U}_k$ and $\boldsymbol{\Lambda}_k$, respectively. The MAF threshold is dependent on the relative sample size between studies to ensure all RVs in the pooled analysis are included in the meta-analysis. Note that for quantitative traits with unrelated samples ($\widehat{\boldsymbol{\Sigma}}_k = \hat{\phi}_k\mathbf{I}_{n_k}$), the sparse LD matrix reduced to $\widetilde{\boldsymbol{G}}_k^T\widehat{\boldsymbol{\Sigma}}_k^{-1}\widetilde{\boldsymbol{G}}_k = \hat{\phi}_k^{-1}\widetilde{\boldsymbol{G}}_k^T\widetilde{\boldsymbol{G}}_k$ which is phenotype-independent (up to a scaling constant $\hat{\phi}_k^{-1}$). Thus, MetaSTAARWorker could further save computation resources in phenome-wide association study (PheWAS) by only storing $\widetilde{\boldsymbol{G}}_k^T\widetilde{\boldsymbol{G}}_k$ under this setting.

To store and share the variance-covariance information of all RVs across the genome, we computed the sparse LD matrices using 500-kb banded windows. The 500-kb banded windows guarantee the LD information of RVs whose distances under 500-kb could be recovered. In practice, user can determine the bandwidth of the sparse LD matrices to be shared.

Meta-analysis of rare variant association tests

We are interested in jointly testing the association between RVs in the genetic set and phenotype via meta-analysis. For simplicity, we assume all variants in the given variant set are RVs and observed in all K studies, so that $M = M_1 = \cdots = M_K$. We denote $\widetilde{\boldsymbol{U}} = \sum_{k=1}^{K} \widetilde{\boldsymbol{U}}_k = \left(\widetilde{U}_{(1)}, \ldots, \widetilde{U}_{(M)}\right)^T$ and hence $\mathrm{Cov}\left(\widetilde{\boldsymbol{U}}\right) = \sum_{k=1}^{K} \mathrm{Cov}\left(\widetilde{\boldsymbol{U}}_k\right) = \sum_{k=1}^{K} \widetilde{\boldsymbol{G}}_k^T \widehat{\boldsymbol{\Sigma}}_k^{-1} \widetilde{\boldsymbol{G}}_k - \widetilde{\boldsymbol{\Lambda}}_k \widetilde{\boldsymbol{\Lambda}}_k^T$. For meta-analysis of burden test using MetaSTAAR, the test statistics is given by

$$Q_{Burden-MS} = \left(\sum_{j=1}^{M} w_j \, \widetilde{U}_{(j)}\right)^2,$$

where w_j is the weight defined as a function of minor allele frequency (MAF) for the jth variant (49). $Q_{Burden-MS}$ asymptotically follows a chi-square distribution with 1 degree of freedom under the null hypothesis, and its P value can be obtained analytically while accounting for linkage disequilibrium (LD) between variants (49, 73).

For meta-analysis of SKAT using MetaSTAAR, the test statistic is given by

$$Q_{SKAT-MS} = \sum_{j=1}^{M} w_j^2 \, \widetilde{U}_{(j)}^2.$$

$Q_{SKAT-MS}$ asymptotically follows a mixture of chi-square distributions under the null hypothesis, and its P value can be obtained analytically while accounting for LD between variants (49, 73).

For meta-analysis of ACAT-V using MetaSTAAR, the test statistic is given by

$$Q_{ACAT-V-MS} = \overline{w^2\mathrm{MAF}(1-\mathrm{MAF})}\tan\big((0.5-p_0)\pi\big)$$
$$+\sum_{j=1}^{M'} w_j^2\mathrm{MAF}_j\big(1-\mathrm{MAF}_j\big)\tan\Big(\big(0.5-p_j\big)\pi\Big),$$

where M' is the number of variants with cumulative minor allele count (cMAC) greater than 10, MAF_j is the minor allele frequency of individual variant j in meta-analysis, and p_j is the association P value of variant j corresponding the individual variant score statistics $\widetilde{U}_{(j)}$ for those variants with cMAC $>$ 10. p_0 is the burden test P value of extremely rare variants with cMAC $\leq$ 10 and $\overline{w^2\mathrm{MAF}(1-\mathrm{MAF})}$ is the average of the weights $w_j^2\mathrm{MAF}_j\big(1-\mathrm{MAF}_j\big)$ among the extremely rare variants with cMAC $\leq$ 10. $Q_{ACAT-V-MS}$ can be well approximated by a Cauchy distribution under the null hypothesis, and its P-value can be obtained analytically while accounting for LD between variants (55).

Given a collection of L annotations, let A_{jl} is the lth annotation for the jth variant. We define the MetaSTAAR-O test statistic as

$$T_{MetaSTAAR-O} = \frac{1}{3|\mathcal{A}|}\sum_{(a_1,a_2)\in\mathcal{A}} T_{MetaSTAAR-B(a_1,a_2)} + T_{MetaSTAAR-S(a_1,a_2)} + T_{MetaSTAAR-A(a_1,a_2)}$$
$$= \frac{1}{3|\mathcal{A}|}\sum_{(a_1,a_2)\in\mathcal{A}}\sum_{l=0}^{L}\frac{tan\{\big(0.5-p_{Burden-MS,l,(a_1,a_2)}\big)\pi\}}{L+1}$$
$$+\frac{tan\{\big(0.5-p_{SKAT-MS,l,(a_1,a_2)}\big)\pi\}}{L+1}+\frac{tan\{\big(0.5-p_{ACAT-V-MS,l,(a_1,a_2)}\big)\pi\}}{L+1},$$

where $p_{Burden-MS,l,(a_1,a_2)}$, $p_{SKAT-MS,l,(a_1,a_2)}$, and $p_{ACAT-V-MS,l,(a_1,a_2)}$ are the P values of

$$Q_{Burden-MS,l,(a_1,a_2)} = \left(\sum_{j=1}^{M} \hat{\pi}_{jl} w_{j,(a_1,a_2)} \widetilde{U}_{(j)}\right)^2,$$

$$Q_{SKAT-MS,l,(a_1,a_2)} = \sum_{j=1}^{M} \hat{\pi}_{jl} w_{j,(a_1,a_2)}^2 \widetilde{U}_{(j)}^2,$$

$$Q_{ACAT-V-MS,l,(a_1,a_2)}$$

$$= \overline{\hat{\pi}_{\cdot l} w_{(a_1,a_2)}^2 \mathrm{MAF}(1-\mathrm{MAF})} \tan\left(\left(0.5 - p_{0,l}\right)\pi\right)$$

$$+ \sum_{j=1}^{M'} \hat{\pi}_{jl} w_{j,(a_1,a_2)}^2 \mathrm{MAF}_j\left(1-\mathrm{MAF}_j\right) \tan\left(\left(0.5 - p_j\right)\pi\right),$$

respectively. Here $\hat{\pi}_{jl} = \frac{rank(A_{jl})}{p}$, where p is the number of variants across the whole genome, $w_{j,(a_1,a_2)} = Beta\left(\mathrm{MAF}_j; a_1, a_2\right)$ with $(a_1, a_2) = (1,25)$ or $(1,1)$, and $\overline{\hat{\pi}_{\cdot l} w_{(a_1,a_2)}^2 \mathrm{MAF}(1-\mathrm{MAF})}$ is the average of the weights $\hat{\pi}_{jl} w_{j,(a_1,a_2)}^2 \mathrm{MAF}_j\left(1-\mathrm{MAF}_j\right)$ among the extremely rare variants with cMAC $\leq$ 10. The P value of $T_{MetaSTAAR-O}$ could be calculated by

$$p_{MetaSTAAR-O} = \frac{1}{2} - \frac{\{arctan(T_{MetaSTAAR-O})\}}{\pi}.$$

MetaSTAAR-O is an omnibus test that has a robust power with respect to the sparsity of causal variants and the directionality of effects of causal variants in a variant set, as well as variant multi-facet functions and MAFs.

In WGS RV meta-analysis, it is very often that some variants may be observed in only a subset of studies but not the others (Supplementary Table 3.1a). If a variant j was not observed in study

k, the jth entry of $\widetilde{\boldsymbol{U}}_k$ and the (j, j') and (j', j)-th entries of $\mathrm{Cov}\left(\widetilde{\boldsymbol{U}}_k\right)$ were set to 0 for all j' in the merged variant list (124, 125).

Conditional meta-analysis using MetaSTAAR

We implemented conditional analysis in MetaSTAAR to perform meta-analysis of RV association tests adjusting for a given list of known variants (130). We first generated the LD matrix between RVs in the variant set and known variants to be adjusted. Following the notations before and let $\boldsymbol{G}_k^{(c)}$ denote the $n_k \times M^{(c)}$ genotype matrix of $M^{(c)}$ known variants to be adjusted for in conditional analysis. The score statistics vector and the corresponding variance-covariance matrix of these adjusted variants were given by $\boldsymbol{U}_k^{(c)} = \boldsymbol{G}_k^{(c)T}\hat{\boldsymbol{e}}_k$ and $\mathrm{Cov}\left(\boldsymbol{U}_k^{(c)}\right) = \boldsymbol{G}_k^{(c)T}\boldsymbol{P}_k\boldsymbol{G}_k^{(c)}$, respectively. The covariance matrix between RVs in the variant set and adjusted variants is given by $\mathrm{Cov}\left(\widetilde{\boldsymbol{U}}_k, \boldsymbol{U}_k^{(c)}\right) = \widetilde{\boldsymbol{G}}_k^T\boldsymbol{P}_k\boldsymbol{G}_k^{(c)}$. MetaSTAAR additionally requires these three components to perform conditional analysis from each study, i.e. $\boldsymbol{U}_k^{(c)}$, $\mathrm{Cov}\left(\boldsymbol{U}_k^{(c)}\right)$, and $\mathrm{Cov}\left(\widetilde{\boldsymbol{U}}_k, \boldsymbol{U}_k^{(c)}\right)$.

To perform conditional meta-analysis of RV association tests, we calculated the adjusted score statistics vector

$$\widetilde{\boldsymbol{U}}_{adj} = \widetilde{\boldsymbol{U}} - \left[\sum_{k=1}^{K} \mathrm{Cov}\left(\widetilde{\boldsymbol{U}}_k, \boldsymbol{U}_k^{(c)}\right)\right]\left[\sum_{k=1}^{K} \mathrm{Cov}\left(\boldsymbol{U}_k^{(c)}\right)\right]^{-1} \sum_{k=1}^{K} \boldsymbol{U}_k^{(c)},$$

and hence

$$\mathrm{Cov}(\widetilde{\boldsymbol{U}}_{adj}) = \mathrm{Cov}(\widetilde{\boldsymbol{U}}) - \left[\sum_{k=1}^{K} \mathrm{Cov}\left(\widetilde{\boldsymbol{U}}_k, \boldsymbol{U}_k^{(c)}\right)\right]\left[\sum_{k=1}^{K} \mathrm{Cov}\left(\boldsymbol{U}_k^{(c)}\right)\right]^{-1}\left[\sum_{k=1}^{K} \mathrm{Cov}\left(\widetilde{\boldsymbol{U}}_k, \boldsymbol{U}_k^{(c)}\right)\right]^{T}.$$

The test statistics of conditional analysis of each test in MetaSTAAR were calculated in the same way as discussed before, with $\widetilde{\boldsymbol{U}}_{adj}$ and $\mathrm{Cov}(\widetilde{\boldsymbol{U}}_{adj})$ instead of $\widetilde{\boldsymbol{U}}$ and $\mathrm{Cov}(\widetilde{\boldsymbol{U}})$.

Meta-analysis of lipid traits in the TOPMed data

The TOPMed WGS data consist of ancestrally diverse and multi-ethnic related samples (79). Race/ethnicity was defined using a combination of self-reported race/ethnicity and study recruitment information. We applied MetaSTAAR to perform RV meta-analysis of four quantitative lipid traits (LDL, HDL, TG and TC) using 14 study cohorts from the TOPMed Freeze 5 WGS data. LDL-C and TC were adjusted for the presence of medications as before (78). For each study, we first fit linear regression model adjusting for age, age^2, sex for each race/ethnicity-specific group. In addition, for Old Order Amish (OOA), we also adjusted for *APOB* p.R3527Q in LDL-C and TC analyses and adjusted for *APOC3* p.R19Ter in TG and HDL-C analyses (78). We performed rank-based inverse normal transformation of the residuals and rescaled these residuals by the standard deviation of the original phenotype within each group. We then fit a heteroscedastic linear mixed model (HLMM) for the rank normalized residuals, adjusting for 10 ancestral PCs, ethnicity group indicators, and a variance component for empirically derived kinship matrix plus separate ancestry-specific residual variance components to account for population structure and relatedness. The output of HLMM was then used to generate summary-level data by MetaSTAARWorker, including summary statistics of all variants and sparse LD matrices of variants whose MAFs are below a study-specific threshold

(Supplementary Table 3.2). We next used MetaSTAAR-O to perform RV meta-analysis based on the summary-level data of the 14 study cohorts, including gene-centric analysis using five variant functional categories (pLoF RVs, missense RVs, synonymous RVs, promoter RVs, and enhancer RVs) for each protein-coding gene, and genetic region analysis using 2-kb sliding windows across the genome with a 1-kb skip length. The WGS RV meta-analysis was performed using the R package MetaSTAAR (version 0.9.6).

Data availability

This chapter used the TOPMed Freeze 5 Whole Genome Sequencing data and lipids phenotype data. The genotype and phenotype data are both available in dbGAP. The discovery phase used the data from the following four study cohorts, where the accession numbers are provided in parenthesis: Framingham Heart Study (phs000974.v1.p1), Old Order Amish (phs000956.v1.p1), Jackson Heart Study (phs000964.v1.p1), and Multi-Ethnic Study of Atherosclerosis (phs001416.v1.p1). The replication phase used the data from the following ten study cohorts: Atherosclerosis Risk in Communities Study (phs001211), Cleveland Family Study (phs000954), Cardiovascular Health Study (phs001368), Diabetes Heart Study (phs001412), Genetic Study of Atherosclerosis Risk (phs001218), Genetic Epidemiology Network of Arteriopathy (phs001345), Genetics of Lipid Lowering Drugs and Diet Network (phs001359), San Antonio Family Heart Study (phs001215), Genome-wide Association Study of Adiposity in Samoans (phs000972) and Women's Health Initiative (phs001237). The sample sizes, ethnicity and phenotype summary statistics of these cohorts are given in Supplementary Table 3.1b.

REFERENCES

1. Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. Nucleic acids research. 2003;31(13):3812-4.

2. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nature methods. 2010;7(4):248.

3. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome research. 2005;15(8):1034-50.

4. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. Genome research. 2010;20(1):110-21.

5. Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS computational biology. 2010;6(12):e1001025.

6. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489(7414):57.

7. Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015;518(7539):317.

8. Forrest AR, Kawaji H, Rehli M, Baillie JK, De Hoon MJ, Haberle V, et al. A promoter-level mammalian expression atlas. Nature. 2014;507(7493):462.

9. Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, et al. An atlas of active enhancers across human cell types and tissues. Nature. 2014;507(7493):455.

10. Kellis M, Wold B, Snyder MP, Bernstein BE, Kundaje A, Marinov GK, et al. Defining functional DNA elements in the human genome. Proceedings of the National Academy of Sciences. 2014;111(17):6131-8.

11. Kircher M, Witten DM, Jain P, O'roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. Nature genetics. 2014;46(3):310.

12. Quang D, Chen Y, Xie X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. Bioinformatics. 2014;31(5):761-3.

13. Ritchie GR, Dunham I, Zeggini E, Flicek P. Functional annotation of noncoding sequence variants. Nature methods. 2014;11(3):294.

14. Shihab HA, Rogers MF, Gough J, Mort M, Cooper DN, Day IN, et al. An integrative approach to predicting the functional effects of non-coding and coding sequence variation. Bioinformatics. 2015;31(10):1536-43.

15. Rogers MF, Shihab HA, Mort M, Cooper DN, Gaunt TR, Campbell C. FATHMM-XF: accurate prediction of pathogenic point mutations via extended features. Bioinformatics. 2017.

16. Ionita-Laza I, McCallum K, Xu B, Buxbaum JD. A spectral approach integrating functional genomic annotations for coding and noncoding variants. Nature genetics. 2016;48(2):214.

17. Lu Q, Hu Y, Sun J, Cheng Y, Cheung K-H, Zhao H. A statistical framework to predict functional non-coding regions in the human genome through integrated analysis of annotation data. Scientific reports. 2015;5:10576.

18. Bodea CA, Mitchell AA, Bloemendal A, Day-Williams AG, Runz H, Sunyaev SR. PINES: phenotype-informed tissue weighting improves prediction of pathogenic noncoding variants. Genome Biology. 2018;19(1):173.

19. Backenroth D, He Z, Kiryluk K, Boeva V, Pethukova L, Khurana E, et al. FUN-LDA: A latent Dirichlet allocation model for predicting tissue-specific functional effects of noncoding variation: methods and applications. The American Journal of Human Genetics. 2018;102(5):920-42.

20. Gulko B, Hubisz MJ, Gronau I, Siepel A. A method for calculating probabilities of fitness consequences for point mutations across the human genome. Nature genetics. 2015;47(3):276.

21. Huang Y-F, Gulko B, Siepel A. Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. Nature genetics. 2017;49(4):618.

22. Tang H, Thomas PD. Tools for predicting the functional impact of nonsynonymous genetic variation. Genetics. 2016;203(2):635-47.

23. Lee PH, Lee C, Li X, Wee B, Dwivedi T, Daly M. Principles and methods of in-silico prioritization of non-coding regulatory variants. Human genetics. 2018;137(1):15-30.

24. Liu X, Wu C, Li C, Boerwinkle E. dbNSFP v3. 0: A one-stop database of functional predictions and annotations for human nonsynonymous and splice-site SNVs. Human mutation. 2016;37(3):235-41.

25. Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic acids research. 2011;39(17):e118-e.

26. Ernst J, Kellis M. Discovery and characterization of chromatin states for systematic annotation of the human genome. Nature biotechnology. 2010;28(8):817.

27. McVicker G, Gordon D, Davis C, Green P. Widespread genomic signatures of natural selection in hominid evolution. PLoS genetics. 2009;5(5):e1000471.

28. Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic acids research. 2013;42(D1):D980-D5.

29. Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, et al. ClinVar: public archive of interpretations of clinically relevant variants. Nucleic acids research. 2015;44(D1):D862-D8.

30. Findlay GM, Daza RM, Martin B, Zhang MD, Leith AP, Gasperini M, et al. Accurate classification of BRCA1 variants with saturation genome editing. Nature. 2018;562(7726):217.

31. Tewhey R, Kotliar D, Park DS, Liu B, Winnicki S, Reilly SK, et al. Direct identification of hundreds of expression-modulating variants using a multiplexed reporter assay. Cell. 2016;165(6):1519-29.

32. Kheradpour P, Ernst J, Melnikov A, Rogov P, Wang L, Zhang X, et al. Systematic dissection of regulatory motifs in 2,000 predicted human enhancers using a massively parallel reporter assay. Genome research. 2013:gr. 144899.112.

33. Rands CM, Meader S, Ponting CP, Lunter G. 8.2% of the human genome is constrained: variation in rates of turnover across functional element classes in the human lineage. PLoS genetics. 2014;10(7):e1004525.

34. Graur D. An upper limit on the functional fraction of the human genome. Genome biology and evolution. 2017;9(7):1880-5.

35. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, et al. GENCODE: the reference human genome annotation for The ENCODE Project. Genome research. 2012;22(9):1760-74.

36. Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. Nature. 2012;491(7422):56-65.

37. He Z, Liu L, Wang K, Ionita-Laza I. A semi-supervised approach for predicting cell-type specific functional consequences of non-coding variation using MPRAs. Nature communications. 2018;9(1):5199.

38. Degner JF, Pai AA, Pique-Regi R, Veyrieras J-B, Gaffney DJ, Pickrell JK, et al. DNase I sensitivity QTLs are a major determinant of human expression variation. Nature. 2012;482(7385):390.

39. Li MJ, Pan Z, Liu Z, Wu J, Wang P, Zhu Y, et al. Predicting regulatory variants with composite statistic. Bioinformatics. 2016;32(18):2729-36.

40. Surakka I, Horikoshi M, Mägi R, Sarin A-P, Mahajan A, Lagou V, et al. The impact of low-frequency and rare variants on lipid levels. Nature genetics. 2015;47(6):589-97.

41. Pruim RJ, Welch RP, Sanna S, Teslovich TM, Chines PS, Gliedt TP, et al. LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics. 2010;26(18):2336-7.

42. Sun R, Xu M, Li X, Gaynor S, Zhou H, Li Z, et al. Integration of multiomic annotation data to prioritize and characterize inflammation and immune-related risk variants in squamous cell lung cancer. Genetic Epidemiology. 2020:1-16.

43. Li X, Li Z, Zhou H, Gaynor SM, Liu Y, Chen H, et al. Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale. Nature genetics. 2020;52(9):969-83.

44. Lawley D, Maxwell A. Factor analysis as a statistical method. Journal of the Royal Statistical Society Series D (The Statistician). 1962;12(3):209-29.

45. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. Journal of the royal statistical society Series B (methodological). 1977:1-38.

46. Li X, Yung G, Zhou H, Sun R, Li Z, Liu Y, et al. A Multi-dimensional Integrative Scoring Framework for Predicting Functional Regions in the Human Genome. bioRxiv. 2020:2021.01. 06.425527.

47. Bansal V, Libiger O, Torkamani A, Schork NJ. Statistical analysis strategies for association studies involving rare variants. Nature Reviews Genetics. 2010;11(11):773.

48. Kiezun A, Garimella K, Do R, Stitziel NO, Neale BM, McLaren PJ, et al. Exome sequencing and the genetic basis of complex traits. Nature genetics. 2012;44(6):623.

49. Lee S, Abecasis GR, Boehnke M, Lin X. Rare-variant association analysis: study designs and statistical tests. The American Journal of Human Genetics. 2014;95(1):5-23.

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

51. Li B, Leal SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. The American Journal of Human Genetics. 2008;83(3):311-21.

52. Madsen BE, Browning SR. A groupwise association test for rare mutations using a weighted sum statistic. PLoS genetics. 2009;5(2):e1000384.

53. Morris AP, Zeggini E. An evaluation of statistical approaches to rare variant analysis in genetic association studies. Genetic epidemiology. 2010;34(2):188-93.

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

55. Liu Y, Chen S, Li Z, Morrison AC, Boerwinkle E, Lin X. ACAT: A Fast and Powerful p Value Combination Method for Rare-Variant Analysis in Sequencing Studies. The American Journal of Human Genetics. 2019;104(3):410-21.

56. Lee S, Wu MC, Lin X. Optimal tests for rare variant effects in sequencing association studies. Biostatistics. 2012;13(4):762-75.

57. Sun J, Zheng Y, Hsu L. A unified mixed-effects model for rare-variant association in sequencing studies. Genetic epidemiology. 2013;37(4):334-44.

58. Pan W, Kim J, Zhang Y, Shen X, Wei P. A powerful and adaptive association test for rare variants. Genetics. 2014:genetics. 114.165035.

59. Kichaev G, Yang W-Y, Lindstrom S, Hormozdiari F, Eskin E, Price AL, et al. Integrating functional data to prioritize causal variants in statistical fine-mapping studies. PLoS genetics. 2014;10(10):e1004722.

60. Kichaev G, Roytman M, Johnson R, Eskin E, Lindström S, Kraft P, et al. Improved methods for multi-trait fine mapping of pleiotropic risk loci. Bioinformatics. 2017;33(2):248-55.

61. Finucane HK, Bulik-Sullivan B, Gusev A, Trynka G, Reshef Y, Loh P-R, et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. Nature genetics. 2015;47(11):1228.

62. Hu Y, Lu Q, Powles R, Yao X, Yang C, Fang F, et al. Leveraging functional annotations in genetic risk prediction for human complex diseases. PLoS computational biology. 2017;13(6):e1005589.

63. Morrison AC, Huang Z, Yu B, Metcalf G, Liu X, Ballantyne C, et al. Practical approaches for whole-genome sequence analysis of heart-and blood-related traits. The American Journal of Human Genetics. 2017;100(2):205-15.

64. Schaid DJ, Chen W, Larson NB. From genome-wide associations to candidate causal variants by statistical fine-mapping. Nature Reviews Genetics. 2018;19(8):491-504.

65. Claussnitzer M, Cho JH, Collins R, Cox NJ, Dermitzakis ET, Hurles ME, et al. A brief history of human disease genetics. Nature. 2020;577(7789):179-89.

66. Frankish A, Diekhans M, Ferreira A-M, Johnson R, Jungreis I, Loveland J, et al. GENCODE reference annotation for the human and mouse genomes. Nucleic acids research. 2019;47(D1):D766-D73.

67. Zuk O, Schaffner SF, Samocha K, Do R, Hechter E, Kathiresan S, et al. Searching for missing heritability: designing rare variant association studies. Proceedings of the National Academy of Sciences. 2014;111(4):E455-E64.

68. Hao X, Zeng P, Zhang S, Zhou X. Identifying and exploiting trait-relevant tissues with multiple functional annotations in genome-wide association studies. PLoS genetics. 2018;14(1):e1007186.

69. He Z, Xu B, Lee S, Ionita-Laza I. Unified Sequence-Based Association Tests Allowing for Multiple Functional Annotations and Meta-analysis of Noncoding Variation in Metabochip Data. The American Journal of Human Genetics. 2017;101(3):340-52.

70. Ma Y, Wei P. FunSPU: a versatile and adaptive multiple functional annotation-based association test of whole-genome sequencing data. PLoS genetics. 2019;15(4):e1008081.

71. Breslow NE, Clayton DG. Approximate inference in generalized linear mixed models. Journal of the American statistical Association. 1993;88(421):9-25.

72. Chen H, Wang C, Conomos MP, Stilp AM, Li Z, Sofer T, et al. Control for population structure and relatedness for binary traits in genetic association studies via logistic mixed models. The American Journal of Human Genetics. 2016;98(4):653-66.

73. Chen H, Huffman JE, Brody JA, Wang C, Lee S, Li Z, et al. Efficient variant set mixed model association tests for continuous and binary traits in large-scale whole-genome sequencing studies. The American Journal of Human Genetics. 2019;104(2):260-74.

74. Gogarten SM, Sofer T, Chen H, Yu C, Brody JA, Thornton TA, et al. Genetic association testing using the GENESIS R/Bioconductor package. Bioinformatics. 2019;35(24):5346-8.

75. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic acids research. 2018;47(D1):D886-D94.

76. Liu Y, Xie J. Cauchy combination test: a powerful test with analytic p-value calculation under arbitrary dependency structures. Journal of the American Statistical Association. 2020;115(529):393-402.

77. Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D. Calibrating a coalescent simulation of human genome sequence variation. Genome research. 2005;15(11):1576-83.

78. Natarajan P, Peloso GM, Zekavat SM, Montasser M, Ganna A, Chaffin M, et al. Deep-coverage whole genome sequences and blood lipids among 16,324 individuals. Nature communications. 2018;9(1):3391.

79. Taliun D, Harris DN, Kessler MD, Carlson J, Szpiech ZA, Torres R, et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. BioRxiv. 2019:563866.

80. Fishilevich S, Nudel R, Rappaport N, Hadar R, Plaschkes I, Iny Stein T, et al. GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. Database. 2017;2017.

81. Dong C, Wei P, Jian X, Gibbs R, Boerwinkle E, Wang K, et al. Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. Human molecular genetics. 2015;24(8):2125-37.

82. Sabatti C, Service SK, Hartikainen A-L, Pouta A, Ripatti S, Brodsky J, et al. Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. Nature genetics. 2009;41(1):35.

83. Kathiresan S, Melander O, Guiducci C, Surti A, Burtt NP, Rieder MJ, et al. Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. Nature genetics. 2008;40(2):189.

84. Huang C-C, Fornage M, Lloyd-Jones DM, Wei GS, Boerwinkle E, Liu K. Longitudinal association of PCSK9 sequence variations with low-density lipoprotein cholesterol levels: the Coronary Artery Risk Development in Young Adults Study. Circulation: Cardiovascular Genetics. 2009;2(4):354-61.

85. Lange LA, Hu Y, Zhang H, Xue C, Schmidt EM, Tang Z-Z, et al. Whole-exome sequencing identifies rare and low-frequency coding variants associated with LDL cholesterol. The American Journal of Human Genetics. 2014;94(2):233-45.

86. Bomba L, Walter K, Soranzo N. The impact of rare and low-frequency genetic variants in common disease. Genome biology. 2017;18(1):77.

87. Ference BA, Majeed F, Penumetcha R, Flack JM, Brook RD. Effect of naturally random allocation to lower low-density lipoprotein cholesterol on the risk of coronary heart disease mediated by polymorphisms in NPC1L1, HMGCR, or both: a 2× 2 factorial Mendelian randomization study. Journal of the American College of Cardiology. 2015;65(15):1552-61.

88. Teslovich TM, Musunuru K, Smith AV, Edmondson AC, Stylianou IM, Koseki M, et al. Biological, clinical and population relevance of 95 loci for blood lipids. Nature. 2010;466(7307):707.

89. Kathiresan S, Willer CJ, Peloso GM, Demissie S, Musunuru K, Schadt EE, et al. Common variants at 30 loci contribute to polygenic dyslipidemia. Nature genetics. 2009;41(1):56.

90. Kamatani Y, Matsuda K, Okada Y, Kubo M, Hosono N, Daigo Y, et al. Genome-wide association study of hematological and biochemical traits in a Japanese population. Nature genetics. 2010;42(3):210.

91. Nagy R, Boutin TS, Marten J, Huffman JE, Kerr SM, Campbell A, et al. Exploration of haplotype research consortium imputation for genome-wide association studies in 20,032 Generation Scotland participants. Genome medicine. 2017;9(1):23.

92. Aulchenko YS, Ripatti S, Lindqvist I, Boomsma D, Heid IM, Pramstaller PP, et al. Loci influencing lipid levels and coronary heart disease risk in 16 European population cohorts. Nature genetics. 2009;41(1):47.

93. Deelen J, Beekman M, Uh HW, Helmer Q, Kuningas M, Christiansen L, et al. Genome-wide association study identifies a single major locus contributing to survival into old age; the APOE locus revisited. Aging cell. 2011;10(4):686-98.

94. Klarin D, Damrauer SM, Cho K, Sun YV, Teslovich TM, Honerlaw J, et al. Genetics of blood lipids among~ 300,000 multi-ethnic participants of the Million Veteran Program. Nature genetics. 2018;50(11):1514-23.

95. Hoffmann TJ, Theusch E, Haldar T, Ranatunga DK, Jorgenson E, Medina MW, et al. A large electronic-health-record-based genome-wide study of serum lipids. Nature genetics. 2018;50(3):401-13.

96. Willer CJ, Schmidt EM, Sengupta S, Peloso GM, Gustafsson S, Kanoni S, et al. Discovery and refinement of loci associated with lipid levels. Nature genetics. 2013;45(11):1274.

97. Cohen JC, Pertsemlidis A, Fahmi S, Esmail S, Vega GL, Grundy SM, et al. Multiple rare variants in NPC1L1 associated with reduced sterol absorption and plasma low-density lipoprotein levels. Proceedings of the National Academy of Sciences. 2006;103(6):1810-5.

98. Myocardial Infarction Genetics Consortium Investigators. Inactivating mutations in NPC1L1 and protection from coronary heart disease. New England Journal of Medicine. 2014;371(22):2072-82.

99. Cooper GM, Goode DL, Ng SB, Sidow A, Bamshad MJ, Shendure J, et al. Single-nucleotide evolutionary constraint scores highlight disease-causing mutations. Nature methods. 2010;7(4):250.

100. Cooper GM, Shendure J. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. Nature Reviews Genetics. 2011;12(9):628.

101. Van Hout CV, Tachmazidou I, Backman JD, Hoffman JD, Liu D, Pandey AK, et al. Exome sequencing and characterization of 49,960 individuals in the UK Biobank. Nature. 2020:1-8.

102. TG, HDL Working Group of the Exome Sequencing Project NH, Lung,, Institute B. Loss-of-function mutations in APOC3, triglycerides, and coronary disease. New England Journal of Medicine. 2014;371(1):22-31.

103. ENCODE Project Consortium. A user's guide to the encyclopedia of DNA elements (ENCODE). PLoS biology. 2011;9(4):e1001046.

104. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, et al. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic acids research. 2018;46(D1):D1062-D7.

105. Davis HR, Veltri EP. Zetia: inhibition of Niemann-Pick C1 Like 1 (NPC1L1) to reduce intestinal cholesterol absorption and treat hyperlipidemia. Journal of atherosclerosis and thrombosis. 2007;14(3):99-108.

106. Klos K, Shimmin L, Ballantyne C, Boerwinkle E, Clark A, Coresh J, et al. APOE/C1/C4/C2 hepatic control region polymorphism influences plasma apoE and LDL cholesterol levels. Human molecular genetics. 2008;17(13):2039-46.

107. Lu Q, Powles RL, Wang Q, He BJ, Zhao H. Integrative tissue-specific functional annotations in the human genome provide novel insights on many complex traits and improve signal prioritization in genome wide association studies. PLoS genetics. 2016;12(4):e1005947.

108. Park J-H, Wacholder S, Gail MH, Peters U, Jacobs KB, Chanock SJ, et al. Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. Nature genetics. 2010;42(7):570.

109. Derkach A, Zhang H, Chatterjee N. Power Analysis for Genetic Association Test (PAGEANT) provides insights to challenges for rare variant association studies. Bioinformatics. 2017;34(9):1506-13.

110. Li Z, Li X, Liu Y, Shen J, Chen H, Zhou H, et al. Dynamic Scan Procedure for Detecting Rare-Variant Association Regions in Whole-Genome Sequencing Studies. The American Journal of Human Genetics. 2019;104(5):802-14.

111. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. The American Journal of Human Genetics. 2011;88(1):76-82.

112. Conomos MP, Reiner AP, Weir BS, Thornton TA. Model-free estimation of recent genetic relatedness. The American Journal of Human Genetics. 2016;98(1):127-48.

113. Dey R, Schmidt EM, Abecasis GR, Lee S. A fast and accurate algorithm to test for binary phenotypes and its application to PheWAS. The American Journal of Human Genetics. 2017;101(1):37-49.

114. Zhou W, Nielsen JB, Fritsche LG, Dey R, Gabrielsen ME, Wolford BN, et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. Nature genetics. 2018;50(9):1335.

115. Karimzadeh M, Ernst C, Kundaje A, Hoffman MM. Umap and Bismap: quantifying genome and methylome mappability. Nucleic acids research. 2018;46(20):e120-e.

116. Rogers MF, Shihab HA, Mort M, Cooper DN, Gaunt TR, Campbell C. FATHMM-XF: accurate prediction of pathogenic point mutations via extended features. Bioinformatics. 2018;34(3):511-3.

117. Liu Y, Xie J. Cauchy combination test: a powerful test with analytic p-value calculation under arbitrary dependency structures. Journal of the American Statistical Association. 2018(just-accepted):1-29.

118. Surakka I, Horikoshi M, Mägi R, Sarin A-P, Mahajan A, Lagou V, et al. The impact of low-frequency and rare variants on lipid levels. Nature genetics. 2015;47(6):589.

119. Regier AA, Farjoun Y, Larson DE, Krasheninina O, Kang HM, Howrigan DP, et al. Functional equivalence of genome sequencing analysis pipelines enables harmonized variant calling across human genetics projects. Nature communications. 2018;9(1):1-8.

120. Gazal S, Finucane HK, Furlotte NA, Loh P-R, Palamara PF, Liu X, et al. Linkage disequilibrium–dependent architecture of human complex traits shows action of negative selection. Nature genetics. 2017;49(10):1421.

121. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nature reviews genetics. 2008;9(5):356-69.

122. Evangelou E, Ioannidis JP. Meta-analysis methods for genome-wide association studies and beyond. Nature Reviews Genetics. 2013;14(6):379-89.

123. Lin D, Zeng D. Meta-analysis of genome-wide association studies: no efficiency gain in using individual participant data. Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society. 2010;34(1):60-6.

124. Liu DJ, Peloso GM, Zhan X, Holmen OL, Zawistowski M, Feng S, et al. Meta-analysis of gene-level tests for rare variant association. Nature genetics. 2014;46(2):200.

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

126. Hu Y-J, Berndt SI, Gustafsson S, Ganna A, Mägi R, Wheeler E, et al. Meta-analysis of gene-level associations for rare variants based on single-variant statistics. The American Journal of Human Genetics. 2013;93(2):236-48.

127. Yang J, Chen S, Abecasis G, IAMDGC. Improved score statistics for meta-analysis in single-variant and gene-level association studies. Genetic epidemiology. 2018;42(4):333-43.

128. Zhou W, Zhao Z, Nielsen JB, Fritsche LG, LeFaive J, Gagliano Taliun SA, et al. Scalable generalized linear mixed model for region-based association tests in large biobanks and cohorts. Nature Genetics. 2020;52(6):634-9.

129. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics. 2010;26(17):2190-1.

130. Quick C, Guan L, Li Z, Li X, Dey R, Liu Y, et al. A versatile toolkit for molecular QTL mapping and meta-analysis at scale. bioRxiv. 2020.

131. Kannel WB, Dawber TR, Kagan A, Revotskie N, Stokes J. Factors of risk in the development of coronary heart disease—six-year follow-up experience: the Framingham Study. Annals of internal medicine. 1961;55(1):33-50.

132. Kannel WB, Feinleib M, McNamara PM, Garrison RJ, Castelli WP. An investigation of coronary heart disease in families: the Framingham Offspring Study. American journal of epidemiology. 1979;110(3):281-90.

133. Splansky GL, Corey D, Yang Q, Atwood LD, Cupples LA, Benjamin EJ, et al. The third generation cohort of the National Heart, Lung, and Blood Institute's Framingham Heart Study: design, recruitment, and initial examination. American journal of epidemiology. 2007;165(11):1328-35.

134. Post W, Bielak LF, Ryan KA, Cheng Y-C, Shen H, Rumberger JA, et al. Determinants of coronary artery and aortic calcification in the Old Order Amish. Circulation. 2007;115(6):717.

135. Mitchell BD, McArdle PF, Shen H, Rampersaud E, Pollin TI, Bielak LF, et al. The genetic response to short-term interventions affecting cardiovascular function: rationale and design of the Heredity and Phenotype Intervention (HAPI) Heart Study. American heart journal. 2008;155(5):823-8.

136. Taylor Jr HA, Wilson JG, Jones DW, Sarpong DF, Srinivasan A, Garrison RJ, et al. Toward resolution of cardiovascular health disparities in African Americans: design and methods of the Jackson Heart Study. Ethn Dis. 2005;15(4 Suppl 6):S6-4.

137. Bild DE, Bluemke DA, Burke GL, Detrano R, Diez Roux AV, Folsom AR, et al. Multi-ethnic study of atherosclerosis: objectives and design. American journal of epidemiology. 2002;156(9):871-81.

138. The ARIC Investigators. The atherosclerosis risk in communities (aric) study: design and objectives. American journal of epidemiology. 1989;129(4):687-702.

139. Redline S, Schluchter MD, Larkin EK, Tishler PV. Predictors of longitudinal change in sleep-disordered breathing in a nonclinic population. Sleep. 2003;26(6):703-9.

140. Larkin EK, Patel SR, Goodloe RJ, Li Y, Zhu X, Gray-McGuire C, et al. A candidate gene study of obstructive sleep apnea in European Americans and African Americans. American journal of respiratory and critical care medicine. 2010;182(7):947-53.

141. Cushman M, Cornell ES, Howard PR, Bovill EG, Tracy RP. Laboratory methods and quality assurance in the Cardiovascular Health Study. Clinical chemistry. 1995;41(2):264-70.

142. Fried LP, Borhani NO, Enright P, Furberg CD, Gardin JM, Kronmal RA, et al. The cardiovascular health study: design and rationale. Annals of epidemiology. 1991;1(3):263-76.

143. Bowden DW, Cox AJ, Freedman BI, Hugenschimdt CE, Wagenknecht LE, Herrington D, et al. Review of the Diabetes Heart Study (DHS) family of studies: a comprehensively examined sample for genetic and epidemiological studies of type 2 diabetes and its complications. The review of diabetic studies: RDS. 2010;7(3):188.

144. Divers J, Palmer ND, Langefeld CD, Brown WM, Lu L, Hicks PJ, et al. Genome-wide association study of coronary artery calcified atherosclerotic plaque in African Americans with type 2 diabetes. BMC genetics. 2017;18(1):105.

145. Vaidya D, Yanek LR, Moy TF, Pearson TA, Becker LC, Becker DM. Incidence of coronary artery disease in siblings of patients with premature coronary artery disease: 10 years of follow-up. The American journal of cardiology. 2007;100(9):1410-5.

146. Faraday N, Becker DM, Yanek LR, Herrera-Galeano JE, Segal JB, Moy TF, et al. Relation between atherosclerosis risk factors and aspirin resistance in a primary prevention population. The American journal of cardiology. 2006;98(6):774-9.

147. FBPP Investigators. Multi-center genetic study of hypertension: the Family Blood Pressure Program (FBPP). Hypertension. 2002;39(1):3-9.

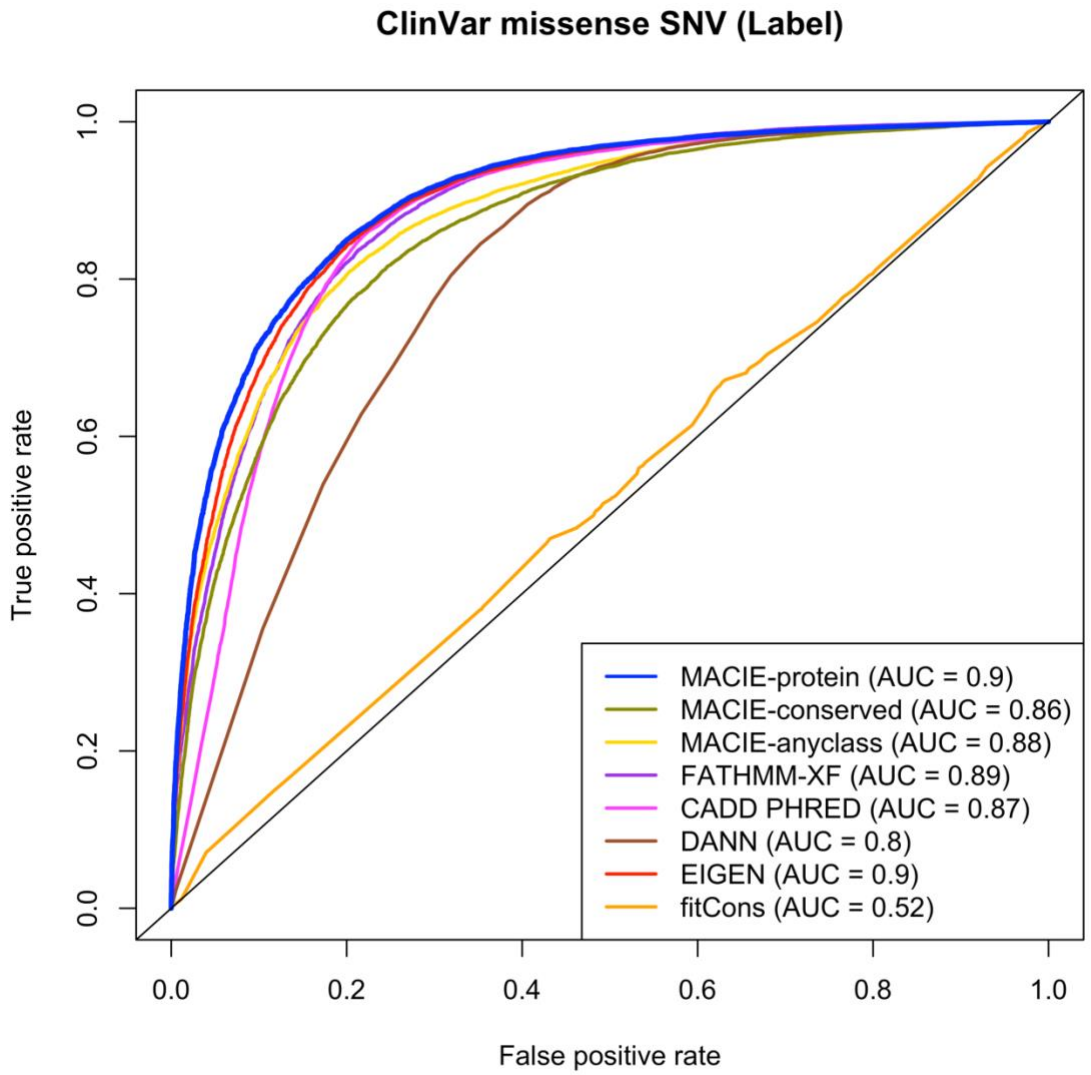
148. Daniels PR, Kardia SL, Hanis CL, Brown CA, Hutchinson R, Boerwinkle E, et al. Familial aggregation of hypertension treatment and control in the Genetic Epidemiology Network of Arteriopathy (GENOA) study. The American journal of medicine. 2004;116(10):676-81.

149. Minster RL, Hawley NL, Su C-T, Sun G, Kershaw EE, Cheng H, et al. A thrifty variant in CREBRF strongly influences body mass index in Samoans. Nature genetics. 2016;48(9):1049.

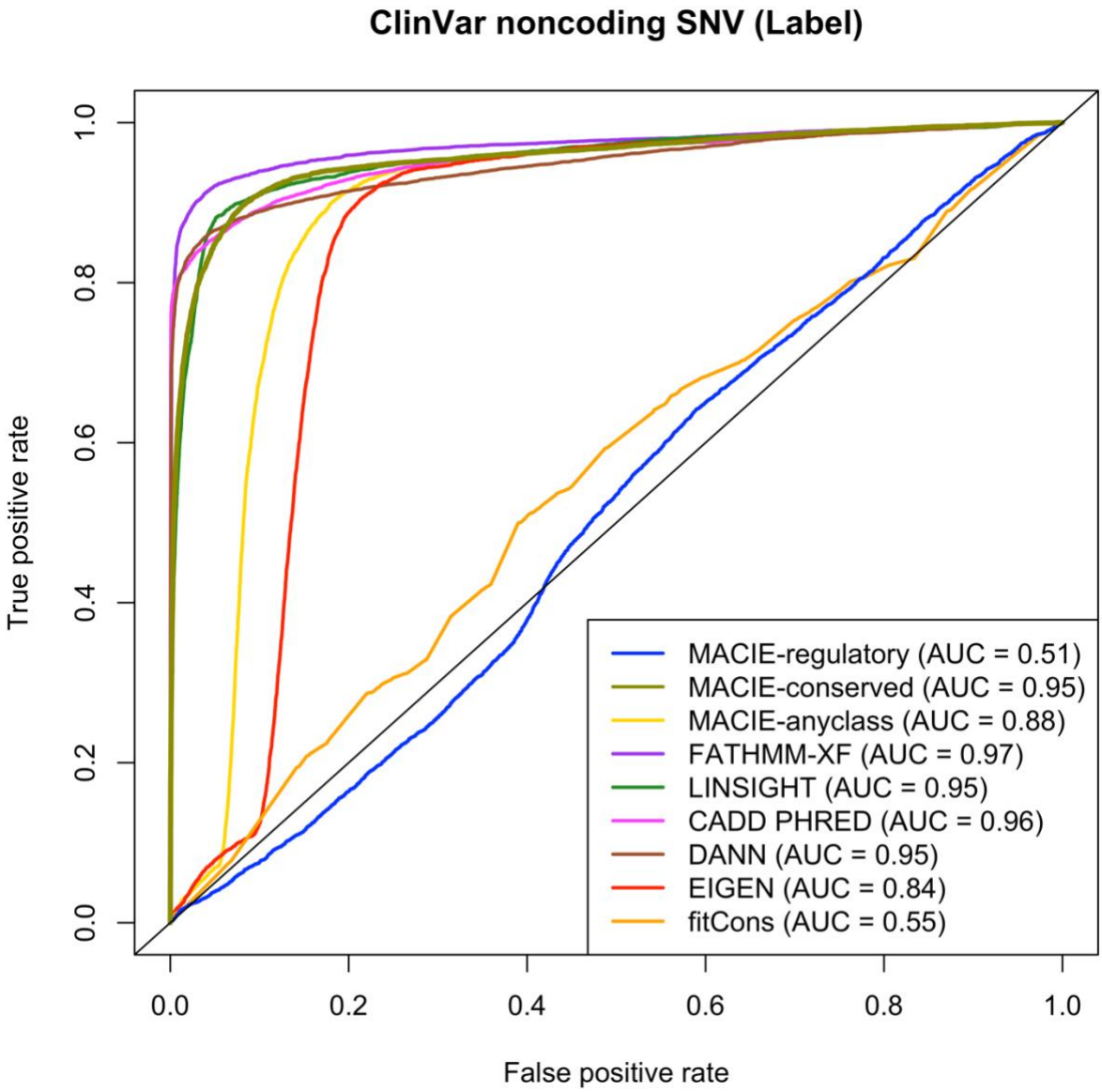
150. Anderson G, Cummings S, Freedman L, Furberg C, Henderson M, Johnson S, et al. Design of the Women's Health Initiative clinical trial and observational study. Controlled clinical trials. 1998;19(1):61-109.

APPENDIX A

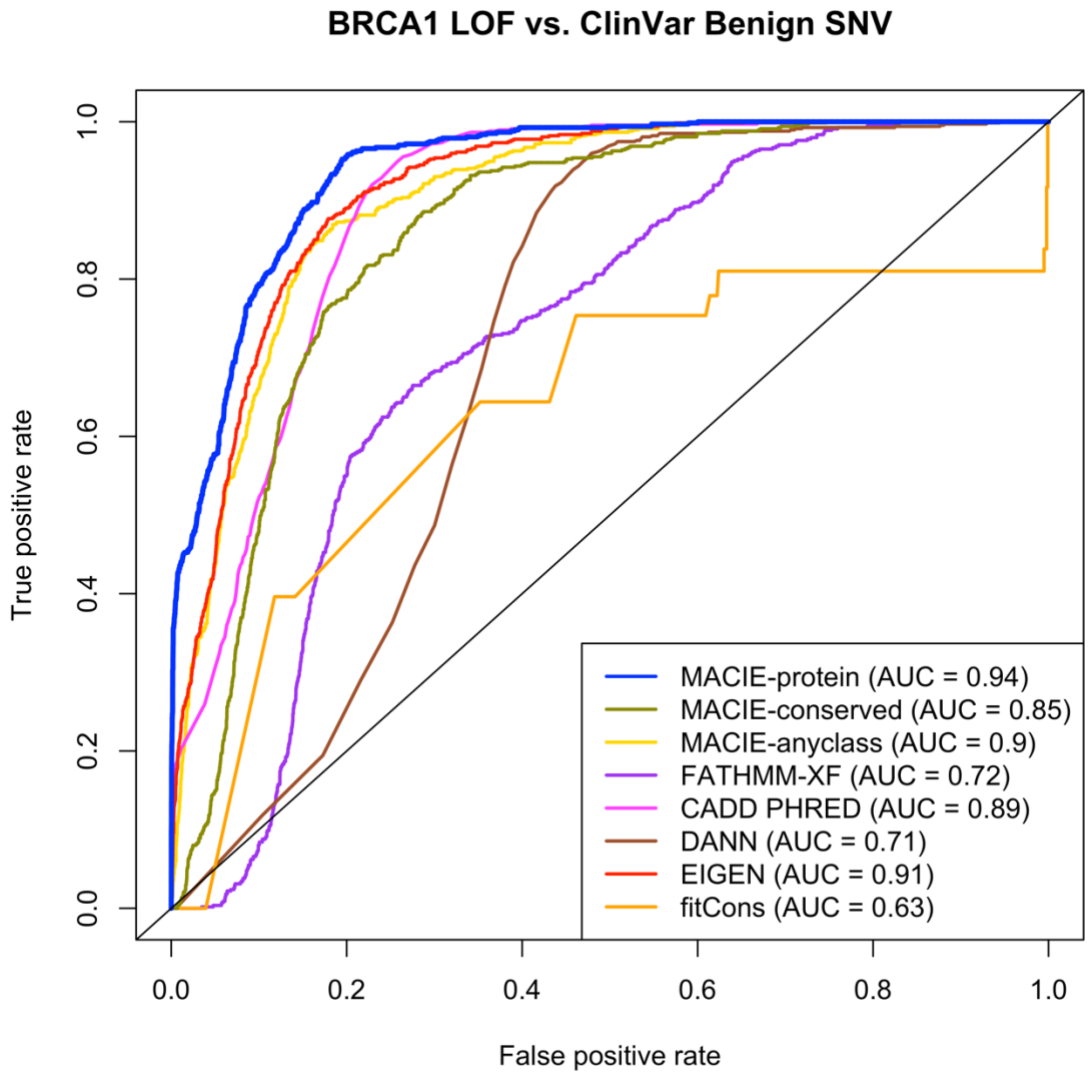
Supplementary 1.1. ROC curves comparing the performances of MACIE and other functional scores in discriminating between ClinVar pathogenic and benign missense variants.



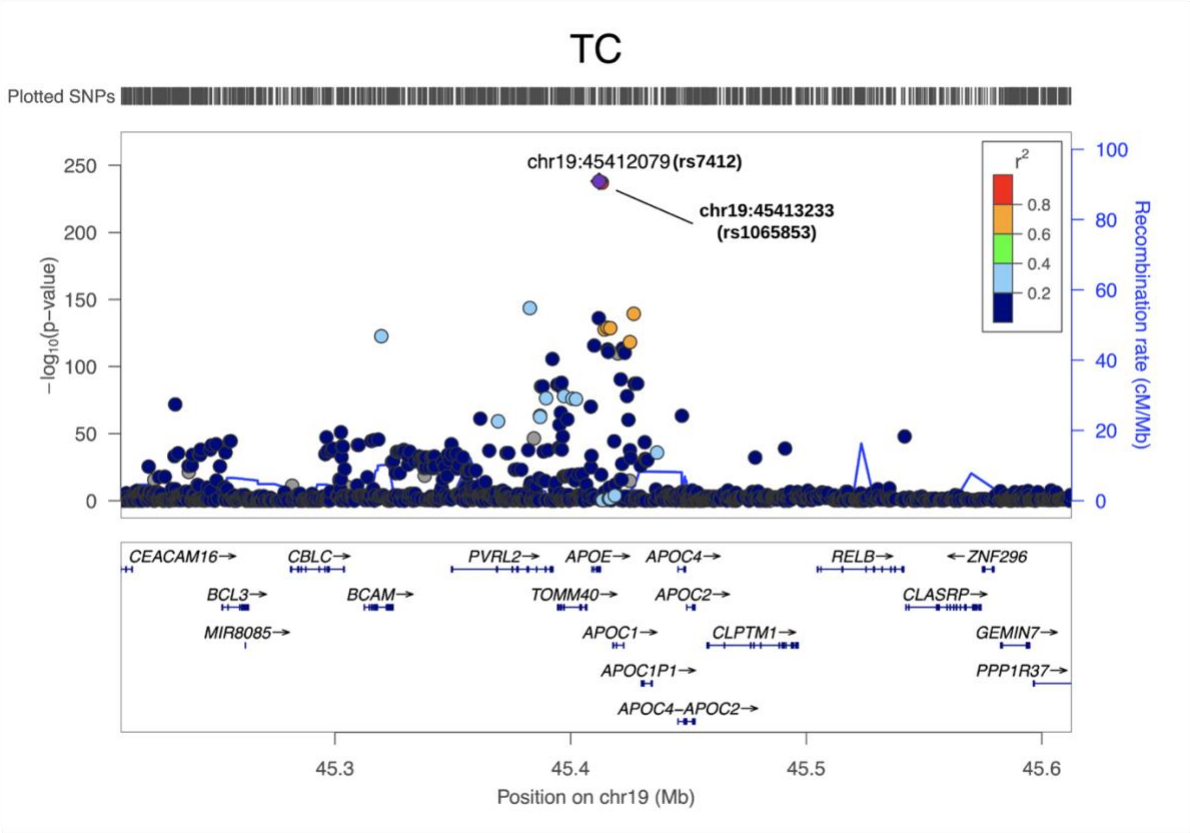
Supplementary Figure 1.2. ROC curves comparing the performances of MACIE and other functional scores in discriminating between ClinVar pathogenic and benign noncoding variants.



Supplementary Figure 1.3. ROC curves comparing the performances of MACIE and other functional scores in discriminating between loss-of-function (LOF) nonsynonymous coding variants within 13 exons that encode functionally critical domains of *BRCA1* (putative functional class) based on saturation genome editing (SGE) data and ClinVar benign nonsynonymous coding variants (putative non-functional class).

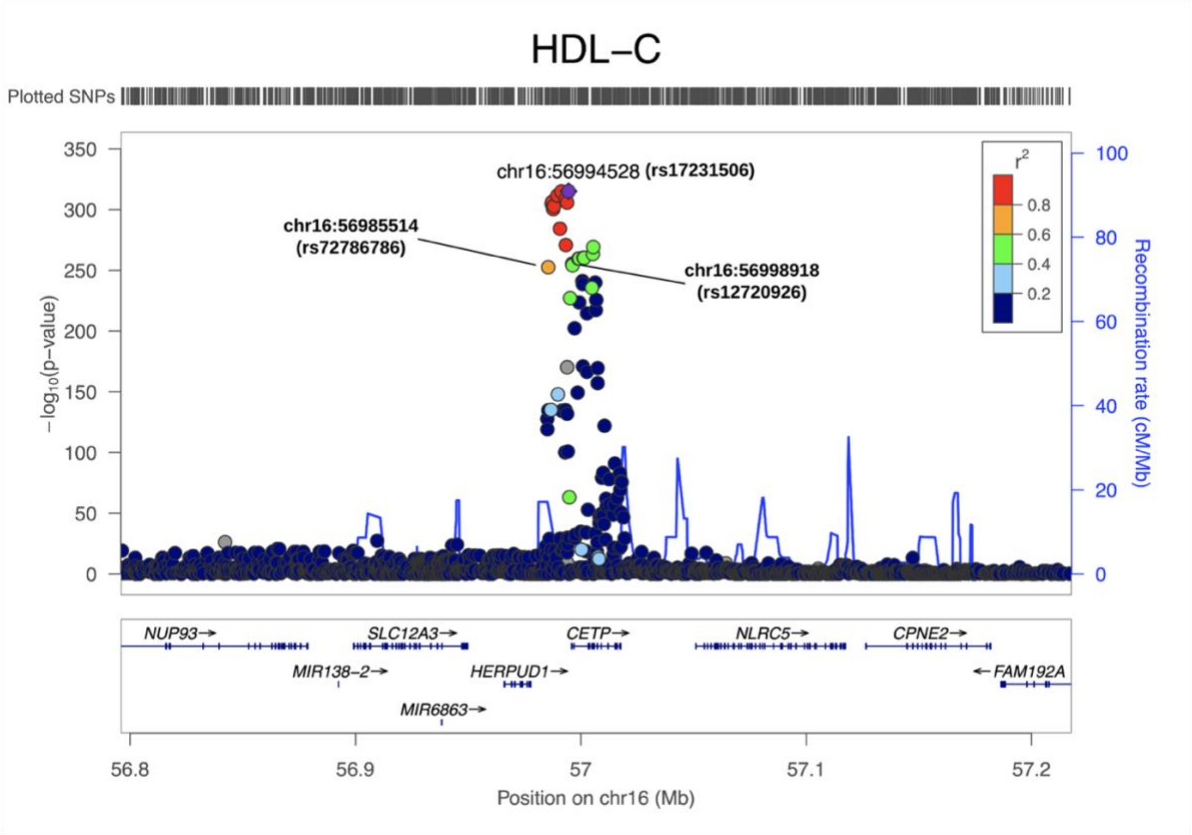


Supplementary 1.4. LocusZoom plot for GWAS associations of TC at the *APOE* locus. The lipids GWAS summary statistics were from the European Network for Genetic and Genomic Epidemiology (ENGAGE) Consortium (*n* = 62,166).



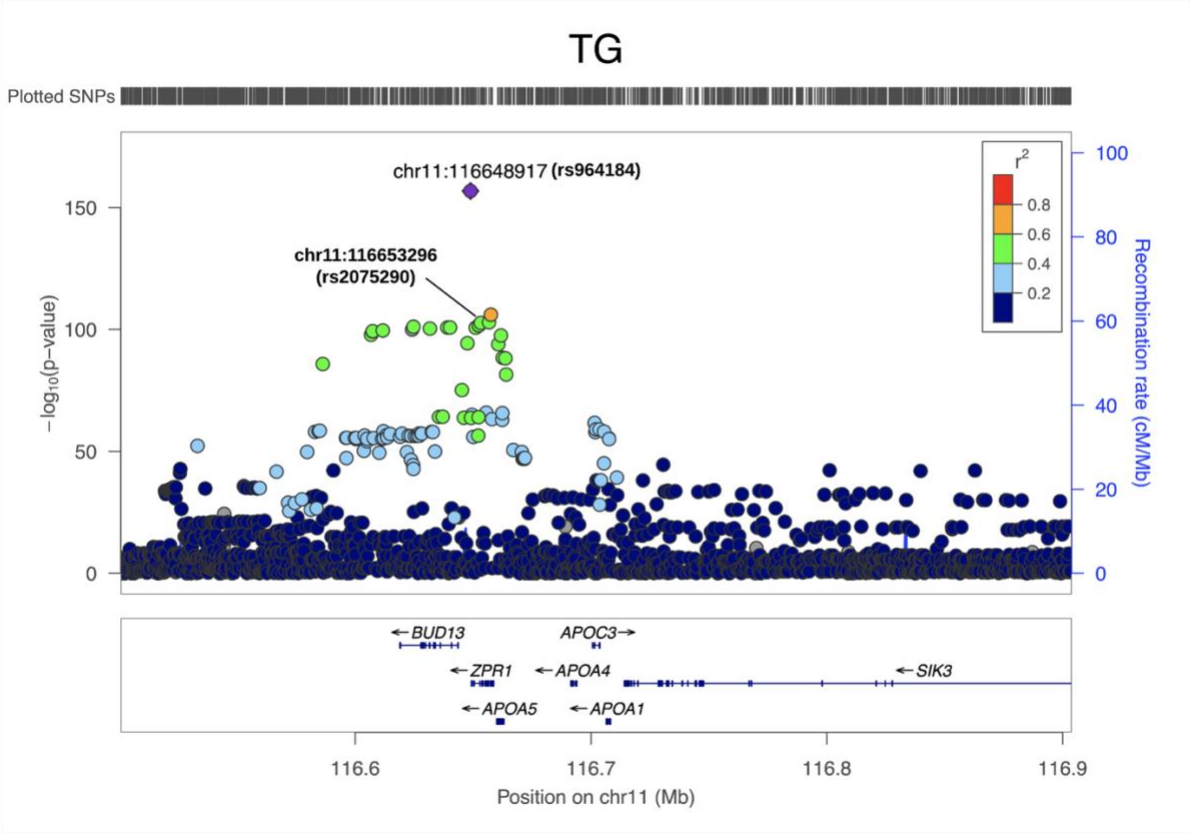
The MACIE-protein and MACIE-conserved scores for rs7412 are 0.96 and 0.97, respectively. The MACIE-conserved and MACIE-regulatory scores for rs1065853 are < 0.01 and > 0.99, respectively.

Supplementary Figure 1.5. LocusZoom plot for GWAS associations of HDL-C at the *CETP* locus. The lipids GWAS summary statistics were from the European Network for Genetic and Genomic Epidemiology (ENGAGE) Consortium (*n* = 60,812).



The MACIE-conserved and MACIE-regulatory scores for rs17231506 are both < 0.01. For both rs72786786 and rs12720926, the MACIE-conserved and MACIE-regulatory scores are < 0.01 and > 0.99, respectively.

Supplementary Figure 1.6. LocusZoom plot for GWAS associations of TG at the *APOC3* locus. The lipids GWAS summary statistics were from the European Network for Genetic and Genomic Epidemiology (ENGAGE) Consortium (*n* = 60,027).

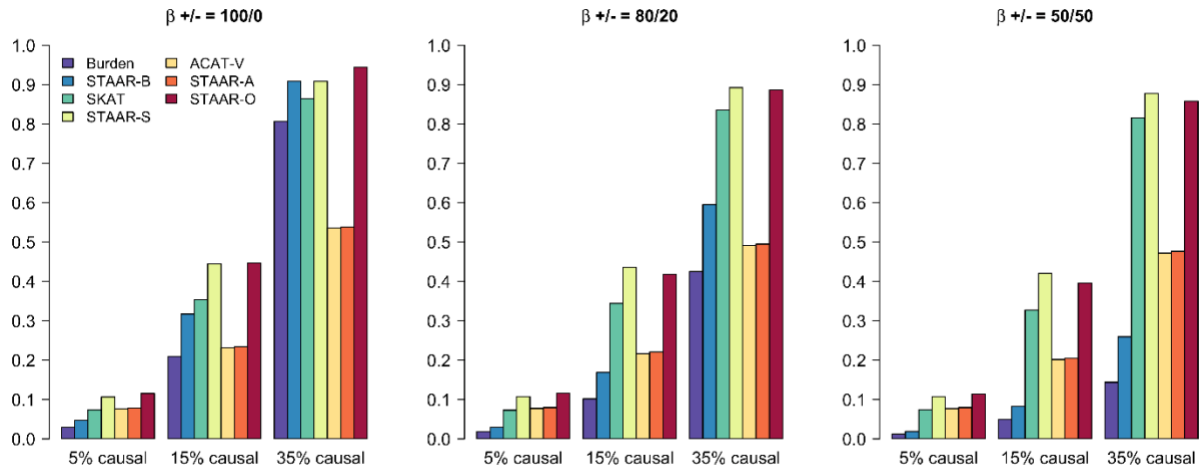


The MACIE-conserved and MACIE-regulatory scores for rs964184 are both < 0.01. The MACIE-conserved and MACIE-regulatory scores for rs2075290 are < 0.01 and 0.88, respectively.

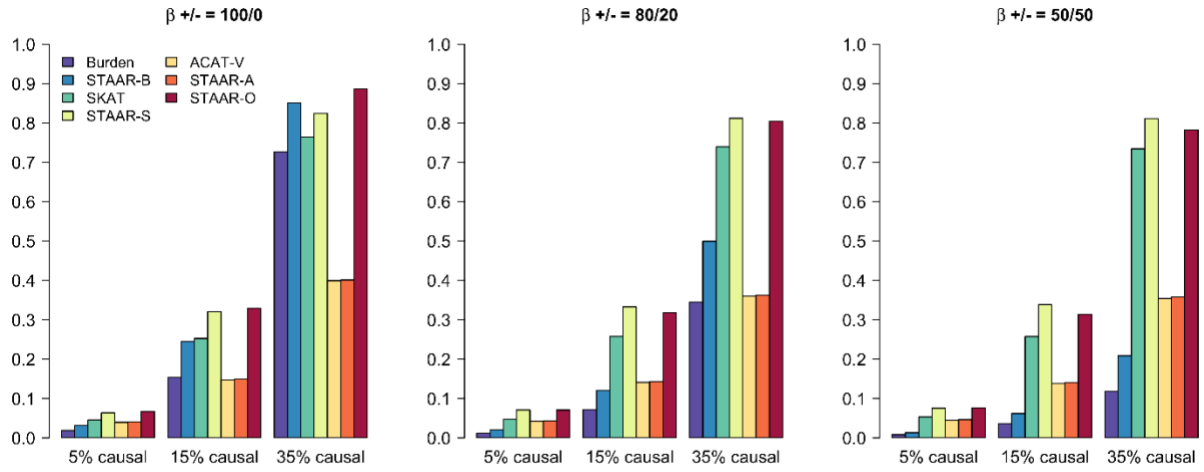
APPENDIX B

Supplementary Figure 2.1. Simulation-study power comparisons of burden test, SKAT, ACAT-V and STAAR methods for continuous and dichotomous traits.

Continuous Trait



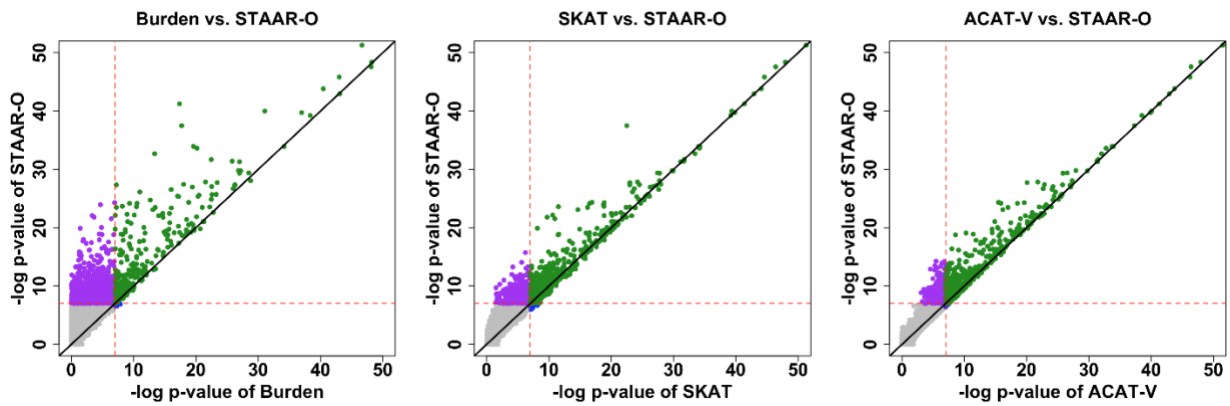
Dichotomous Trait



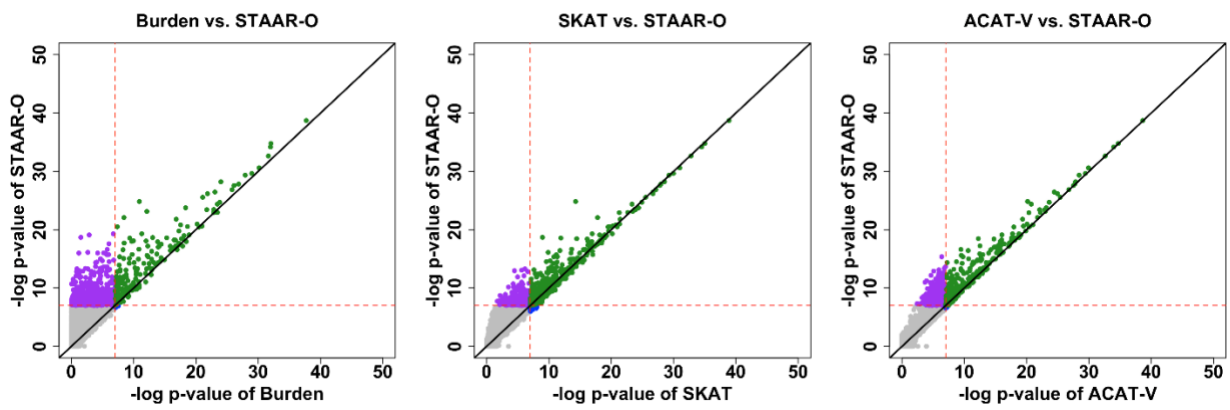
In each simulation replicate, a 5-kb region was randomly selected as the signal region. Within each signal region, variants were randomly generated to be causal based on the multivariate logistic model and on average there were 5%, 15% or 35% causal variants in the signal region. The effect sizes of causal variants were $\beta_j = c_0 |\log_{10} MAF_j|$, where c_0 was set to be 0.13 for continuous traits and 0.255 for dichotomous traits, which gives an odds ratio of 3 for a variant with a MAF of 5×10^{-5}. Power was estimated as the proportion of the p-values less than $\alpha = 10^{-7}$ based on 10^4 replicates. Total sample sizes considered were 10,000. For each setting, seven statistical tests were compared: burden test, STAAR-B, SKAT, STAAR-S, ACAT-V, STAAR-A, and STAAR-O (Methods).

Supplementary Figure 2.2. Scatterplot of *P* values comparing STAAR-O to conventional variant-set tests (Burden, SKAT, ACAT-V) for continuous and dichotomous traits when 5% of rare variants are causal variants.

Continuous Trait



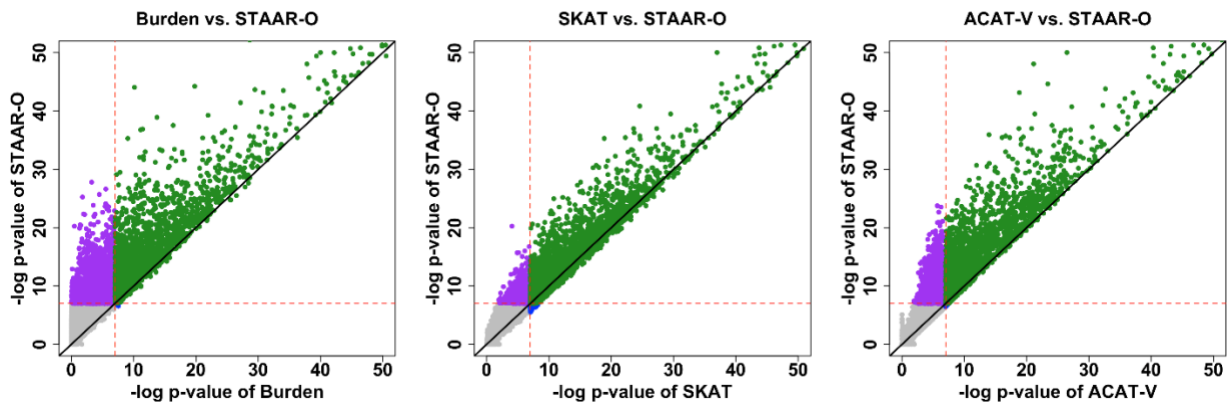
Dichotomous Trait



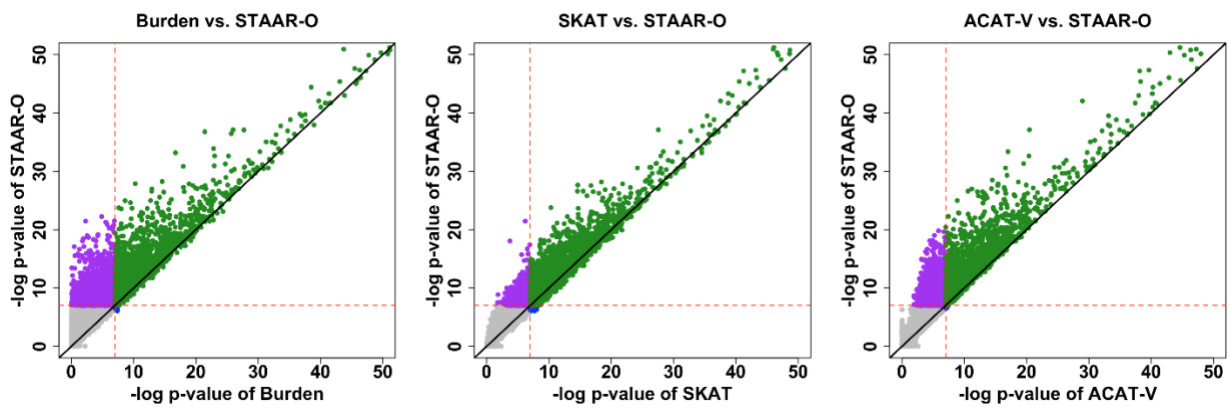
In each simulation replicate, a 5-kb region was randomly selected as the signal region. Within each signal region, variants were randomly generated to be causal based on the multivariate logistic model and on average there were 5% causal variants in the signal region. The effect sizes of causal variants were $\beta_j = c_0 |\log_{10} MAF_j|$. For continuous traits, $c_0 = 0.13$; for dichotomous traits, $c_0 = 0.255$. All causal variants had positive effect sizes. Power was estimated as the proportion of the *P* values less than $\alpha = 10^{-7}$ based on 10^4 replicates. Total sample sizes considered were 10,000.

Supplementary Figure 2.3. Scatterplot of *P* values comparing STAAR-O to conventional variant-set tests (Burden, SKAT, ACAT-V) for continuous and dichotomous traits when 15% of rare variants are causal variants.

Continuous Trait



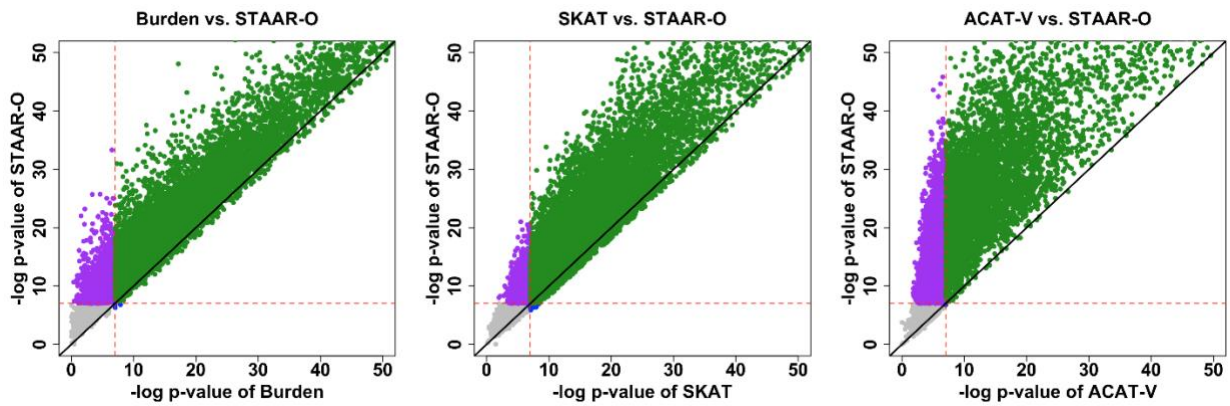
Dichotomous Trait



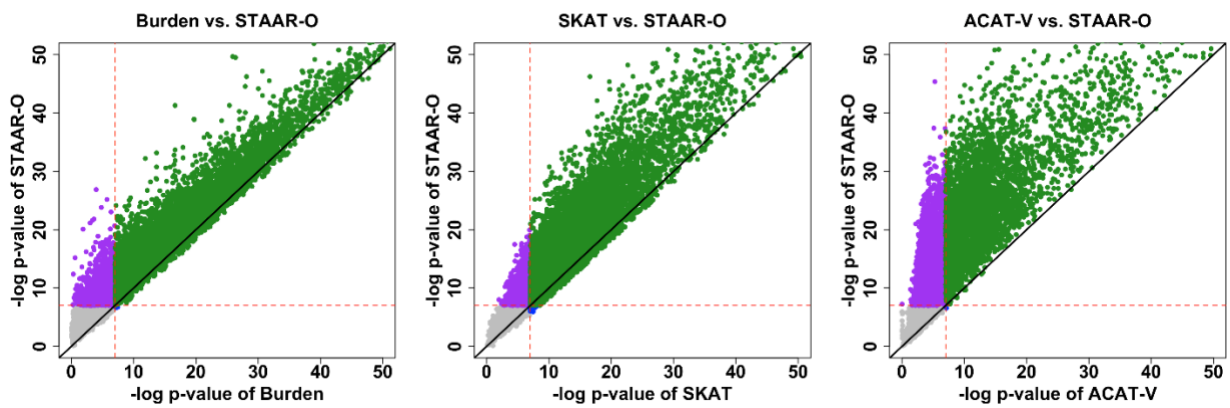
In each simulation replicate, a 5-kb region was randomly selected as the signal region. Within each signal region, variants were randomly generated to be causal based on the multivariate logistic model and on average there were 15% causal variants in the signal region. The effect sizes of causal variants were $\beta_j = c_0 |\log_{10} MAF_j|$. For continuous traits, $c_0 = 0.13$; for dichotomous traits, $c_0 = 0.255$. All causal variants had positive effect sizes. Power was estimated as the proportion of the *P* values less than $\alpha = 10^{-7}$ based on 10^4 replicates. Total sample sizes considered were 10,000.

Supplementary Figure 2.4. Scatterplot of *P* values comparing STAAR-O to conventional variant-set tests (Burden, SKAT, ACAT-V) for continuous and dichotomous traits when 35% of rare variants are causal variants.

Continuous Trait

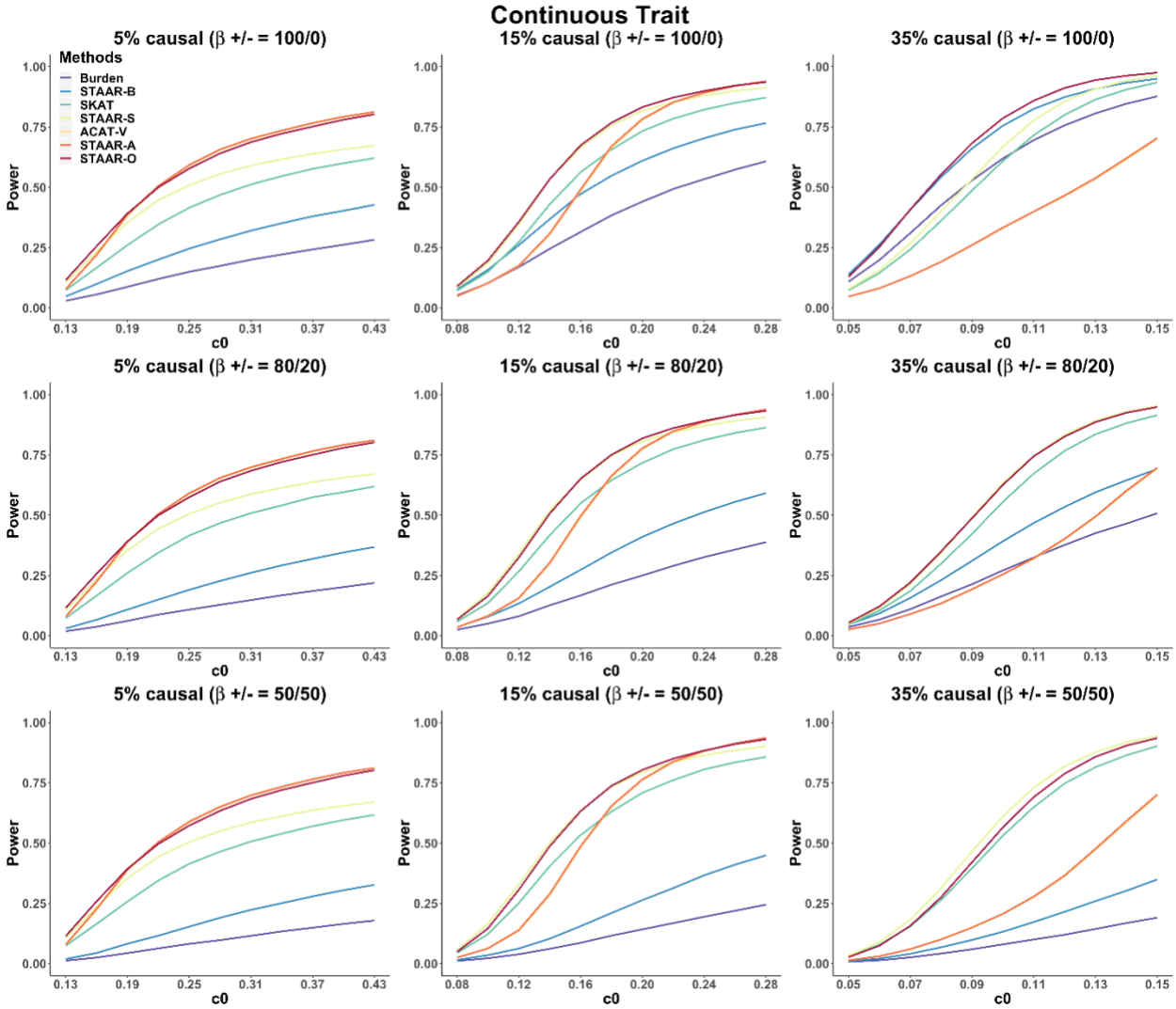


Dichotomous Trait



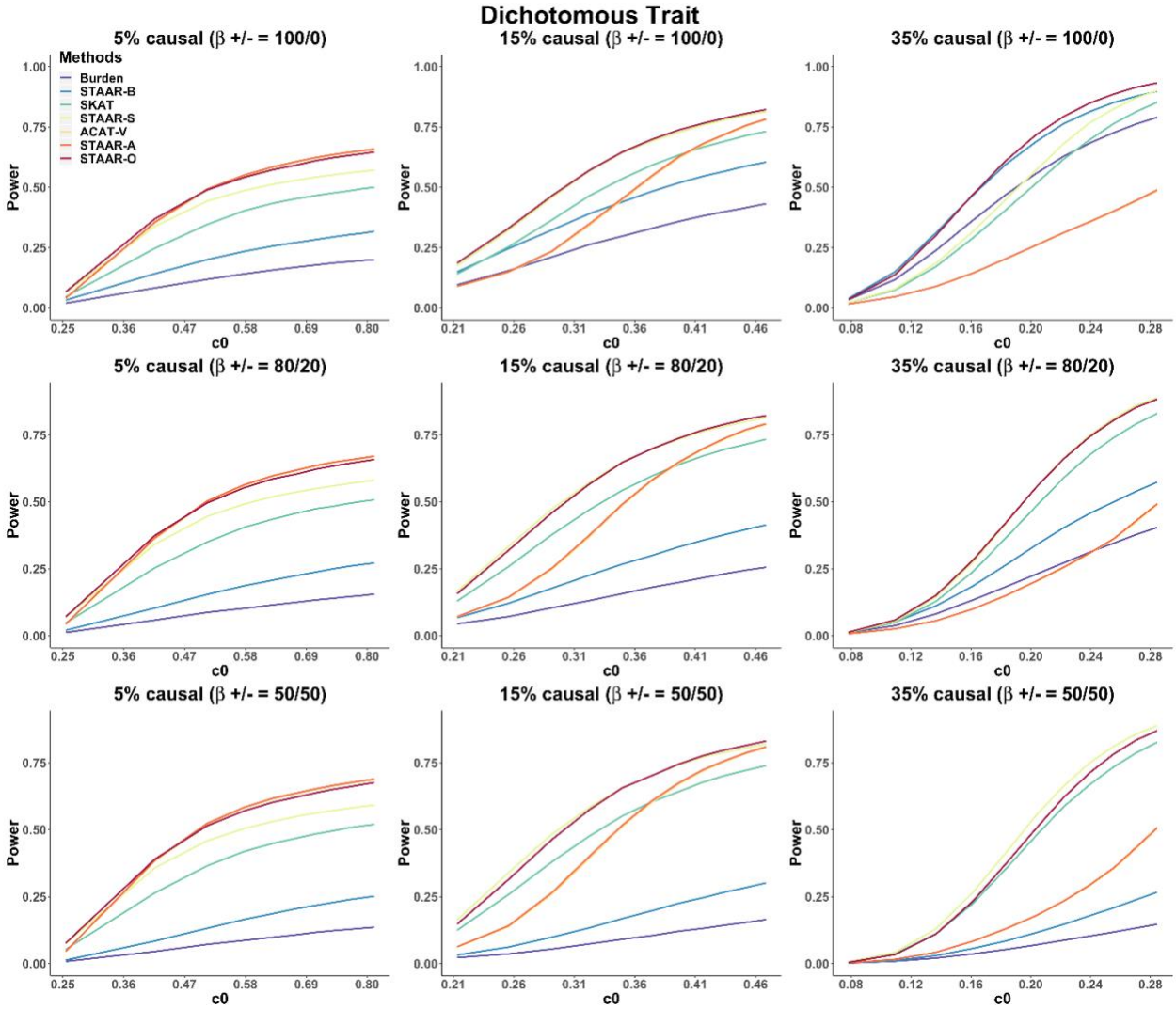
In each simulation replicate, a 5-kb region was randomly selected as the signal region. Within each signal region, variants were randomly generated to be causal based on the multivariate logistic model and on average there were 35% causal variants in the signal region. The effect sizes of causal variants were $\beta_j = c_0 |\log_{10} MAF_j|$. For continuous traits, $c_0 = 0.13$; for dichotomous traits, $c_0 = 0.255$. All causal variants had positive effect sizes. Power was estimated as the proportion of the *P* values less than $\alpha = 10^{-7}$ based on 10^4 replicates. Total sample sizes considered were 10,000.

Supplementary Figure 2.5. Simulation-study power comparisons of burden test, SKAT, ACAT-V and STAAR for continuous traits with different effect sizes (c_0) and different proportions of effect size directions.



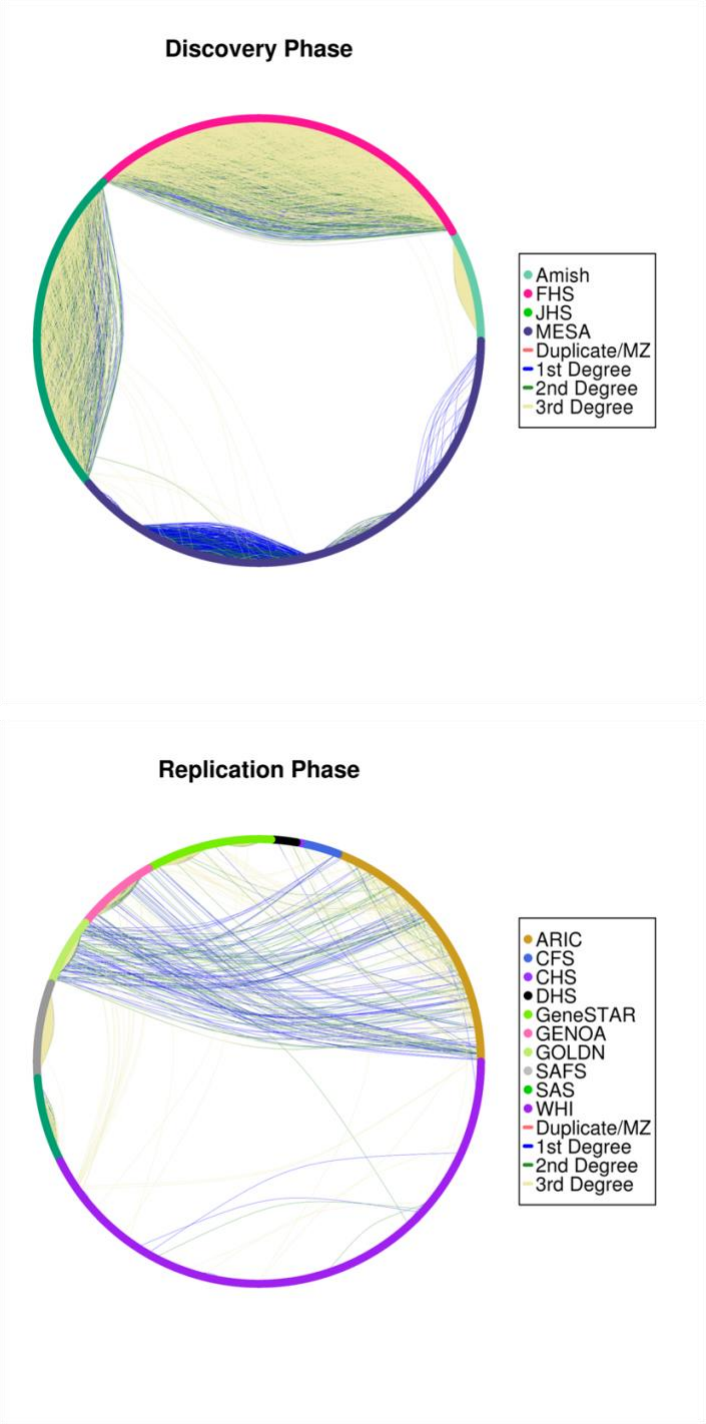
In each simulation replicate, a 5-kb region was randomly selected as the signal region. Within each signal region, variants were randomly generated to be causal based on the multivariate logistic model and on average there were 5%, 15% or 35% causal variants in the signal region. The effect sizes of causal variants were $\beta_j = c_0 |\log_{10} MAF_j|$. Power was estimated as the proportion of the p-values less than $\alpha = 10^{-7}$ based on 10^4 replicates. Total sample sizes considered were 10,000. For each setting, seven statistical tests were compared: burden test, STAAR-B, SKAT, STAAR-S, ACAT-V, STAAR-A, and STAAR-O (Methods).

Supplementary Figure 2.6. Simulation-study power comparisons of Burden, SKAT, ACAT-V and STAAR for dichotomous traits with different effect sizes (c_0) and different proportions of effect size directions.



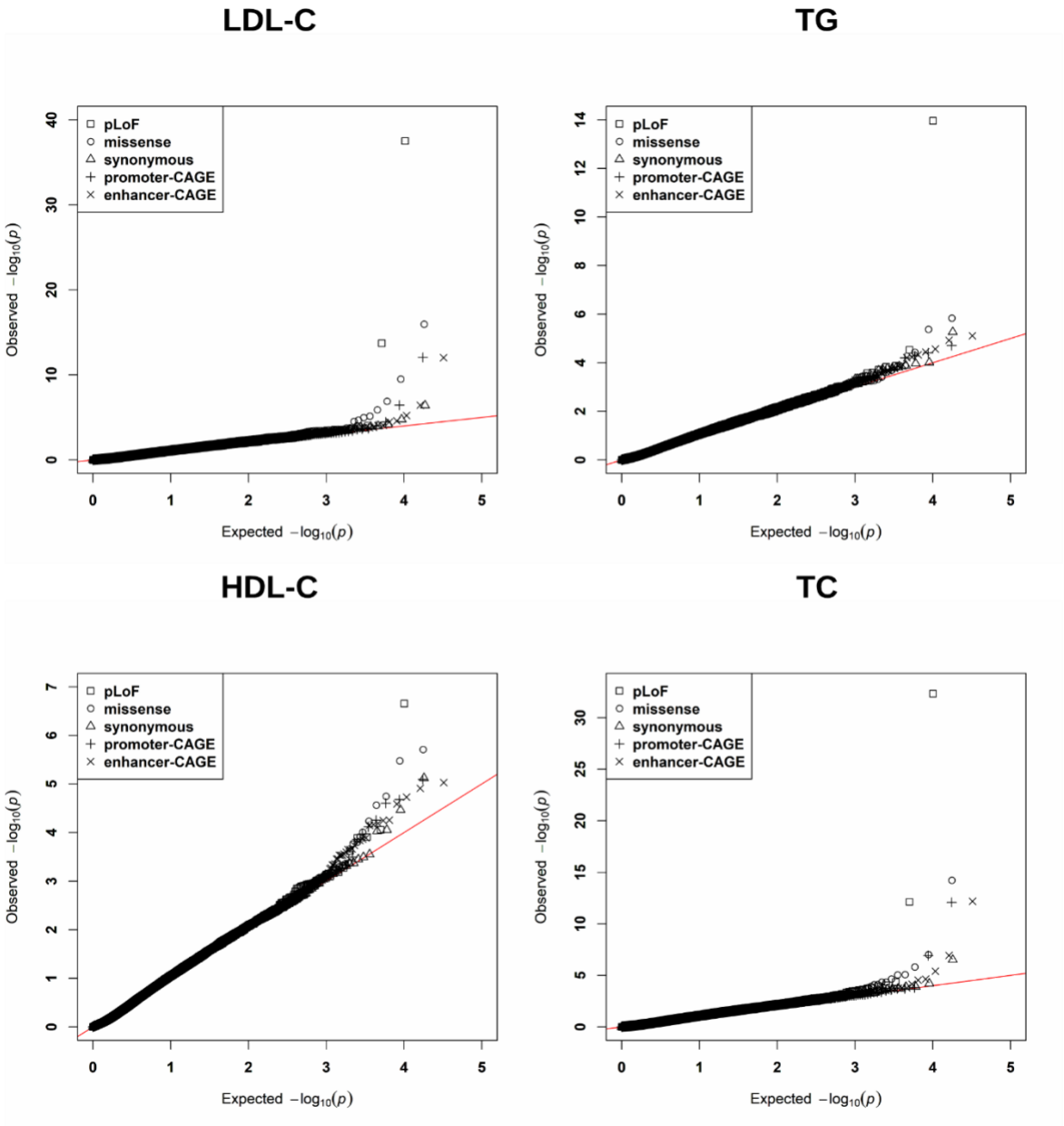
In each simulation replicate, a 5-kb region was randomly selected as the signal region. Within each signal region, variants were randomly generated to be causal based on the multivariate logistic model, and on average there were 5%, 15% or 35% causal variants in the signal region. The effect sizes of causal variants were $\beta_j = c_0 |\log_{10} MAF_j|$. Power was estimated as the proportion of the p-values less than $\alpha = 10^{-7}$ based on 10^4 replicates. Total sample sizes considered were 10,000. For each setting, seven statistical tests were compared: burden test, STAAR-B, SKAT, STAAR-S, ACAT-V, STAAR-A, and STAAR-O (Methods).

Supplementary Figure 2.7. Relatedness of subjects within and across studies in the discovery and replication samples of the TOPMed lipid study.



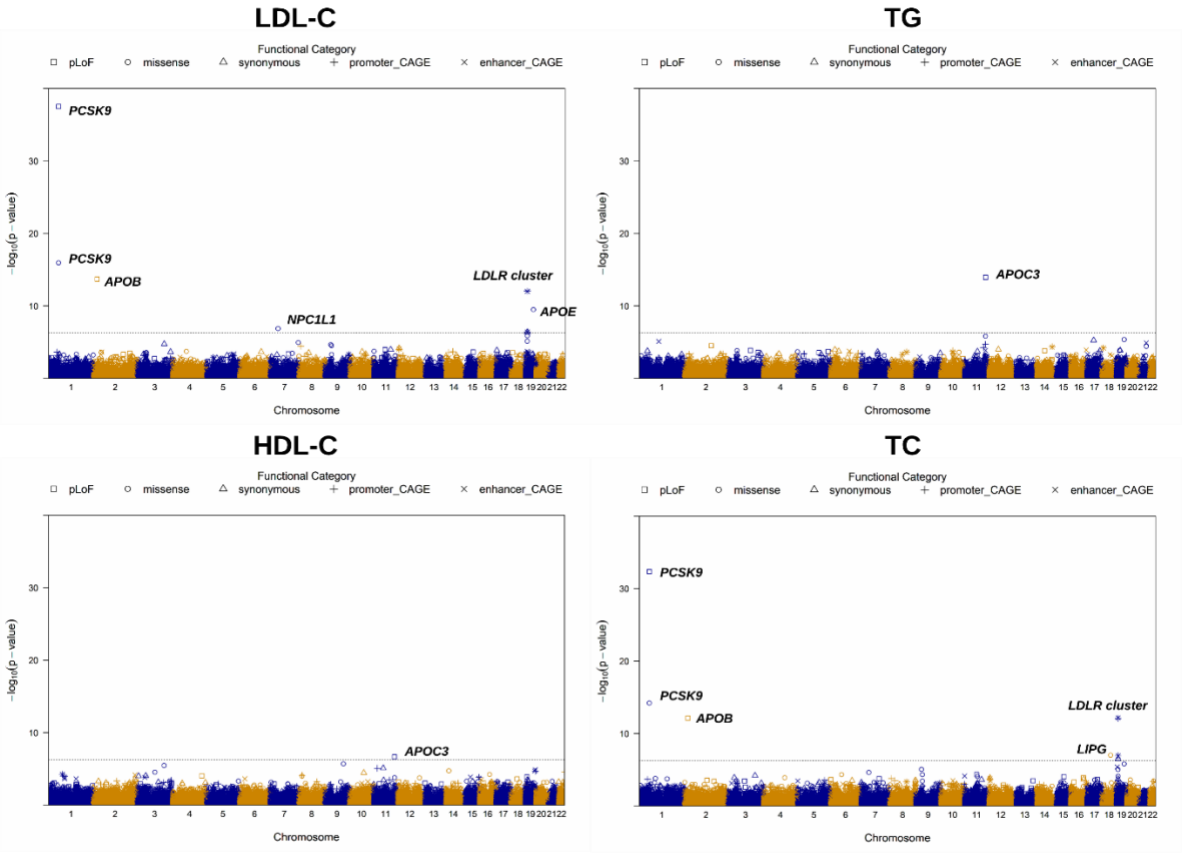
See Supplementary Note for study abbreviations.

Supplementary Figure 2.8. Quantile-quantile plots for gene-centric unconditional analysis of lipid traits LDL-C, HDL-C, TG and TC in discovery phase using the TOPMed cohort (*n* = 12,316).



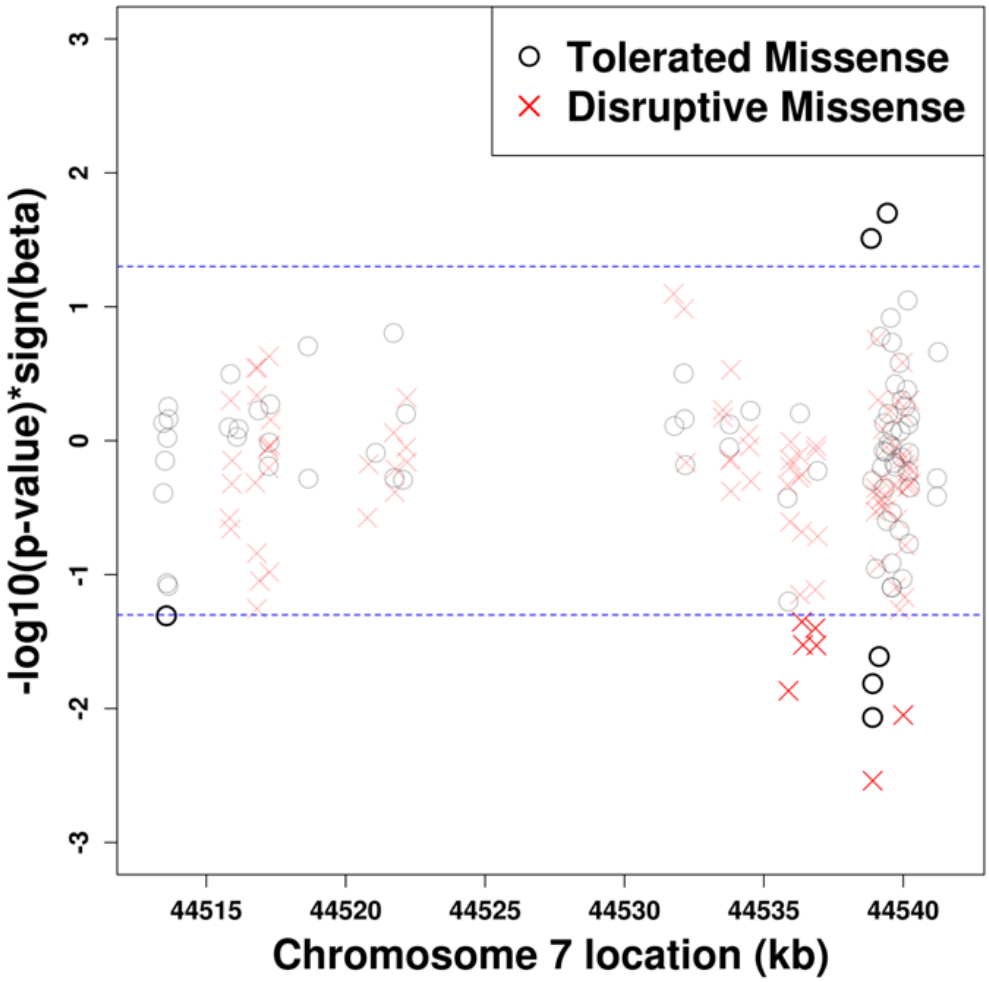
Different symbols represent the STAAR-O *P* value of the gene using different functional categories (pLoF, missense, synonymous, promoter, and enhancer). Promoter and enhancer are the promoter and the GeneHancer region with overlap of DNase hypersensitivity sites for a given gene. Four lipid traits were analyzed using linear mixed models (Methods): LDL-C, low-density lipoprotein cholesterol); HDL-C, high-density lipoprotein cholesterol; TG, triglycerides; TC, total cholesterol.

Supplementary Figure 2.9. Manhattan plots for gene-centric unconditional analysis of lipid traits LDL-C, HDL-C, TG and TC in discovery phase using the TOPMed cohort (*n* = 12,316).



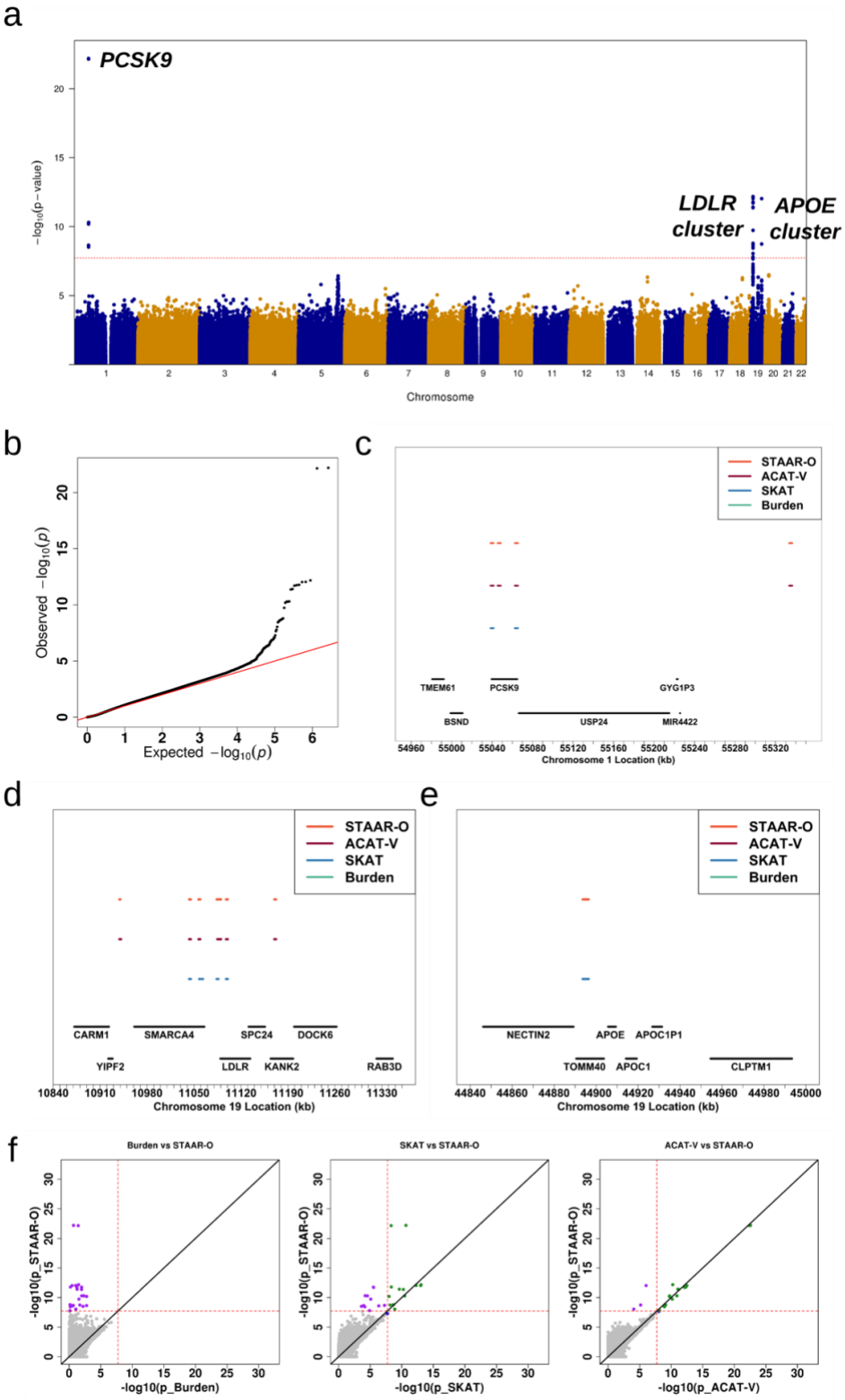
The horizontal line indicates a genome-wide STAAR-O *P* value threshold of 5.00×10^{-7}. Different symbols represent the STAAR-O *P* value of the gene using different functional categories (pLoF, missense, synonymous, promoter, and enhancer). Promoter and enhancer are the promoter and the GeneHancer region with overlap of DNase hypersensitivity sites for a given gene, respectively. Four lipid traits were analyzed **(**Methods**)**: LDL-C, low-density lipoprotein cholesterol; HDL-C, high-density lipoprotein cholesterol; TG, triglycerides; TC, total cholesterol.

Supplementary Figure 2.10. Individual variant unconditional *P*-values associated with LDL-C for missense RVs in gene *NPC1L1* on chromosome 7 in discovery phase using the TOPMed cohort (*n* = 12,316).



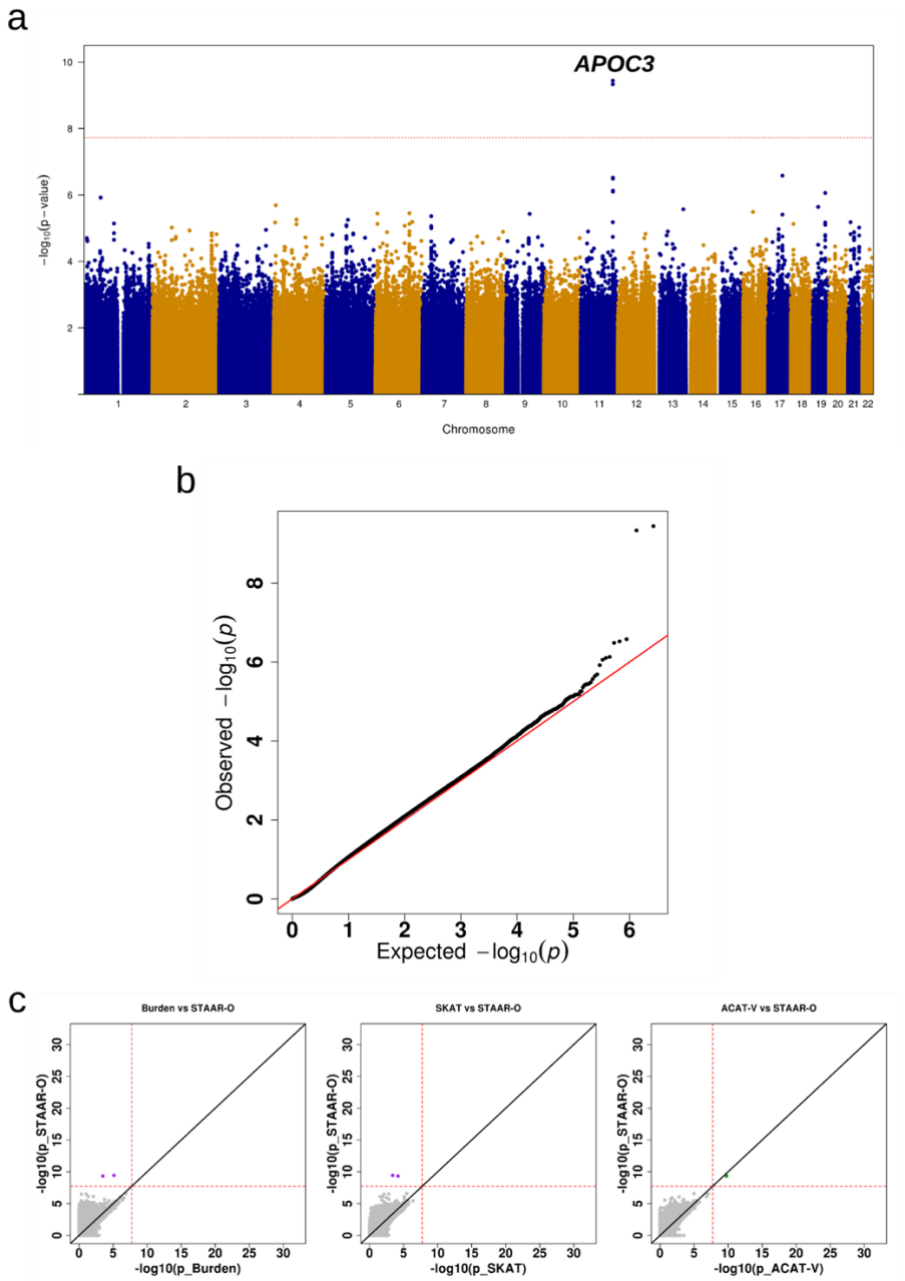
Each dot represents a variant with x-axis label being the physical position on build hg38 and y-axis label being the sign of effect size times $-\log_{10}(P)$. The *P* values were calculated using the individual variant score test. Different symbols represent different types of missense variants, including disruptive missense variant (MetaSVM="D") and tolerated missense variant (MetaSVM="T").

Supplementary Figure 2.11. Genetic region (2-kb sliding window) unconditional analysis results of TC in discovery phase using the TOPMed cohort.



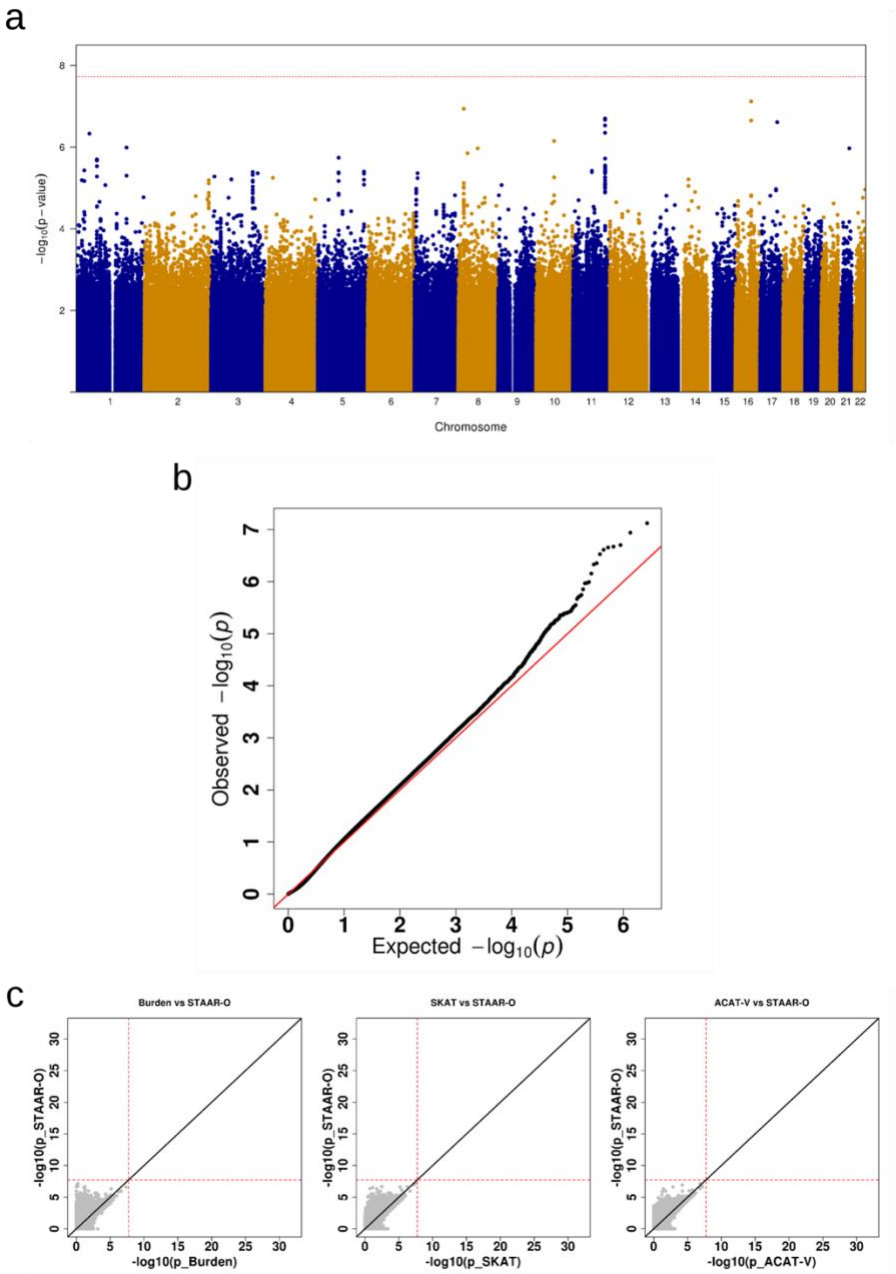
a, Manhattan plot showing the associations of 2.66 million 2-kb sliding windows for TC (total cholesterol) versus $-\log_{10}(P)$ of STAAR-O. The horizontal line indicates a genome-wide *P*-value threshold of 1.88×10^{-8} (*n* = 12,316). b, Quantile-quantile plot of 2-kb sliding window STAAR-O *P* values for TC (*n* = 12,316). c, Genetic landscape of the windows significantly associated with TC that are located in the 500-kb region on chromosome 1. Four statistical tests were compared: Burden, SKAT, ACAT-V and STAAR-O. A dot indicates that the sliding window at this location is significant using the statistical test that the color of the dot represents (*n* = 12,316). d, Genetic landscape of the windows significantly associated with TC that are located in the 200-kb region on chromosome 19. Four statistical tests were compared: Burden, SKAT, ACAT-V and STAAR-O. A dot indicates that the sliding window at this location is significant using the statistical test that the color of the dot represents (*n* = 12,316). e, Genetic landscape of the windows significantly associated with TC that are located in the 150-kb region on chromosome 19. Four statistical tests were compared: Burden, SKAT, ACAT-V and STAAR-O. A dot indicates that the sliding window at this location is significant using the statistical test that the color of the dot represents (*n* = 12,316). f, Scatterplot of *P* values for the 2-kb sliding windows comparing STAAR-O with Burden, SKAT and ACAT-V tests. Each dot represents a sliding window with x-axis label being the $-\log_{10}(P)$ of the conventional test and y-axis label being the $-\log_{10}(P)$ of STAAR-O (*n* = 12,316).

Supplementary Figure 2.12. Genetic region (2-kb sliding window) unconditional analysis results of TG in discovery phase using the TOPMed cohort.



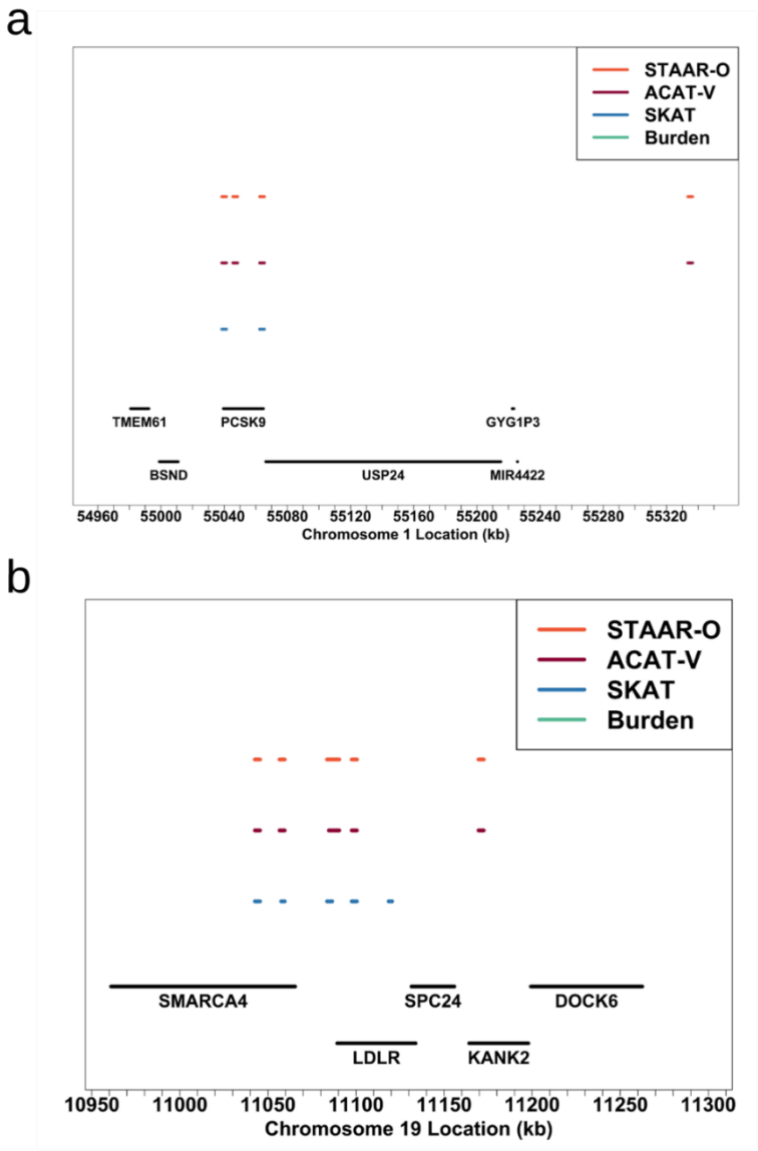
a, Manhattan plot showing the associations of 2.66 million 2-kb sliding windows for TG (triglycerides) versus $-\log_{10}(P)$ of STAAR-O. The horizontal line indicates a genome-wide *P*-value threshold of 1.88×10^{-8} (*n* = 12,316). b, Quantile-quantile plot of 2-kb sliding window STAAR-O *P*-values for TG (*n* = 12,316). c, Scatterplot of *P*-values for the 2-kb sliding windows comparing STAAR-O with Burden, SKAT and ACAT-V tests. Each dot represents a sliding window with x-axis label being the $-\log_{10}(P)$ of the conventional test and y-axis label being the $-\log_{10}(P)$ of STAAR-O (*n* = 12,316).

Supplementary Figure 2.13. Genetic region (2-kb sliding window) unconditional analysis results of HDL-C in discovery phase using the TOPMed cohort.



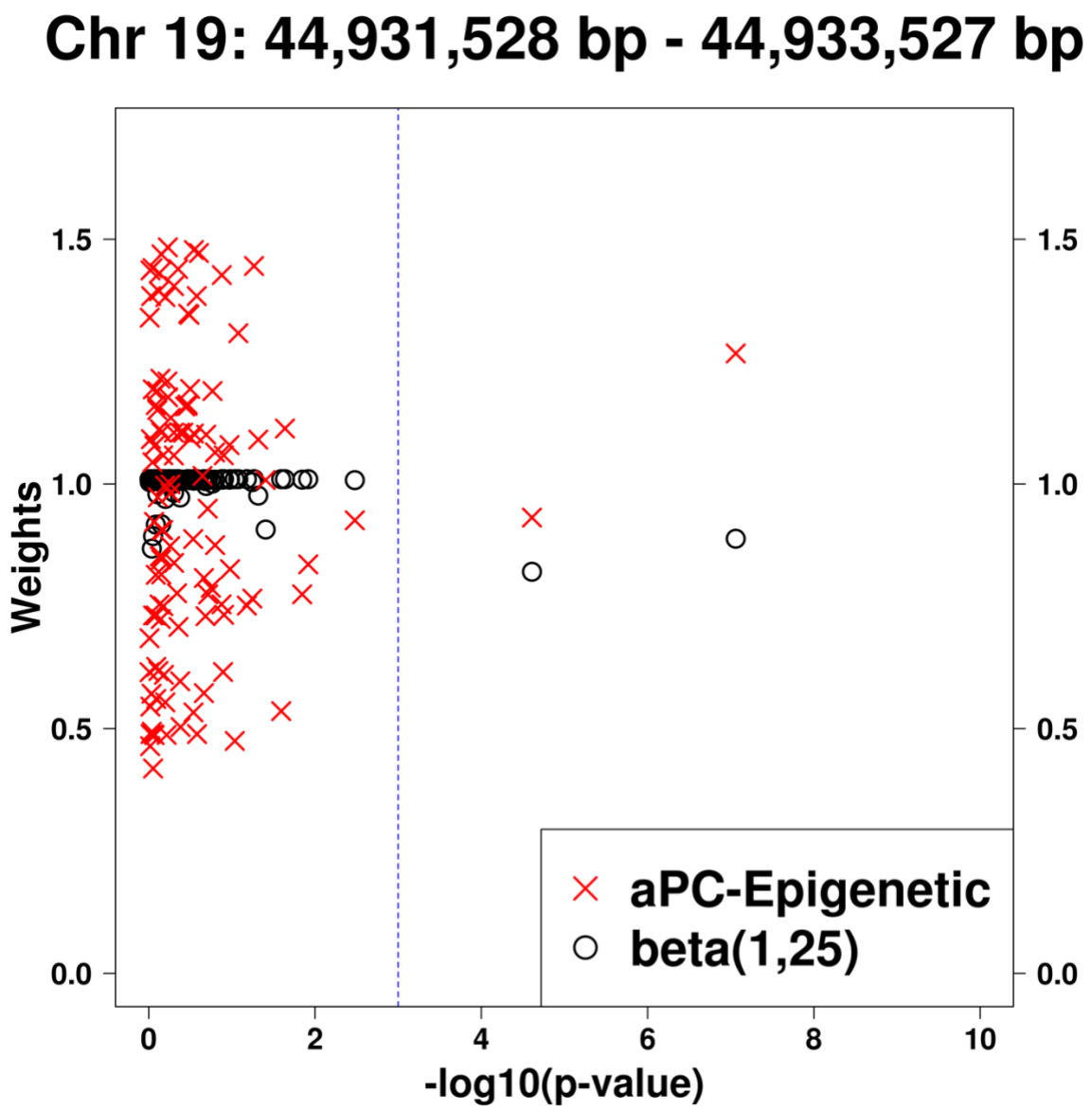
a, Manhattan plot showing the associations of 2.66 million 2-kb sliding windows for HDL-C (high-density lipoprotein cholesterol) versus $-\log_{10}(P)$ of STAAR-O. The horizontal line indicates a genome-wide *P*-value threshold of 1.88×10^{-8} (n = 12,316). b, Quantile-quantile plot of 2-kb sliding window STAAR-O *P*-values for HDL-C (n = 12,316). c, Scatterplot of *P*-values for the 2-kb sliding windows comparing STAAR-O with Burden, SKAT and ACAT-V tests. Each dot represents a sliding window with x-axis label being the $-\log_{10}(P)$ of the conventional test and y-axis label being the $-\log_{10}(P)$ of STAAR-O (n = 12,316).

Supplementary Figure 2.14. Genetic landscape of the sliding windows significantly associated with LDL-C in unconditional analysis using different methods in discovery phase using the TOPMed cohort.



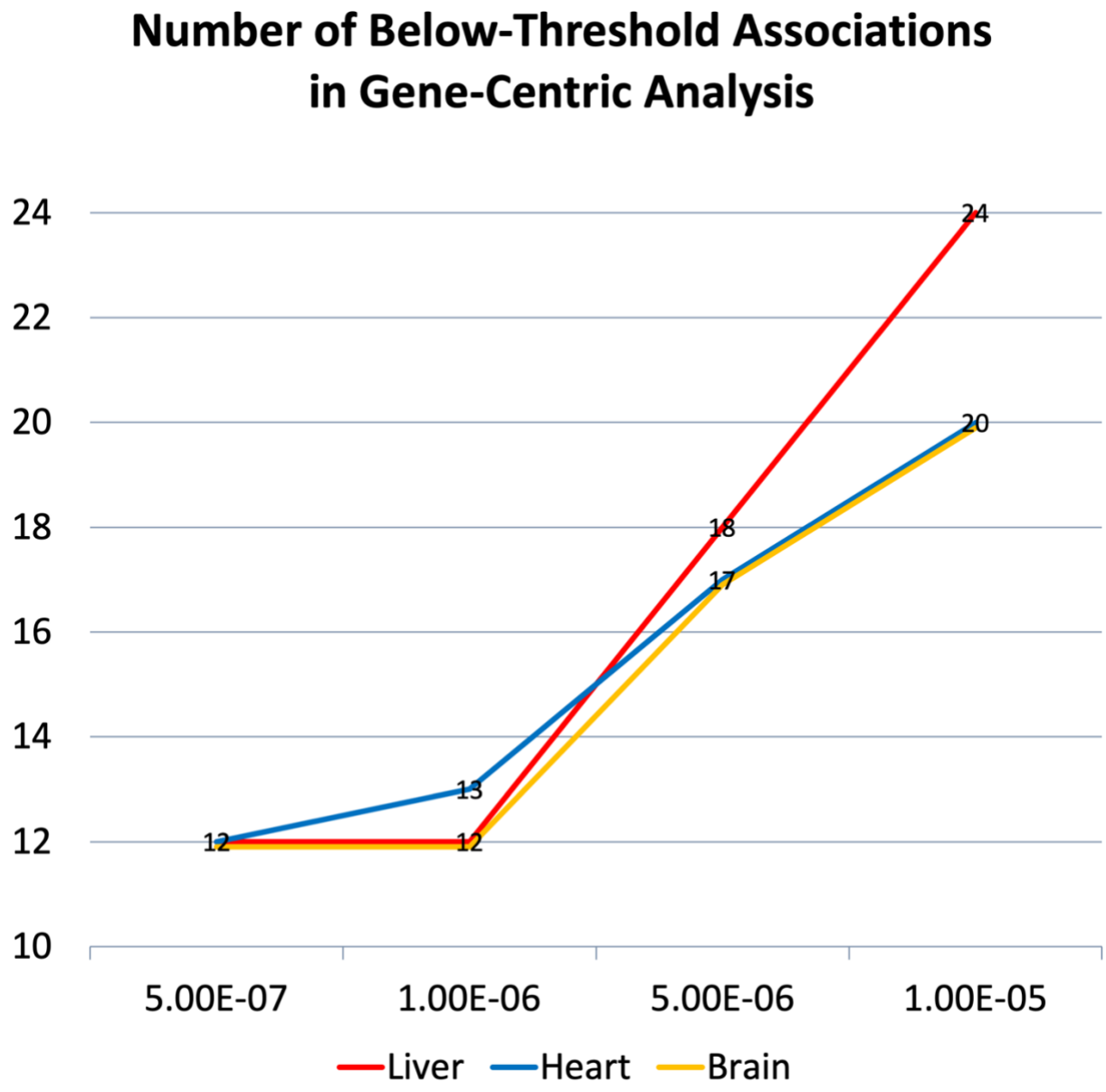
a, Genetic landscape of the windows significantly associated with LDL-C that are located in the 500-kb region on chromosome 1. Four statistical tests were compared: Burden, SKAT, ACAT-V and STAAR-O. A dot indicates that the sliding window at this location is significant using the statistical test that the color of the dot represents (n = 12,316). b, Genetic landscape of the windows significantly associated with LDL-C that are located in the 150-kb region on chromosome 19. Four statistical tests were compared: Burden, SKAT, ACAT-V and STAAR-O. A dot indicates that the sliding window at this location is significant using the statistical test that the color of the dot represents (n = 12,316).

Supplementary Figure 2.15. Individual variant unconditional *P* values associated with LDL-C for RVs in a sliding window near gene *APOC1L1* in discovery phase using the TOPMed cohort (*n* = 12,316).



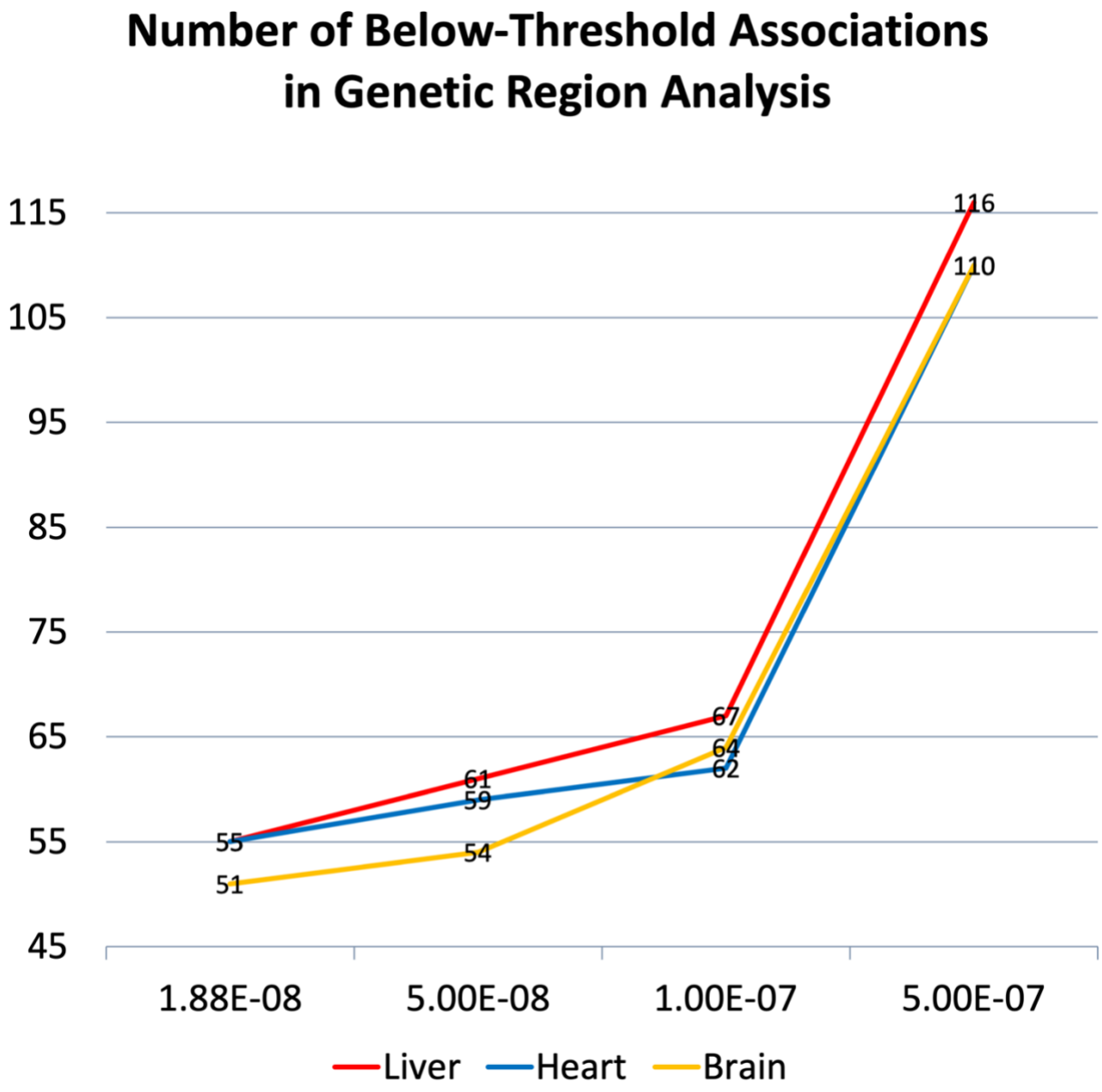
The x-axis label is the $-\log_{10}(P)$. The *P*-values were calculated using the individual variant score test. The y-axis label is the standardized Beta(MAF;1,25) weights and the standardized aPC-Epigenetic weights for individual RVs in the sliding window. The sliding window is located from 44,931,528 bp to 44,933,527 bp on chromosome 19. The physical positions are on build hg38.

Supplementary Figure 2.16. Number of below-threshold associations by incorporating tissue-specific aPCs in gene-centric analysis in discovery phase using the TOPMed cohort (*n* = 12,316).



Various levels of unconditional STAAR-O *P* value thresholds ($\alpha = 5.00 \times 10^{-7}, 1.00 \times 10^{-6}, 5.00 \times 10^{-6}, 1.00 \times 10^{-5}$) using discovery phase are compared.

Supplementary Figure 2.17. Number of below-threshold associations by incorporating tissue-specific aPCs in genetic region analysis in discovery phase using the TOPMed cohort (*n* = 12,316).



Various levels of unconditional STAAR-O *P* value thresholds ($\alpha = 1.88 \times 10^{-8}, 5.00 \times 10^{-8}, 1.00 \times 10^{-7}, 5.00 \times 10^{-7}$) using discovery phase are compared.

TOPMed study participants and acknowledgements

Discovery phase (n = 12,316)

Framingham Heart Study (FHS)

The FHS is a three generational prospective cohort that has been described in detail previously (131). Individuals were initially recruited in 1948 in Framingham, USA to evaluate cardiovascular disease risk factors. The second generation cohort (5,124 offspring of the original cohort) was recruited between 1971 and 1975 (132, 133). The third generation cohort (4,095 grandchildren of the original cohort) was collected between 2002 and 2005. Fasting lipid levels were measured at exam 1 of the Offspring (1971-1975) and third generation (2002-2005) cohorts, using standard LRC protocols.

Whole genome sequencing (WGS) for the Trans-Omics in Precision Medicine (TOPMed) program was supported by the National Heart, Lung and Blood Institute (NHLBI). WGS for "NHLBI TOPMed: Whole Genome Sequencing and Related Phenotypes in the Framingham Heart Study" (phs000974.v1.p1) was performed at the Broad Institute of MIT and Harvard (HHSN268201500014C).

The Framingham Heart Study (FHS) acknowledges the support of contracts NO1-HC-25195, HHSN268201500001I and 75N92019D00031 from the National Heart, Lung and Blood Institute and grant supplement R01 HL092577-06S1 for this research. We also acknowledge the dedication of the FHS study participants without whom this research would not be possible. Dr. Vasan is supported in part by the Evans Medical Foundation and the Jay and Louis Coffman Endowment from the Department of Medicine, Boston University School of Medicine.

Old Order Amish (OOA)

The Old Order Amish individuals included in this study were participants of several ongoing studies of cardiovascular health carried out at the University of Maryland among relatively healthy volunteers from the Old Order Amish community of Lancaster County, PA and their family members (134, 135).

Whole genome sequencing (WGS) for the Trans-Omics in Precision Medicine (TOPMed) program was supported by the National Heart, Lung and Blood Institute (NHLBI). WGS for "NHLBI TOPMed: Genetics of Cardiometabolic Health in the Amish" (phs000956.v1.p1) was performed at the Broad Institute of MIT and Harvard (3R01HL121007-01S1).

Jackson Heart Study (JHS)

The JHS is a large, population-based observational study evaluating the etiology of cardiovascular, renal, and respiratory diseases among African Americans residing in the three

counties (Hinds, Madison, and Rankin) that make up the Jackson, Mississippi metropolitan area (136). Data and biologic materials have been collected from 5,301 participants, including a nested family cohort of 1,498 members of 264 families. The age at enrollment for the unrelated cohort was 35-84 years; the family cohort included related individuals >21 years old. Participants provided extensive medical and social history, had an array of physical and biochemical measurements and diagnostic procedures, and provided genomic DNA during a baseline examination (2000-2004) and two follow-up examinations (2005-2008 and 2009-2012). The study population is characterized by a high prevalence of diabetes, hypertension, obesity, and related disorders. Annual follow-up interviews and cohort surveillance are ongoing.

Whole genome sequencing (WGS) for the Trans-Omics in Precision Medicine (TOPMed) program was supported by the National Heart, Lung and Blood Institute (NHLBI). WGS for "NHLBI TOPMed: The Jackson Heart Study" (phs000964.v1.p1) was performed at the University of Washington Northwest Genomics Center (HHSN268201100037C).

The Jackson Heart Study (JHS) is supported and conducted in collaboration with Jackson State University (HHSN268201800013I), Tougaloo College (HHSN268201800014I), the Mississippi State Department of Health (HHSN268201800015I/HHSN26800001) and the University of Mississippi Medical Center (HHSN268201800010I, HHSN268201800011I and HHSN268201800012I) contracts from the National Heart, Lung, and Blood Institute (NHLBI) and the National Institute for Minority Health and Health Disparities (NIMHD). The authors also wish to thank the staffs and participants of the JHS.

Multi-Ethnic Study of Atherosclerosis (MESA)

The Multi-Ethnic Study of Atherosclerosis is a National Heart, Lung and Blood Institute-sponsored, population-based investigation of subclinical cardiovascular disease and its progression (137). A total of 6,814 individuals, aged 45 to 84 years, were recruited from six US communities (Baltimore City and County, MD; Chicago, IL; Forsyth County, NC; Los Angeles County, CA; New York, NY; and St. Paul, MN) between July 2000 and August 2002. Participants were excluded if they had physician-diagnosed cardiovascular disease prior to enrollment, including angina, myocardial infarction, heart failure, stroke or TIA, resuscitated cardiac arrest or a cardiovascular intervention (e.g., CABG, angioplasty, valve replacement, or pacemaker/defibrillator placement). Pre-specified recruitment plans identified four racial/ethnic groups (White European-American, African-American, Hispanic-American, and Chinese-American) for enrollment, with targeted oversampling of minority groups to enhance statistical power.

Whole genome sequencing (WGS) for the Trans-Omics in Precision Medicine (TOPMed) program was supported by the National Heart, Lung and Blood Institute (NHLBI). WGS for "NHLBI TOPMed: Multi-Ethnic Study of Atherosclerosis (MESA)" (phs001416.v1.p1) was performed at the Broad Institute of MIT and Harvard (3U54HG003067-13S1). Centralized read

mapping and genotype calling, along with variant quality metrics and filtering were provided by the TOPMed Informatics Research Center (3R01HL-117626-02S1; contract HHSN268201800002I). Phenotype harmonization, data management, sample-identity QC, and general study coordination, were provided by the TOPMed Data Coordinating Center (3R01HL-120393-02S1; contract HHSN268201800001I).

MESA and the MESA SHARe projects are conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with MESA investigators. Support for MESA is provided by contracts 75N92020D00001, HHSN268201500003I, N01-HC-95159, 75N92020D00005, N01-HC-95160, 75N92020D00002, N01-HC-95161, 75N92020D00003, N01-HC-95162, 75N92020D00006, N01-HC-95163, 75N92020D00004, N01-HC-95164, 75N92020D00007, N01-HC-95165, N01-HC-95166, N01-HC-95167, N01-HC-95168, N01-HC-95169, UL1-TR-000040, UL1-TR-001079, and UL1-TR-001420, UL1TR001881, and DK063491.

Replication phase (n = 17,822)

Atherosclerosis Risk in Communities Study (ARIC)

The ARIC study is a population-based prospective cohort study of cardiovascular disease sponsored by the National Heart, Lung, and Blood Institute (NHLBI). ARIC included 15,792 individuals, predominantly European American and African American, aged 45-64 years at baseline (1987-89), chosen by probability sampling from four US communities. Cohort members completed three additional triennial follow-up examinations, a fifth exam in 2011-2013, and a sixth exam in 2016-2017. The ARIC study has been described in detail previously (138).

Whole genome sequencing (WGS) for the Trans-Omics in Precision Medicine (TOPMed) program was supported by the National Heart, Lung and Blood Institute (NHLBI). WGS for "NHLBI TOPMed: Atherosclerosis Risk in Communities (ARIC)" (phs001211) was performed at the Baylor College of Medicine Human Genome Sequencing Center (HHSN268201500015C and 3U54HG003273-12S2) and the Broad Institute of MIT and Harvard (3R01HL092577-06S1). Centralized read mapping and genotype calling, along with variant quality metrics and filtering were provided by the TOPMed Informatics Research Center (3R01HL-117626-02S1; contract HHSN268201800002I). Phenotype harmonization, data management, sample-identity QC, and general study coordination, were provided by the TOPMed Data Coordinating Center (3R01HL-120393-02S1; contract HHSN268201800001I). We gratefully acknowledge the studies and participants who provided biological samples and data for TOPMed.

The Atherosclerosis Risk in Communities study has been funded in whole or in part with Federal funds from the National Heart, Lung, and Blood Institute, National Institutes of Health, Department of Health and Human Services (contract numbers HHSN268201700001I, HHSN268201700002I, HHSN268201700003I, HHSN268201700004I and

HHSN268201700005I). The authors thank the staff and participants of the ARIC study for their important contributions.

Cleveland Family Study (CFS)

The CFS is a family-based longitudinal study that includes participants with laboratory diagnosed sleep apnea, their family members and neighborhood control families followed between 1990 and 2006. Four examinations over 16 years provided measurements of sleep apnea with overnight polysomnography, anthropometry, and other related phenotypes, as detailed previously (139, 140). After an overnight fast, blood was collected which was assayed for lipid levels at the University of Vermont Laboratory for Clinical Biochemistry Research. Lipids (triglycerides, HDL cholesterol) from fasted blood serum were measured by enzymatic methods using Centers for Disease Control and Prevention guidelines (141).

Whole genome sequencing (WGS) for the Trans-Omics in Precision Medicine (TOPMed) program was supported by the National Heart, Lung and Blood Institute (NHLBI). WGS for "NHLBI TOPMed: Cleveland Family Study" (phs000954) was performed at the University of Washington Northwest Genomics Center (3R01HL098433-05S1).

This research was supported by grants HL 046389; HL113338;1R35HL135818 from the National Heart, Lung, and Blood Institute (NHLBI).

Cardiovascular Health Study (CHS)

The Cardiovascular Health Study is a prospective population-based cohort study of risk factors for CHD and stroke in adults 65 years and older (142). The main objective is to identify factors related to the onset and course of heart disease and stroke. The four Field Centers are located in Forsyth County, NC; Sacramento County, CA; Washington County, MD; and Pittsburgh, PA. The original cohort of 5201 elderly were recruited in 1989-1990; and in 1992-1993, 687 additional minority participants were recruited and examined. Each community sample was obtained from random samples of the Medicare eligibility lists of the Health Care Financing Administration (HCFA). Eligible to participate were persons living in the household of each sampled individual who were: 1) 65 yr or older; 2) non-institutionalized; 3) expected to remain in the area for 3 yr; and 4) able to give informed consent. Excluded were those wheelchair-bound, receiving hospice care or cancer treatment. The minority cohort was recruited using similar methods. Participants were eligible whether or not they had clinically apparent cardiovascular disease. Subjects were followed with semi-annual contacts, alternating between telephone calls and surveillance clinic visits.

Whole genome sequencing (WGS) for the Trans-Omics in Precision Medicine (TOPMed) program was supported by the National Heart, Lung and Blood Institute (NHLBI). WGS for

“NHLBI TOPMed: Cardiovascular Health Study” (phs001368) was performed at the Baylor College of Medicine Human Genome Sequencing Center (HHSN268201500015C).

This research was supported by contracts HHSN268201200036C, HHSN268200800007C, HHSN268201800001C, N01HC55222, N01HC85079, N01HC85080, N01HC85081, N01HC85082, N01HC85083, N01HC85086, and grants U01HL080295 and U01HL130114 from the National Heart, Lung, and Blood Institute (NHLBI), with additional contribution from the National Institute of Neurological Disorders and Stroke (NINDS). Additional support was provided by R01AG023629 from the National Institute on Aging (NIA). A full list of principal CHS investigators and institutions can be found at CHS-NHLBI.org.

Diabetes Heart Study (DHS)

The Diabetes Heart Study (DHS) began as a family-based study enriched for type 2 diabetes (T2D). The initial cohort included 1443 European American and African American participants from 564 families with multiple cases of type 2 diabetes recruited between 1998 and 2006 (143). As an ancillary study, the African American Diabetes Heart Study (AA-DHS) expanded the total number of African Americans to 691 by recruiting additional unrelated participants with type 2 diabetes from 2007 and 2010 (144). All participants were extensively phenotyped for measures of subclinical CVD and other known CVD risk factors. Primary outcomes were quantified burden of vascular calcified plaque in the coronary artery, carotid artery, and abdominal aorta all determined from non-contrast computed tomography scans. For TOPMed, DHS and AA-DHS African American participants with CAC were selected for WGS, prioritizing the inclusion of families.

Whole genome sequencing (WGS) for the Trans-Omics in Precision Medicine (TOPMed) program was supported by the National Heart, Lung and Blood Institute (NHLBI). WGS for “NHLBI TOPMed: Diabetes Heart Study” (phs001412) was performed at the Broad Institute of MIT and Harvard (HHSN268201500014C).

This work was supported by R01 HL92301, R01 HL67348, R01 NS058700, R01 AR48797, R01 DK071891, the General Clinical Research Center of the Wake Forest University School of Medicine (M01 RR07122, F32 HL085989), the American Diabetes Association, and a pilot grant from the Claude Pepper Older Americans Independence Center of Wake Forest University Health Sciences (P60 AG10484).

Genetic Study of Atherosclerosis Risk (GeneSTAR)

GeneSTAR is an ongoing family-based prospective study designed to determine environmental, phenotypic, and genetic causes of premature cardiovascular disease. GeneSTAR was originally conducted in healthy adult European- and African-American siblings of probands with documented early onset coronary disease under 60 years of age at the time of hospitalization in

any of 10 Baltimore area hospitals from 1982-2006. Participants were screened for traditional coronary disease and stroke risk factors and have been followed regularly to ascertain incident cardiovascular disease (145). Commencing in 2003, the siblings, their offspring, and the coparent of the offspring who were free of cardiovascular disease participated in a 2 week trial of aspirin 81 mg/day with pre and post ex vivo platelet function assessed using multiple agonists and were screened for traditional coronary disease and stroke risk factors (146). Of the total 3949 participants, 1786 were selected for TOPMed prioritized on complete platelet function measures and largest family size.

Whole genome sequencing (WGS) for the Trans-Omics in Precision Medicine (TOPMed) program was supported by the National Heart, Lung and Blood Institute (NHLBI). WGS for "NHLBI TOPMed: Genetic Study of Atherosclerosis Risk" (phs001218) was performed at the Microgen Corp. and the Broad Institute of MIT and Harvard (HHSN268201500014C).

GeneSTAR was supported by grants from the National Institutes of Health/National Heart, Lung, and Blood Institute (U01 HL72518, HL087698, HL49762, HL59684, HL58625, HL071025, HL112064), by a grant from the National Institutes of Health/National Institute of Nursing Research (NR0224103), and by a grant from the National Institutes of Health/National Center for Research Resources (M01-RR000052) to the Johns Hopkins General Clinical Research Center.

Genetic Epidemiology Network of Arteriopathy (GENOA)

The Genetic Epidemiology Network of Arteriopathy (GENOA) is one of four networks in the NHLBI Family-Blood Pressure Program (FBPP) (147). GENOA's long-term objective is to elucidate the genetics of target organ complications of hypertension, including both atherosclerotic and arteriolosclerotic complications involving the heart, brain, kidneys, and peripheral arteries (148). The longitudinal GENOA Study recruited European-American and African-American sibships with at least 2 individuals with clinically diagnosed essential hypertension before age 60 years. All other members of the sibship were invited to participate regardless of their hypertension status. Participants were diagnosed with hypertension if they had either 1) a previous clinical diagnosis of hypertension by a physician with current anti-hypertensive treatment, or 2) an average systolic blood pressure ≥ 140 mm Hg or diastolic blood pressure ≥ 90 mm Hg based on the second and third readings at the time of their clinic visit. Only participants of the African-American Cohort were sequenced through TOPMed.

During the first exam (Phase 1; 1996-2000), 1,583 European-Americans from Rochester, MN and 1,854 African-Americans from Jackson, MS were examined. Between 2000 and 2004 (Phase 2), 1,241 participants of the European-American Cohort and 1,482 participants of the African-American cohort returned for a second examination. The second examination of the European-American cohort included computed tomography scans for coronary artery calcification while the second examination of the African-American cohort included an echocardiogram. Between

2009 and 2011, an examination that included computed tomography scans for coronary artery calcification (CAC Study) was conducted on 752 participants of the African-American Cohort.

Every participant with an echocardiogram was selected for whole genome sequencing (WGS) through TOPMed. We then selected 106 African-American participants who had a computed tomography scan for coronary artery calcification but not an echocardiogram or were a sibling of someone already selected for WGS. Finally, we excluded individuals whom we knew were already being whole genome sequenced through TOPMed or another sequencing effort (GENOA participants who overlap with ARIC or JHS participants).

Support for GENOA was provided by the National Heart, Lung and Blood Institute (HL054457, HL054464, HL054481, HL119443, HL085571, and HL087660) of the National Institutes of Health. WGS for "NHLBI TOPMed: Genetic Epidemiology Network of Arteriopathy" (phs001345) was performed at the Mayo Clinic Genotyping Core, the DNA Sequencing and Gene Analysis Center at the University of Washington (3R01HL055673-18S1), and the Broad Institute (HHSN268201500014C) for their genotyping and sequencing services. We would like to thank the GENOA participants.

Genetics of Lipid Lowering Drugs and Diet Network (GOLDN)

GOLDN is a family-based study of European descent individuals recruited in Minneapolis and Salt Lake City (two of the NHLBI Family Heart Study sites). It aims to uncover genetic predictors of variability in lipid phenotypes, which include both fasting and postprandial lipids quantified using traditional methods, NMR, and high-throughput lipidomics. During the initial screening of ~1,350 individuals, the following criteria were used for exclusion: age < 18 years; fasting triglycerides ≥1500 mg/dL; recent history of myocardial infarction, coronary bypass surgery, or coronary angioplasty; self-report of a positive history of liver, kidney, pancreas, or gallbladder disease, or a history of nutrient malabsorption; current use of insulin; abnormal liver or kidney function; in women of childbearing potential, pregnancy, breastfeeding, not using an acceptable form of contraception. Of those who enrolled, 1,048 individuals consented to the use of their DNA in research; 893 participants with data on all exposures, outcomes, and covariates were included in the current study.

GOLDN biospecimens, baseline phenotype data, and intervention phenotype data were collected with funding from National Heart, Lung and Blood Institute (NHLBI) grant U01 HL072524. Whole-genome sequencing in GOLDN was funded by NHLBI grant R01 HL104135-04S1.

Whole genome sequencing (WGS) for the Trans-Omics in Precision Medicine (TOPMed) program was supported by the National Heart, Lung and Blood Institute (NHLBI). WGS for "NHLBI TOPMed: Genetics of Lipid Lowering Drugs and Diet Network" (phs001359) was performed at the University of Washington Northwest Genomics Center (3R01HL104135-04S1).

San Antonio Family Heart Study (SAFS)

The SAFHS began in 1991, and included 1,431 individuals in 42 extended families at baseline. Probands were 40 to 60 year old low-income Mexican Americans selected at random without regard to presence or absence of disease, almost exclusively from Mexican American census tracts in San Antonio, Texas. All first, second, and third degree relatives of the proband and of the proband's spouse, aged 16 years or above, were eligible to participate in the study. As part of our ongoing studies, we have recruited new family members from the original families, expanding the cohort to almost 3,099 individuals primarily from 73 families. Our study is a mixed longitudinal design. Subjects have been seen between 1 and 4 times with an average of 1.95 examinations.

Whole genome sequencing (WGS) for the Trans-Omics in Precision Medicine (TOPMed) program was supported by the National Heart, Lung and Blood Institute (NHLBI). WGS for "NHLBI TOPMed: San Antonio Family Heart Study" (phs001215) was performed at the Illumina Genomic Services (3R01HL113323-03S1).

Collection of the San Antonio Family Study data was supported in part by National Institutes of Health (NIH) grants R01 HL045522, MH078143, MH078111 and MH083824; and whole genome sequencing of SAFS subjects was supported by U01 DK085524 and R01 HL113323. We are very grateful to the participants of the San Antonio Family Study for their continued involvement in our research programs.

Genome-wide Association Study of Adiposity in Samoans (SAS)

The parent Samoan study is a population-based genome-wide association study (GWAS) of adiposity and cardiometabolic phenotypes among adults, 25-65 years of age, from the independent nation of Samoa in the South Pacific. The research goal of this study is to identify genetic variation that increases susceptibility to obesity and cardiometabolic phenotypes. Biomarker and questionnaire data were collected to assess cardiometabolic phenotypes. DNA was collected and the Affymetrix 6.0 chip used for SNP genotyping. After quality control checks on genotyping and excluding individuals with key missing data we have a final sample of 3,122 adults with high-quality genome-wide marker data (149). Participation in TOPMed provided whole genome sequence data for 1,222 individuals from the GWAS sample chosen for maximal informativity for our Samoan-specific imputation panel.

Whole genome sequencing (WGS) for the Trans-Omics in Precision Medicine (TOPMed) program was supported by the National Heart, Lung and Blood Institute (NHLBI). WGS for "NHLBI TOPMed: Genome-wide Association Study of Adiposity in Samoans" (phs000972) was performed at the University of Washington Northwest Genomics Center (HHSN268201100037C) and the New York Genome Center (HHSN268201500016C).

Date collection was funded by NIH grant R01-HL093093. We thank the Samoan participants of the study and local village authorities. We acknowledge the support of the Samoan Ministry of Health and the Samoa Bureau of Statistics for their support of this research.

Women's Health Initiative (WHI)

The Women's Health Initiative (WHI) is a large study of postmenopausal women's health investigating risk factors for cancer, CVD, age-related fractures and chronic disease [ref]. It began in 1993 as a set of randomized controlled clinical trials (CT) and an observational study (OS). Specifically, the CT (n=68,132) included three overlapping components: The Hormone Therapy (HT) Trials (n=27,347), Dietary Modification (DM) Trial (n=48,835), and Calcium and Vitamin D (CaD) Trial (n=36,282). Eligible women could be randomized into as many as all three CTs components. Women who were ineligible or unwilling to join the CT were then invited to join the OS (n=93,676) (150).

Whole genome sequencing (WGS) for the Trans-Omics in Precision Medicine (TOPMed) program was supported by the National Heart, Lung and Blood Institute (NHLBI). WGS for "NHLBI TOPMed: Women's Health Initiative" (phs001237) was performed at the Broad Institute of MIT and Harvard (HHSN268201500014C).

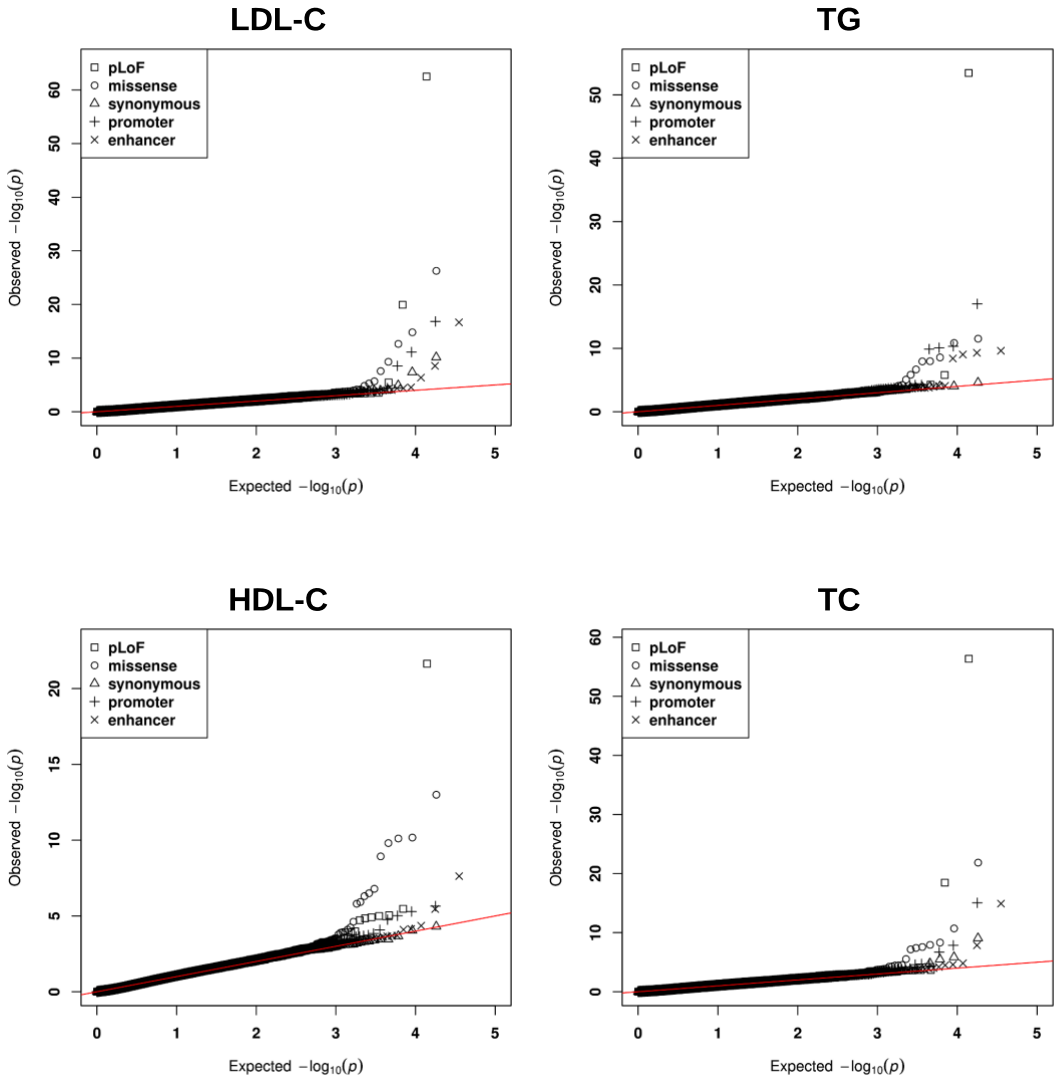
The WHI program is funded by the National Heart, Lung, and Blood Institute, National Institutes of Health, U.S. Department of Health and Human Services through contracts HHSN268201600018C, HHSN268201600001C, HHSN268201600002C, HHSN268201600003C, and HHSN268201600004C. The authors thank the WHI investigators and staff for their dedication, and the study participants for making the program possible. A full listing of WHI investigators can be found at: http://www.whi.org/researchers/Documents%20%20Write%20a%20Paper/WHI%20Investigator%20Long%20List.pdf.

UK Biobank (external to TOPMed)

The UK Biobank analyses were conducted using the UK Biobank resource under application 7089.

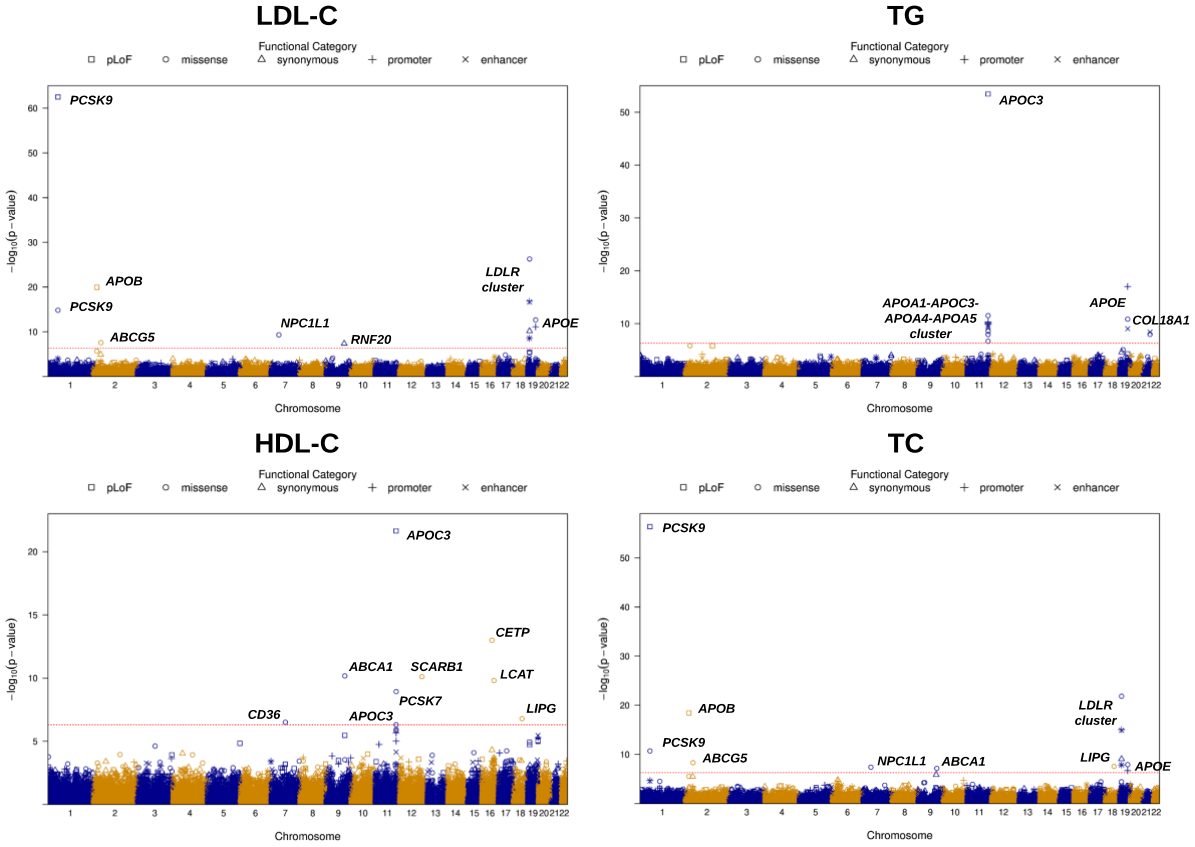
APPENDIX C

Supplementary Figure 3.1. Quantile-quantile plots for gene-centric unconditional meta-analysis of lipid traits LDL-C, HDL-C, TG and TC using the TOPMed data (*n* = 30,138).



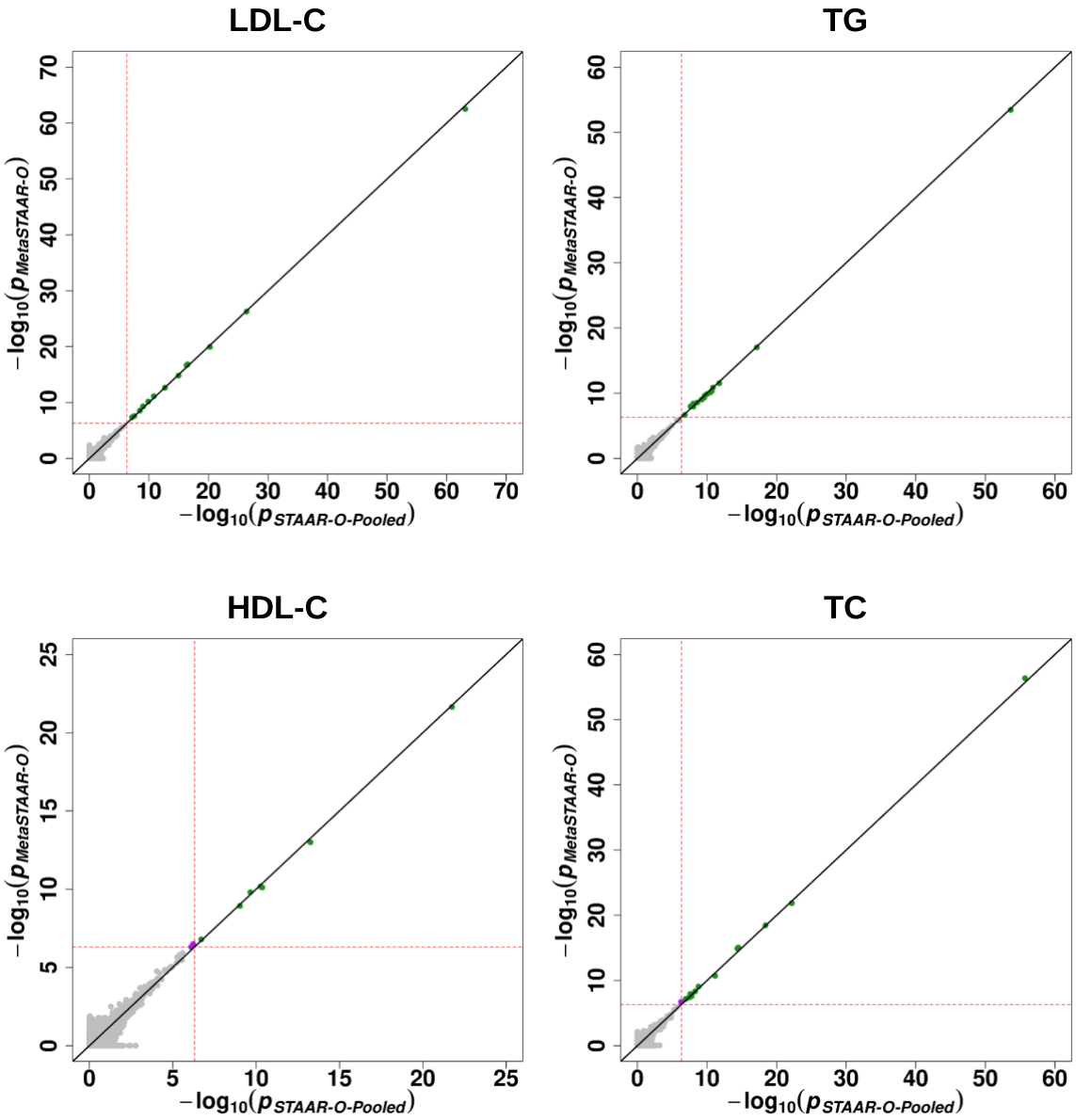
Different symbols represent the MetaSTAAR-O *P* values of the gene using different functional categories (putative loss-of-function, missense, synonymous, promoter and enhancer). Promoter and enhancer are the promoter and the GeneHancer region with overlap of CAGE sites for a given gene, respectively (Methods). Four lipid traits were analyzed using MetaSTAAR: LDL-C, low-density lipoprotein cholesterol; HDL-C, high-density lipoprotein cholesterol; TG, triglycerides; TC, total cholesterol.

Supplementary Figure 3.2. Manhattan plots for gene-centric unconditional meta-analysis of lipid traits LDL-C, HDL-C, TG and TC using the TOPMed data (*n* = 30,138).



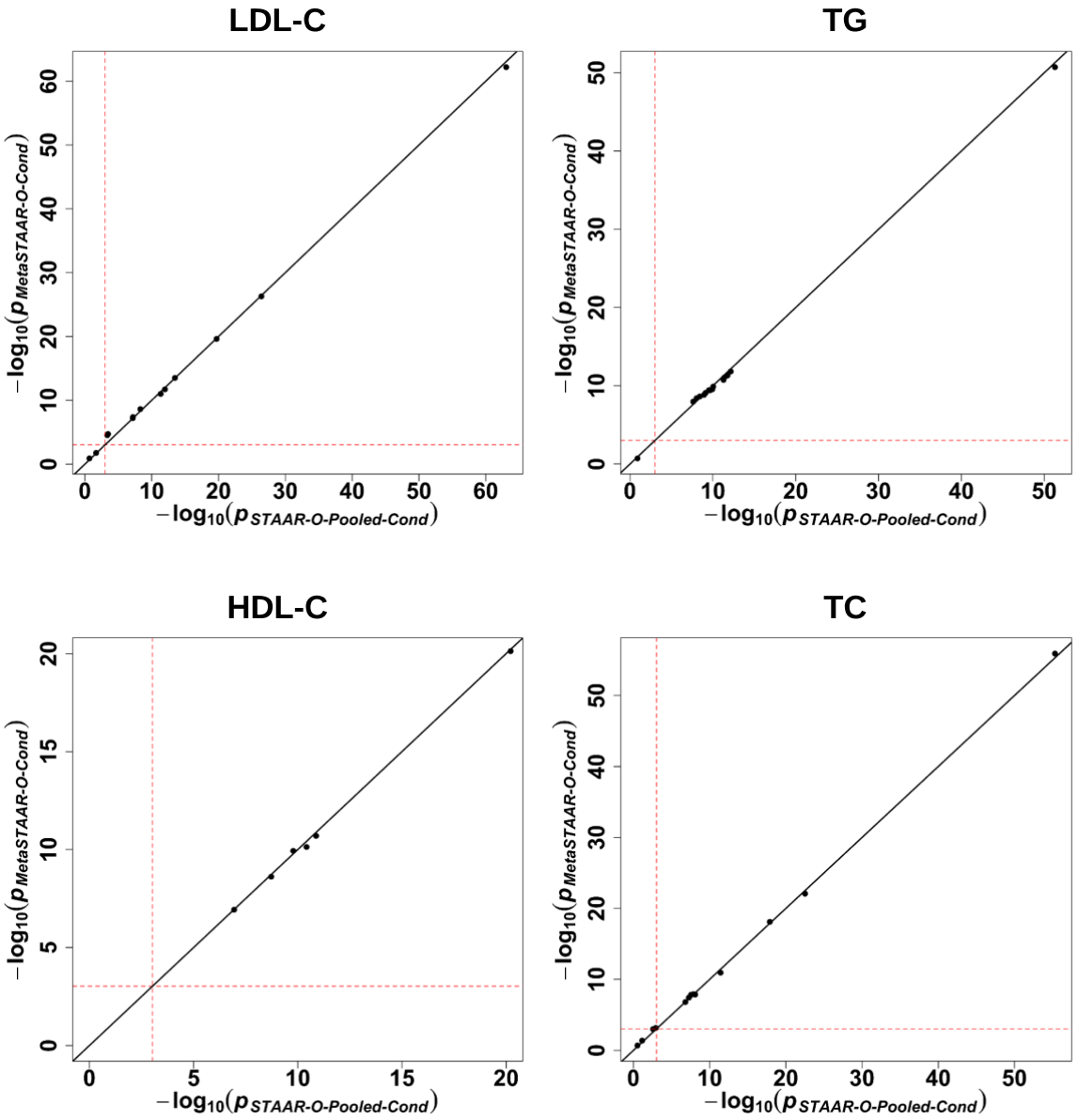
The horizontal line indicates a genome-wide MetaSTAAR-O *P* value threshold of 5.00×10^{-7}. Different symbols represent the MetaSTAAR-O *P* values of the gene using different functional categories (putative loss-of-function, missense, synonymous, promoter and enhancer). Promoter and enhancer are the promoter and the GeneHancer region with overlap of CAGE sites for a given gene, respectively (Methods). Four lipid traits were analyzed using MetaSTAAR: LDL-C, low-density lipoprotein cholesterol; HDL-C, high-density lipoprotein cholesterol; TG, triglycerides; TC, total cholesterol.

Supplementary Figure 3.3. Scatterplots comparing gene-centric unconditional meta-analysis *P* values from MetaSTAAR-O with STAAR-O from the joint analysis of pooled individual-level data (STAAR-O-Pooled) of lipid traits LDL-C, HDL-C, TG and TC using the TOPMed data (*n* = 30,138).



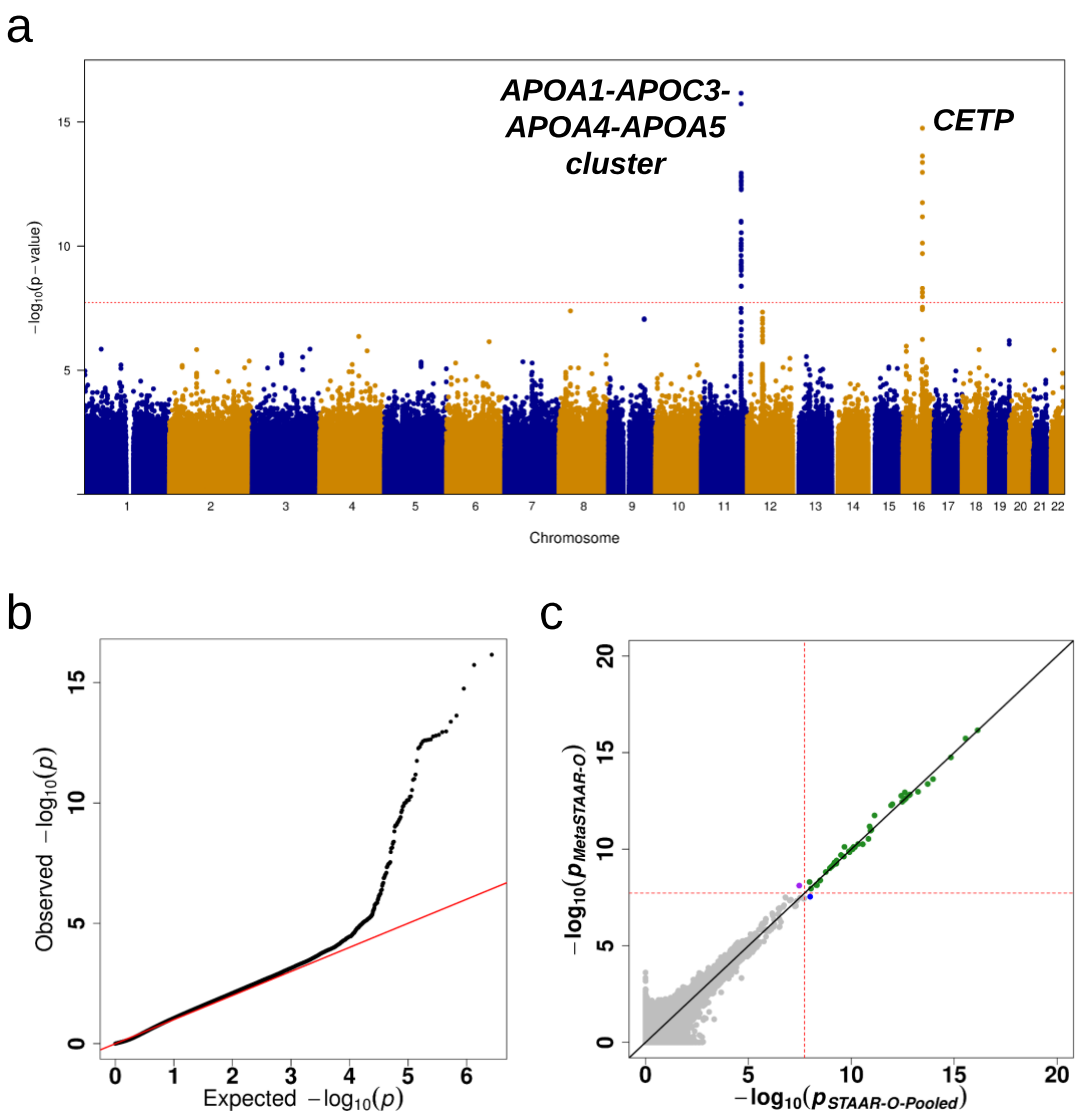
Each dot represents a functional category of a gene with x-axis label being the $-\log_{10}(P)$ of STAAR-O-Pooled and y-axis label being the $-\log_{10}(P)$ of MetaSTAAR-O (*n* = 30,138).

Supplementary Figure 3.4. Scatterplots comparing gene-centric conditional meta-analysis *P* values from MetaSTAAR-O (MetaSTAAR-O-Cond) with STAAR-O from the joint analysis of pooled individual-level data (STAAR-O-Pooled-Cond) of lipid traits LDL-C, HDL-C, TG and TC using the TOPMed data (*n* = 30,138).



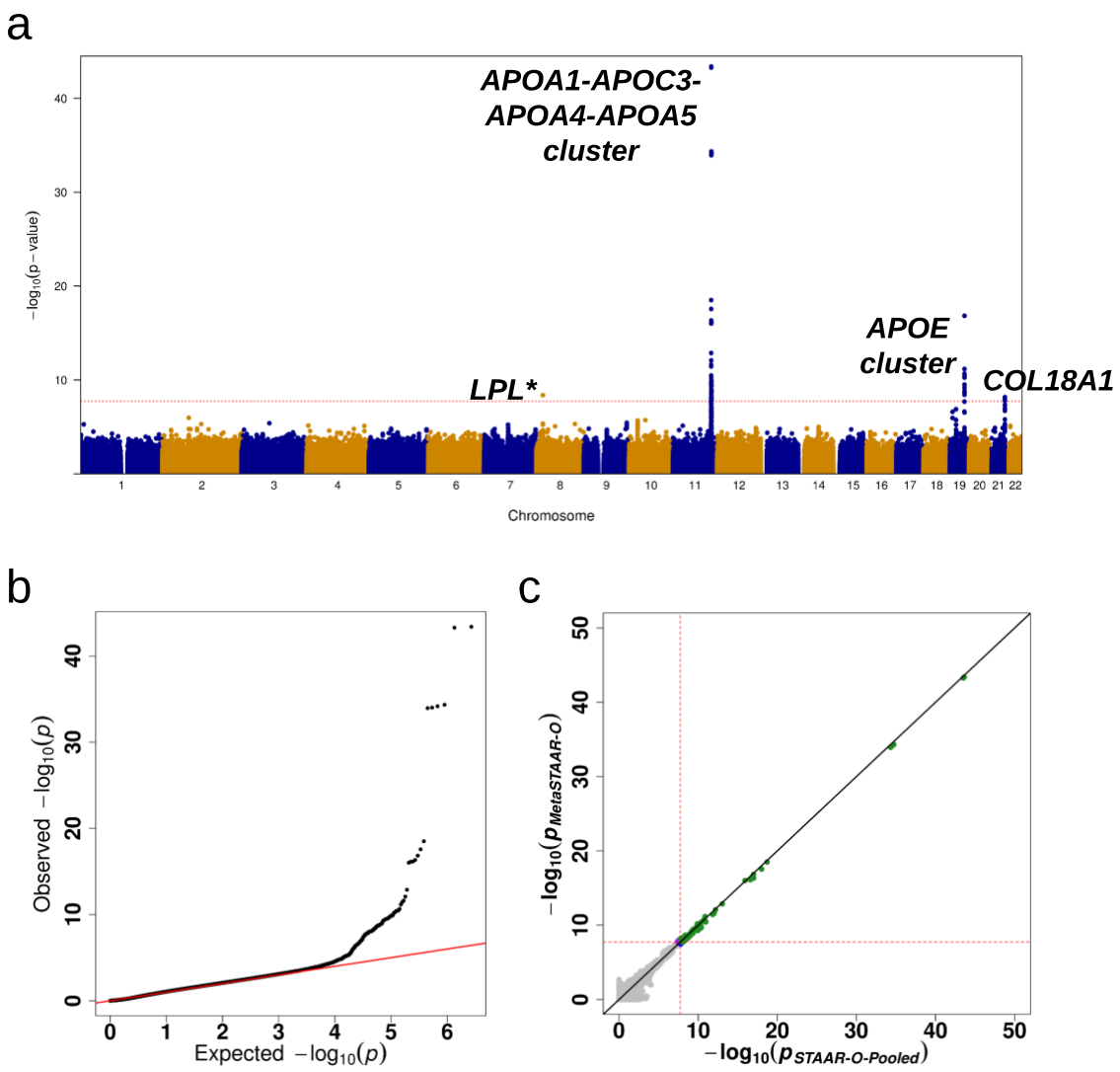
Significant associations in pooled analysis were used in the comparison (unconditional STAAR-O-Pooled $P < 5.00 \times 10^{-7}$). Each dot represents a functional category of a gene with x-axis label being the $-\log_{10}(P)$ of STAAR-O-Pooled and y-axis label being the $-\log_{10}(P)$ of MetaSTAAR-O (*n* = 30,138).

Supplementary Figure 3.5. Genetic region (2-kb sliding window) unconditional meta-analysis results of HDL-C using the TOPMed data.



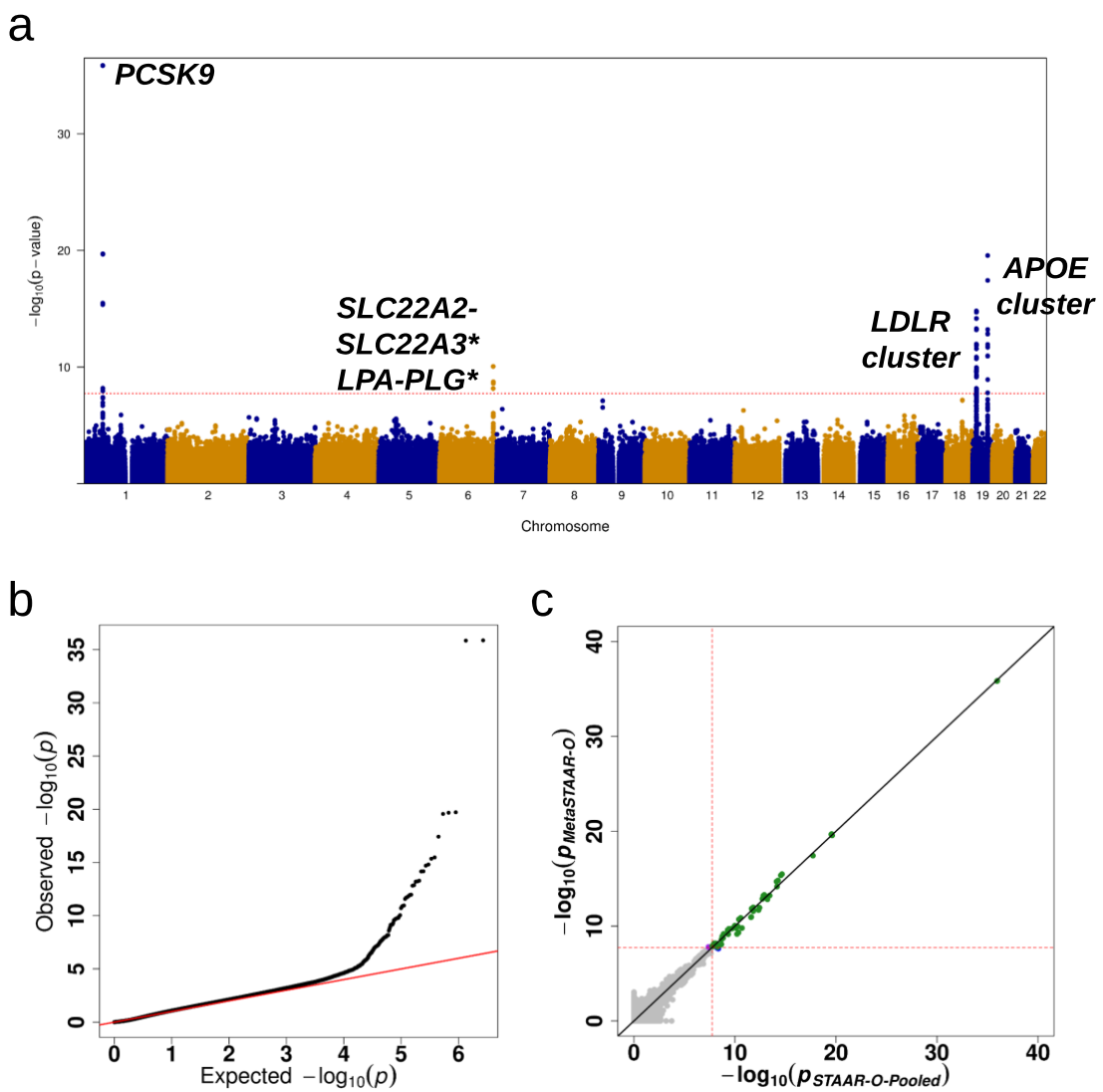
a, Manhattan plot showing the associations of 2.68 million 2-kb sliding windows for HDL-C (high-density lipoprotein cholesterol) versus $-\log_{10}(P)$ of MetaSTAAR-O. The horizontal line indicates a genome-wide *P* value threshold of 1.86×10^{-8} (*n* = 30,138). b, Quantile-quantile plot of 2-kb sliding window MetaSTAAR-O *P* values for HDL-C (*n* = 30,138). c, Scatterplot of *P* values for 2-kb sliding windows comparing MetaSTAAR-O with STAAR-O from the joint analysis of pooled individual-level data (STAAR-O-Pooled). Each dot represents a sliding window with x-axis label being the $-\log_{10}(P)$ of STAAR-O-Pooled and y-axis label being the $-\log_{10}(P)$ of MetaSTAAR-O (*n* = 30,138).

Supplementary Figure 3.6. Genetic region (2-kb sliding window) unconditional meta-analysis results of TG using the TOPMed data.



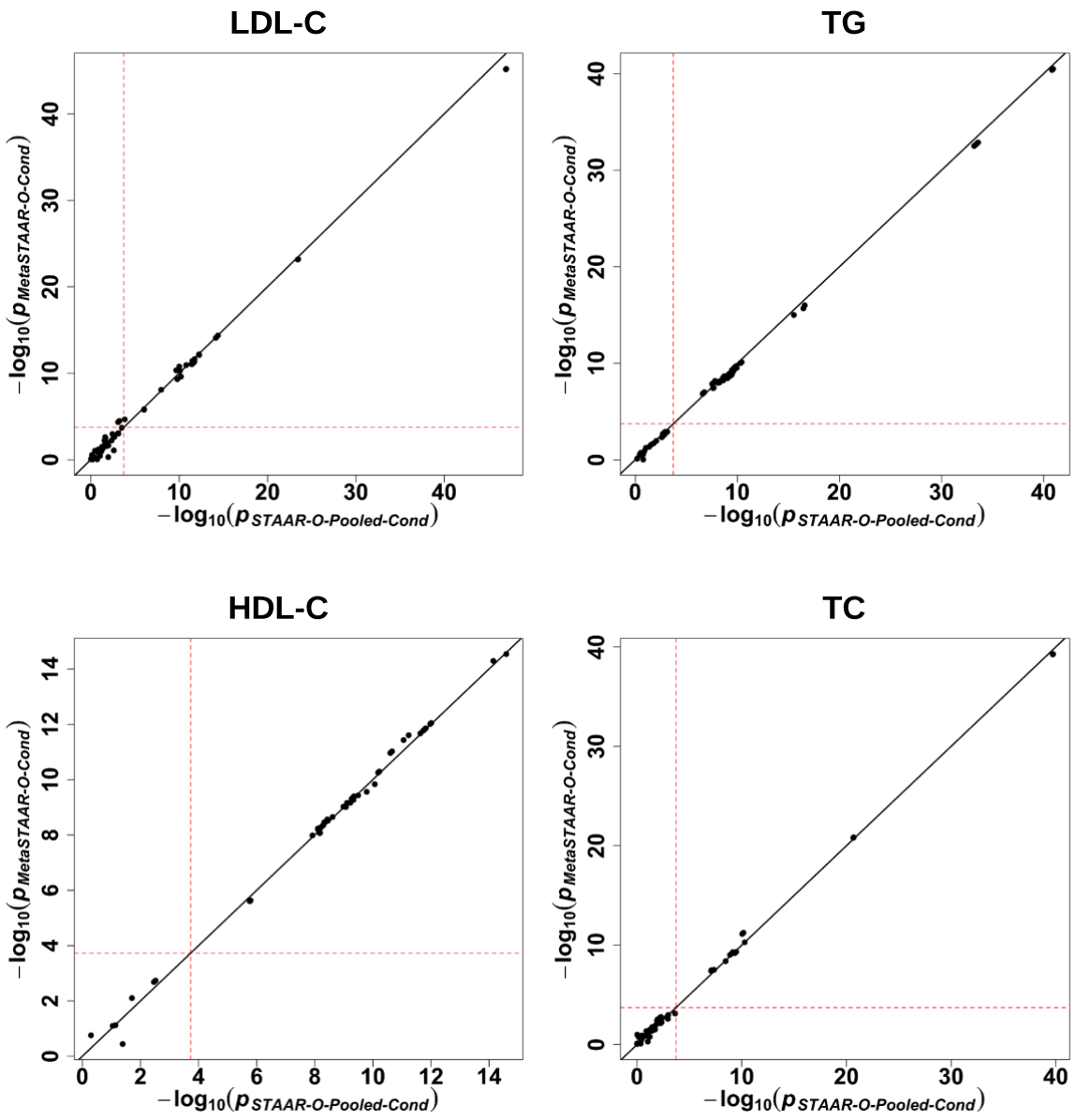
a, Manhattan plot showing the associations of 2.68 million 2-kb sliding windows for TG (triglycerides) versus $-\log_{10}(P)$ of MetaSTAAR-O. The horizontal line indicates a genome-wide *P* value threshold of 1.86×10^{-8} (*n* = 30,138). b, Quantile-quantile plot of 2-kb sliding window MetaSTAAR-O *P* values for TG (*n* = 30,138). c, Scatterplot of *P* values for 2-kb sliding windows comparing MetaSTAAR-O with STAAR-O from the joint analysis of pooled individual-level data (STAAR-O-Pooled). Each dot represents a sliding window with x-axis label being the $-\log_{10}(P)$ of STAAR-O-Pooled and y-axis label being the $-\log_{10}(P)$ of MetaSTAAR-O (*n* = 30,138). *Intergenic sliding window.

Supplementary Figure 3.7. Genetic region (2-kb sliding window) unconditional meta-analysis results of TC using the TOPMed data.



a, Manhattan plot showing the associations of 2.68 million 2-kb sliding windows for TC (total cholesterol) versus $-\log_{10}(P)$ of MetaSTAAR-O. The horizontal line indicates a genome-wide *P* value threshold of 1.86×10^{-8} (*n* = 30,138). b, Quantile-quantile plot of 2-kb sliding window MetaSTAAR-O *P* values for TC (*n* = 30,138). c, Scatterplot of *P* values for 2-kb sliding windows comparing MetaSTAAR-O with STAAR-O from the joint analysis of pooled individual-level data (STAAR-O-Pooled). Each dot represents a sliding window with x-axis label being the $-\log_{10}(P)$ of STAAR-O-Pooled and y-axis label being the $-\log_{10}(P)$ of MetaSTAAR-O (*n* = 30,138). *Intergenic sliding window.

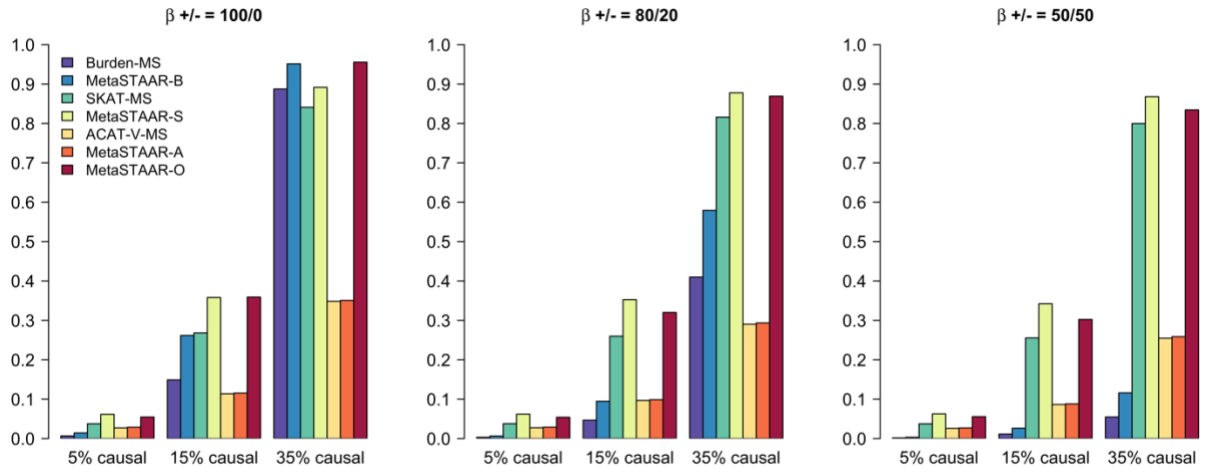
Supplementary Figure 3.8. Scatterplots comparing genetic region (2-kb sliding window) conditional meta-analysis *P* values from MetaSTAAR-O (MetaSTAAR-O-Cond) with STAAR-O from the joint analysis of pooled individual-level data (STAAR-O-Pooled-Cond) of lipid traits LDL-C, HDL-C, TG and TC using the TOPMed data (*n* = 30,138).



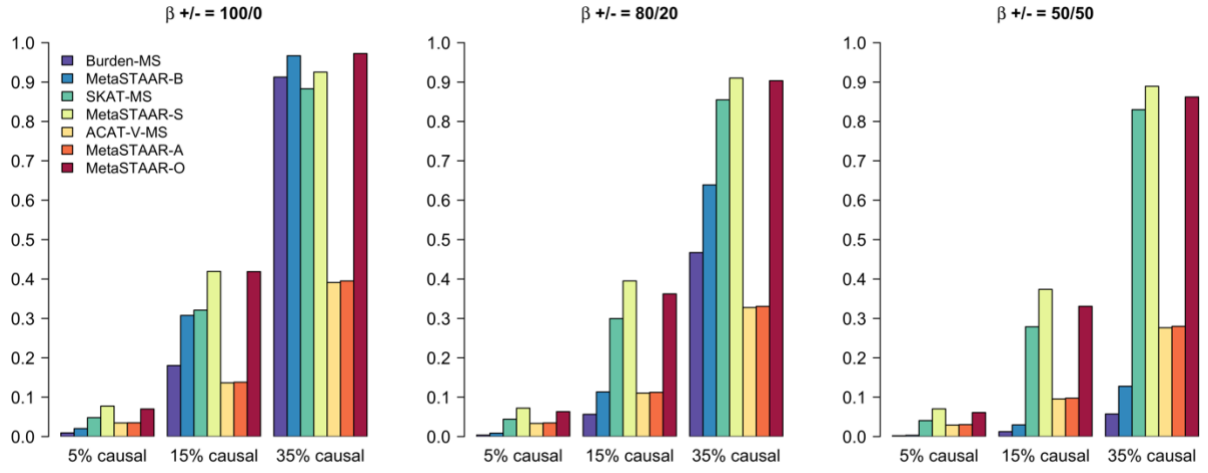
Significant associations in pooled analysis were used in the comparison (unconditional STAAR-O-Pooled $P < 1.86 \times 10^{-8}$). Each dot represents a functional category of a gene with x-axis label being the $-\log_{10}(P)$ of STAAR-O-Pooled and y-axis label being the $-\log_{10}(P)$ of MetaSTAAR-O (*n* = 30,138).

Supplementary Figure 3.9. Power comparisons of Burden, SKAT, ACAT-V and STAAR methods implemented in MetaSTAAR for quantitative and dichotomous traits.

Quantitative Trait



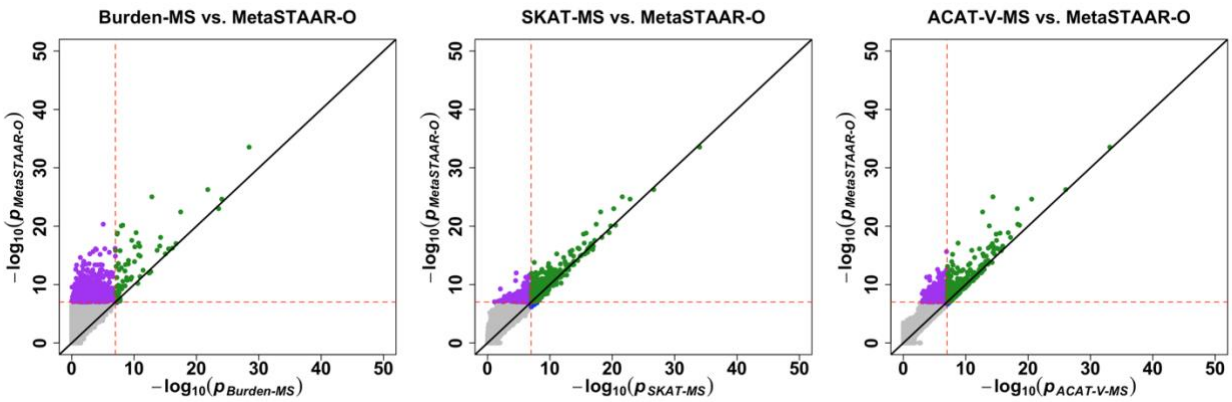
Dichotomous Trait



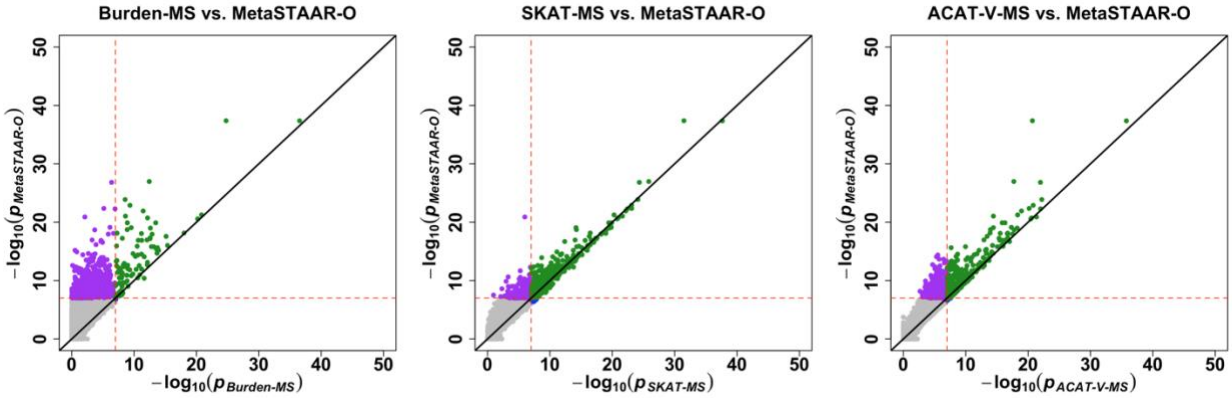
Meta-analysis of Burden, SKAT and ACAT-V implemented in MetaSTAAR are denoted by Burden-MS, SKAT-MS and ACAT-V-MS (MS for short). Meta-analysis of STAAR methods incorporating ten functional annotations are denoted by MetaSTAAR-B, MetaSTAAR-S, MetaSTAAR-A and MetaSTAAR-O. In each simulation replicate, a 2-kb region was randomly selected as the signal region. Within each signal region, variants were randomly generated to be causal based on the multivariate logistic model and on average there were 5%, 15% or 35% causal variants in the signal region. The effect sizes of causal variants were $\beta_j = c_0 |\log_{10} MAF_j|$, where c_0 was set to be 0.07 for quantitative traits and 0.11 for dichotomous traits, which gives an odds ratio of 1.6 for a variant with a MAF of 5×10^{-5}. Power was estimated as the proportion of the P values less than $\alpha = 10^{-7}$ based on 10^4 replicates. Five studies were included in the meta-analysis, each with a sample size of 10,000.

Supplementary Figure 3.10. Scatterplot of *P* values comparing MetaSTAAR-O to Burden-MS, SKAT-MS and ACAT-V-MS (MS is short of MetaSTAAR) for quantitative and dichotomous traits when 5% of rare variants are causal variants.

Quantitative Trait



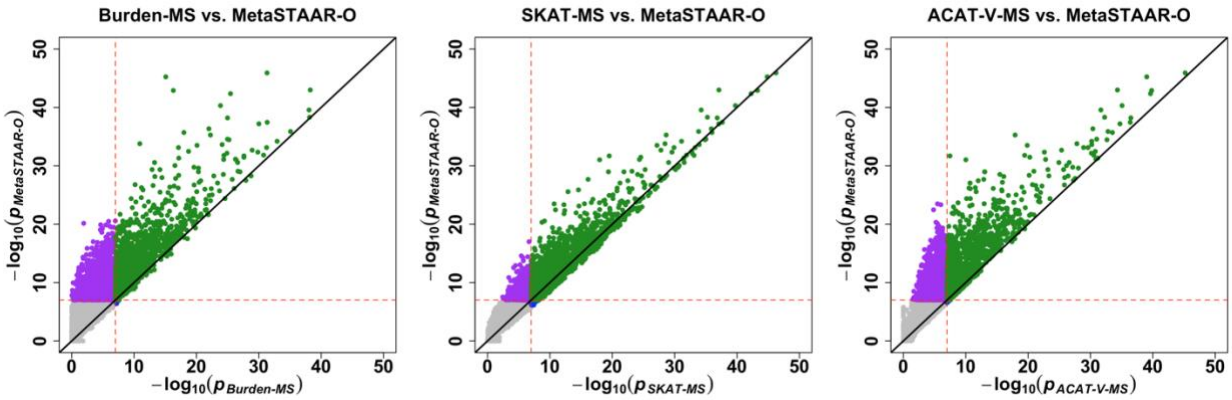
Dichotomous Trait



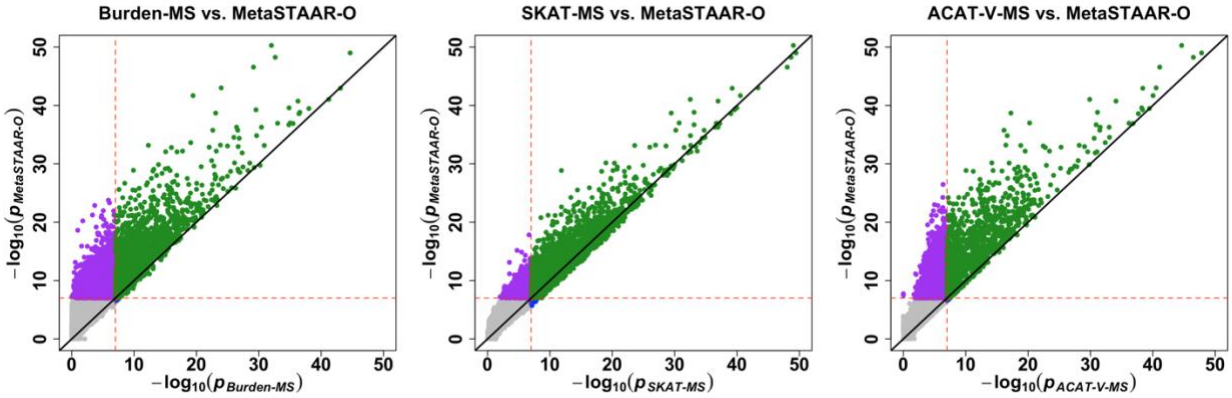
In each simulation replicate, a 2-kb region was randomly selected as the signal region. Within each signal region, variants were randomly generated to be causal based on the multivariate logistic model and on average there were 5% causal variants in the signal region. The effect sizes of causal variants were $\beta_j = c_0 |\log_{10} MAF_j|$. For quantitative traits, $c_0 = 0.07$; for dichotomous traits, $c_0 = 0.11$. All causal variants had positive effect sizes. Power was estimated as the proportion of the *P* values less than $\alpha = 10^{-7}$ based on 10^4 replicates. Five studies were included in the meta-analysis, each with a sample size of 10,000.

Supplementary Figure 3.11. Scatterplot of *P* values comparing MetaSTAAR-O to Burden-MS, SKAT-MS and ACAT-V-MS (MS is short of MetaSTAAR) for quantitative and dichotomous traits when 15% of rare variants are causal variants.

Quantitative Trait



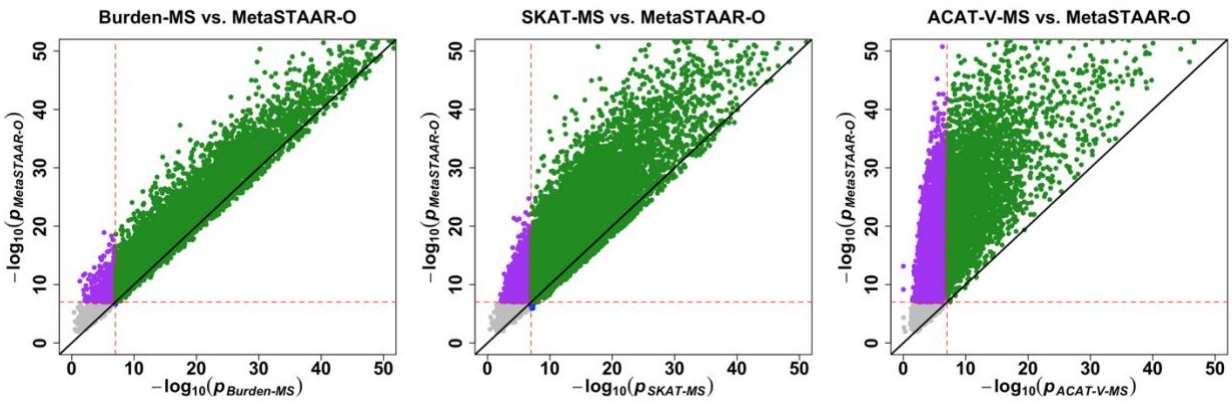
Dichotomous Trait



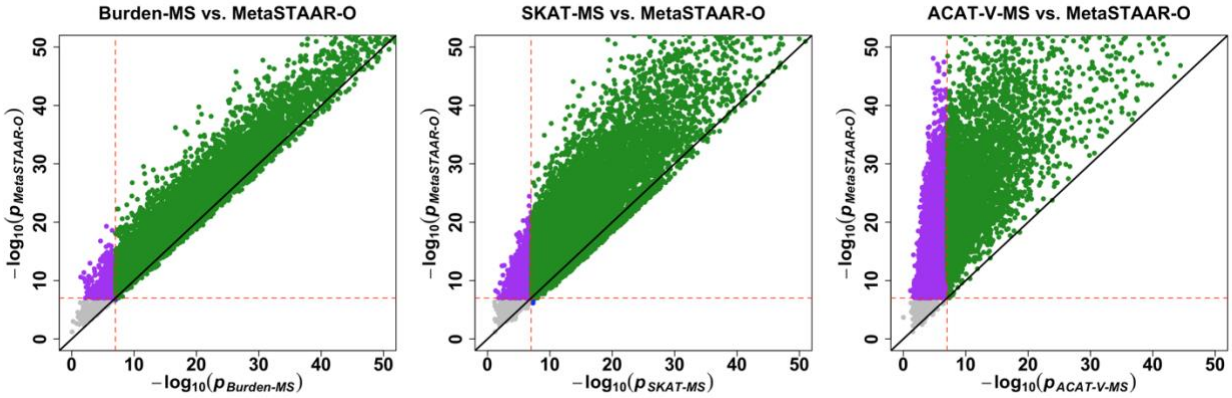
In each simulation replicate, a 2-kb region was randomly selected as the signal region. Within each signal region, variants were randomly generated to be causal based on the multivariate logistic model and on average there were 15% causal variants in the signal region. The effect sizes of causal variants were $\beta_j = c_0 |\log_{10} MAF_j|$. For quantitative traits, $c_0 = 0.07$; for dichotomous traits, $c_0 = 0.11$. All causal variants had positive effect sizes. Power was estimated as the proportion of the *P* values less than $\alpha = 10^{-7}$ based on 10^4 replicates. Five studies were included in the meta-analysis, each with a sample size of 10,000.

Supplementary Figure 3.12. Scatterplot of *P* values comparing MetaSTAAR-O to Burden-MS, SKAT-MS and ACAT-V-MS (MS is short of MetaSTAAR) for quantitative and dichotomous traits when 35% of rare variants are causal variants.

Quantitative Trait



Dichotomous Trait



In each simulation replicate, a 2-kb region was randomly selected as the signal region. Within each signal region, variants were randomly generated to be causal based on the multivariate logistic model and on average there were 35% causal variants in the signal region. The effect sizes of causal variants were $\beta_j = c_0 |\log_{10} MAF_j|$. For quantitative traits, $c_0 = 0.07$; for dichotomous traits, $c_0 = 0.11$. All causal variants had positive effect sizes. Power was estimated as the proportion of the *P* values less than $\alpha = 10^{-7}$ based on 10^4 replicates. Five studies were included in the meta-analysis, each with a sample size of 10,000.

Supplementary Note. Data simulation.

Type I error simulations

We performed extensive simulation studies to evaluate whether the proposed MetaSTAAR framework preserves the desired type I error rate. We generated continuous traits from a linear model defined as

$$Y_i = 0.5X_{1i} + 0.5X_{2i} + \epsilon_i,$$

where $X_{1i} \sim N(0,1), X_{2i} \sim \text{Bernoulli}(0.5)$, and $\epsilon_i \sim N(0,1)$. Dichotomous traits were generated from a logistic model defined as

$$\text{logit } P(Y_i = 1) = \alpha_0 + 0.5X_{1i} + 0.5X_{2i},$$

where X_{1i} and X_{2i} were defined the same as continuous traits and α_0 was determined to set the prevalence to 1%. In this setting, we used a balanced case-control design. We then generated five participating studies in the meta-analysis using the above model, each with a sample size of 10,000. For each study, we generated genotypes by simulating 20,000 sequences for 20 different regions each spanning 1 Mb. The data were generated to mimic the LD structure of an African American population by using the calibration coalescent model (COSI)(77). In each simulation replicate, 10 annotations were generated as $A_1, \ldots, A_{10}$ i.i.d. $N(0,1)$ for each variant, and we randomly selected 2-kb regions from these 20-Mb regions for type I error simulations. We applied MetaSTAAR-B, MetaSTAAR-S, MetaSTAAR-A and MetaSTAAR-O by incorporating MAFs and the 10 annotations and repeated the procedure with 10^9 replicates to examine the type I error rate at $\alpha = 10^{-5}, 10^{-6}, 10^{-7}$ levels.

Empirical power simulations

Next, we carried out simulation study under a variety of configurations to assess the power gain of MetaSTAAR-O by incorporating multiple functional annotations compared to the burden, SKAT, and ACAT-V tests implemented in MetaSTAAR. In each simulation replicate, we randomly selected 2-kb regions from a 1-Mb region for power simulations. We considered five participating studies in the meta-analysis, each with a sample size of 10,000. Then for each of the five participating study, we generated the phenotype in the meta-analysis using the following model. We first generated causal variants according to a logistic model defined as

$$\text{logit } P\big(c_j = 1\big) = \delta_0 + \delta_{l_1} A_{j,l_1} + \delta_{l_2} A_{j,l_2} + \delta_{l_3} A_{j,l_3} + \delta_{l_4} A_{j,l_4} + \delta_{l_5} A_{j,l_5},$$

where $\{l_1, \cdots, l_5\} \subset \{1, \cdots, 10\}$ were randomly sampled for each region. For different regions, causality of variants was allowed to be dependent on different sets of annotations. We set $\delta_{l_.} = \log(5)$ for all annotations and varied the proportions of causal variants in the signal region by setting $\delta_0 = \text{logit}(0.0015)$, $\text{logit}(0.015)$, and $\text{logit}(0.18)$ for averaging 5%, 15% and 35% causal variants in the signal region, respectively.

We generated continuous traits from a linear model given by

$$Y_i = 0.5X_{1i} + 0.5X_{2i} + \beta_1 G_{1j} + \cdots + \beta_s G_{sj} + \epsilon_i,$$

where $X_{1i}, X_{2i}, \epsilon_i$ were defined the same as the type I error simulations, $G_{1j}, \dots, G_{sj}$ were the genotypes of the s causal variants in the signal region, and $\beta_1, \dots, \beta_s$ were the corresponding effect sizes of causal variants. Dichotomous traits were generated from a logistic model given by

$$\text{logit}\, P(Y_i = 1) = 0.5X_{1i} + 0.5X_{2i} + \beta_1 G_{1j} + \cdots + \beta_s G_{sj},$$

where α_0, X_{1i}, X_{2i} were defined the same as the type I error simulations, $G_{1j}, \dots, G_{sj}$ were the genotypes of the s causal variants in the signal region, and $\beta_1, \dots, \beta_s$ were the corresponding log ORs of the s causal variants.

Under both settings, we model the effect sizes of causal variants using $\beta_j = \gamma_j = c_0 |\log_{10} MAF_j|$. The effect size of causal variant was therefore a decreasing function of MAF. For continuous traits, c_0 was set to be 0.07. For dichotomous traits, c_0 was set to be 0.11, which gives an odds ratio of 1.6 for a variant with MAF of 5×10^{-5}. For each setting, we additionally varied the proportions of causal variant effect size directions by setting 100%, 80%, and 50% variants to have positive effects. We applied MetaSTAAR-B, MetaSTAAR-S, MetaSTAAR-A, and MetaSTAAR-O using MAFs and all 10 annotations together with the burden-MS, SKAT-MS, and ACAT-V-MS (MS is short of MetaSTAAR), and repeated the procedure with 10^4 replicates to examine the powers at $\alpha = 10^{-7}$ level.