UCLA
UCLA Previously Published Works

Title

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

Permalink

https://escholarship.org/uc/item/1bb12264

Journal

Nature Genetics, 55(1)

ISSN

1061-4036

Authors

Li, Xihao
Quick, Corbin
Zhou, Hufeng
et al.

Publication Date

2023

DOI

10.1038/s41588-022-01225-6

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

Nat Genet. Author manuscript; available in PMC 2023 April 10.

Published in final edited form as:

Nat Genet. 2023 January ; 55(1): 154–164. doi:10.1038/s41588-022-01225-6.

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

A full list of authors and affiliations appears at the end of the article.

Abstract

Meta-analysis of whole-genome/exome sequencing (WGS/WES) studies provides an attractive solution to the problem of collecting large sample sizes for discovering rare variants associated with complex phenotypes. Existing rare variant meta-analysis approaches are not scalable to biobank-scale WGS data. Here we present MetaSTAAR, a powerful and resource-efficient rare variant meta-analysis framework for large-scale WGS/WES studies. MetaSTAAR accounts for relatedness and population structure, can analyze both quantitative and dichotomous traits, and boosts the power of rare variant tests by incorporating multiple variant functional annotations. Through meta-analysis of four lipid traits in 30,138 ancestrally diverse samples from 14 studies of the Trans-Omics for Precision Medicine (TOPMed) Program, we show that MetaSTAAR performs rare variant meta-analysis at scale and produces results comparable to using pooled data. Additionally, we identified several conditionally significant rare variant associations with lipid traits. We further demonstrate that MetaSTAAR is scalable to biobank-scale cohorts through meta-analysis of TOPMed WGS data and UK Biobank WES data of ~200,000 samples.

Ongoing large-scale whole-genome/exome sequencing (WGS/WES) studies, such as the Trans-Omics for Precision Medicine (TOPMed) Program of the National Heart, Lung and Blood Institute (NHLBI)[1], the Genome Sequencing Program (GSP) of the National Human Genome Research Institute, and UK Biobank WES Program[2,3], provide valuable insights into the genetic contributions of rare variants (minor allele frequency (MAF) < 1%) to many complex diseases and traits[4–7]. Because single-variant analyses are typically underpowered to identify rare variant associations[8], variant set tests have been proposed to jointly analyze

*Correspondence should be addressed to Z.L. (li@hsph.harvard.edu) and X. Lin (xlin@hsph.harvard.edu).

Author contributions

X. Li, C.Q., G.M.P., Z.L. and X. Lin designed the experiments. X. Li, C.Q., Z.L. and X. Lin performed the experiments. X. Li, C.Q., H.Z., S.M.G., Y.L., H.C., M.S.S., R.S., R.D., D.K.A., L.F.B., J.C.B., J.B., E.B., D.W.B., J.A.B., B.E.C., A.C., L.A.C., J.E.C., P.S.d.V., R.D., B.I.F., H.H.H.G., X.G., J.H., R.R.K., C.K., B.G.K., L.A.L., A.M., L.W.M., S.T.M., B.D.M., M.E.M., A.C.M., T.N., J.R.O., N.D.P., P.A.P., B.M.P., L.M.R., S.R., A.P.R., M.S.R., K.M.R., S.S.R., C.M.S, J.A.S., K.D.T., R.S.V., C.J.W., J.G.W., L.R.Y., W.Z., J.I.R., P.N., G.M.P., Z.L. and X. Lin acquired, analyzed or interpreted data. G.M.P., P.N. and the NHLBI TOPMed Lipids Working Group provided administrative, technical or material support. X. Li, Z.L. and X. Lin drafted the manuscript and revised it according to suggestions by the coauthors. All authors critically reviewed the manuscript, suggested revisions as needed and approved the final version.

Code availability

MetaSTAAR is implemented as an open source R package available at https://github.com/xihaoli/MetaSTAAR and https://content.sph.harvard.edu/xlin/software.html. Data analysis was performed in R (3.5.1). STAAR v0.9.6 and MetaSTAAR v0.9.6 were used in simulation and real data analysis and implemented as open-source R packages available at https://github.com/xihaoli/STAAR[52] and https://github.com/xihaoli/MetaSTAAR[53]. The scripts used to generate the results have been archived on *Zenodo* using https://doi.org/10.5281/zenodo.6668274[54]. RareMetal v4.15.1 (https://github.com/statgen/raremetal) and GMMAT v1.3.2 (https://cran.r-project.org/web/packages/GMMAT/index.html) were used for comparison. The assembled functional annotation data were downloaded from FAVOR using Wget (https://www.gnu.org/software/wget/wget.html).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

the effects of multiple rare variants to improve power[9–13]. In addition, pooling rare variants across multiple studies can boost association analysis power[14]. As such, meta-analysis of data from multiple WGS/WES studies provides a natural and cost-effective solution to augment sample size[15].

Compared to the joint analysis of pooled individual-level data, meta-analysis requires each participating study to share only summary statistics, which have much smaller file sizes than the individual-level data, and which protect the data privacy of study participants and reduce the challenges and the burden in sharing and integrating large subject-level data. Summary statistics are also increasingly available in public repositories, such as the GWAS catalog[16], based on which single-variant meta-analysis for GWAS can be readily performed. Compared to meta-analysis of each variant individually in GWAS, meta-analysis of rare variants in sequencing studies focuses on variant sets as analysis units to mitigate limited power of single-variant analysis. The statistical power of meta-analysis is (under plausible conditions) asymptotically equivalent to that of pooled analysis[17–19], making meta-analysis an essential tool for analyzing rare variant associations in large-scale WGS/WES studies, especially when individual-level data across studies cannot be shared.

Several methods are currently used for meta-analysis of rare variants in genetic association studies, including MetaSKAT, RareMetal and SMMAT[19–24]. MetaSKAT allows for linear and logistic models for continuous and binary traits, respectively, while RareMetal allows for linear mixed models. Neither MetaSKAT nor RareMetal provide logistic mixed models for binary traits[25], but SMMAT does provide logistic mixed models[24]. However, SMMAT requires pre-specifying variant sets at the study design stage and stores the summary statistics of those pre-specified variant sets. Furthermore, these existing rare variant meta-analysis methods do not allow the incorporation of multiple variant functional annotations. The STAAR method[26] boosts the power of rare variant association tests by incorporating multiple variant functional annotations and using ACAT[27] to combine the P values of the rare variant association test statistics calculated using different functional annotations as the weights. However, STAAR requires individual-level data and is not applicable for meta-analysis of studies through summary statistics.

For meta-analysis of common variants in GWAS, the only summary statistics needed are the individual variant test statistics and their variances, for each study. Meta-analysis of rare variants also requires storing the covariances of individual variant test statistics, which can be costly. Existing methods require $O(M^2)$ storage for a participating study summary statistics, where M is the total number of rare variants in a genetic region, a capacity which is not scalable for large cohort and biobank WGS/WES studies. For example, RareMetal would require more than 50 terabytes to store summary statistics of 250 million variants for the 30,000 individuals in TOPMed's current WGS data.

To address these issues, we propose Meta-analysis of variant-Set Test for Association using Annotation infoRmation (MetaSTAAR), a general framework for rare variant meta-analysis of large-scale WGS studies and biobanks with hundreds of millions of rare variants across the genome, by (1) compactly storing study-specific variant summary statistics and (2) dynamically incorporating multiple variant functional annotations. MetaSTAAR also

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

accounts for relatedness and population structure for both quantitative and dichotomous traits through fitting null Generalized Linear Mixed Models (GLMMs) using ancestry PCs and sparse genetic relatedness matrices (GRMs)[26,28].

By calculating and storing a new form of storage-efficient rare variant summary statistics within each study, including sparse weighted linkage disequilibrium (LD) matrices and low-rank matrices capturing the effects of covariates in the null GLMM, MetaSTAAR is computationally scalable and resource-efficient for rare variant meta-analysis of large-scale WGS data, requiring approximately $O(M)$ for storage of summary statistics. MetaSTAAR can be applied to any rare variant analysis unit, including gene-centric analysis by grouping variants into functional categories for each gene and genetic region analysis using sliding windows[26]. MetaSTAAR also enables approximate conditional analysis to, for example, identify rare variant 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. We applied MetaSTAAR to perform WGS rare Single Nucleotide Variant (SNV) meta-analysis of 30,138 ancestrally diverse participants from 14 studies of four quantitative lipid traits from the NHLBI TOPMed consortium: circulating low-density lipoprotein cholesterol (LDL-C), high-density lipoprotein cholesterol (HDL-C), triglycerides (TG), and total cholesterol (TC) levels. We show that MetaSTAAR is computationally scalable and resource-efficient for large-scale WGS rare variant meta-analysis, requiring over 100x less storage and computation time than existing methods such as MetaSKAT, RareMetal and SMMAT, all of which cannot currently handle large cohorts and biobanks with WGS data. Furthermore, MetaSTAAR provides comparable rare variant results to those using pooled data, and identifies several conditionally significant rare variant associations with lipids, after adjusting for known lipid-associated common variants. We also performed meta-analysis of the TOPMed lipid traits data and UK Biobank WES data (with ~200,000 samples) demonstrating that MetaSTAAR is scalable to large biobanks and cohorts.

Results

Overview of the methods

There are two main steps in the MetaSTAAR framework: (i) preparing variant summary statistics of each study using MetaSTAARWorker, and (ii) testing for associations between each variant set and phenotypes via meta-analysis using MetaSTAAR by combining the variant summary statistics across studies and incorporating multiple variant functional annotations[26] (Figure 1).

The first step, implemented in MetaSTAARWorker, generates and efficiently stores variant summary statistics for each study, including individual variant score statistics and their variance-covariance matrices, using sparse weighted LD matrices as well as low-rank matrices to capture the covariate effects. Specifically, for each participating study, MetaSTAARWorker first fits null GLMMs, accounting for relatedness and population structure[24,29] using linear and logistic mixed models for (respectively) quantitative and

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

dichotomous traits. It uses a sparse GRM and allows for study-specific covariates (for example, ancestral principal components) when fitting the null GLMM to ensure computational efficiency while also preserving accuracy[28]. MetaSTAARWorker then calculates individual variant score statistics and their estimated variances for all polymorphic variants in the study, which can be used to perform single-variant meta-analysis[30].

For meta-analysis of rare variants, one of the most time-consuming and resource-intensive components is generating and storing the variance-covariance matrices of individual variant score statistics by accounting for the LD structure among rare variants and covariate effects, such as ancestral principal components. To address this issue, MetaSTAARWorker decomposes the variance-covariance matrix of individual rare variant score statistics as the difference between the sparse weighted LD matrix and the cross product of a low-rank dense matrix, which captures the covariate effects (Methods). The weighted LD matrix is defined as the cross-product of the genotype matrix weighted by the inverse phenotypic variance-covariance matrix, which is the inverse variance-covariance matrix of phenotype for quantitative traits and the inverse variance-covariance matrix of the working vector of phenotype for dichotomous traits (Methods). By using the sparse genetic relatedness matrix, the inverse phenotypic variance-covariance matrix is sparse. Furthermore, given that the genotype matrix is also sparse for rare variants, the weighted LD matrix is sparse. Therefore, MetaSTAARWorker stores the weighted LD matrix in a sparse matrix format by taking advantage of sparse genotype dosages of rare variants and sparse GRM, and stores the low-rank dense projection matrix along with the individual variant summary statistics (Methods).

By storing these two matrices separately, MetaSTAARWorker only requires approximate $O(M)$ storage, which indicates that the storage, in practice, is approximately linear in sample size (Supplementary Figure 1). Compared with existing methods, such as MetaSKAT, RareMetal (RareMetalWorker) and SMMAT, which require $O(M^2)$ storage[19–21], MetaSTAARWorker can efficiently reduce the rare variant summary statistics storage, while still being able to reconstruct the variance-covariance matrix of rare variants. This efficient sparse matrix-based approach for efficiently storing rare variant summary statistics makes our approach feasible in rare variant meta-analysis of large-scale cohort and biobank WGS studies.

After collecting the rare variant summary statistics from each participating study, MetaSTAAR combines study-specific rare variant summary statistics into a merged variant list for any user-specified variant set. MetaSTAAR then uses the rare variant summary statistics from each study to calculate the aggregated score statistics and their variance-covariance matrices that correspond to all rare variants in the merged variant list. Since the vast majority of rare variants sequenced across the genome are extremely rare variants, a considerable number of rare variants are study-specific for WGS/WES meta-analysis (Supplementary Table 1). As such, if a genetic variant is monomorphic in a study, MetaSTAAR will set the variant score statistic and the corresponding row and column in the variance-covariance matrix to 0 for that study[19–21]. For a given variant set, the variance-covariance matrix of the score statistics is calculated using the sparse weighted LD matrix and the low-rank matrix of the covariate effects (Methods). With the aggregated score

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

statistics and the efficiently stored variance-covariance matrices, MetaSTAAR performs rare variant meta-analysis in large cohort and biobank WGS studies by incorporating multiple functional annotations in the weighting scheme using the STAAR framework. It then outputs the meta-analysis STAAR-O (MetaSTAAR-O) *P* value for the variant set[26]. In addition, MetaSTAAR permits rare variant association signals conditional on a set of known variants (Methods). It provides rare variant analysis results nearly identical to those from a pooled analysis.

Application to the TOPMed lipids WGS data

We applied MetaSTAAR to identify rare variant associations with four quantitative lipid traits (LDL-C, HDL-C, TG and TC) by meta-analysis of 14 study cohorts in the TOPMed Freeze 5 WGS data, which consists of 30,138 individuals (Supplementary Note). The sample sizes of the 14 studies range from 49 to 7,596 individuals. LDL-C and TC were adjusted for the usage of lipid-lowering medication[27] (Methods), and DNA samples from whole blood were sequenced at >30x target coverage. Sample- and variant-level quality control were performed separately for each participating study[1,27].

Race/ethnicity was defined using a combination of self-reported race/ethnicity from participant questionnaires and study recruitment information (Supplementary Note)[31]. Of the 30,138 multi-ethnic related samples, 8,114 (26.9%) were Black or African American individuals, 17,928 (59.5%) were White, 675 (2.2%) were Asian American, 2,318 (7.7%) were Hispanic/Latino American and 1,103 (3.7%) were 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 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 rare variants (MAF < 1%). The study-specific demographics, summaries of lipid levels and variant number distributions are given in Supplementary Table 1.

Computational and storage cost of MetaSTAARWorker

The key features of MetaSTAAR compared to other rare variant meta-analysis methods are presented in Table 1. They demonstrate that MetaSTAAR is the only rare variant meta-analysis method currently scalable for large WGS studies, which incorporates multiple variant functional annotations and accounts for relatedness and population structure for both quantitative and dichotomous traits.

We first 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 heteroskedastic linear mixed models with sparse GRMs plus ancestry-specific residual variance components (Methods). We then used MetaSTAARWorker to generate and store the score statistics and variances of all variants, and sparse weighted LD matrices of variants whose MAFs are below a user-specified threshold (Supplementary Table 2). MetaSTAARWorker required 300 CPU hours using a 2.10 GHz computing core with 12 GB memory on average to generate

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

the rare variant summary statistics for each TOPMed study and each trait. These calculations can be done in parallel. Each trait required 590 GB on average to store the rare variant summary statistics of all 14 cohorts (Supplementary Table 2).

We then considered multiple subsets of individuals from the TOPMed Freeze 5 WGS data with lipids and compared the computational performances of MetaSTAARWorker and the existing method RareMetalWorker. RareMetalWorker does not support heteroskedastic linear mixed models, which allow different residual variances in different subgroups of a given study. Therefore, we used a linear model to appropriately compare the two methods. MetaSTAARWorker requires hundreds of times less storage and computation time than RareMetalWorker (Table 2). In addition, the ratio of both storage space and computation time between RareMetalWorker and MetaSTAARWorker increases as the sample size increases, due to the difference in computation complexity and storage of the two methods (Table 1 and Supplementary Table 3). The estimated storage of RareMetalWorker is more than 50 terabytes for rare variant summary statistics of WGS data with 30,000 individuals. Hence, RareMetalWorker would require more than 2 petabytes to store the summary statistics of WGS data with 200,000 individuals.

To demonstrate the scalability of MetaSTAARWorker for biobank-size data, we also generated and stored the summary statistics of the four lipid traits using UK Biobank WES data with 190,476 related samples (Methods). MetaSTAARWorker required 300 CPU hours using a 2.10 GHz computing core with 12 GB memory on average to generate the rare variant summary statistics for each trait. Each trait required 2.68 GB on average to store the rare variant summary statistics (Supplementary Table 4).

Gene-centric meta-analysis of rare variants of TOPMed data

We applied MetaSTAAR-O to perform gene-centric meta-analysis of coding, promoter, and enhancer rare variants of genes with lipid traits in TOPMed. Rare variants (combined MAF < 1%) from five functional categories (masks) of each gene were aggregated, separately, and analyzed for each of the four lipid traits, including (i) putative loss-of-function (i.e., stop gain, stop loss and splice) rare variants, (ii) missense rare variants, (iii) synonymous rare variants, (iv) promoter rare variants overlapping Cap Analysis of Gene Expression (CAGE) sites[32], and (v) enhancer rare variants overlapping CAGE sites[33,34], where each mask was defined in Methods. We incorporated 10 annotation principal components (aPCs) including 1 liver-specific aPC[26,35], CADD[36], LINSIGHT[37], FATHMM-XF[38] and MetaSVM[39] (for missense rare variants only) along with the two MAF weights[8] in MetaSTAAR-O. Overall, the distributions of MetaSTAAR-O *P* values were well-calibrated for all four lipid phenotypes (Extended Data Figure 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 5 and Extended Data Figure 2). After conditioning on known lipids-associated variants[16,35], 40 out of the 53 associations remained significant at the Bonferroni-corrected threshold of $\alpha = 5.00 \times 10^{-7}$ (Table 3).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

We then compared the rare variant meta-analysis results of the 14 cohorts obtained from MetaSTAAR-O with the results from the joint analysis of pooled data using STAAR-O. All conditionally significant findings using STAAR-O analysis of the pooled data were detected by MetaSTAAR-O (Table 3). Furthermore, the $\log_{10}$-transformed P values from MetaSTAAR-O and STAAR-O unconditional and conditional pooled analysis were highly concordant (Pearson $r^2 > 0.99$) among significant and suggestive significant masks defined by various levels of unconditional P value thresholds ($\alpha = 2.5 \times 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}$) for each lipid phenotype (Supplementary Table 6, Extended Data Figure 3).

The computation time required for MetaSTAAR-O to perform WGS rare variant meta-analysis, including gene-centric analysis of three coding and two noncoding masks, on 30,138 related samples from 14 study cohorts using the TOPMed data was 500 CPU hours using a 2.10 GHz computing core with 12 GB memory on average for each lipid trait, which is also comparable to the pooled analysis.

Rare variant meta-analysis of TOPMed and UK Biobank data

We applied MetaSTAAR to meta-analyze the TOPMed data and UK Biobank WES data using the rare variant summary statistics generated by MetaSTAARWorker, with three coding masks (putative loss-of-function rare variants, missense rare variants, synonymous rare variants) for protein-coding genes (Methods) annotated using FAVOR[26].

Among the 31 conditionally significant coding masks detected by MetaSTAAR-O using 14 study cohorts in TOPMed Freeze 5 WGS data, 21 were replicated at the Bonferroni-corrected threshold of $\alpha = 5.00 \times 10^{-7}$ in conditional analysis using the UK Biobank WES data, and using meta-analysis of the two studies 20 of these 21 associations were at least one order of magnitude more significant than the TOPMed-only meta-analysis (Supplementary Table 7, Methods). Performing rare variant gene-centric meta-analysis of three coding masks using MetaSTAAR required 300 CPU hours using a 2.10 GHz computing core with 12 GB memory on average for each lipid trait.

Simulation studies

We performed simulation studies to evaluate the type I error rate and power of MetaSTAAR in 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 (Supplementary Note). For each study, genotypes were generated by simulating 20,000 sequences for 20 megabase (Mb) to mimic the LD structure of an African American population using the calibration coalescent model (COSI)[40]. We randomly selected 2-kilobase (kb) regions from the 20-Mb region as variant sets for testing in simulation studies.

Type I error rate simulations

For rare variant meta-analyses of both quantitative and dichotomous traits, we performed 10^9 simulations using MetaSTAAR and evaluated the empirical type I error rates for the meta-analyses of burden[9–11], SKAT[12], ACAT-V[13] and STAAR-O[26] tests at $\alpha = 10^{-5}, 10^{-6}, 10^{-7}$ (Supplementary Table 8). The results show that these four tests from MetaSTAAR provided

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

good control of the type I error rates for both continuous and dichotomous traits at all evaluated α levels.

Empirical power simulations

We then examined the empirical power of MetaSTAAR-O in a variety of configurations and compared it with existing methods. MAF and ten annotations were incorporated, of which five were informative and the other five were non-informative. 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 (an average of 5, 15 and 35%) in the signal region and allowed the causality of variants to be dependent on different sets of five annotations through a logistic model (Supplementary Note). The results show that across different proportions of effect size directions MetaSTAAR-O, when incorporating annotations, had more power to detect signal regions than use of the burden, SKAT, and ACAT-V tests implemented in MetaSTAAR (Extended Data Figure 4, Supplementary Figures 2–4). Our simulations indicate that MetaSTAAR-O has notably better power than existing rare variant meta-analysis methods, through its incorporation of multiple relevant annotations, and that MetaSTAAR's power advantage is robust to the presence of subsets of non-informative annotations.

Discussion

In this study, we propose MetaSTAAR as a computationally-scalable and resource-efficient framework to perform rare variant association meta-analysis of large WGS/WES studies incorporating multiple variant functional annotations and accounting for population structure and relatedness for both quantitative and dichotomous traits.

We first highlighted MetaSTAARWorker, the preliminary step of MetaSTAAR that generates and efficiently stores rare variant summary statistics, including variant score statistics and their variance-covariance matrices, for each participating study. Existing methods, such as MetaSKAT, RareMetal and SMMAT, store the full variance-covariance matrix of rare variants and require $O(M^2)$ storage, which is not scalable for large-scale WGS/WES studies. In contrast, MetaSTAARWorker stores the sparse weighted LD matrix and low-rank matrix that captures the covariate effects separately, and hence only requires approximately $O(M)$ storage without information loss. The computational complexity of generating summary statistics using MetaSTAARWorker is also substantially improved by taking advantage of sparse matrix computation (Table 2, Supplementary Tables 3 and 9). MetaSTAARWorker was benchmarked to improve both computation speed and storage requirements by more than 100-fold, using the 30,138 samples of the TOPMed lipids WGS data. With our method's linear growth with sample size, we expect a more than 700x gain for a sample size of 200,000 whole genomes, meaning that MetaSTAARWorker can facilitate large-scale WGS rare variant association meta-analysis.

The second feature of the MetaSTAAR framework is how it dynamically uses multiple functional annotations, improving power over existing rare variant meta-analysis methods. MetaSTAAR also provides conditional analysis to identify novel rare variant association signals independent of known variants. Our gene-centric meta-analyses of WGS rare

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

variants using MetaSTAAR-O, using TOPMed Freeze 5 data from 14 study cohorts, identified 40 conditionally-significant associations with lipid traits. These associations included *NPC1L1* missense rare variants and LDL-C; *CD36*, *APOC3*, *SCARB1* missense rare variants and HDL-C; and *NPC1L1* missense rare variants and TC, that were missed by meta-analytic burden, SKAT and ACAT-V tests that did not incorporate functional annotations (Supplementary Table 5).

The third feature of MetaSTAAR is how it can analyze variant sets in the genome without pre-defining them. MetaSTAAR generates and stores the rare variant summary statistics for all variants only once, and using them can perform meta-analysis of any variant set. This is particularly useful for WGS rare variant meta-analysis, especially for the noncoding genome, as it remains challenging to functionally interpret noncoding rare variants. Users may therefore want to define their own masks of interest after study-specific rare variant summary statistics are publicly released[41]. For example, in addition to the two noncoding masks defined in gene-centric analysis (promoters and enhancers of individual genes), another practical strategy to analyze noncoding rare variant associations is using sliding windows with fixed length[42] or dynamic windows with flexible locations and sizes[43]. MetaSTAAR could be adapted for such analysis.

As demonstrated through our TOPMed meta-analysis, for detecting rare variant association signals the MetaSTAAR framework delivers almost identical statistical power to joint analysis of pooled individual-level WGS data. This is achieved while bypassing cumbersome data sharing and harmonization across studies. In addition, MetaSTAAR generates phenotype-independent sparse weighted LD matrices for unrelated samples, hence further saving computational resources in phenome-wide association studies (Methods).

Overall, the proposed MetaSTAAR framework is fast, scalable, highly resource-efficient, and provides competitive levels of power for meta-analysis of large WGS/WES studies and biobanks. It is of particular appeal for analysis of datasets with hundreds of millions of variants measured on millions of multi-ethnic whole genomes, now being rapidly generated in studies, including TOPMed[1], GSP, UK Biobank[2,3], All of Us[44], and the Million Veteran Program[45].

Methods

Ethics statement

This study relied on analyses of genetic data from TOPMed cohorts. The study has been approved by the TOPMed Publications Committee, TOPMed Lipids Working Group and all the participating cohorts, including Framingham Heart Study (phs000974.v1.p1), Old Order Amish (phs000956.v1.p1), Jackson Heart Study (phs000964.v1.p1), Multi-Ethnic Study of Atherosclerosis (phs001416.v1.p1), 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

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Women's Health Initiative (phs001237), where the accession numbers are provided in parenthesis. The use of human genetics data from TOPMed cohorts was approved by the Harvard T.H. Chan School of Public Health IRB (IRB13–0353).

Key steps of MetaSTAAR

(1) Pre-fitting step by fitting null models using MetaSTAAR. Each participating study fits a generalized linear mixed model to account for population structure and relatedness using ancestry PCs and a sparse GRM, which can be calculated using standard approaches[28]. (2) Efficient generation and storage of summary statistics using MetaSTAARWorker. MetaSTAARWorker constructs "sparse weighted LD matrices" by taking advantage of sparse genotype dosages of rare variants and sparse GRM. By using "sparse weighted LD matrices", MetaSTAARWorker significantly overcomes the computation and resource limitation of rare variant meta-analysis of large-scale WGS data, and its storage and computation time are tens to hundreds of times smaller than existing methods[19,21,24] (Table 2 and Supplementary Table 9). (3) Rare variant meta-analysis using summary statistics by incorporating multi-faceted variant functional annotations using MetaSTAAR. MetaSTAAR allows for the incorporation of multiple variant functional annotations as weights in calculating rare variant meta-analysis test statistics using summary statistics, to increase the power of rare variant association tests. In contrast, existing methods do not allow for the incorporation of multiple variant functional annotations and may therefore be subject to power loss[19,21,24]. Although STAAR[26] incorporates multiple variant functional annotations in rare variant association tests, it requires individual level data. Specifically, for a set of variant functional annotations, e.g., annotation PCs[26], MetaSTAAR calculates meta-analysis rare variant association test statistics using summary statistics weighted by each functional annotation and combines the resulting annotation-weight-specific rare variant meta-analysis P values using ACAT[27], thereby providing a robust and powerful rare variant meta-analysis test.

Notation 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 $Y_k = (Y_{1,k}, \ldots, Y_{n_k,k})^T$ denote a continuous or dichotomous trait vector with mean $\hat{\boldsymbol{\mu}}_k = (\hat{\mu}_{1,k}, \ldots, \hat{\mu}_{n_k,k})^T$; $\boldsymbol{X}_k$ 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 $\hat{\boldsymbol{e}}_k = (\hat{e}_{1,k}, \ldots, \hat{e}_{n_k,k})^T$ 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)

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

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

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, $\boldsymbol{I}_{n_k}$ is a column vector of 1's with length n_k, $\boldsymbol{\alpha}_k = (\alpha_{1,k}, \ldots, \alpha_{q_k,k})^T$ is a vector of regression coefficients for $\boldsymbol{X}_k$. We calculate

$$\widehat{\boldsymbol{\Sigma}}_k = \widehat{\boldsymbol{R}}_k, \tag{2}$$

with $\widehat{\boldsymbol{R}}_k = \hat{\phi}_k \mathbf{I}_{n_k}$ for linear models, where $\hat{\phi}_k$ is an estimator of the residual variance ϕ_k, $\mathbf{I}_{n_k}$ is the identity matrix of dimension $n_k \times n_k$; and $\widehat{\boldsymbol{R}}_k = \text{diag}(1/(\hat{\mu}_{i,k}(1-\hat{\mu}_{i,k})))$ for logistic models, where $\hat{\mu}_{i,k}$ is the fitted value for individual i under the null model (1), and obtain $\hat{\boldsymbol{e}}_k = (\boldsymbol{Y}_k - \hat{\boldsymbol{\mu}}_k)/\hat{\phi}_k$.

When the data consist of related samples, we consider the following null Generalized Linear Mixed Models (GLMMs)[24,29,46]

$$g(\boldsymbol{\mu}_k) = \mathbf{1}_{n_k}\alpha_{0,k} + \boldsymbol{X}_k\boldsymbol{\alpha}_k + \boldsymbol{b}_k, \tag{3}$$

where the random effects $\boldsymbol{b}_k$ account for relatedness and remaining population structure unaccounted by ancestral PCs. We assume that $\boldsymbol{b}_k = (b_{1,k}, \ldots, b_{n_k,k})^T \sim N(\mathbf{0}, \theta_k\boldsymbol{\Phi}_k)$ with variance component θ_k and a family relatedness matrix $\boldsymbol{\Phi}_k$. If pedigree information is available, $\boldsymbol{\Phi}_k$ is a pedigree-based kinship matrix which is sparse by nature. However, in practice, pedigree information is often unavailable or incomplete. In this case, $\boldsymbol{\Phi}_k$ can be estimated using a sparse genetic relatedness matrix, which was justified in previous studies[28,47,48]. The remaining variables are defined in the same way as those in the GLM (1). We calculate

$$\widehat{\boldsymbol{\Sigma}}_k = \widehat{\boldsymbol{R}}_k + \hat{\theta}_k\boldsymbol{\Phi}_k, \tag{4}$$

with $\widehat{\boldsymbol{R}}_k = \hat{\phi}_k \mathbf{I}_{n_k}$ for linear mixed models; and $\widehat{\boldsymbol{R}}_k = \text{diag}(1/(\hat{\mu}_{i,k}(1-\hat{\mu}_{i,k})))$ for logistic mixed models, where $\hat{\mu}_{i,k}$ is the fitted value for individual i under the null model (3), and obtain $\hat{\boldsymbol{e}}_k = (\boldsymbol{Y}_k - \hat{\boldsymbol{\mu}}_k)/\hat{\phi}_k$. Note that we allow for heteroskedastic models with group-specific residual variance components in both linear models and linear mixed models for quantitative traits.

Rare variant summary statistics stored by MetaSTAARWorker

We describe the rare variant summary statistics that are stored by MetaSTAARWorker, including individual variant score statistics $\boldsymbol{U}_k$, sparse weighted LD matrices $\boldsymbol{G}_k^T\widehat{\boldsymbol{\Sigma}}_k^{-1}\boldsymbol{G}_k$ and low-rank matrices $\boldsymbol{\Lambda}_k$ that account for covariate effects, as defined below.

For the kth study, let $\boldsymbol{G}_k$ denote the genotype matrix of rare variants below a user-specified MAF threshold, and $\widehat{\boldsymbol{\Sigma}}_k$ is defined by (2) and (4) for GLM and GLMM, respectively. MetaSTAARWorker computes and shares a vector of score statistics $\boldsymbol{U}_k = \boldsymbol{G}_k^T\hat{\boldsymbol{e}}_k$, a sparse weighted LD matrix $\boldsymbol{G}_k^T\widehat{\boldsymbol{\Sigma}}_k^{-1}\boldsymbol{G}_k$, and 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. In WGS/WES data, more than 97% of the variants have MAF < 0.01 and around 46% of variants are singletons[1], hence the genotype matrix $\boldsymbol{G}_k$ is highly sparse. As the sparse weighted LD matrix $\boldsymbol{G}_k^T\widehat{\boldsymbol{\Sigma}}_k^{-1}\boldsymbol{G}_k$ is calculated with $\boldsymbol{G}_k$ in sparse format, its storage cost is benchmarked to be approximately $\mathcal{O}(M_k)$. $\boldsymbol{\Lambda}_k$ has the same number of

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

rows as $\boldsymbol{U}_k$ and the same number of columns as $\boldsymbol{X}_k$, which implies that $\boldsymbol{\Lambda}_k$ is a low-rank matrix and can be stored efficiently as $\mathcal{O}(M_k)$. Hence the variance-covariance matrix of $\boldsymbol{U}_k$ can be calculated by $\mathrm{Cov}(\boldsymbol{U}_k) = \boldsymbol{G}_k^T\widehat{\boldsymbol{\Sigma}}_k^{-1}\boldsymbol{G}_k - \boldsymbol{\Lambda}_k\boldsymbol{\Lambda}_k^T$ and shared efficiently as $\mathcal{O}(M_k)$. Note that MetaSKAT, RareMetalWorker and SMMAT directly store the variance-covariance $\mathrm{Cov}(\boldsymbol{U}_k)$ with dimension $M_k \times M_k$ in the dense matrix format. Hence the storage cost of these methods is $\mathcal{O}\left(M_k^2\right)$. The MAF threshold can be specified based on the relative sample size between studies to ensure that all rare variants in the pooled analysis are included in the meta-analysis. Additionally, for unrelated samples ($\widehat{\boldsymbol{\Sigma}}_k = \hat{\phi}_k\mathbf{I}_{n_k}$), the sparse weighted LD matrix reduces to $\boldsymbol{G}_k^T\widehat{\boldsymbol{\Sigma}}_k^{-1}\boldsymbol{G}_k = \hat{\phi}_k^{-1}\boldsymbol{G}_k^T\boldsymbol{G}_k$ which is phenotype-independent (up to a scaling constant $\hat{\phi}_k^{-1}$). Thus, MetaSTAARWorker could further save computation resources in phenome-wide association studies by only storing $\boldsymbol{G}_k^T\boldsymbol{G}_k$ under this setting.

To store and share the variance-covariance information of all rare variants across the genome, we computed the sparse weighted LD matrix for each consecutive region of 500 kb in length. In each region, any non-zero value in the 500-kb banded window (defined by a parallelogram with a side length of 500 kb) was stored (Figure 1b). The 500-kb banded windows guarantee the LD information of rare variants whose distances within 500 kb could be recovered. In practice, users can determine the bandwidth of the sparse weighted LD matrices to be shared.

Meta-analysis of rare variant association tests

We are interested in testing the association between rare variants in a variant set and phenotype via meta-analysis. For a given variant set, let M be the total number of rare variants, defined by the combined MAF in the meta-analysis of all K studies. In WGS/WES rare variant meta-analysis, some variants may often be observed in only a subset of studies but not in others (Supplementary Table 1). As such, for the kth study, let $\widetilde{\boldsymbol{U}}_k = \left(\widetilde{U}_{k,(1)}, \ldots, \widetilde{U}_{k,(M)}\right)^T$ denote the extended vector of $\boldsymbol{U}_k = \left(U_{k,1}, \ldots, U_{k,M_k}\right)^T$, where $\widetilde{U}_{k,(i)} = U_{ki}$ if variant i is observed in the kth study and $\widetilde{U}_{k,(i)} = 0$ otherwise. Let $\widetilde{\boldsymbol{\Lambda}}_k = \left(\widetilde{\boldsymbol{\Lambda}}_{k,(1)}, \ldots, \widetilde{\boldsymbol{\Lambda}}_{k,(M)}\right)^T$ denote the extended matrix of $\boldsymbol{\Lambda}_k = \left(\boldsymbol{\Lambda}_{k,1}, \ldots, \boldsymbol{\Lambda}_{k,M_k}\right)^T$, where $\widetilde{\boldsymbol{\Lambda}}_{k,(i)} = \boldsymbol{\Lambda}_{k,i}$ if variant i is observed in the kth study and $\widetilde{\boldsymbol{\Lambda}}_{k,(i)} = \boldsymbol{0}$ otherwise. Let $\widetilde{\boldsymbol{G}}_k^T\widehat{\boldsymbol{\Sigma}}_k^{-1}\widetilde{\boldsymbol{G}}_k = \left(\tilde{\sigma}_{(i),(j)}\right)_{M\times M}$ denote the extend matrix of $\boldsymbol{G}_k^T\widehat{\boldsymbol{\Sigma}}_k^{-1}\boldsymbol{G}_k = \left(\sigma_{ij}\right)_{M_k\times M_k}$, where $\tilde{\sigma}_{(i),(j)} = \sigma_{ij}$ if both variant i and variant j are observed in kth study and $\tilde{\sigma}_{(i),(j)} = 0$ otherwise. Note that a variant is removed from the meta-analysis if it fails the quality control in any of the studies. 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 statistic 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 the combined minor allele frequency, for the jth variant[8,26]. $Q_{Burden-MS}$ asymptotically follows a chi-square distribution with 1 degree

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

of freedom under the null hypothesis, and its P value can be obtained analytically while accounting for LD between variants[8,24].

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[8,24].

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

$$Q_{ACAT-V-MS} = \overline{w^2\mathrm{MAF}(1-\mathrm{MAF})}\tan((0.5-p_{(0)})\pi) + \sum_{j=1}^{M'} w_{(j)}^2 \mathrm{MAF}_{(j)}(1-\mathrm{MAF}_{(j)})\tan((0.5-p_{(j)})\pi),$$

where M' is the number of variants with the combined cumulative minor allele count (cMAC) greater than 10, $\mathrm{MAF}_{(j)}$ is the combined 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 combined cMAC > 10. $p_{(0)}$ is the burden test P value of extremely rare variants with combined cMAC ≤ 10 and $\overline{w^2\mathrm{MAF}(1-\mathrm{MAF})}$ is the average of the weights $w_{(j)}^2\mathrm{MAF}_{(j)}(1-\mathrm{MAF}_{(j)})$ among the extremely rare variants with combined cMAC ≤ 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[13].

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

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

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

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

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

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

$$Q_{ACAT-V-MS,l,(a_1,a_2)} = \overline{\hat{\pi}_{\cdot l} w^2_{(a_1,a_2)} \mathrm{MAF}(1-\mathrm{MAF})} \tan((0.5 - p_{(0),l})\pi) + \sum_{j=1}^{M'} \hat{\pi}_{(j),l} w^2_{j,(a_1,a_2)} \mathrm{MAF}_{(j)}(1-\mathrm{MAF}_{(j)}) \tan((0.5 - p_{(j)})\pi),$$

respectively, and their calculations use the study-specific rare variant summary statistics: the score statistics $\widetilde{\boldsymbol{U}}_k$, the sparse weighted LD matrix $\widetilde{\boldsymbol{G}}_k^T \widehat{\boldsymbol{\Sigma}}_k^{-1} \widetilde{\boldsymbol{G}}_k$ and the low-rank matrix $\widetilde{\boldsymbol{\Lambda}}_k$ that accounts for covariate effects. Here $\hat{\pi}_{(j),l} = \frac{\mathrm{rank}(A_{(j),l})}{m}$, where m is the number of variants across the whole genome, $w_{(j),(a_1,a_2)} = Beta(\mathrm{MAF}_{(j)}; a_1, a_2)$ with $(a_1, a_2) = (1, 25)$ or $(1, 1)$, and $\overline{\hat{\pi}_{\cdot l} w^2_{(a_1,a_2)} \mathrm{MAF}(1-\mathrm{MAF})}$ is the average of the weights $\hat{\pi}_{(j),l} w^2_{(j),(a_1,a_2)} \mathrm{MAF}_{(j)}(1-\mathrm{MAF}_{(j)})$ among the extremely rare variants with combined cMAC ≤ 10. The P value of $T_{MetaSTAAR-O}$ can 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.

Conditional meta-analysis using MetaSTAAR

We implemented conditional analysis in MetaSTAAR to perform meta-analysis of rare variant association tests adjusting for a given list of known variants[49]. MetaSTAARWorker generates the score statistics vector of the known variants and the variance-covariance matrix between rare variants in the variant set and known variants. Note that the known variants are not subject to the MAF threshold. Following the notations before, 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 are 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, where $\boldsymbol{P}_k = \widehat{\boldsymbol{\Sigma}}_k^{-1} - \widehat{\boldsymbol{\Sigma}}_k^{-1} \boldsymbol{X}_k \left(\boldsymbol{X}_k^T \widehat{\boldsymbol{\Sigma}}_k^{-1} \boldsymbol{X}_k\right)^{-1} \boldsymbol{X}_k^T \widehat{\boldsymbol{\Sigma}}_k^{-1}$. The covariance matrix between rare variants in the variant set and adjusted variants is given by $\mathrm{Cov}\left(\boldsymbol{U}_k, \boldsymbol{U}_k^{(c)}\right) = \boldsymbol{G}_k^T \boldsymbol{P}_k \boldsymbol{G}_k^{(c)}$. 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 $\mathrm{Cov}\left(\boldsymbol{U}_k, \boldsymbol{U}_k^{(c)}\right)$, which can be stored in MetaSTAARWorker.

To perform conditional meta-analysis of rare variant association tests, we calculate the adjusted score statistics vector

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

and hence

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

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

where $\widetilde{\boldsymbol{U}}_k$, $\mathrm{Cov}\left(\widetilde{\boldsymbol{U}}_k^{(c)}\right)$, and $\mathrm{Cov}\left(\widetilde{\boldsymbol{U}}_k,\widetilde{\boldsymbol{U}}_k^{(c)}\right)$ are the extended vector and matrix of $\boldsymbol{U}_k^{(c)}$, $\mathrm{Cov}\left(\boldsymbol{U}_k^{(c)}\right)$, and $\mathrm{Cov}\left(\boldsymbol{U}_k,\boldsymbol{U}_k^{(c)}\right)$ are defined in the same way as discussed before. The test statistics of conditional analysis of each test in MetaSTAAR are calculated in the same way as discussed before, with $\widetilde{\boldsymbol{U}}_{adj}$ and $\mathrm{Cov}\left(\widetilde{\boldsymbol{U}}_{adj}\right)$ instead of $\widetilde{\boldsymbol{U}}$ and $\mathrm{Cov}\left(\widetilde{\boldsymbol{U}}\right)$.

Lipid Traits

Conventionally measured plasma lipids, including total cholesterol, LDL-C, HDL-C, and triglycerides, were included for analysis. LDL-C was either calculated by the Friedewald equation when triglycerides were <400 mg/dl or directly measured. Given the average effect of statins, when statins were present, total cholesterol was adjusted by dividing by 0.8 and LDL-C by dividing by 0.7. Triglycerides were natural log transformed for analysis. Phenotypes were harmonized by each cohort and deposited into the dbGaP TOPMed Exchange Area (https://www.ncbi.nlm.nih.gov/gap).

Meta-analysis of lipid traits in the TOPMed WGS data

The TOPMed WGS data consist of multi-ethnic related samples[1]. Race/ethnicity was defined using a combination of self-reported race/ethnicity from participant questionnaires and study recruitment information (Supplementary Note)[31]. We applied MetaSTAAR to perform rare variant meta-analysis of four quantitative lipid traits (LDL-C, HDL-C, 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[27]. For each study, we first fit a 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[27].

We performed rank-based inverse normal transformation of the residuals and rescaled these residuals by the standard deviation of the original phenotype within each race/ethnicity-specific group. We then fit a heteroskedastic linear mixed model (HLMM) for the rank normalized residuals, adjusting for 10 ancestral principal components, ethnicity group indicators, and a variance component for empirically derived sparse 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 rare variant summary statistics by MetaSTAARWorker (Supplementary Table 2).

We next applied MetaSTAAR-O to perform rare variant meta-analysis based on the rare variant summary statistics of the 14 study cohorts, including gene-centric analysis using five variant functional categories (putative loss-of-function rare variants, missense rare variants, synonymous rare variants, promoter rare variants and enhancer rare variants) for each protein-coding gene. The WGS rare variant meta-analysis was performed using the R package MetaSTAAR (version 0.9.6, https://github.com/xihaoli/MetaSTAAR). The WGS

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

rare variant pooled analysis was performed using the R package STAAR (version 0.9.6, https://github.com/xihaoli/STAAR).

Meta-analysis of lipid traits in TOPMed and UK Biobank data

For TOPMed lipids data consisting of 30,138 samples, we generated the rare variant summary statistics using MetaSTAARWorker but treated all samples as one study cohort.

We downloaded VCF format files for WES data of 200,643 UK Biobank participants[3]. Quality control measures were performed in the following steps[2]. We first removed the variants with Hardy-Weinberg Equilibrium $P < 1 \times 10^{-15}$. Second, any SNV genotype with read depth less than seven reads (DP < 7) and indel genotype with DP < 15 was changed to a no-call. Third, any heterozygous genotype was changed to a no-call if any of the conditions are satisfied: (1) genotype quality < 20; (2) allele balance < 0.15 for SNV and allele balance < 0.20 for indel; (3) binomial test on allelic balance using allelic depth $P < 1 \times 10^{-3}$. We finally excluded the variants with more than 10% missing genotypes.

We harmonized four lipid traits (LDL-C, HDL-C, TG and TC) of the UK Biobank WES data. TC was adjusted by dividing the value by 0.8 among individuals reporting lipid lowering medication use or statin use at any time point. For LDL-C, we excluded individuals with LDL-C < 10 mg/dl or TG > 400 mg/ml. LDL-C was then adjusted by dividing the value by 0.7 among individuals reporting lipid lowering medication use or statin use at any time point. TG levels were natural logarithm transformed. A total of 185,712, 175,109, 190,262, and 190,415 individuals had data on LDL-C, HDL-C, TG, and TC, respectively.

We fit a linear mixed model adjusting 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. We next fit a linear mixed model (LMM) for the rank normalized residuals, adjusting for age, age^2, sex, and the 10 ancestral principal components, and a variance component for an empirically-derived sparse kinship matrix to account for population structure and relatedness. The output of LMM was then used to generate rare variant summary statistics, including score statistics and sparse weighted LD matrices, by MetaSTAARWorker.

We performed meta-analysis based on the rare variant summary statistics of the TOPMed data and the UK Biobank WES data, including gene-centric meta-analysis using three variant functional categories (putative loss-of-function rare variants, missense rare variants and synonymous rare variants) for each protein-coding gene using MetaSTAAR. We incorporated the two MAF weights[8] and MetaSVM[39] (for missense rare variants only) as annotations in MetaSTAAR-O. We additionally performed rare variant set gene-centric analysis of the three coding masks based on the rare variant summary statistics of the UK Biobank WES data. The rare variant meta-analysis was performed using the R package MetaSTAAR (version 0.9.6).

Genome build

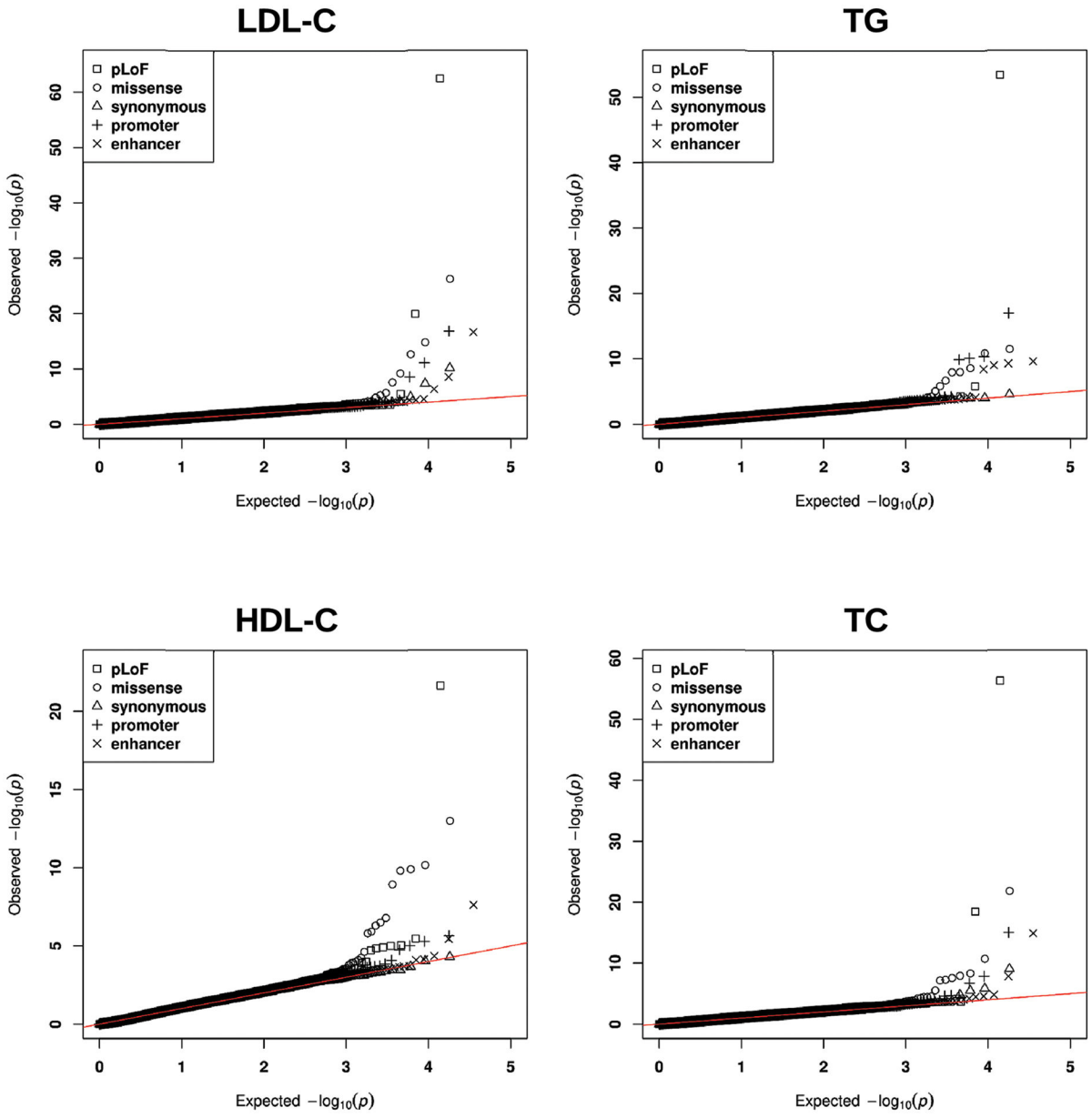
All genome coordinates are given in NCBI GRCh38/UCSC hg38.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Statistics and reproducibility

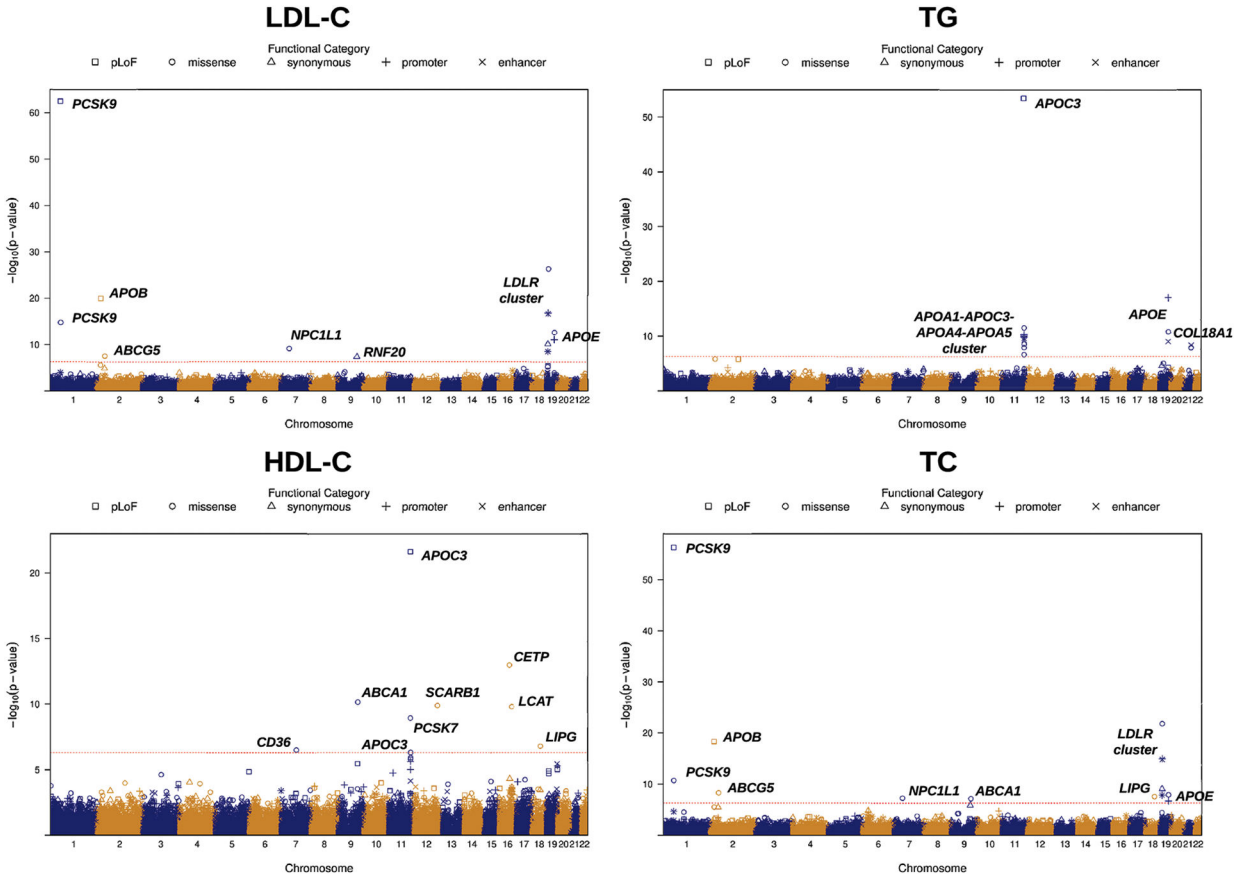
No statistical method was used to predetermine sample size. The meta-analysis consists of fourteen study cohorts of TOPMed Freeze 5 and had 30,138 samples with lipid traits. The UK Biobank whole-exome sequencing data had 190,476 samples with lipid traits. We did not use any study design that required randomization or blinding.

Extended Data



Extended Data Fig. 1 |. Quantile-quantile plots for gene-centric unconditional meta-analysis of lipid traits LDL-C, HDL-C, TG and TC using TOPMed WGS data (n = 30,138).
MetaSTAAR-O is a two-sided test. Different symbols represent the MetaSTAAR-O *P* values of different functional categories of individual genes (putative loss-of-function,

missense, synonymous, promoter and enhancer). The promoter and enhancer of a gene are the promoter and the GeneHancer region that overlap with CAGE sites for a given gene, respectively (Methods). Four lipid traits were analyzed using MetaSTAAR-O: LDL-C, low-density lipoprotein cholesterol; HDL-C, high-density lipoprotein cholesterol; TG, triglycerides; and TC, total cholesterol.

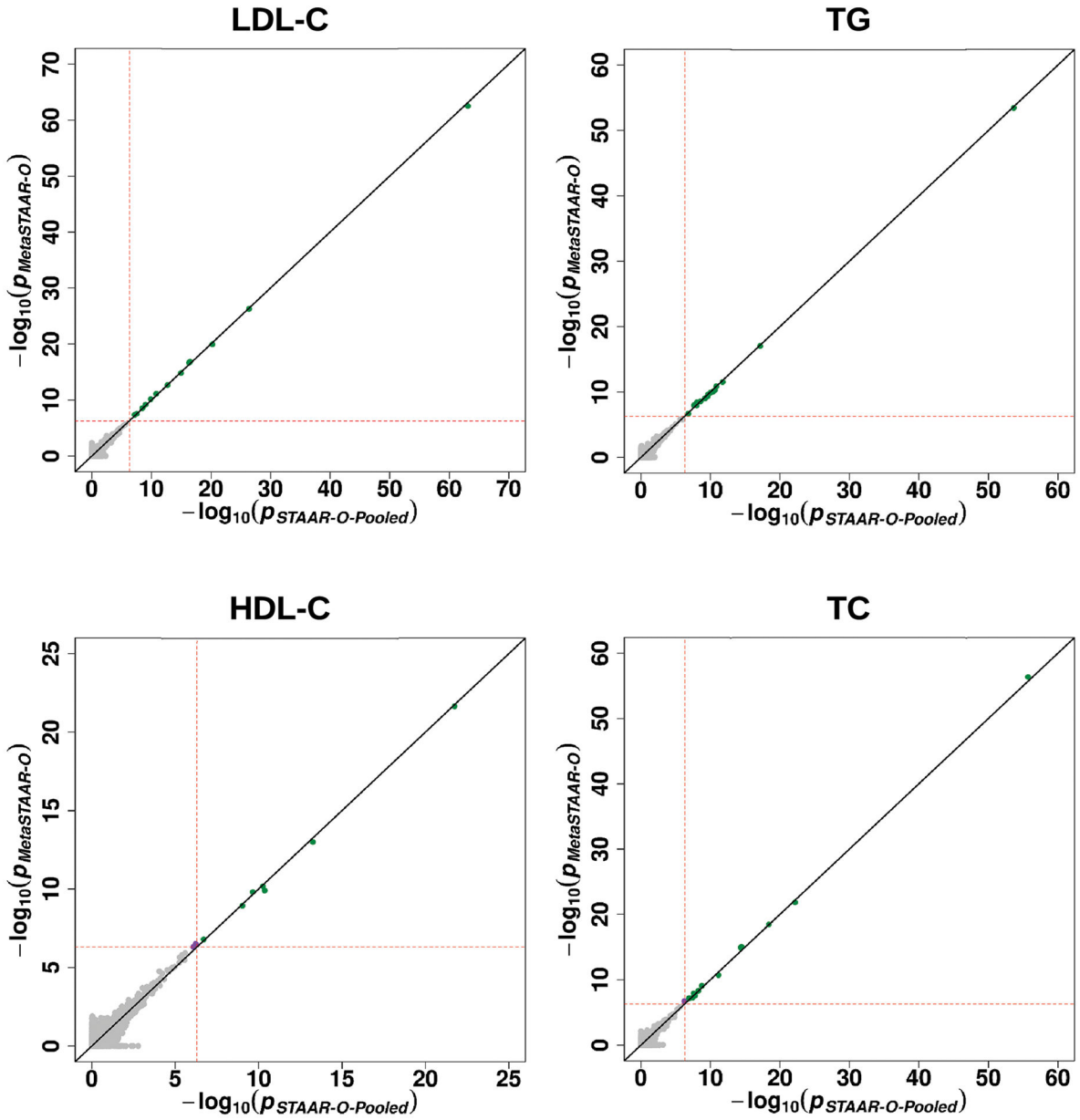


Extended Data Fig. 2 |. Manhattan plots for gene-centric unconditional meta-analysis of lipid traits LDL-C, HDL-C, TG and TC using TOPMed WGS data (n = 30,138).

The horizontal line indicates the genome-wide MetaSTAAR-O *P* value threshold of 5.00×10^{-7}. The significant threshold is defined by multiple comparisons using the Bonferroni correction ($0.05/(20,000 \times 5) = 5.00 \times 10^{-7}$). MetaSTAAR-O is a two-sided test. Different symbols represent the MetaSTAAR-O P values of different functional categories of individual genes (putative loss-of-function, missense, synonymous, promoter and enhancer). The promoter and enhancer of a gene are the promoter and the GeneHancer region that overlap with CAGE sites for a given gene, respectively (Methods). Four lipid traits were analyzed using MetaSTAAR-O: LDL-C, low-density lipoprotein cholesterol; HDL-C, high-density lipoprotein cholesterol; TG, triglycerides; TC, total cholesterol.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

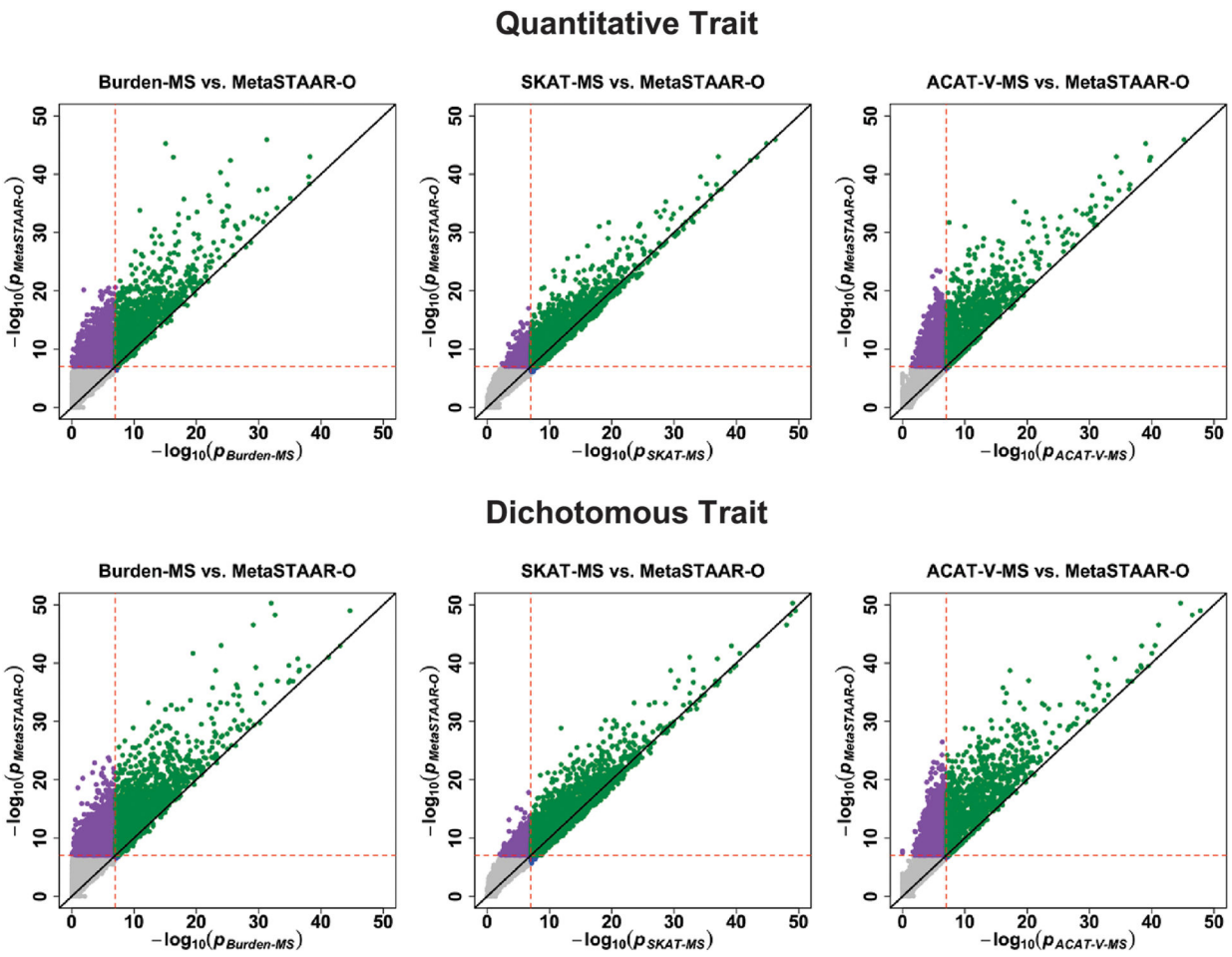
Author Manuscript Author Manuscript Author Manuscript Author Manuscript



Extended Data Fig. 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 TOPMed WGS 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 ($n = 30{,}138$). The horizontal and vertical lines indicate the genome-wide P value threshold of 5.00×10^{-7}. The significant threshold is defined by multiple comparisons using the Bonferroni correction ($0.05/(20{,}000 \times 5) = 5.00 \times 10^{-7}$). Both MetaSTAAR and STAAR are two-sided tests. LDL-C, low-density lipoprotein cholesterol; HDL-C, high-density lipoprotein cholesterol; TG, triglycerides; TC, total cholesterol.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript



Extended Data Fig. 4 |. Scatterplot of *P* values comparing MetaSTAAR-O to Burden-MS, SKAT-MS and ACAT-V-MS (MS is short for MetaSTAAR) for quantitative and dichotomous traits when 15% of rare variants are causal variants.

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 a multiple 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. Burden-MS, SKAT-MS, ACAT-V-MS and MetaSTAAR-O are two-sided tests. Five studies were included in meta-analysis, each with a sample size of 10,000.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Authors

Xihao Li[1], Corbin Quick[1], Hufeng Zhou[1], Sheila M. Gaynor[1], Yaowu Liu[2], Han Chen[3,4], Margaret Sunitha Selvaraj[5,6,7], Ryan Sun[8], Rounak Dey[1], Donna

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

K. Arnett[9], Lawrence F. Bielak[10], Joshua C. Bis[11], John Blangero[12], Eric Boerwinkle[3,13], Donald W. Bowden[14], Jennifer A. Brody[11], Brian E. Cade[6,15,16], Adolfo Correa[17], L. Adrienne Cupples[18,19], Joanne E. Curran[12], Paul S. de Vries[3], Ravindranath Duggirala[12], Barry I. Freedman[20], Harald H. H. Göring[12], Xiuqing Guo[21], Jeffrey Haessler[22], Rita R. Kalyani[23], Charles Kooperberg[22], Brian G. Kral[23], Leslie A. Lange[24], Ani Manichaikul[25], Lisa W. Martin[26], Stephen T. McGarvey[27], Braxton D. Mitchell[28,29], May E. Montasser[30], Alanna C. Morrison[3], Take Naseri[31], Jeffrey R. O'Connell[28], Nicholette D. Palmer[14], Patricia A. Peyser[10], Bruce M. Psaty[11,32,33], Laura M. Raffield[34], Susan Redline[15,16,35], Alexander P. Reiner[22,32], Muagututi'a Sefuiva Reupena[36], Kenneth M. Rice[37], Stephen S. Rich[25], Colleen M. Sitlani[11], Jennifer A. Smith[10,38], Kent D. Taylor[21], Ramachandran S. Vasan[19,39], Cristen J. Willer[40,41,42], James G. Wilson[43], Lisa R. Yanek[23], Wei Zhao[10],

NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium,

TOPMed Lipids Working Group,

Jerome I. Rotter[21], Pradeep Natarajan[5,6,7], Gina M. Peloso[18], Zilin Li[1,44,*], Xihong Lin[1,6,45,*]

Affiliations

[1]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA.

[2]School of Statistics, Southwestern University of Finance and Economics, Chengdu, Sichuan, China.

[3]Human Genetics Center, Department of Epidemiology, Human Genetics, and Environmental Sciences, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX, USA.

[4]Center for Precision Health, School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX, USA.

[5]Center for Genomic Medicine and Cardiovascular Research Center, Massachusetts General Hospital, Boston, MA, USA.

[6]Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA, USA.

[7]Department of Medicine, Harvard Medical School, Boston, MA, USA.

[8]Department of Biostatistics, University of Texas MD Anderson Cancer Center, Houston, TX, USA.

[9]University of Kentucky, College of Public Health, Lexington, KY, USA.

[10]Department of Epidemiology, School of Public Health, University of Michigan, Ann Arbor, MI, USA.

[11]Cardiovascular Health Research Unit, Department of Medicine, University of Washington, Seattle, WA, USA.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

[12]Department of Human Genetics and South Texas Diabetes and Obesity Institute, School of Medicine, The University of Texas Rio Grande Valley, Brownsville, TX, USA.

[13]Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX, USA.

[14]Department of Biochemistry, Wake Forest University School of Medicine, Winston-Salem, NC, USA.

[15]Division of Sleep and Circadian Disorders, Brigham and Women's Hospital, Boston, MA, USA.

[16]Division of Sleep Medicine, Harvard Medical School, Boston, MA, USA.

[17]Jackson Heart Study, Department of Medicine, University of Mississippi Medical Center, Jackson, MS, USA.

[18]Department of Biostatistics, Boston University School of Public Health, Boston, MA, USA.

[19]Framingham Heart Study, National Heart, Lung, and Blood Institute and Boston University, Framingham, MA, USA.

[20]Department of Internal Medicine, Nephrology, Wake Forest University School of Medicine, Winston-Salem, NC, USA.

[21]The Institute for Translational Genomics and Population Sciences, Department of Pediatrics, The Lundquist Institute for Biomedical Innovation at Harbor-UCLA Medical Center, Torrance, CA, USA.

[22]Division of Public Health Sciences, Fred Hutchinson Cancer Center, Seattle, WA, USA.

[23]GeneSTAR Research Program, Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, USA.

[24]Division of Biomedical Informatics and Personalized Medicine, Department of Medicine, University of Colorado Anschutz Medical Campus, Aurora, CO, USA.

[25]Center for Public Health Genomics, University of Virginia, Charlottesville, VA, USA.

[26]Division of Cardiology, George Washington School of Medicine and Health Sciences, Washington, DC, USA.

[27]Department of Epidemiology, International Health Institute, Department of Anthropology, Brown University, Providence, RI, USA.

[28]Department of Medicine, University of Maryland School of Medicine, Baltimore, MD, USA.

[29]Geriatrics Research and Education Clinical Center, Baltimore VA Medical Center, Baltimore, MD, USA.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

[30]Division of Endocrinology, Diabetes, and Nutrition, Program for Personalized and Genomic Medicine, University of Maryland School of Medicine, Baltimore, MD, USA.

[31]Ministry of Health, Government of Samoa, Apia, Samoa.

[32]Departments of Epidemiology, University of Washington, Seattle, WA, USA.

[33]Department of Health Systems and Population Health, University of Washington, Seattle, WA, USA.

[34]Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA.

[35]Division of Pulmonary, Critical Care, and Sleep Medicine, Beth Israel Deaconess Medical Center, Boston, MA, USA.

[36]Lutia I Puava Ae Mapu I Fagalele, Apia, Samoa.

[37]Department of Biostatistics, University of Washington, Seattle, WA, USA.

[38]Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, MI, USA.

[39]Department of Medicine, Boston University School of Medicine, Boston, MA, USA.

[40]Department of Internal Medicine, University of Michigan, Ann Arbor, MI, USA.

[41]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA.

[42]Department of Human Genetics, University of Michigan, Ann Arbor, MI, USA.

[43]Division of Cardiology, Beth Israel Deaconess Medical Center, Boston, MA, USA.

[44]Department of Biostatistics and Health Data Science, Indiana University School of Medicine, Indianapolis, IN, USA.

[45]Department of Statistics, Harvard University, Cambridge, MA, USA.

Acknowledgments

This work was supported by grants R35-CA197449, U19-CA203654, U01-HG012064, and U01-HG009088 (X. Lin), NHLBI BioData Catalyst Fellowship (Z.L.), R01-HL142711 and R01-HL127564 (P.N. and G.M.P.), 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, UL1-TR-001420, UL1-TR001881, DK063491, R01-HL071051, R01-HL071205, R01-HL071250, R01-HL071251, R01-HL071258, R01-HL071259, and UL1-RR033176 (J.I.R. and X.G.), R35-HL135824 (C.J.W.), U01-HL72518, HL087698, HL49762, HL59684, HL58625, HL071025, HL112064, NR0224103, and M01-RR000052 (to the Johns Hopkins General Clinical Research Center), NO1-HC-25195, HHSN268201500001I, 75N92019D00031, and R01-HL092577-06S1 (R.S.V. and L.A.C.), the Evans Medical Foundation and the Jay and Louis Coffman Endowment from the Department of Medicine, Boston University School of Medicine (R.S.V.), HHSN2682018000001I and U01-HL137162 (K.M.R.), R01-HL093093 and R01-HL133040 (S.T.M.), R35-HL135818, R01-HL113338, and HL436801 (S.R.), KL2TR002490 (L.M.R.), R01-HL92301, R01-HL67348, R01-NS058700, R01-AR48797, and R01-AG058921 (N.D.P. and D.W.B.), R01-DK071891 (N.D.P., B.I.F., and D.W.B.), M01-RR07122 and F32-HL085989 (to the General Clinical Research Center of the Wake Forest University School of Medicine), the American Diabetes Association, P60-AG10484 (to the Claude Pepper Older Americans Independence Center of Wake Forest University Health Sciences), U01-HL137181 (J.R.O.), HHSN268201600018C, HHSN268201600001C, HHSN268201600002C,

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

HHSN268201600003C, and HHSN268201600004C (C.K.), R01-HL113323, U01-DK085524, R01-HL045522, R01-MH078143, R01-MH078111, and R01-MH083824 (H.H.H.G., R.D., J.E.C., and J.B.), 18CDA34110116 from American Heart Association (P.S.d.V.), HHSN268201800010I, HHSN268201800011I, HHSN268201800012I, HHSN268201800013I, HHSN268201800014I, and HHSN268201800015I (A.C.), R01-HL153805, R03-HL154284 (B.E.C.), HHSN268201700001I, HHSN268201700002I, HHSN268201700003I, HHSN268201700005I, and HHSN268201700004I (E.B.), U01-HL072524, R01-HL104135-04S1, U01-HL054472, U01-HL054473, U01-HL054495, U01-HL054509, and R01-HL055673-18S1 (D.K.A.). Molecular data for the Trans-Omics in Precision Medicine (TOPMed) program was supported by the National Heart, Lung and Blood Institute (NHLBI). Core support including centralized genomic 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). Core support including phenotype harmonization, data management, sample-identity QC, and general program coordination were provided by the TOPMed Data Coordinating Center (R01HL-120393; U01HL-120393; contract HHSN268201800001I). We gratefully acknowledge the studies and participants who provided biological samples and data for TOPMed and UK Biobank. The full study specific acknowledgements and NHLBI BioData Catalyst acknowledgement are detailed in Supplementary Note.

Competing interests

S.M.G. is now an employee of Regeneron Genetics Center. For B.D.M.: The Amish Research Program receives partial support from Regeneron Pharmaceuticals. M.E.M. reports grant from Regeneron Pharmaceutical unrelated to the present work. B.M.P. serves on the Steering Committee of the Yale Open Data Access Project funded by Johnson & Johnson. L.M.R. is a consultant for the TOPMed Admistrative Coordinating Center (through Westat). For S.R.: Jazz Pharma, Eli Lilly, Apnimed, unrelated to the present work. The spouse of C.J.W. works at Regeneron Pharmaceuticals. P.N. reports investigator-initiated grants from Amgen, Apple, AstraZeneca, Boston Scientific, and Novartis, personal fees from Apple, AstraZeneca, Blackstone Life Sciences, Foresite Labs, Novartis, Roche / Genentech, is a co-founder of TenSixteen Bio, is a shareholder of geneXwell and TenSixteen Bio, and spousal employment at Vertex, all unrelated to the present work. X. Lin is a consultant of AbbVie Pharmaceuticals and Verily Life Sciences. The remaining authors declare no competing interests.

Data availability

This paper used the TOPMed Freeze 5 WGS data and lipids phenotype data. Genotype and phenotype data are both available in database of Genotypes and Phenotypes. The TOPMed WGS data were from the following fourteen study cohorts (accession numbers provided in parentheses): 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); 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, ancestry and phenotype summary statistics of these cohorts are given in Supplementary Table 1. The UK Biobank analyses were conducted using the UK Biobank resource under application 52008.

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); FANTOM5 CAGE (https://fantom.gsc.riken.jp/5/data); GeneCards (https://www.genecards.org; v4.7 for hg38); and Umap/Bismap (https://bismap.hoffmanlab.org; 'before March 2020' version). In addition, recombination rate and nucleotide diversity were obtained from Gazal et al[50]. The whole-genome individual functional annotation data was assembled from a variety of sources and

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

the computed annotation principal components are available at the Functional Annotation of Variant-Online Resource (FAVOR) site (https://favor.genohub.org) and the FAVOR database (https://doi.org/10.7910/DVN/1VGTJI)[51]. The tissue-specific functional annotations were downloaded from ENCODE (https://www.encodeproject.org/report/?type=Experiment).

NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium

Namiko Abe[46], Gonçalo Abecasis[47], Francois Aguet[48], Christine Albert[49], Laura Almasy[50], Alvaro Alonso[51], Seth Ament[52], Peter Anderson[53], Pramod Anugu[54], Deborah Applebaum-Bowden[55], Kristin Ardlie[48], Dan Arking[56], Allison Ashley-Koch[57], Stella Aslibekyan[58], Tim Assimes[59], Paul Auer[60], Dimitrios Avramopoulos[56], Najib Ayas[61], Adithya Balasubramanian[62], John Barnard[63], Kathleen Barnes[64], R. Graham Barr[65], Emily Barron-Casella[56], Lucas Barwick[66], Terri Beaty[56], Gerald Beck[67], Diane Becker[68], Lewis Becker[56], Rebecca Beer[69], Amber Beitelshees[52], Emelia Benjamin[70], Takis Benos[71], Marcos Bezerra[72], Thomas Blackwell[47], Nathan Blue[73], Russell Bowler[74], Ulrich Broeckel[75], Jai Broome[53], Deborah Brown[76], Karen Bunting[46], Esteban Burchard[77], Carlos Bustamante[78], Erin Buth[79], Jonathan Cardwell[80], Vincent Carey[81], Julie Carrier[82], April Carson[83], Cara Carty[84], Richard Casaburi[85], Juan P. Casas Romero[86], James Casella[56], Peter Castaldi[87], Mark Chaffin[48], Christy Chang[52], Yi-Cheng Chang[88], Daniel Chasman[89], Sameer Chavan[80], Bo-Juen Chen[46], Wei-Min Chen[90], Yii-Der Ida Chen[91], Michael Cho[81], Seung Hoan Choi[48], Lee-Ming Chuang[92], Mina Chung[93], Ren-Hua Chung[94], Clary Clish[95], Suzy Comhair[96], Matthew Conomos[79], Elaine Cornell[97], Carolyn Crandall[85], James Crapo[98], Jeffrey Curtis[99], Brian Custer[100], Coleen Damcott[52], Dawood Darbar[101], Sean David[102], Colleen Davis[53], Michelle Daya[80], Mariza de Andrade[103], Lisa de las Fuentes[104], Michael DeBaun[105], Ranjan Deka[106], Dawn DeMeo[81], Scott Devine[52], Huyen Dinh[62], Harsha Doddapaneni[107], Qing Duan[108], Shannon Dugan-Perez[62], Jon Peter Durda[97], Susan K. Dutcher[109], Charles Eaton[110], Lynette Ekunwe[54], Adel El Boueiz[111], Patrick Ellinor[112], Leslie Emery[53], Serpil Erzurum[63], Charles Farber[90], Jesse Farek[62], Tasha Fingerlin[113], Matthew Flickinger[47], Myriam Fornage[114], Nora Franceschini[115], Chris Frazar[53], Mao Fu[52], Stephanie M. Fullerton[53], Lucinda Fulton[116], Stacey Gabriel[48], Weiniu Gan[69], Shanshan Gao[80], Yan Gao[54], Margery Gass[117], Heather Geiger[118], Bruce Gelb[119], Mark Geraci[120], Soren Germer[46], Robert Gerszten[121], Auyon Ghosh[81], Richard Gibbs[62], Chris Gignoux[59], Mark Gladwin[71], David Glahn[122], Stephanie Gogarten[53], Da-Wei Gong[52], Sharon Graw[123], Kathryn J. Gray[124], Daniel Grine[80], Colin Gross[47], C. Charles Gu[116], Yue Guan[52], Namrata Gupta[48], Michael Hall[125], Yi Han[62], Patrick Hanly[126], Daniel Harris[127], Nicola L. Hawley[128], Jiang He[129], Ben Heavner[79], Susan Heckbert[130], Ryan Hernandez[77], David Herrington[131], Craig Hersh[132], Bertha Hidalgo[58], James Hixson[114], Brian Hobbs[81], John Hokanson[80], Elliott Hong[52], Karin Hoth[133], Chao (Agnes) Hsiung[134], Jianhong Hu[62], Yi-Jen Hung[135], Haley Huston[136], Chii Min Hwu[137], Marguerite Ryan Irvin[58], Rebecca Jackson[138], Deepti Jain[53], Cashell Jaquish[139], Jill Johnsen[140], Andrew Johnson[69], Craig Johnson[53], Rich Johnston[51], Kimberly Jones[56], Hyun Min Kang[141], Robert Kaplan[142], Sharon Kardia[47], Shannon Kelly[143], Eimear Kenny[119], Michael Kessler[52], Alyna Khan[53], Ziad Khan[62], Wonji Kim[144], John Kimoff[145], Greg Kinney[146], Barbara Konkle[147], Holly Kramer[148], Christoph Lange[149], Ethan Lange[80], Cathy Laurie[53], Cecelia Laurie[53], Meryl LeBoff[81], Jiwon Lee[81], Sandra Lee[62], Wen-Jane Lee[137], Jonathon LeFaive[47],

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

David Levine[53], Dan Levy[69], Joshua Lewis[52], Xiaohui Li[91], Yun Li[108], Henry Lin[91], Honghuang Lin[150], Simin Liu[151], Yongmei Liu[152], Yu Liu[153], Ruth J.F. Loos[154], Steven Lubitz[112], Kathryn Lunetta[155], James Luo[69], Ulysses Magalang[156], Michael Mahaney[157], Barry Make[56], Alisa Manning[158], JoAnn Manson[81], Melissa Marton[118], Susan Mathai[80], Rasika Mathias[56], Susanne May[79], Patrick McArdle[52], Merry-Lynn McDonald[159], Sean McFarland[144], Daniel McGoldrick[160], Caitlin McHugh[79], Becky McNeil[161], Hao Mei[54], James Meigs[162], Vipin Menon[62], Luisa Mestroni[123], Ginger Metcalf[62], Deborah A. Meyers[163], Emmanuel Mignot[164], Julie Mikulla[69], Nancy Min[54], Mollie Minear[165], Ryan L. Minster[71], Matt Moll[87], Zeineen Momin[62], Courtney Montgomery[166], Donna Muzny[62], Josyf C. Mychaleckyj[90], Girish Nadkarni[119], Rakhi Naik[56], Sergei Nekhai[167], Sarah C. Nelson[79], Bonnie Neltner[80], Caitlin Nessner[62], Deborah Nickerson[168], Osuji Nkechinyere[62], Kari North[108], Tim O'Connor[52], Heather Ochs-Balcom[169], Geoffrey Okwuonu[62], Allan Pack[170], David T. Paik[171], James Pankow[172], George Papanicolaou[69], Cora Parker[173], Juan Manuel Peralta[174], Marco Perez[59], James Perry[52], Ulrike Peters[175], Lawrence S Phillips[51], Jacob Pleiness[47], Toni Pollin[52], Wendy Post[176], Julia Powers Becker[177], Meher Preethi Boorgula[80], Michael Preuss[119], Pankaj Qasba[69], Dandi Qiao[81], Zhaohui Qin[51], Nicholas Rafaels[178], Mahitha Rajendran[62], D.C. Rao[116], Laura Rasmussen-Torvik[179], Aakrosh Ratan[90], Robert Reed[52], Catherine Reeves[180], Elizabeth Regan[98], Rebecca Robillard[181], Nicolas Robine[118], Dan Roden[182], Carolina Roselli[48], Ingo Ruczinski[56], Alexi Runnels[118], Pamela Russell[80], Sarah Ruuska[136], Kathleen Ryan[52], Ester Cerdeira Sabino[183], Danish Saleheen[184], Shabnam Salimi[185], Sejal Salvi[62], Steven Salzberg[56], Kevin Sandow[186], Vijay G. Sankaran[187], Jireh Santibanez[62], Karen Schwander[116], David Schwartz[80], Frank Sciurba[71], Christine Seidman[188], Jonathan Seidman[189], Frédéric Sériès[190], Vivien Sheehan[191], Stephanie L. Sherman[192], Amol Shetty[52], Aniket Shetty[80], Wayne Hui-Heng Sheu[137], M. Benjamin Shoemaker[193], Brian Silver[194], Edwin Silverman[81], Robert Skomro[195], Albert Vernon Smith[196], Josh Smith[53], Nicholas Smith[130], Tanja Smith[46], Sylvia Smoller[142], Beverly Snively[197], Michael Snyder[59], Tamar Sofer[81], Nona Sotoodehnia[53], Adrienne M. Stilp[53], Garrett Storm[198], Elizabeth Streeten[52], Jessica Lasky Su[199], Yun Ju Sung[116], Jody Sylvia[81], Adam Szpiro[53], Daniel Taliun[47], Hua Tang[200], Margaret Taub[56], Matthew Taylor[123], Simeon Taylor[52], Marilyn Telen[57], Timothy A. Thornton[53], Machiko Threlkeld[201], Lesley Tinker[202], David Tirschwell[53], Sarah Tishkoff[203], Hemant Tiwari[204], Catherine Tong[205], Russell Tracy[206], Michael Tsai[172], Dhananjay Vaidya[56], David Van Den Berg[207], Peter VandeHaar[47], Scott Vrieze[172], Tarik Walker[80], Robert Wallace[133], Avram Walts[80], Fei Fei Wang[53], Heming Wang[208], Jiongming Wang[209], Karol Watson[85], Jennifer Watt[62], Daniel E. Weeks[210], Joshua Weinstock[141], Bruce Weir[53], Scott T. Weiss[211], Lu-Chen Weng[112], Jennifer Wessel[212], Kayleen Williams[79], L. Keoki Williams[213], Carla Wilson[81], Lara Winterkorn[118], Quenna Wong[53], Joseph Wu[171], Huichun Xu[52], Ivana Yang[80], Ketian Yu[47], Seyedeh Maryam Zekavat[48], Yingze Zhang[214], Snow Xueyan Zhao[98], Xiaofeng Zhu[215], Elad Ziv[216], Michael Zody[46], Sebastian Zoellner[47]

46 - New York Genome Center, New York, New York, 10013, US; 47 - University of Michigan, Ann Arbor, Michigan, 48109, US; 48 - Broad Institute, Cambridge, Massachusetts, 2142, US; 49 - Cedars Sinai, Boston, Massachusetts, 2114, US; 50 - Children's Hospital of Philadelphia, University of Pennsylvania, Philadelphia, Pennsylvania, 19104, US; 51 - Emory University, Atlanta, Georgia, 30322, US; 52 - University of

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Maryland, Baltimore, Maryland, 21201, US; 53 - University of Washington, Seattle, Washington, 98195, US; 54 - University of Mississippi Medical Center, Jackson, Mississippi, 39216, US; 55 - National Institutes of Health, Bethesda, Maryland, 20892, US; 56 - Johns Hopkins University, Baltimore, Maryland, 21218, US; 57 - Duke University, Durham, North Carolina, 27708, US; 58 - University of Alabama, Birmingham, Alabama, 35487, US; 59 - Stanford University, Stanford, California, 94305, US; 60 - Medical College of Wisconsin, Milwaukee, Wisconsin, 53211, US; 61 - Providence Health Care, Medicine, Vancouver, CA; 62 - Baylor College of Medicine Human Genome Sequencing Center, Houston, Texas, 77030, US; 63 - Cleveland Clinic, Cleveland, Ohio, 44195, US; 64 - Tempus, University of Colorado Anschutz Medical Campus, Aurora, Colorado, 80045, US; 65 - Columbia University, New York, New York, 10032, US; 66 - The Emmes Corporation, LTRC, Rockville, Maryland, 20850, US; 67 - Cleveland Clinic, Quantitative Health Sciences, Cleveland, Ohio, 44195, US; 68 - Johns Hopkins University, Medicine, Baltimore, Maryland, 21218, US; 69 - National Heart, Lung, and Blood Institute, National Institutes of Health, Bethesda, Maryland, 20892, US; 70 - Boston University, Massachusetts General Hospital, Boston University School of Medicine, Boston, Massachusetts, 2114, US; 71 - University of Pittsburgh, Pittsburgh, Pennsylvania, 15260, US; 72 - Fundação de Hematologia e Hemoterapia de Pernambuco - Hemope, Recife, 52011–000, BR; 73 - University of Utah, Obstetrics and Gynecology, Salt Lake City, Utah, 84132, US; 74 - National Jewish Health, National Jewish Health, Denver, Colorado, 80206, US; 75 - Medical College of Wisconsin, Pediatrics, Milwaukee, Wisconsin, 53226, US; 76 - University of Texas Health at Houston, Pediatrics, Houston, Texas, 77030, US; 77 - University of California, San Francisco, San Francisco, California, 94143, US; 78 - Stanford University, Biomedical Data Science, Stanford, California, 94305, US; 79 - University of Washington, Biostatistics, Seattle, Washington, 98195, US; 80 - University of Colorado at Denver, Denver, Colorado, 80204, US; 81 - Brigham & Women’s Hospital, Boston, Massachusetts, 2115, US; 82 - University of Montreal, US; 83 - University of Mississippi, Medicine, Jackson, Mississippi, 39213, US; 84 - Washington State University, Pullman, Washington, 99164, US; 85 - University of California, Los Angeles, Los Angeles, California, 90095, US; 86 - Brigham & Women’s Hospital, US; 87 - Brigham & Women’s Hospital, Medicine, Boston, Massachusetts, 2115, US; 88 - National Taiwan University, Taipei, 10617, TW; 89 - Brigham & Women’s Hospital, Division of Preventive Medicine, Boston, Massachusetts, 2215, US; 90 - University of Virginia, Charlottesville, Virginia, 22903, US; 91 - Lundquist Institute, Torrance, California, 90502, US; 92 - National Taiwan University, National Taiwan University Hospital, Taipei, 10617, TW; 93 - Cleveland Clinic, Cleveland Clinic, Cleveland, Ohio, 44195, US; 94 - National Health Research Institute Taiwan, Miaoli County, 350, TW; 95 - Broad Institute, Metabolomics Platform, Cambridge, Massachusetts, 2142, US; 96 - Cleveland Clinic, Immunity and Immunology, Cleveland, Ohio, 44195, US; 97 - University of Vermont, Burlington, Vermont, 5405, US; 98 - National Jewish Health, Denver, Colorado, 80206, US; 99 - University of Michigan, Internal Medicine, Ann Arbor, Michigan, 48109, US; 100 - Vitalant Research Institute, San Francisco, California, 94118, US; 101 - University of Illinois at Chicago, Chicago, Illinois, 60607, US; 102 - University of Chicago, Chicago, Illinois, 60637, US; 103 - Mayo Clinic, Health Quantitative Sciences Research, Rochester, Minnesota, 55905, US; 104 - Washington University in St Louis, Department of Medicine, Cardiovascular Division, St. Louis, Missouri, 63110,

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

US; 105 - Vanderbilt University, Nashville, Tennessee, 37235, US; 106 - University of Cincinnati, Cincinnati, Ohio, 45220, US; 107 - Baylor College of Medicine Human Genome Sequencing Center, Houston, Texas, 77030; 108 - University of North Carolina, Chapel Hill, North Carolina, 27599, US; 109 - Washington University in St Louis, Genetics, St Louis, Missouri, 63110, US; 110 - Brown University, Providence, Rhode Island, 2912, US; 111 - Harvard University, Channing Division of Network Medicine, Cambridge, Massachusetts, 2138, US; 112 - Massachusetts General Hospital, Boston, Massachusetts, 2114, US; 113 - National Jewish Health, Center for Genes, Environment and Health, Denver, Colorado, 80206, US; 114 - University of Texas Health at Houston, Houston, Texas, 77225, US; 115 - University of North Carolina, Epidemiology, Chapel Hill, North Carolina, 27599, US; 116 - Washington University in St Louis, St Louis, Missouri, 63130, US; 117 - Fred Hutchinson Cancer Research Center, Seattle, Washington, 98109, US; 118 - New York Genome Center, New York City, New York, 10013, US; 119 - Icahn School of Medicine at Mount Sinai, New York, New York, 10029, US; 120 - University of Pittsburgh, Pittsburgh, Pennsylvania, US; 121 - Beth Israel Deaconess Medical Center, Boston, Massachusetts, 2215, US; 122 - Boston Children's Hospital, Harvard Medical School, Department of Psychiatry, Boston, Massachusetts, 2115, US; 123 - University of Colorado Anschutz Medical Campus, Aurora, Colorado, 80045, US; 124 - Mass General Brigham, Obstetrics and Gynecology, Boston, Massachusetts, 2115, US; 125 - University of Mississippi, Cardiology, Jackson, Mississippi, 39216, US; 126 - University of Calgary, Medicine, Calgary, CA; 127 - University of Maryland, Genetics, Philadelphia, Pennsylvania, 19104, US; 128 - Yale University, Department of Chronic Disease Epidemiology, New Haven, Connecticut, 6520, US; 129 - Tulane University, New Orleans, Louisiana, 70118, US; 130 - University of Washington, Epidemiology, Seattle, Washington, 98195, US; 131 - Wake Forest Baptist Health, Winston-Salem, North Carolina, 27157, US; 132 - Brigham & Women's Hospital, Channing Division of Network Medicine, Boston, Massachusetts, 2115, US; 133 - University of Iowa, Iowa City, Iowa, 52242, US; 134 - National Health Research Institute Taiwan, Institute of Population Health Sciences, NHRI, Miaoli County, 350, TW; 135 - Tri-Service General Hospital National Defense Medical Center, TW; 136 - Blood Works Northwest, Seattle, Washington, 98104, US; 137 - Taichung Veterans General Hospital Taiwan, Taichung City, 407, TW; 138 - Oklahoma State University Medical Center, Internal Medicine, DIvision of Endocrinology, Diabetes and Metabolism, Columbus, Ohio, 43210, US; 139 - National Heart, Lung, and Blood Institute, National Institutes of Health, NHLBI, Bethesda, Maryland, 20892, US; 140 - Blood Works Northwest, Research Institute, Seattle, Washington, 98104, US; 141 - University of Michigan, Biostatistics, Ann Arbor, Michigan, 48109, US; 142 - Albert Einstein College of Medicine, New York, New York, 10461, US; 143 - University of California, San Francisco, San Francisco, California, 94118, US; 144 - Harvard University, Cambridge, Massachusetts, 2138, US; 145 - McGill University, Montréal, QC H3A 0G4, CA; 146 - University of Colorado at Denver, Epidemiology, Aurora, Colorado, 80045, US; 147 - Blood Works Northwest, Medicine, Seattle, Washington, 98104, US; 148 - Loyola University, Public Health Sciences, Maywood, Illinois, 60153, US; 149 - Harvard School of Public Health, Biostats, Boston, Massachusetts, 2115, US; 150 - Boston University, University of Massachusetts Chan Medical School, Worcester, Massachusetts, 1655, US; 151 - Brown University, Epidemiology and Medicine, Providence, Rhode Island, 2912, US; 152 - Duke University,

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Cardiology, Durham, North Carolina, 27708, US; 153 - Stanford University, Cardiovascular Institute, Stanford, California, 94305, US; 154 - Icahn School of Medicine at Mount Sinai, The Charles Bronfman Institute for Personalized Medicine, New York, New York, 10029, US; 155 - Boston University, Boston, Massachusetts, 2215, US; 156 - Ohio State University, Division of Pulmonary, Critical Care and Sleep Medicine, Columbus, Ohio, 43210, US; 157 - University of Texas Rio Grande Valley School of Medicine, Brownsville, Texas, 78520, US; 158 - Broad Institute, Harvard University, Massachusetts General Hospital; 159 - University of Alabama, University of Alabama at Birmingham, Birmingham, Alabama, 35487, US; 160 - University of Washington, Genome Sciences, Seattle, Washington, 98195, US; 161 - RTI International, US; 162 - Massachusetts General Hospital, Medicine, Boston, Massachusetts, 2114, US; 163 - University of Arizona, Tucson, Arizona, 85721, US; 164 - Stanford University, Center For Sleep Sciences and Medicine, Palo Alto, California, 94304, US; 165 - National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, Maryland, 20892, US; 166 - Oklahoma Medical Research Foundation, Genes and Human Disease, Oklahoma City, Oklahoma, 73104, US; 167 - Howard University, Washington, District of Columbia, 20059, US; 168 - University of Washington, Department of Genome Sciences, Seattle, Washington, 98195, US; 169 - University at Buffalo, Buffalo, New York, 14260, US; 170 - University of Pennsylvania, Division of Sleep Medicine/Department of Medicine, Philadelphia, Pennsylvania, 19104–3403, US; 171 - Stanford University, Stanford Cardiovascular Institute, Stanford, California, 94305, US; 172 - University of Minnesota, Minneapolis, Minnesota, 55455, US; 173 - RTI International, Biostatistics and Epidemiology Division, Research Triangle Park, North Carolina, 27709–2194, US; 174 - University of Texas Rio Grande Valley School of Medicine, Edinburg, Texas, 78539, US; 175 - Fred Hutchinson Cancer Research Center, Fred Hutch and UW, Seattle, Washington, 98109, US; 176 - Johns Hopkins University, Cardiology/Medicine, Baltimore, Maryland, 21218, US; 177 - University of Colorado at Denver, Medicine, Denver, Colorado, 80204, US; 178 - University of Colorado at Denver, CCPM, Denver, Colorado, 80045, US; 179 - Northwestern University, Chicago, Illinois, 60208, US; 180 - New York Genome Center, New York Genome Center, New York City, New York, 10013, US; 181 - University of Ottawa, Sleep Research Unit, University of Ottawa Institute for Mental Health Research, Ottawa, ON K1Z 7K4, CA; 182 - Vanderbilt University, Medicine, Pharmacology, Biomedicla Informatics, Nashville, Tennessee, 37235, US; 183 - Universidade de Sao Paulo, Faculdade de Medicina, Sao Paulo, 1310000, BR; 184 - Columbia University, New York, New York, 10027, US; 185 - University of Maryland, Pathology, Seattle, Washington, 98195, US; 186 - Lundquist Institute, TGPS, Torrance, California, 90502, US; 187 - Harvard University, Division of Hematology/Oncology, Boston, Massachusetts, 2115, US; 188 - Harvard Medical School, Genetics, Boston, Massachusetts, 2115, US; 189 - Harvard Medical School, Boston, Massachusetts, 2115, US; 190 - Université Laval, Quebec City, G1V 0A6, CA; 191 - Emory University, Pediatrics, Atlanta, Georgia, 30307, US; 192 - Emory University, Human Genetics, Atlanta, Georgia, 30322, US; 193 - Vanderbilt University, Medicine/Cardiology, Nashville, Tennessee, 37235, US; 194 - UMass Memorial Medical Center, Worcester, Massachusetts, 1655, US; 195 - University of Saskatchewan, Saskatoon, SK S7N 5C9, CA; 196 - University of Michigan; 197 - Wake Forest Baptist Health, Biostatistical Sciences, Winston-Salem, North Carolina, 27157, US; 198 - University of Colorado at Denver, Genomic Cardiology, Aurora, Colorado,

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

80045, US; 199 - Brigham & Women's Hospital, Channing Department of Medicine, Boston, Massachusetts, 2115, US; 200 - Stanford University, Genetics, Stanford, California, 94305, US; 201 - University of Washington, University of Washington, Department of Genome Sciences, Seattle, Washington, 98195, US; 202 - Fred Hutchinson Cancer Research Center, Cancer Prevention Division of Public Health Sciences, Seattle, Washington, 98109, US; 203 - University of Pennsylvania, Genetics, Philadelphia, Pennsylvania, 19104, US; 204 - University of Alabama, Biostatistics, Birmingham, Alabama, 35487, US; 205 - University of Washington, Department of Biostatistics, Seattle, Washington, 98195, US; 206 - University of Vermont, Pathology & Laboratory Medicine, Burlington, Vermont, 5405, US; 207 - University of Southern California, USC Methylation Characterization Center, University of Southern California, California, 90033, US; 208 - Brigham & Women's Hospital, Mass General Brigham, Boston, Massachusetts, 2115, US; 209 - University of Michigan, US; 210 - University of Pittsburgh, Department of Human Genetics, Pittsburgh, Pennsylvania, 15260, US; 211 - Brigham & Women's Hospital, Channing Division of Network Medicine, Department of Medicine, Boston, Massachusetts, 2115, US; 212 - Indiana University, Epidemiology, Indianapolis, Indiana, 46202, US; 213 - Henry Ford Health System, Detroit, Michigan, 48202, US; 214 - University of Pittsburgh, Medicine, Pittsburgh, Pennsylvania, 15260, US; 215 - Case Western Reserve University, Department of Population and Quantitative Health Sciences, Cleveland, Ohio, 44106, US; 216 - University of California, San Francisco, US

References

1. Taliun D et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. Nature 590, 290–299 (2021). [PubMed: 33568819]
2. Van Hout CV et al. Exome sequencing and characterization of 49,960 individuals in the UK Biobank. Nature 586, 749–756 (2020). [PubMed: 33087929]
3. Szustakowski JD et al. Advancing human genetics research and drug discovery through exome sequencing of the UK Biobank. Nature Genetics 53, 942–948 (2021). [PubMed: 34183854]
4. Hindy G et al. Rare coding variants in 35 genes associate with circulating lipid levels—A multi-ancestry analysis of 170,000 exomes. The American Journal of Human Genetics 109, 81–96 (2022). [PubMed: 34932938]
5. Flannick J et al. Exome sequencing of 20,791 cases of type 2 diabetes and 24,440 controls. Nature 570, 71–76 (2019). [PubMed: 31118516]
6. Jurgens SJ et al. Analysis of rare genetic variation underlying cardiometabolic diseases and traits among 200,000 individuals in the UK Biobank. Nature Genetics 54, 240–250 (2022). [PubMed: 35177841]
7. Wainschtein P et al. Assessing the contribution of rare variants to complex trait heritability from whole-genome sequence data. Nature Genetics 54, 263–273 (2022). [PubMed: 35256806]
8. Lee S, Abecasis Gonçalo R., Boehnke M & Lin X Rare-Variant Association Analysis: Study Designs and Statistical Tests. The American Journal of Human Genetics 95, 5–23 (2014). [PubMed: 24995866]
9. 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 83, 311–321 (2008). [PubMed: 18691683]
10. Madsen BE & Browning SR A Groupwise Association Test for Rare Mutations Using a Weighted Sum Statistic. PLOS Genetics 5, e1000384 (2009). [PubMed: 19214210]
11. Morris AP & Zeggini E An evaluation of statistical approaches to rare variant analysis in genetic association studies. Genetic Epidemiology 34, 188–193 (2010). [PubMed: 19810025]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

12. Wu Michael C. et al. Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test. The American Journal of Human Genetics 89, 82–93 (2011). [PubMed: 21737059]
13. Liu Y et al. ACAT: a fast and powerful p value combination method for rare-variant analysis in sequencing studies. The American Journal of Human Genetics 104, 410–421 (2019). [PubMed: 30849328]
14. McCarthy MI et al. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nature Reviews Genetics 9, 356–369 (2008).
15. Evangelou E & Ioannidis JPA Meta-analysis methods for genome-wide association studies and beyond. Nature Reviews Genetics 14, 379–389 (2013).
16. Buniello A et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Research 47, D1005–D1012 (2018).
17. Lin DY & Zeng D Meta-analysis of genome-wide association studies: no efficiency gain in using individual participant data. Genetic Epidemiology 34, 60–66 (2010). [PubMed: 19847795]
18. Lin DY & Zeng D On the relative efficiency of using summary statistics versus individual-level data in meta-analysis. Biometrika 97, 321–332 (2010). [PubMed: 23049122]
19. Liu DJ et al. Meta-analysis of gene-level tests for rare variant association. Nature Genetics 46, 200–204 (2014). [PubMed: 24336170]
20. Feng S, Liu D, Zhan X, Wing MK & Abecasis GR RAREMETAL: fast and powerful meta-analysis for rare variants. Bioinformatics 30, 2828–2829 (2014). [PubMed: 24894501]
21. Lee S, Teslovich Tanya M., Boehnke M & Lin X General Framework for Meta-analysis of Rare Variants in Sequencing Association Studies. The American Journal of Human Genetics 93, 42–53 (2013). [PubMed: 23768515]
22. Hu Y-J et al. Meta-analysis of Gene-Level Associations for Rare Variants Based on Single-Variant Statistics. The American Journal of Human Genetics 93, 236–248 (2013). [PubMed: 23891470]
23. Yang J, Chen S, Abecasis G & IAMDGC. Improved score statistics for meta-analysis in single-variant and gene-level association studies. Genetic Epidemiology 42, 333–343 (2018). [PubMed: 29696691]
24. Chen H 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 104, 260–274 (2019). [PubMed: 30639324]
25. Chen M-H, Pitsillides A & Yang Q An evaluation of approaches for rare variant association analyses of binary traits in related samples. Scientific Reports 11, 3145 (2021). [PubMed: 33542345]
26. Li X 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 52, 969–983 (2020). [PubMed: 32839606]
27. Natarajan P et al. Deep-coverage whole genome sequences and blood lipids among 16,324 individuals. Nature Communications 9, 3391 (2018).
28. Gogarten SM et al. Genetic association testing using the GENESIS R/Bioconductor package. Bioinformatics 35, 5346–5348 (2019). [PubMed: 31329242]
29. Chen H 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 98, 653–666 (2016). [PubMed: 27018471]
30. Willer CJ, Li Y & Abecasis GR METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics 26, 2190–2191 (2010). [PubMed: 20616382]
31. Stilp AM et al. A System for Phenotype Harmonization in the National Heart, Lung, and Blood Institute Trans-Omics for Precision Medicine (TOPMed) Program. American Journal of Epidemiology (2021).
32. Forrest AR et al. A promoter-level mammalian expression atlas. Nature 507, 462 (2014). [PubMed: 24670764]
33. Andersson R et al. An atlas of active enhancers across human cell types and tissues. Nature 507, 455–461 (2014). [PubMed: 24670763]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

34. Fishilevich S et al. GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. Database 2017(2017).
35. Li Z et al. A framework for detecting noncoding rare variant associations of large-scale whole-genome sequencing studies. bioRxiv, 2021.11.05.467531 (2021).
36. Kircher M et al. A general framework for estimating the relative pathogenicity of human genetic variants. Nature Genetics 46, 310–315 (2014). [PubMed: 24487276]
37. Huang Y-F, Gulko B & Siepel A Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. Nature Genetics 49, 618–624 (2017). [PubMed: 28288115]
38. Rogers MF et al. FATHMM-XF: accurate prediction of pathogenic point mutations via extended features. Bioinformatics 34, 511–513 (2017).
39. Dong C et al. Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. Human Molecular Genetics 24, 2125–2137 (2014). [PubMed: 25552646]
40. Schaffner SF et al. Calibrating a coalescent simulation of human genome sequence variation. Genome Research 15, 1576–1583 (2005). [PubMed: 16251467]
41. Lee PH et al. Principles and methods of in-silico prioritization of non-coding regulatory variants. Human Genetics 137, 15–30 (2018). [PubMed: 29288389]
42. Morrison AC et al. Practical approaches for whole-genome sequence analysis of heart-and blood-related traits. The American Journal of Human Genetics 100, 205–215 (2017). [PubMed: 28089252]
43. Li Z et al. Dynamic Scan Procedure for Detecting Rare-Variant Association Regions in Whole-Genome Sequencing Studies. The American Journal of Human Genetics 104, 802–814 (2019). [PubMed: 30982610]
44. The “All of Us” Research Program. New England Journal of Medicine 381, 668–676 (2019). [PubMed: 31412182]
45. Klarin D et al. Genetics of blood lipids among ~300,000 multi-ethnic participants of the Million Veteran Program. Nature Genetics 50, 1514–1523 (2018). [PubMed: 30275531]
46. Breslow NE & Clayton DG Approximate Inference in Generalized Linear Mixed Models. Journal of the American Statistical Association 88, 9–25 (1993).
47. Jiang L et al. A resource-efficient tool for mixed model association analysis of large-scale data. Nature Genetics 51, 1749–1755 (2019). [PubMed: 31768069]
48. Jiang L, Zheng Z, Fang H & Yang J A generalized linear mixed model association tool for biobank-scale data. Nature Genetics 53, 1616–1621 (2021). [PubMed: 34737426]
49. Quick C et al. A versatile toolkit for molecular QTL mapping and meta-analysis at scale. bioRxiv, 2020.12.18.423490 (2020).
50. Gazal S et al. Linkage disequilibrium–dependent architecture of human complex traits shows action of negative selection. Nature Genetics 49, 1421–1427 (2017). [PubMed: 28892061]
51. Zhou H, Arapoglou T, Li X, Li Z & Lin X FAVOR Essential Database. V1 edn (Harvard Dataverse, 2022).
52. Li X, Li Z & Chen H xihaoli/STAAR: STAAR_v0.9.6. Version 0.9.6 10.5281/zenodo.6960622 (2022).
53. Li X & Li Z xihaoli/MetaSTAAR: MetaSTAAR_v0.9.6. Version 0.9.6 10.5281/zenodo.6960606 (2022).
54. Li X, Li Z & Lin X MetaSTAAR. Version 1 10.5281/zenodo.6668274 (2022).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

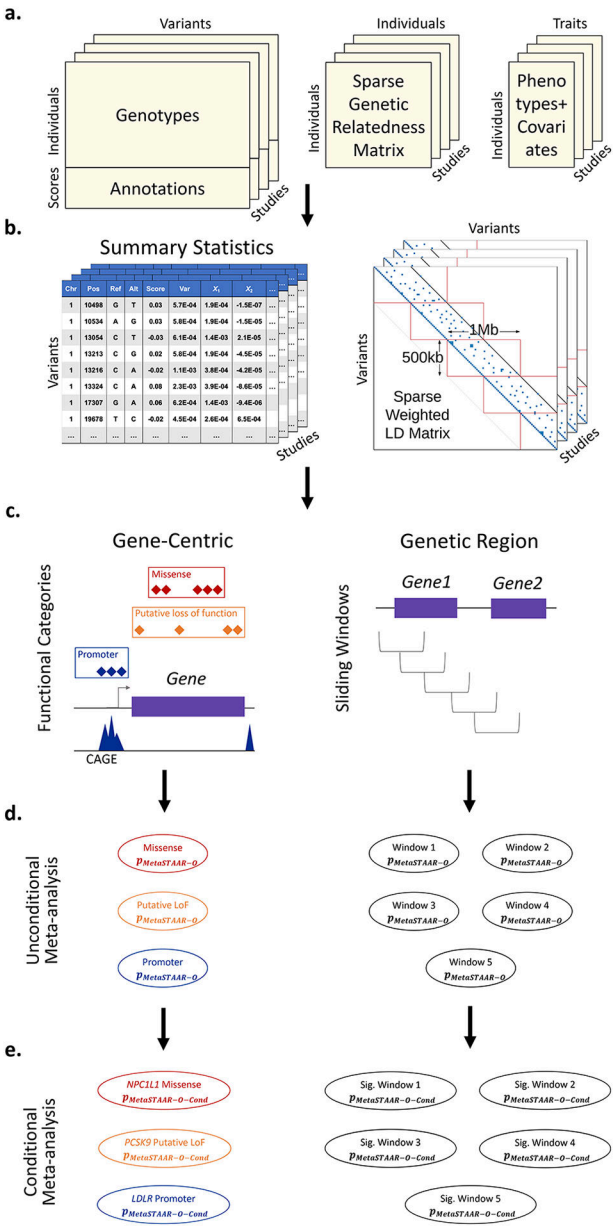


Figure 1 |. MetaSTAAR workflow.

a, Input data of MetaSTAAR for each study, including genotypes, phenotypes, covariates and sparse genetic relatedness matrix are prepared. **b**, Summary statistics, including individual variant score statistics, sparse weighted LD matrices and low-rank projection matrices accounting for covariate effects for each study are generated using MetaSTAARWorker. **c**, All rare variants in the merged variant list are functionally annotated 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.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 1 |

Comparison of different rare variant meta-analysis methods

Methods / Features	Quantitative traits			Binary traits		Does not require pre-specified variant sets	Allows for incorporating multiple functional annotations	Feasible for large sample sizes	Storage complexity
	Linear model	Linear mixed model	Accounts for heteroskedastic variance	Logistic model	Logistic mixed model				
MetaSTAAR	✓	✓	✓	✓	✓	✓	✓	✓	$O(M)$
RareMetal	✓	✓				✓			$O(M^2)$
MetaSKAT	✓			✓					$O(M^2)$
SMMAT	✓	✓	✓	✓	✓				$O(M^2)$

M, total number of rare variants in a genetic region.

Nat Genet. Author manuscript; available in PMC 2023 April 10.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 2 |

Comparison of computation time and storage of MetaSTAARWorker and RareMetalWorker.

Region	Sample size	No. of SNVs	MetaSTAARWorker		RareMetalWorker (RMW)		RMW/ MetaSTAARWorker	
			CPU hours (h)	Storage (GB)	CPU hours (h)	Storage (GB)	CPU hours Ratio	Storage Ratio
chromosome 6: 160 Mb – 161 Mb	4,791	35,993	0.22	0.01	2.05	1.77	10.4	158.2
	12,316	52,853	0.30	0.02	10.47	3.77	39.5	180.8
	30,138	88,845	0.40	0.04	69.94	10.14	195.2	227.2
chromosome 16: 0 Mb −12 Mb	4,791	666,256	2.58	0.30	80.28	65.34	33.4	220.2
	12,316	978,314	3.83	0.56	358.04 *	123.94 *	94.7	223.3
	30,138	1,617,138	5.92	1.11	2303.78 *	328.79 *	402.2	296.8

Runtime and storage of MetaSTAARWorker v0.9.6 (linear model) and RareMetalWorker v4.15.1 (linear model) to generate summary statistics, 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 genetic regions were considered in this test: all uncommon variants (MAF < 5%) from 160 Mb to 161 Mb on chromosome 6 and all uncommon variants from 0 Mb to 12 Mb on chromosome 16. The sparse-weighted LD matrices were computed using 500-kb banded windows. 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. No. of SNVs, number of uncommon variants (MAF < 5%) in the region. MESA, Multi-Ethnic Study of Atherosclerosis; FHS, Framingham Heart Study; JHS, Jackson Heart Study; OOA, Old Order Amish.

* Predicted numbers based on partial results.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 3 |

Gene-centric meta-analysis results of both unconditional analysis and analysis conditional on known common and low-frequency variants using MetaSTAAR and pooled analysis.

Trait	Gene	Chr. no.	Category	No. of SNVs	MetaSTAAR-O (Unconditional)	STAAR-O-Pooled (Unconditional)	MetaSTAAR-O (Conditional)	STAAR-O-Pooled (Conditional)	Variants (adjusted)
LDL-C	*PCSK9*	1	Putative loss of function	9	3.07E-63	7.28E-64	6.42E-63	8.81E-64	rs11591147, rs28362263, rs505151, rs12117661, rs472495
	APOB	2	Putative loss of function	16	1.14E-20	5.79E-21	2.42E-20	1.95E-20	rs1367117, rs563290, rs533617
	PCSK9	1	Missense	167	1.55E-15	1.11E-15	3.11E-14	3.38E-14	rs11591147, rs28362263, rs505151, rs12117661, rs472495
	ABCG5	2	Missense	148	2.66E-08	2.87E-08	6.28E-08	6.84E-08	rs4245791
	NPC1L1	7	Missense	293	6.53E-10	9.71E-10	3.23E-09	5.04E-09	rs217381
	LDLR	19	Missense	192	5.33E-27	4.00E-27	5.16E-27	3.84E-27	rs12151108, rs688, rs6511720
	APOE	19	Missense	88	2.24E-13	2.08E-13	1.83E-12	1.04E-12	rs7412, rs429358, rs35136575
	RNF20	9	Synonymous	58	4.25E-08	6.55E-08	4.25E-08	6.55E-08	n/a
	APOE	19	Promoter	102	7.52E-12	1.57E-11	9.98E-12	4.60E-12	rs7412, rs429358, rs35136575
HDL-C	*APOC3*	11	Putative loss of function	7	2.26E-22	1.78E-22	7.49E-21	6.13E-21	rs964184, rs12269901
	CD36	7	Missense	237	3.18E-07	5.93E-07	4.03E-08	8.11E-08	rs3211938
	ABCA1	9	Missense	346	6.72E-11	5.46E-11	2.00E-11	1.32E-11	rs4149310, rs1883025, rs11789603
	PCSK7	11	Missense	116	1.17E-09	9.32E-10	2.43E-09	1.89E-09	rs964184, rs12269901
	SCARB1	12	Missense	120	1.24E-10	4.30E-11	1.12E-10	3.81E-11	rs10773112, rs4765127
	LCAT	16	Missense	63	1.55E-10	2.21E-10	1.16E-10	1.66E-10	rs1109166
	LIPG	18	Missense	101	1.62E-07	1.96E-07	1.18E-07	1.14E-07	rs8086351, rs9958734
TG	*APOC3*	11	Putative loss of function	7	3.57E-54	2.25E-54	1.88E-51	5.05E-52	rs964184, rs9804646, rs3135506, rs2266788
	APOA5	11	Missense	64	2.14E-07	1.55E-07	2.37E-09	3.86E-09	rs964184, rs9804646, rs3135506, rs2266788
	APOA4	11	Missense	118	1.15E-08	9.22E-09	8.05E-10	7.52E-10	rs964184, rs9804646, rs3135506, rs2266788

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Trait	Gene	Chr. no.	Category	No. of SNVs	MetaSTAAR-O (Unconditional)	STAAR-O-Pooled (Unconditional)	MetaSTAAR-O (Conditional)	STAAR-O-Pooled (Conditional)	Variants (adjusted)
	APOC3	11	Missense	18	3.03E-12	1.73E-12	1.52E-12	6.99E-13	rs964184, rs9804646, rs3135506, rs2266788
	PAFAH1B2	11	Missense	31	2.71E-09	2.61E-09	3.76E-10	3.03E-10	rs964184, rs9804646, rs3135506, rs2266788
	APOE	19	Missense	89	1.43E-11	1.44E-11	1.30E-10	9.16E-11	rs12721054, rs5112, rs429358
	COL18A1	21	Missense	588	1.07E-08	2.36E-08	1.07E-08	2.36E-08	n/a
	APOA5	11	Promoter	15	1.32E-10	9.86E-11	4.74E-12	1.70E-12	rs964184, rs9804646, rs3135506, rs2266788
	APOA4	11	Promoter	198	8.12E-11	3.70E-11	1.45E-09	1.16E-09	rs964184, rs9804646, rs3135506, rs2266788
	APOC3	11	Promoter	62	4.72E-11	2.09E-11	1.80E-11	5.20E-12	rs964184, rs9804646, rs3135506, rs2266788
	APOE	19	Promoter	104	9.50E-18	7.03E-18	3.83E-10	2.28E-10	rs12721054, rs5112, rs429358
	APOA5	11	Enhancer	13	2.38E-10	2.09E-10	8.90E-12	4.17E-12	rs964184, rs9804646, rs3135506, rs2266788
	APOA1	11	Enhancer	357	5.04E-10	2.69E-10	2.87E-10	1.19E-10	rs964184, rs9804646, rs3135506, rs2266788
	COL18A1	21	Enhancer	312	3.97E-09	9.35E-09	3.97E-09	9.35E-09	n/a
TC	*PCSK9*	1	Putative loss of function	9	4.46E-57	1.96E-56	1.23E-56	4.85E-56	rs11591147, rs28362263, rs505151, rs12117661, rs2495477
	APOB	2	Putative loss of function	16	3.52E-19	3.98E-19	7.85E-19	1.29E-18	rs1367117, rs10692845, rs533617
	PCSK9	1	Missense	169	1.94E-11	6.89E-12	1.15E-11	3.77E-12	rs11591147, rs28362263, rs505151, rs12117661, rs2495477
	ABCG5	2	Missense	157	4.74E-09	5.24E-09	1.21E-08	1.41E-08	rs4245791
	NPC1L1	7	Missense	301	5.19E-08	3.59E-08	2.08E-07	1.54E-07	rs217381
	ABCA1	9	Missense	346	6.89E-08	1.18E-07	3.72E-08	5.27E-08	rs1800978, rs4149310, rs3847302
	LIPG	18	Missense	101	2.69E-08	1.71E-08	1.39E-08	7.84E-09	rs9958734
	LDLR	19	Missense	200	1.41E-22	6.53E-23	8.33E-23	3.04E-23	rs73015024, rs688, rs2278426, rs6511720
	APOE	19	Missense	90	1.18E-08	2.58E-08	1.46E-08	2.79E-08	rs7412, rs429358, rs12721054

Nat Genet. Author manuscript; available in PMC 2023 April 10.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Trait	Gene	Chr. no.	Category	No. of SNVs	MetaSTAAR-O (Unconditional)	STAAR-O-Pooled (Unconditional)	MetaSTAAR-O (Conditional)	STAAR-O-Pooled (Conditional)	Variants (adjusted)
	APOE	19	Promoter	105	1.92E-07	5.58E-07	7.20E-08	1.36E-07	rs7412, rs429358, rs12721054

A total of 30,138 samples from 14 study cohorts in TOPMed program were considered in the meta-analysis using MetaSTAAR-O and pooled analysis using STAAR-O (STAAR-O-Pooled). Results for the conditionally significant genes (unconditional MetaSTAAR-O $P < 5.00 \times 10^{-7}$; conditional MetaSTAAR-O $P < 5.00 \times 10^{-7}$) are presented in the table. The significant threshold was defined by the multiple comparisons using the Bonferroni correction $(0.05/(20,000 \times 5) = 5.00 \times 10^{-7})$. Both MetaSTAAR-O and STAAR-O-Pooled are two-sided tests. Chr. no., chromosome number; category, functional category; no. of SNVs, number of rare variants (pooled MAF < 1%) of the particular functional category in the gene; MetaSTAAR-O, MetaSTAAR-O *P* value; STAAR-O-Pooled, STAAR-O *P* value from the joint analysis of pooled individual-level data; LDL-C, low-density lipoprotein cholesterol; HDL-C, high-density lipoprotein cholesterol; TG, triglycerides; TC, total cholesterol; Variants (adjusted), adjusted variants in the conditional analysis; n/a, no variant adjusted in the conditional analysis. MetaSTAAR-O results and pooled analysis results using STAAR-O are very similar.