Title

A framework for detecting noncoding rare-variant associations of large-scale whole-genome sequencing studies

Journal

Nature Methods, 19(12)

ISSN

1548-7091

Authors

Li, Zilin

Li, Xihao

Zhou, Hufeng

et al.

Publication Date

2022-12-01

DOI

10.1038/s41592-022-01640-x

Peer reviewed

HHS Public Access

Author manuscript

Nat Methods. Author manuscript; available in PMC 2023 March 12.

Published in final edited form as:

Nat Methods. 2022 December ; 19(12): 1599–1611. doi:10.1038/s41592-022-01640-x.

A framework for detecting noncoding rare variant associations of large-scale whole-genome sequencing studies

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

Abstract

Large-scale whole-genome sequencing (WGS) studies have enabled analysis of noncoding rare variant (RV) associations with complex human diseases and traits. Variant set analysis is a powerful approach to study RV association. However, existing methods have limited ability in analyzing the noncoding genome. We propose a computationally efficient and robust noncoding RV association-detection framework, STAARpipeline, to automatically annotate a WGS study and perform flexible noncoding RV association analysis, including gene-centric analysis and fixed-window and dynamic-window-based non-gene-centric analysis by incorporating variant functional annotations. In gene-centric analysis, STAARpipeline uses STAAR to group noncoding variants based on functional categories of genes and incorporate multiple functional annotations. In non-gene-centric analysis, STAARpipeline uses SCANG-STAAR to incorporate dynamic window sizes and multiple functional annotations. We apply STAARpipeline to identify noncoding RV sets associated with four lipid traits in 21,015 discovery samples from the Trans-Omics for Precision

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

Author contributions

Z. L., X. Li and X. Lin designed the experiments., Z.L., X. Li, H.Z. and X. Lin performed the experiments. Z. L., X. Li, H.Z., S.M.G., M.S.S., T.A., C.Q., Y.L., H.C., R.S., R.D., D.K.A., L.F.B., J.C.B., T.W.B, J.B., E.B., D.W.B., J.A.B., B.E.C., M.P.C., A.C., L.A.C., J.E.C., P.S.d.V., R.D., B.I.F., H.H.H.G., X.G., R.R.K., C.L.K., B.G.K., L.A.L., A.W.M., L.W.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., J.A.S., K.D.T., R.S.V., D.E.W., J.G.W., L.R.Y., W.Z., J.I.R., C.J.W., P.N., G.M.P. and X. Lin acquired, analyzed or interpreted data. G.M.P., P.N. and NHLBI TOPMed Lipids Working Group provided administrative, technical or material support. Z.L., X. Li, S.M.G. and X. Lin drafted the manuscript and revised according to co-authors' suggestions. All authors critically reviewed the manuscript, suggested revisions as needed, and approved the final version.

A full list of consortium members appears at the Supplementary Note.

Competing interests

S.M.G. is now an employee of Regeneron Genetics Center. J.B.M. is an Academic Associate for Quest Diagnostics R&D. 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.

Code availability

STAARpipeline is implemented as an open-source R package available at https://github.com/xihaoli/STAARpipeline[56] and https://content.sph.harvard.edu/xlin/software.html. *STAARpipelineSummary* is implemented as an open-source R package available at https://github.com/xihaoli/STAARpipelineSummary[57] and https://content.sph.harvard.edu/xlin/software.html. The scripts used to generate the results have been archived on Zenodo using https://doi.org/10.5281/zenodo.6871408[58].

Data analysis was performed in R (3.6.1). STAAR v0.9.6, STAARpipeline v0.9.6, and STAARpipelineSummary v0.9.6 were used in simulation and real data analysis, seqMeta v1.6.7 was used in simulation. Wget v1.14 was used for downloading the annotation data. FAVORannotator v1.0.0 (https://github.com/xihaoli/STAARpipeline-Tutorial) was used for functionally annotate the whole-genome data.

Medicine (TOPMed) program and replicate several of them in additional 9,123 TOPMed samples. We also analyze five non-lipid TOPMed traits.

Introduction

Genome-wide association studies (GWASs) have successfully identified thousands of common genetic variants for complex diseases and traits; however, these common variants only explain a small fraction of heritability[1]. Recent studies suggest that the missing heritability of complex traits and diseases and causal variants may be accounted for in part by RVs (minor allele frequency (MAF) < 1%)[2–4]. Although whole-exome sequencing (WES) studies have identified exome-wide significant RV associations for complex diseases and traits[5, 6], more than 98% of the genetic variants are located in the noncoding genome[6]. Many common variants identified by GWAS as being associated with phenotypes are located in noncoding regions[7–9]. Further, the ENCODE project shows that a significant fraction of noncoding regions are functionally active[10, 11], indicating that rare noncoding regions may have an effect on diseases or traits.

An increasing number of whole-genome sequencing (WGS) association studies, such as the Genome Sequencing Program (GSP) of the National Human Genome Research Institute (NHGRI), the Trans-Omics for Precision Medicine (TOPMed) Program of the National Heart, Lung, and Blood Institute (NHLBI), and UK Biobank provide an opportunity to study the genetic contributions of noncoding RVs to complex traits and diseases. It is of substantial interest to use these rich WGS data to explore the role of noncoding RVs in the genetic underpinning of common human diseases.

Single-variant analyses are not appropriate for analysis of rare variants as they lack sufficient power[12–14]. To improve power, variant set tests have been proposed that assess the effects of sets of multiple RVs jointly. These include burden tests, SKAT, and most recently STAAR (variant-set test for association using annotation information), which incorporates multiple functional annotations for genetic variants to boost the power[15–17]. A key challenge of these approaches is the selection of RVs to form variant sets. Several methods have been proposed to create coding and noncoding variant sets for analysis of WGS/WES studies[17–22]. However, these methods have limited ability to define analysis units in the noncoding genome[23]. For example, for gene-centric analysis, STAAR uses two noncoding functional categories (masks) of regulatory regions: promoters and enhancers in GeneHancer[24] overlaid with Cap Analysis of Gene Expression (CAGE) sites[25, 26]; for non-gene-centric analysis, STAAR uses fixed-size sliding windows to scan the genome.

As signal regions (variant-phenotype-association regions) are unknown in practice and their sizes vary across the genome, the fixed-size sliding window approach is likely to lead to power loss when the prespecified window sizes are too big or too small compared with the actual sizes of signal regions. Furthermore, it is often knowledge- and effort-intensive to functionally annotate variants in a WGS/WES study of interest using the existing resources. Limited tools exist for multi-faceted functional annotation and analytic integration of WGS/WES data for rare variant association tests (RVATs). Finally, there are few robust pipelines that perform scalable and comprehensive noncoding RV association analysis in

large-scale WGS data with hundreds of millions of noncoding RVs across the genome. Much uncertainty remains on the best practices for performing computationally efficient RV analysis of large scale WGS studies.

To address these issues, we propose a computationally efficient and robust noncoding rare variant association-detection framework for WGS data. We make three new contributions toward automatically selecting interpretable and powerful variant sets for noncoding RV analysis. First, in gene-centric analysis, we propose additional strategies for grouping noncoding variants based on functional annotations, including untranslated regions, upstream regions, downstream regions, promoters, enhancers of protein-coding genes, and long noncoding RNA genes within STAAR. For promoters and enhancers, we offer additional options of overlaying promoters and GeneHancer-based enhancers with not only CAGE sites but also with DNase Hypersensitivity (DHS) sites[10]. Second, in non-gene-centric analysis, instead of using fixed-size sliding windows in STAAR, we propose SCANG-STAAR, a flexible data-adaptive window size RVAT method that extends the SCANG (scan the genome) method[19] by incorporating multiple functional annotations through STAAR[17], while accounting for both relatedness and population structure through the generalized linear mixed model (GLMM) framework[27] for quantitative and dichotomous traits[28, 29]. Third, we develop *STAARpipeline*, a pipeline that (1) functionally annotates both noncoding and coding variants of a WGS study and builds an annotated genotype dataset using the multi-faceted functional annotation database FAVOR[17, 30] (Functional Annotations of Variants - Online Resource), through FAVORannotator; and (2) performs RVATs using the proposed methods for both gene-centric analysis and non-gene-centric analysis.

We applied the proposed framework to detect noncoding RVs associated with four quantitative lipid traits: low-density lipoprotein cholesterol (LDL-C); high-density lipoprotein cholesterol (HDL-C); triglycerides (TG) and total cholesterol (TC) using 21,015 discovery samples and 9,123 replication samples from the NHLBI TOPMed Freeze 5 WGS data. We performed conditional analysis by conditioning on known lipids-associated variants and identified several novel replicated RVs sets associated with lipids. We also applied the proposed framework to identify RV associations in the noncoding genome for five additional non-lipid traits in TOPMed Freeze 5: C-reactive protein (CRP), estimated glomerular filtration rate (eGFR), fasting glucose (FG), fasting insulin (FI) and telomere length (TL).

Results

Overview of Noncoding RVATs

We propose a computationally efficient and robust noncoding RVAT framework for phenotype-genotype association analyses of whole-genome sequencing data, focusing on rare variant association analysis in the noncoding genome. This regression-based framework allows adjusting for covariates, population structure, and relatedness by fitting linear and logistic mixed models for quantitative and dichotomous traits[28, 29]. A central component of our approach is the development of strategies to aggregate noncoding rare variants using both flexible gene-centric and non-gene-centric approaches to empower RVATs. For the gene-centric approach, we group noncoding RVs for each gene using eight functional

categories of regulatory regions provided by functional annotations and apply STAAR, which incorporates multiple *in-silico* variant functional annotation scores that prioritize functional variants using multi-dimensional variant biological functions[17]. For the non-gene-centric analysis, instead of using sliding windows with fixed sizes, we propose SCANG-STAAR, a procedure using dynamic windows with data-adaptive sizes and incorporating multi-dimensional functional annotations. We also perform analytical follow-up to dissect RV association signals independent of a given set of known variants via conditional analysis (Fig. 1).

Gene-centric analysis of the noncoding genome

In gene-centric analysis of noncoding variants, we provide eight functional categories of regulatory regions to aggregate noncoding rare variants: (1) promoter RVs overlaid with CAGE sites, (2) promoter RVs overlaid with DHS sites, (3) enhancer RVs overlaid with CAGE sites, (4) enhancer RVs overlaid with DHS sites, (5) untranslated region (UTR) RVs, (6) upstream region RVs, (7) downstream region RVs and (8) noncoding RNA (ncRNA) RVs. The promoter RVs are defined as RVs in the +/− 3-kilobase (kb) window of transcription start sites with the overlap of CAGE sites or DHS sites. The enhancer RVs are defined as RVs in GeneHancer predicted regions with the overlap of CAGE sites or DHS sites[10, 24–26]. We define the UTR, upstream, downstream, and ncRNA RVs by GENCODE Variant Effect Predictor (VEP) categories[31, 32]. For the UTR mask, we include RVs in both 5' and 3' UTR regions. For the ncRNA mask, we include the exonic and splicing ncRNA RVs. We consider the protein-coding gene for the first seven categories provided by Ensembl[33] and the ncRNA genes provided by GENCODE[31, 32].

For each noncoding mask, we calculate its *P* value using the STAAR method that empowers RVATs by incorporating multiple variant functional annotation scores[17]. Functional annotations consist of diverse biological information of genomic elements. Incorporating this external biological information provided by functional annotations can increase the association analysis power[34]. For example, annotation principal components (aPCs) provide multi-dimensional summaries of variant annotations and capture the multi-faceted biological impact. The aPCs are calculated using the first principal component of the set of individual functional annotation scores measuring similar biological functionality[17]. We incorporate nine aPCs and three integrative scores (CADD[35], LINSIGHT[36], and FATHMM-XF[37]) as weights in constructing STAAR statistics[17]. We additionally incorporate a liver-tissue-specific aPC for lipids analysis. Details of these 13 functional annotations are given in Supplementary Table 1.

Specifically, we calculate the *P* value of each variant set using STAAR-O[17], an omnibus test aggregating multiple annotation-weighted burden test[15], SKAT[16], and ACAT-V[38] in the STAAR framework.

Non-gene-centric analysis using dynamic windows with SCANG-STAAR

We improve the STAAR-based fixed-size sliding window RVAT[17, 18] by proposing a dynamic window based SCANG-STAAR method, which extends the SCANG[19] procedure by incorporating multi-dimensional functional annotations to flexibly detect the locations

and the sizes of signal windows across the genome. As the locations of regions associated with a disease or trait are often unknown in advance and their sizes may vary across the genome, the use of a pre-specified fixed-size sliding window for RVAT can lead to power loss, if the pre-specified window sizes do not align with the true locations of the signals.

Specifically, we extend the SCANG-SKAT (SCANG-S) procedure to SCANG-STAAR-S by calculating the STAAR-SKAT (STAAR-S) P value in each overlapping window by incorporating multiple variant functional annotations, instead of using just the MAF-weight-based SKAT P value. In SCANG-STAAR-S, we first calculate a threshold that controls the genome-wise type I error rate at a given α level, based on the minimum value of the STAAR-S P value from all moving windows of different sizes in a range of windows (Online Methods). The procedure then selects the candidate significant windows whose set-based P value passes that threshold. When this results in multiple overlapping windows, we localize the detected significant window as the window whose P value is smaller than both the threshold and any window that overlaps with it. We then calculate the genome-wide P value of the detected windows by accounting for multiple comparisons of overlapping windows and controlling the corresponding genome-wise (family-wise) error rate (Online Methods).

Besides the SCANG-STAAR-S method, we also provide the SCANG-STAAR-B procedure, based on the STAAR-Burden P value. Compared with SCANG-STAAR-B, SCANG-STAAR-S has two advantages in detecting noncoding associations using dynamic windows in practice. First, the effects of causal variants in a neighborhood in the noncoding genome tend to be in different directions, especially in intergenic regions. Second, due to the different correlation structures of the two test statistics for overlapping windows, the genome-wide significance threshold of SCANG-STAAR-B is lower than that of SCANG-STAAR-S. For example, to control the genome-wise error rate at 0.05 level in our analysis of LDL-C, the P value thresholds for SCANG-STAAR-S and SCANG-STAAR-B are 3.80×10^{-9} and 2.31×10^{-10}, respectively. We additionally provide the SCANG-STAAR-O procedure, which is based on an omnibus P value of SCANG-STAAR-S and SCANG-STAAR-B calculated by the ACAT method[36]. However, different from STAAR-O, we do not incorporate the ACAT-V test in the omnibus test, since the ACAT-V test is designed for sparse alternatives. Hence, it tends to detect the region with the smallest size that contains the most significant variant in the dynamic window procedure.

Analytical follow-up using conditional analysis

We perform follow-up conditional analysis to identify RV association signals that are independent of known single variant associations. We first select a list of known variants by including the previously identified trait-associated variants, for example, variants indexed in the GWAS Catalog[39]. We then perform stepwise selection to select the subset of independent variants from the known variants list to be used in the conditional analysis. We perform iterative conditional association analysis until the P values of all variants in the known variant list are larger than a cut-off (1×10^{-4}, Online Methods). Instead of adjusting for all known trait-associated variants in the entire chromosome, we adjust for variants in an extended region of the specific variant, for example, a +/− 1-megabase (Mb) window

beyond the variant of interest. Finally, we perform conditional analysis of each variant set by fitting the regression model adjusting for the selected known variants near the variant set (for example, in a +/− 1-Mb window).

STAARpipeline and computation cost

Our R package *STAARpipeline* performs scalable phenotype-genotype association analyses of functionally annotated WGS data using the developed RVAT methods. An additional package, *STAARpipelineSummary* summarizes the rare variant findings generated by *STAARpipeline*, including the results of both unconditional and conditional analysis and visualization of analysis results.

Specifically, to perform RVATs for a given WGS study, we first need to functionally annotate the variants and create variant sets. To achieve this, we use FAVORannotator, a workflow that annotates the variants of a given WGS study using the FAVOR database and generates annotated genotype files for use in *STAARpipeline*. Across the genome, *STAARpipeline* runs gene-centric noncoding and sliding window tests using STAAR and dynamic window analysis using SCANG-STAAR. *STAARpipeline* can also perform RV analysis of coding variants and single variant analysis of common and low-frequency variants (Discussion).

All analyses can be computed with attractive time and memory resources, even for large-scale WGS/WES datasets such as TOPMed, GSP and UK Biobank. We benchmarked *STAARpipeline*'s WGS association analysis of n=30,138 pooled related TOPMed lipids samples including both discovery and replication data in: 15 hours using 200 2.10 GHz computing cores with 11 Gb memory of gene-centric noncoding analysis; or 11 hours using 200 cores with 11 Gb memory of sliding window analysis; or 20 hours using 800 cores with 15 Gb memory of dynamic window analysis (including SCANG-STAAR-S, SCANG-STAAR-B and SCANG-STAAR-O). *STAARpipelineSummary* summarizes the results from *STAARpipeline* and provides analytical follow-up via conditional analysis. Summarizing the genome-wide TOPMed results took 24 hours using one core with 25 Gb memory.

Rare variant association analysis of lipid traits in the TOPMed WGS data

We applied *STAARpipeline* to identify RV-sets associated with four quantitative lipid traits (LDL-C, HDL-C, TG and TC) using TOPMed WGS data[4, 17, 21]. DNA samples were sequenced at the >30X target coverage[4]. The discovery phase consisted of six study cohorts with 21,015 samples sequenced in TOPMed Freeze 5. The replication phase consisted of eight remaining study cohorts with 9,123 samples in TOPMed Freeze 5 (Supplementary Note, Supplementary Table 2). Sample-level and variant-level quality control (QC) procedures were performed[4, 21]. Race/ethnicity was defined using a combination of self-reported race/ethnicity and study recruitment information[40]. The discovery cohorts consisted of 5,849 (27.8%) Black or African American, 12,313 (58.6%) White, 675 (3.2%) Asian American, 1,075 (5.1%) Hispanic/Latino American, and 1,103 (5.3%) Samoan participants. Among all samples in the discovery phase, 3,610 (17.2%) had first degree relatedness, 546 (2.6%) had second degree relatedness, and 472 (2.2%) had third degree relatedness (Supplementary Fig. 1). There were 215 million single-nucleotide variants (SNVs) observed

in the discovery phase, and 205 million (94.9%) were rare variants (MAF < 1%). Among these 205 million rare variants, 202 million (98.8%) were noncoding variants defined by GENCODE VEP. Details of the study-specific demographics, summaries of lipid levels, and variant number distributions are given in Supplementary Tables 2–3 and Extended Data Fig. 1.

For each phenotype, we applied rank-based inverse normal transformation of the phenotype. We adjusted for age, age^2, sex, race/ethnicity, study, and the first 10 ancestral PCs, and controlled for relatedness through heteroscedastic linear mixed models with sparse genetic relatedness matrices (GRMs) plus study-race/ethnicity-specific group-specific residual variance components (Online Methods). We accounted for the presence of medications of LDL-C and TC as before[21]. We tested for an association between lipid traits and RVs (MAF < 1%) in each variant set. In gene-centric analysis, we defined the eight analysis units as previously described: seven noncoding functional categories of protein-coding genes and one category for ncRNA genes. In non-gene-centric analysis, we performed a 2-kb sliding window analysis with 1-kb skip length, and a dynamic window analysis using SCANG-STAAR-S of all moving windows containing 40 to 300 variants[19]. In unconditional analysis we used Bonferroni-corrected genome-wide significance thresholds of $\alpha = 0.05/(20{,}000 \times 7) = 3.57 \times 10^{-7}$ accounting for 7 different noncoding masks across protein-coding genes; $\alpha = 0.05/20{,}000 = 2.50 \times 10^{-6}$ accounting for ncRNA genes, and $\alpha = 0.05/(2.66 \times 10^6) = 1.88 \times 10^{-8}$ accounting for 2.66 million 2-kb sliding windows across the genome. We controlled the genome-wise (family-wise) error rate for SCANG-STAAR-S dynamic window analysis at $\alpha = 0.05$ level[19]. We selected individual variants to be adjusted for in conditional analysis from the list of phenotype-associated common and low-frequency variants (MAF ≥ 1%) indexed in the GWAS Catalog[39]. Then we obtained the independent known variants using the algorithm described before in the analytical follow-up via conditional analysis section (Online Methods, Supplementary Table 4).

In gene-centric noncoding unconditional analysis of the discovery samples, *STAARpipeline* identified 43 genome-wide significant associations with at least one of the four lipid levels (Supplementary Table 5, Extended Data Figs. 2a–d, 3a–d, 4a–d, 5a–d). After conditioning on known lipid-associated variants, 14 out of the 43 associations remained significant at the Bonferroni-corrected level $\alpha = 0.05/43 = 1.16 \times 10^{-3}$ (Table 1). In the replication data, after adjusting for known lipid-associated variants, 4 of these 14 associations achieved significance at Bonferroni-corrected level $\alpha = 0.05/14 = 3.57 \times 10^{-3}$. These included enhancer DHS RVs in *APOA1* and HDL-C, promoter CAGE RVs in *APOE* and TG, and enhancer CAGE or DHS RVs in *APOE* and TG. After further adjustment for known individual rare variants (minor allele count, MAC ≥ 20, Supplementary Table 6), none of the associations remained significant at the same significance level of 3.57×10^{-3} (Supplementary Table 7).

In unconditional analysis of the discovery samples, using the 2-kb sliding window procedure we identified 140 windows as genome-wide significant (Supplementary Table 8, Extended Data Figs. 2e–f, 3e–f, 4e–f, 5e–f). Among these 140 significant sliding windows, 14 were located in noncoding regions and, after conditioning on known lipid-associated variants, all remained significant at the Bonferroni-corrected level $\alpha = 0.05/140 = 3.57 \times 10^{-4}$ (Table 2).

In replication data, 9 of the 14 associations were significant at the Bonferroni-corrected level $\alpha = 0.05/14 = 3.57 \times 10^{-3}$ after adjusting for known phenotype-specific variants. When we further adjusted these 9 associations for known individual variants (MAC ≥ 20), associations for two intronic sliding windows (*PAFAH1B2* and TG) remained significant at the same level of 3.57×10^{-3} (Supplementary Table 9).

We further compared the unconditional *P* values of different tests using the sliding window procedures[41], including burden, SKAT and ACAT-V with only MAFs as the weights, and STAAR-O, which incorporates multiple variant functional annotations. Overall, by dynamically incorporating multiple functional annotations that captures different aspects of variant functions, STAAR-O detected more significant sliding windows, and showed consistently smaller *P* values for top sliding windows compared to existing RVATs without incorporating functional annotations (Supplementary Figs. 2–5). These results suggest that incorporating multiple functional annotations using the STAAR framework can boost the power for WGS RV association analysis.

In unconditional analysis of the discovery samples using the dynamic window procedure SCANG-STAAR-S, we identified 90 genome-wide significant associations (Supplementary Table 10). Among them, 10 were located in noncoding regions and remained significant at the Bonferroni-corrected level $\alpha = 0.05/90 = 5.56 \times 10^{-4}$ after conditioning on known lipid-associated variants (Table 3). In the replication data, after adjusting for known phenotype-specific variants, 7 were significant at the Bonferroni-corrected level $\alpha = 0.05/10 = 5 \times 10^{-3}$. After further adjustment for known individual rare variants (MAC ≥ 20), 3 associations remained significant, including RVs in an intronic region of *PAFAH1B2* and TG, RVs in an intronic region of *SIDT2* and TG, and RVs in an intronic region of *CEP164* and TG (Supplementary Table 11).

Rare variant analysis of five non-lipid traits in the TOPMed WGS data

We further applied *STAARpipeline* to analyzing a broader spectrum of five phenotypes in the TOPMed Freeze 5 WGS data: CRP (n = 22,775)[42], eGFR (n = 23,732)[43], FG (n = 23,859)[44], FI (n = 21,900)[44] and TL (n = 39,742)[45]. Similar to the lipids analysis, for each phenotype, we performed gene-centric analysis, 2-kb sliding window analysis, and dynamic window analysis to detect RV associations in the noncoding genome (Online Methods).

In gene-centric noncoding unconditional analysis, *STAARpipeline* identified 6 genome-wide significant associations, and all 6 associations remained significant at the Bonferroni-corrected level $\alpha = 0.05/6 = 8.33 \times 10^{-3}$ after conditioning on known phenotype-specific variants (Supplementary Table 12, Supplementary Figs. 6a–d, 7a–d, 8a–d, 9a–d, 10a–d). After further adjustment for known individual rare variants using conditional analysis, although the strengths of 5 associations were reduced, all 6 associations remained significant at the same significance level of $\alpha = 8.33 \times 10^{-3}$ (Supplementary Table 12). In 2-kb sliding window unconditional analysis, we identified 19 genome-wide significant associations and 12 of them were in the noncoding genome (Supplementary Table 13, Supplementary Figs. 6e–f, 7e–f, 8e–f, 9e–f, 10e–f). After adjusting for known phenotype-specific variants, all the 12 associations remained significant at the Bonferroni-corrected level $\alpha = 0.05/19 = 2.63 \times$

10^{-3}, and 8 of 12 associations remained significant after further adjusting for known rare variants with MAC ≥ 20 (Supplementary Table 13).

In dynamic window unconditional analysis, we identified 17 genome-wide significant associations and 11 of them were in the noncoding genome (Supplementary Table 14). These 11 associations included 7 non-overlapping noncoding significant associations detected by the sliding window procedure, and 4 associations that were missed by the sliding window procedure. After adjusting for known phenotype-specific variants, all 11 associations remained significant at the Bonferroni-corrected level $\alpha = 0.05/17 = 2.94 \times 10^{-3}$, and 8 of 11 associations remained significant after further adjusting for known rare variants (Supplementary Table 14).

Simulation studies

We performed simulation studies to evaluate the type I error rate and power of SCANG-STAAR in a variety of configurations. We generated sequence data by simulating 100,000 chromosomes in a 10-Mb region using the calibration coalescent model (COSI)[46] that mimics the linkage disequilibrium (LD) structure of samples from African Americans. The simulation studies used the 10-Mb sequence to mimic whole genome sequencing data and focused on rare variants (MAF < 1%). We considered the total sample sizes $n = 50{,}000$ in all simulations. Quantitative and dichotomous phenotypes were generated by following the steps described in Data simulation (Online Methods).

Type I error simulations

For both quantitative and dichotomous traits, we performed 10,000 simulations using SCANG-STAAR-S, SCANG-STAAR-B and SCANG-STAAR-O to analyze a 10-Mb genome, and evaluated the empirical genome-wise (family-wise) type I error rates at nominal $\alpha = 0.05$ and 0.01 (Supplementary Table 15). The results show that all the three tests based on SCANG-STAAR provide a good control of the type I error rates for both continuous and dichotomous traits at the two α levels.

Empirical power simulations

We then compared the empirical power of SCANG-STAAR with the existing methods, including the sliding window procedures using burden[15], SKAT[16], SKAT-O[47] and STAAR[17], and the dynamic window procedure using SCANG[19] at genome-wise (family-wise) error rate $\alpha = 0.01$ level with 1,000 replicates. The genome-wise (family-wise) type I error rate was controlled using the empirical threshold for SCANG-STAAR and SCANG, and the Bonferroni correction for the sliding window procedures. We randomly selected two signal regions (variant-phenotype association regions) across the 10-Mb genome in each replicate. The lengths of the signal regions were randomly selected from lengths of 1 kb, 1.5 kb and 2 kb. We considered the proportions of causal variants is 15% on average among the signal regions, and the probability that variants are causal was allowed to be dependent on different sets of annotations through a logistic model, of which five were informative and the other five were non-informative. All the 10 annotations were used in SCANG-STAAR and STAAR. In order to evaluate power, we considered two criteria, causal variant detection rate and signal region detection rate[19] (Online Methods).

For both quantitative and dichotomous traits, SCANG-STAAR had a higher power than the 1-kb and 2-kb sliding window procedure using burden, SKAT, SKAT-O and STAAR in terms of both causal variant detection rate and signal region detection rate across different proportions of effect size directions (Supplementary Figs. 11–16). Our simulation studies indicate that SCANG-STAAR improves power by flexibly detecting the locations and the sizes of signal regions. In addition, SCANG-STAAR had a higher power than SCANG for both causal variant detection rate and signal region detection rate (Supplementary Figs. 17–18). Our simulation studies indicate that SCANG-STAAR improves power by incorporating informative variant functional annotations.

Discussion

We developed a comprehensive association analysis framework for detecting noncoding rare variant set associations in large-scale WGS studies by defining a variety of noncoding variant sets and incorporating multi-faceted variant functional annotations. Our approach allows for analyzing both continuous and binary traits and accounts for both population structure and relatedness using generalized linear mixed models in gene-centric analysis and non-gene-centric analysis. It could further account for the stratification of recent population structure using the principal components calculated from RVs through the regression framework[48]. For gene-centric analysis, we proposed several strategies to define analysis units of RVs in the noncoding genome, including seven functional categories of regulatory regions for protein-coding genes, ncRNA genes, and performed RVATs of each noncoding mask using STAAR. For non-gene-centric analysis, to overcome the limitations of fixed-size sliding windows, we proposed SCANG-STAAR, a data-adaptive-size dynamic window scan procedure that incorporates multi-faceted functional annotations. We proposed *STAARpipeline* to perform RVATs using these methods for both noncoding and coding variants in unconditional analysis and conditional analyses, which provides an analytical follow-up to distinguish novel RV association signals independent of known variants.

STAARpipeline is a fast and resource-efficient tool for RV association analysis of WGS data that scales linearly on hundreds of thousands of samples.

STAARpipeline allows researchers to conveniently functionally annotate a WGS/WES study using the variant functional annotation database FAVOR and the FAVORannotator workflow. *STAARpipeline* optimizes computational feasibility of RV association analysis in two steps. First, *STAARpipeline* reduces the computation burden of fitting the null mixed model using the estimated sparse GRM[17, 49]. Second, *STAARpipeline* performs the RV association tests by taking advantage of sparse genotype dosages of RVs[50].

We demonstrated the power gain of *STAARpipeline* over the existing approaches in the data analysis of 9 traits from TOPMed Freeze 5. First, *STAARpipeline* detected 49 significant associations in gene-centric noncoding analysis, and 35 associations (71.4%) were detected by the 6 newly proposed noncoding masks (Supplementary Tables 16). Second, the proposed dynamic window analysis procedure SCANG-STAAR detected 43 non-overlapped significant noncoding associations in the noncoding genome, which was 19.4% more than the existing 2-kb sliding window procedure that detected 36 non-

overlapped significant noncoding associations (Supplementary Tables 17). In addition, SCANG-STAAR only missed one non-overlapped associations detected by 2-kb sliding window procedure (Supplementary Tables 8, 10, 13–14).

In the WGS RV analysis of lipid traits in TOPMed, we identified and replicated using our *STAARpipeline* several conditional associations in the noncoding genome, including RVs in an intronic region of *PAFAH1B2* and TG, RVs in an intronic region of *SIDT2* and TG, and RVs in an intronic region of *CEP164* and TG, which were not detected by previous analysis of TOPMed Freeze 3[17, 21]. Several coding rare variants in *PAFAH1B2* have been previously detected associated with TG[51], our findings detected additionally significant RV association in the noncoding region of *PAFAH1B2*. Two intronic common variants in *SIDT2* have been reported associated with TG[52], additional intronic rare variant association in *SIDT2* was detected using *STAARpipeline*.

Since SCANG-STAAR considers many more overlapping windows than the sliding window procedure, the genome-wide significance threshold is smaller than that of the sliding window procedure. For example, to control the genome-wise error rate at 0.05 level in our analysis of LDL-C, the *P* value threshold of SCANG-STAAR-S was 3.80×10^{-9} while the Bonferroni-corrected threshold of the 2-kb sliding window procedure was 1.88×10^{-8}. When the window size of the signal region is close to the sliding window size, the sliding window procedure may detect associations missed by the dynamic window procedure because of this gap of the *P* value thresholds. In *STAARpipeline* we pragmatically provide both procedures.

In addition to noncoding rare variants association analysis, *STAARpipeline* also provides single variant analysis for common and low-frequency variants and gene-centric analysis for coding rare variants. The single variant analysis in *STAARpipeline* provides individual *P* values of variants given a MAF or MAC cut-off, for example, MAC ≥ 20. The gene-centric coding analysis provides five functional categories to aggregate coding rare variants of each protein-coding gene: (1) putative loss of function (stop gain, stop loss and splice) RVs, (2) missense RVs, (3) disruptive missense RVs, (4) putative loss of function and disruptive missense RVs, and (5) synonymous RVs. The putative loss of function, missense, and synonymous RVs are defined by GENCODE VEP categories[29,30]. The disruptive variants are further defined by MetaSVM[53], which measures the deleteriousness of missense mutations. As in the noncoding RV association analysis, single variant and gene-centric coding analyses also scale well in computation time and memory for large-scale WGS data. Using 30,138 related TOPMed samples these two analyses respectively took 3 hours and 5 hours for 100 cores with 6 Gb memory. Thus, *STAARpipeline* provides an efficient and comprehensive analysis tool for both coding and noncoding variant association discovery in large-scale sequencing studies.

With the emergence of large-scale WGS data, there is a pressing need to identify genetic components of complex traits in the noncoding genome. Here we introduce a powerful and scalable framework, *STAARpipeline*, for noncoding RV association detection across the genome. *STAARpipeline* provides several strategies to aggregate noncoding rare variants to empower RV association analysis in the noncoding region. We demonstrate the

computational efficiency of *STAARpipeline* in application to the WGS association analysis of a range of traits up to ~40,000 TOPMed samples. The optimization approaches of *STAARpipeline* make it scalable for even larger data sets. Thus, our framework provides an essential solution for noncoding RV association detection in large-scale WGS data analysis and dissects the genetic contribution of noncoding rare variants to complex diseases.

Methods

Notations and model

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

We consider the Generalized Linear Model for unrelated samples,

$$g(\mu_i) = \alpha_0 + \boldsymbol{X}_i^T\boldsymbol{\alpha} + \boldsymbol{G}_i^T\boldsymbol{\beta}, \quad (1)$$

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

We consider the following Generalized Linear Mixed Model[27, 28, 54] for related samples,

$$g(\mu_i) = \alpha_0 + \boldsymbol{X}_i^T\boldsymbol{\alpha} + \boldsymbol{G}_i^T\boldsymbol{\beta} + b_i, \quad (2)$$

where the random effects b_i account for remaining population structure unaccounted by ancestral principal components and relatedness. Let $\mathbf{b} = (b_1, \ldots, b_n)^T \sim N(\mathbf{0}, \theta\boldsymbol{\Phi})$ with variance components θ and a genetic relatedness matrix $\boldsymbol{\Phi}$[17, 49]. Our goal is testing the null hypothesis of whether the variant-set is associated with the phenotype, adjusting for covariates and relatedness, which corresponds to H_0: $\boldsymbol{\beta} = \mathbf{0}$, that is, $\beta_1 = \beta_2 = \cdots = \beta_p = 0$.

Variant set test using STAAR

The *STAARpipeline* calculates the variant set P value of each analysis unit using the STAAR method that incorporates multiple variant functional annotation scores[17]. Assume there are K annotations and $\hat{\pi}_{jk} = \frac{rank(A_{jk})}{M}$, where A_{jk} is the kth annotation for the jth variant ($k = 1, \cdots, K$; $j = 1, \cdots, p$). For $k = 0$, we assume $\hat{\pi}_{j0} = 1$. Assume $w_{jl} = Beta(\text{MAF}_j; a_{1l}, a_{2l})$, where $(a_{11}, a_{21}) = (1,25)$, $(a_{12}, a_{22}) = (1,1)$ and MAF_j is the MAF of the jth variant ($j = 1, \cdots, p$). The burden test statistic using kth variant functional annotation and lth beta density as the weight is given by

$$Q_{Burden,l,k} = \left(\sum_{j=1}^{p} \hat{\pi}_{jk} w_{jl} S_j\right)^2.$$

The SKAT test statistic using kth variant functional annotation and lth beta density as the weight is given by

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

($k = 0, \cdots, K$; $l = 1,2$). The ACAT-V test statistic using kth variant functional annotation and lth beta density as the weight is given by

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

where $\overline{\hat{\pi}_{\cdot k} w_{\cdot l}^2 \mathrm{MAF}(1-\mathrm{MAF})}$ is the average of the weights $\hat{\pi}_{jk} w_{jl}^2 \mathrm{MAF}_j(1-\mathrm{MAF}_j)$ among the extremely rare variants with MAC ≤ 10, and p' is the number of variants with MAC > 10 in the variant set.

Let $p_{Burden,l,k}$ be the P value of $Q_{Burden,l,k}$, $p_{SKAT,l,k}$ be the P value of $Q_{SKAT,l,k}$, and $p_{ACAT-V,l,k}$ be the P value of $Q_{ACAT-V,l,k}$ ($k = 0, \cdots, K$; $l = 1,2$). We define STAAR-Burden (STAAR-B), STAAR-SKAT (STAAR-S), and STAAR-ACAT-V (STAAR-A) as $T_{STAAR-test} = \sum_{l=1}^{2} \sum_{k=0}^{K} \frac{tan\{(0.5 - p_{test,l,k})\pi\}}{2(K+1)}$, and the corresponding P value is calculated by $p_{STAAR-test} \approx \frac{1}{2} - \frac{\{arctan(T_{STAAR-test})\}}{\pi}$, where $test \in \{Burden, SKAT, ACAT-V\}$. The STAAR-O test statistic is defined as

$$T_{STAAR-O} = \frac{1}{3}[tan\{(0.5 - p_{STAAR-Burden})\pi\} + tan\{(0.5 - p_{STAAR-SKAT})\pi\} + tan\{(0.5 - p_{STAAR-ACAT-V})\pi\}]$$

and the corresponding P value is calculated by

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

In gene-centric and sliding window analysis, we use the STAAR-O test for each analysis unit.

Dynamic window analysis using SCANG-STAAR

The *STAARpipeline* performs dynamic window analysis using the SCANG-STAAR procedure, which extends the dynamic window rare variant test procedure SCANG by incorporating multiple variant functional annotations using the STAAR method. Under the global null hypothesis, there is no variant associated with the phenotype across the

genome. Under the alternative hypothesis, there exists at least one region associated with the phenotype. SCANG-STAAR procedure provides a valid test by using the minimum value of the P value of all candidate moving windows of different sizes

$$p_{min} = \min_{L_{min} \le |I| \le L_{max}} p(I),$$

where $p(I)$ is the P value of region I, $|I|$ is the number of variants in a window I, and L_{min} and L_{max} are the smallest and largest number of variants in the searching windows, respectively. For SCANG-STAAR-S and SCANG-STAAR-B procedures, $p(I)$ is the STAAR-S and STAAR-B P value of window I, respectively. For SCANG-STAAR-O, (I) is the omnibus P value of STAAR-S and STAAR-B calculated by ACAT method[38]. Similar to the SCANG procedure, SCANG-STAAR controls the genome-wise type I error rate at a given α level by using the αth quantile of the empirical distribution of p_{min} as an empirical threshold $h(\alpha, p_{min}, L_{min}, L_{max})$[19]. We reject the null hypothesis if the P value of any window is smaller than $h(\alpha, p_{min}, L_{min}, L_{max})$. If this results in only one window, the detected window is $\hat{I} = \text{argmin}_{L_{min} \le |I| \le L_{max}} p(I)$. If this results in multiple overlapping windows, we localize the signals as the window whose P value is smaller than both the threshold and the windows that overlap with it.

Conditional analysis

The *STAARpipeline* performs conditional analysis to identify RV association independent of known variants. We first select a list of known variants by including the trait-associated variants identified in literature, for example, variants indexed in GWAS Catalog[39] or significant variants in large-scale GWAS. The significant variants detected in individual analysis using the same data could also be added into the known variants list to ensure the RV signals are not captured by the significant individual variants. We then use the following stepwise selection strategy to select a subset of independent variants representing the known variant list as the variants adjusted in the conditional analysis:

1. Calculate the individual P value of all variants in the known variants list and select the most significant variant.
2. For each step, calculate the P values of all the remaining variants conditional on the variant(s) that have already been selected. For each variant, we only condition on the selected variants within a specified region of that variant, such as the +/− 1-Mb window.
3. Select the variant with minimum conditional P value that is lower than the cutoff P value, for example, 1×10^{-4}.
4. Repeat steps 2–3 until no variants can be selected.

Finally, we calculate the conditional P value of each significant RV analysis unit by adjusting for the selected variants residing in an extended region (for example, +/− 1-Mb window) of the analysis unit.

Statistical analysis of lipid traits in the TOPMed data

The TOPMed WGS data consist of ancestrally diverse and multi-ethnic related samples[4]. Race/ethnicity was defined using a combination of self-reported race/ethnicity and study recruitment information (Supplementary Note)[40]. The discovery cohorts consist of 5,849 (27.8%) Black or African American, 12,313 (58.6%) White, 675 (3.2%) Asian American, 1,075 (5.1%) Hispanic/Latino American and 1,103 (5.3%) Samoans. The replication cohorts consist of 2,265 (24.8%) Black or African American, 5,615 (61.5%) White, and 1,243 (13.6%) Hispanic/Latino American.

We applied *STAARpipeline* to identify RV sets associated with four quantitative lipid traits (LDL-C, HDL-C, TG and TC) using the TOPMed WGS data. LDL-C and TC were adjusted for the presence of medications as before[21]. Linear regression model adjusting for age, age^2, sex was first fit for each study-race/ethnicity-specific group. In addition, for Old Order Amish, 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[21]. The residuals were rank-based inverse normal transformed and rescaled by the standard deviation of the original phenotype within each group. We then fit a heteroscedastic linear mixed model (HLMM) for the rank normalized residuals, adjusting for 10 ancestral PCs, study-ethnicity group indicators, and a variance component for empirically derived kinship matrix plus separate group-specific residual variance components to account for population structure and relatedness. The output of HLMM was then used to perform following variant set analyses for rare variants (MAF < 1%) by scanning the genome, including gene-centric analysis using seven variant categories (promoter RVs overlaid with CAGE sites, promoter RVs overlaid with DHS sites, enhancer RVs overlaid with CAGE sites, enhancer RVs overlaid with DHS sites, UTR RVs, upstream RVs and downstream RVs) for each protein coded gene, ncRNA RVs, 2-kb sliding windows with 1-kb skip length, and dynamic windows with variants number between 40 and 300. The WGS RVAT analysis was performed using R packages STAAR (version 0.9.6), STAARpipeline (version 0.9.6) and STAARpipelineSummary (version 0.9.6).

Rare variant association analysis of CRP, eGFR, FG, FI, and TL in the TOPMed data

We applied *STAARpipeline* to identify RV sets associated with five non-lipid traits from the 14 cohorts in TOPMed Freeze 5, including CRP of 22,775 individuals, eGFR of 23,732 individuals, FG of 23,859 individuals, FI of 21,900 individuals and TL of 39,742 individuals (Supplementary Note). These five traits were defined the same as in the previous studies[42–45]. For CRP and FI, we additionally performed log-transformation of the trait in the analysis[42, 44]. For each trait, we first fit a linear regression model adjusting for age and sex for each study-race/ethnicity group, with additional adjustment of age^2 for CRP, age^2 and body mass index (BMI) for FG and FI, and sequencing center and 10 ancestral PCs for TL[42–45]. The residuals were transformed using the rank-based inverse normal transformation and rescaled by the standard deviation of the original phenotype within each study-race/ethnicity group. We then fit a heteroscedastic linear mixed model (HLMM) for the rank normalized residuals, adjusting for 10 ancestral PCs, study-ethnicity group indicators, and a variance component for empirically derived kinship matrix plus separate group-specific residual variance components to account for population structure and

relatedness. We additionally adjusted for age, sex, and sequencing center for TL. The output of HLMM was then used in the RV association analysis of the *STAARpipeline*.

In gene-centric noncoding analysis, *STAARpipeline* identified 6 conditionally significant associations with at least one of the five traits compared with the previous analyses[42–45]. These included promoter CAGE or enhancer DHS RVs in associating *CRP* and CRP, ncRNA RVs in *CTC-523E23.15* and FI, enhancer CAGE or DHS RVs in *TINF2* and TL, enhancer DHS RVs in *MRVI1* and TL (Supplementary Table 12).

In non-gene-centric noncoding analysis using 2-kb sliding windows, we identified 8 conditionally significant associations with at least one of the five traits. These included associations for 2 intergenic sliding windows near *CRP* and CRP, 2 intergenic sliding windows near *AC073409.1* and TL, an intronic sliding window in *TERT* and TL, an intergenic sliding window near *RNGTT* and TL, and 2 intronic sliding windows in *ZGPAT* and TL (Supplementary Table 13). We also identified 8 conditionally significant associations in the noncoding genome with at least one of the five traits in dynamic window analysis. These included associations for 2 intergenic regions near *CRP* and CRP, an intergenic region near *AC073409.1* and TL, an intronic region in *MLIP* and TL, an intronic region in *TERT* and TL, an intergenic region near *RNGTT* and TL, an intronic region in *NOS1* and TL, and an intronic region in *ZGPAT* and TL (Supplementary Table 14). Note that the associations between RVs in the intronic region of *MLIP* or *NOS1* with TL are missed by the 2-kb sliding window analysis.

Data simulation

Type I error rate simulations—We performed extensive simulation studies to show that the proposed SCANG-STAAR method controls the genome-wise (family-wise) type I error rate. We generated genotypes by simulating 100,000 sequencing chromosomes for a 10-Mb region that represent the whole genome. The data were generated to mimic the LD structure of an African American population by using the calibration coalescent model (COSI)[46]. We considered the total sample sizes $n = 50{,}000$ in all simulations. We generated continuous traits from a linear model

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

where $X_{1i} \sim N(0,1)$, $X_{2i} \sim$ Bernoulli(0.5), and $\epsilon_i \sim N(0,1)$. We generated dichotomous traits from a logistic model

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

where X_{1i} and X_{2i} were defined the same as continuous traits and α_0 was determined to set the prevalence to 1%. We used case-control sampling. In each simulation replicate, 10 annotations were generated as $A_1, \ldots, A_{10}$ i.i.d. $N(0,1)$ for each variant. For the size simulation, under the null, these annotations are not associated with the regression coefficients of the phenotype models. We applied SCANG-STAAR-S, SCANG-STAAR-B, SCANG-STAAR-O by incorporating the MAF and the 10 annotations as weights and

repeated the procedure with 10,000 replicates to examine the genome-wise (family-wise) type I error rates for both continuous and dichotomous traits at α = 0.05 and 0.01 levels.

Empirical power simulations—Next, we carried out simulation studies to assess the power gain of SCANG-STAAR compared to the existing methods, including the sliding window procedures using burden[15], SKAT[16], SKAT-O[47] and STAAR[17], and the dynamic window procedure using SCANG[19], with 1,000 replicates. In each simulation replicate, we randomly selected two signal regions (variant-phenotype association regions) across the 10-Mb genome for power simulations, where the length of the signal regions was randomly selected from 1 kb, 1.5 kb and 2 kb. For each signal region, causal variants were generated according to a logistic model

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

where five annotations $\{k_1, \cdots, k_5\} \subset \{1, \cdots, 10\}$ were randomly sampled for each region. This assumes the probability of a variant being causal is a function of five randomly selected annotations. Note that we generated 10 functional annotations for each variant and used 5 to determine the probability of causal variants. For RVATs using SCANG-STAAR, we used all annotations, including 5 informative annotations and 5 non-informative annotations. For different regions, causality of variants was allowed to be dependent on different sets of annotations. We set $\delta_{k_l} = \log(5)$ for all annotations and $\delta_0 = \text{logit}(0.015)$, resulting in 15% causal variants on average in signal regions.

We generated continuous traits from a linear model given by

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

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

$$\text{logit}P(Y_i = 1) = \alpha_0 + 0.5X_{1i} + 0.5X_{2i} + \beta_1 G_{1j} + \cdots + \beta_s G_{sj},$$

where α_0, X_{1i}, X_{2i} were defined as in the type I error rate simulations, $G_{1j}, \ldots, G_{sj}$ were the genotypes of the s randomly selected causal variants in the signal region, and $\beta_1, \ldots, \beta_s$ were the corresponding log odds ratios (ORs) of the s causal variants. We used case-control sampling. For both models, we set the effect sizes of causal variants as a decreasing function of MAFs, $c_0 = c_0|\log_{10}MAF_j|$; for continuous trait, $c_0 = 0.10$, and for dichotomous traits, $c_0 = 0.14$, which gives an odds ratio of 2 for a variant with MAF 1×10^{-5}. For each region, we varied the proportions of causal variant effect size directions by setting 50%, 80% or 100% variants with positive effects.

We applied SCANG-STAAR, SCANG, and 1-kb and 2-kb sliding window methods using burden, SKAT, SKAT-O and STAAR. We repeated the procedure with 1,000 replicates to

examine power at genome-wise (family-wise) error rate $\alpha = 0.01$ level. The genome-wise (family-wise) type I error rate was controlled using the empirical threshold for SCANG-STAAR and SCANG, and the Bonferroni correction for the sliding window procedures. For SCANG-STAAR and sliding window methods using STAAR, we incorporated all 10 annotations in the weighting scheme, including the 5 annotations that are associated with the variant being causal and the 5 annotations that are not.

To evaluate power, we considered two criteria, causal variant detection rate and signal region detection rate. The causal variant detection rate can be regarded as the power of causal variants detection. The signal region detection rate can be regarded as the power of signal regions detection. The causal variant detection rate is defined as

$$Causal\ Variant\ Detection\ Rate = \frac{Number\ of\ \text{detected}\ causal\ \text{variants}}{Total\ number\ of\ causal\ \text{variants}}.$$

We define a causal variant as “detected” if it is in one of the detected signal regions. The signal region detection rate is defined as

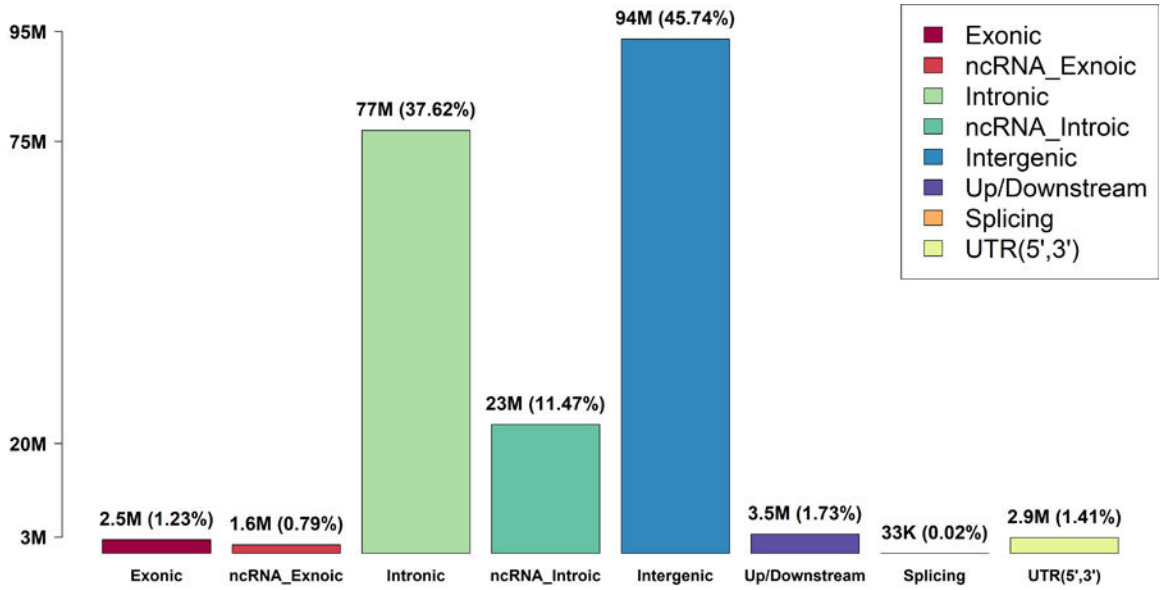
$$Signal\ Region\ Detection\ Rate = \frac{Number\ of\ detected\ signal\ regions}{Total\ number\ of\ signal\ regions}.$$

We define the signal region as “detected” if it overlaps with one of the detected signal regions. Both the causal variant detection rate and the signal region detection rate can be regarded as measures of the test’s power.

We also did a sensitivity analysis where all ten annotations were uninformative. Specifically, 15% of variants within each signal region were randomly chosen as causal variants without using the annotation information. SCANG-STAAR had a similar performance to SCANG, and had a higher power than the sliding window methods with fixed window sizes in terms of both causal variant detection rate and signal region detection rate (Supplementary Figs. 19–20). Our simulation results indicate that SCANG-STAAR is robust to the noninformative annotations and improves power by flexibly detecting the locations and the sizes of signal regions when functional annotations are informative.

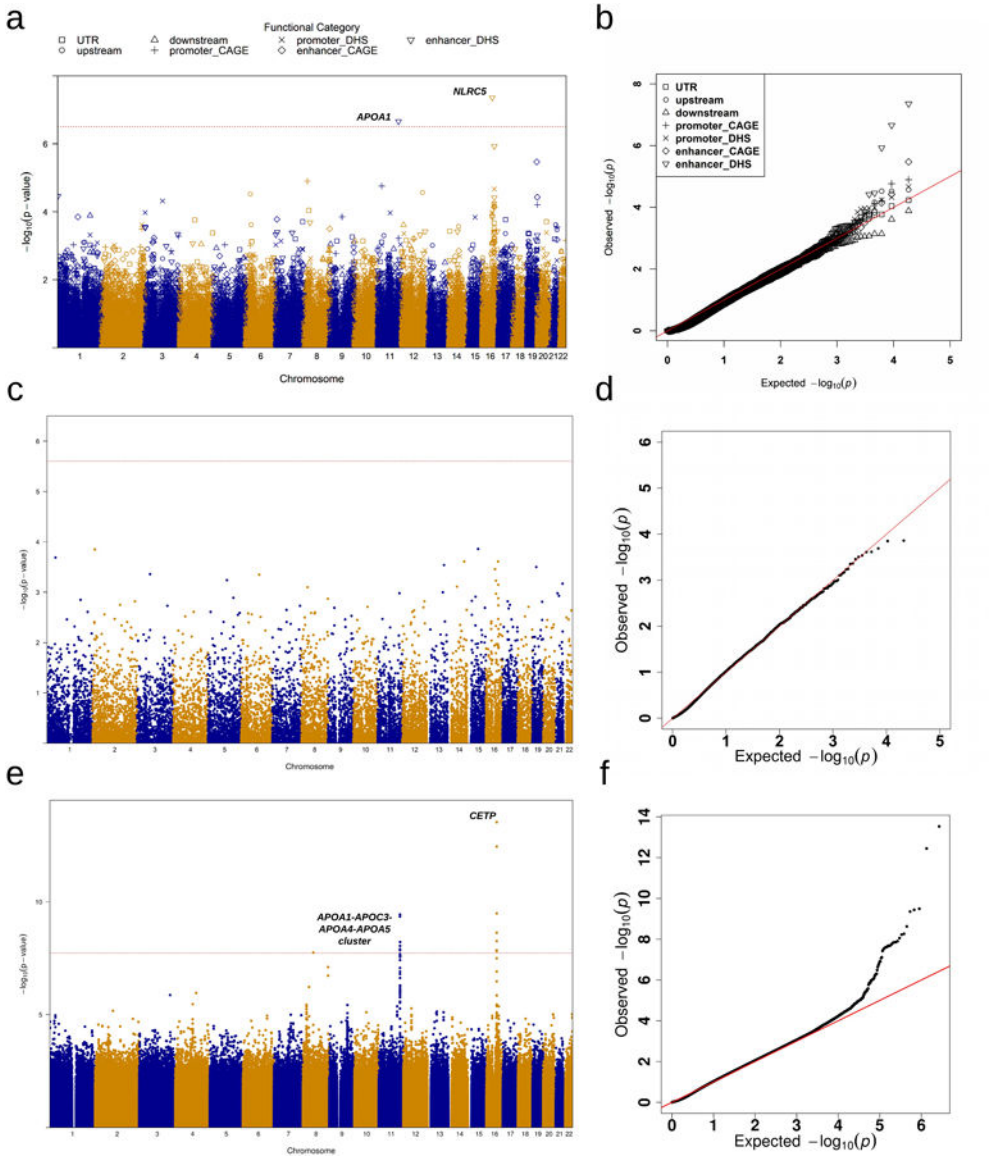
Genome build—All genome coordinates are given in NCBI GRCh38/UCSC hg38.

Extended Data



Extended Data Fig. 1|. Rare variant (MAF < 0.01) distribution in the discovery phase using TOPMed cohorts (n=21,015).

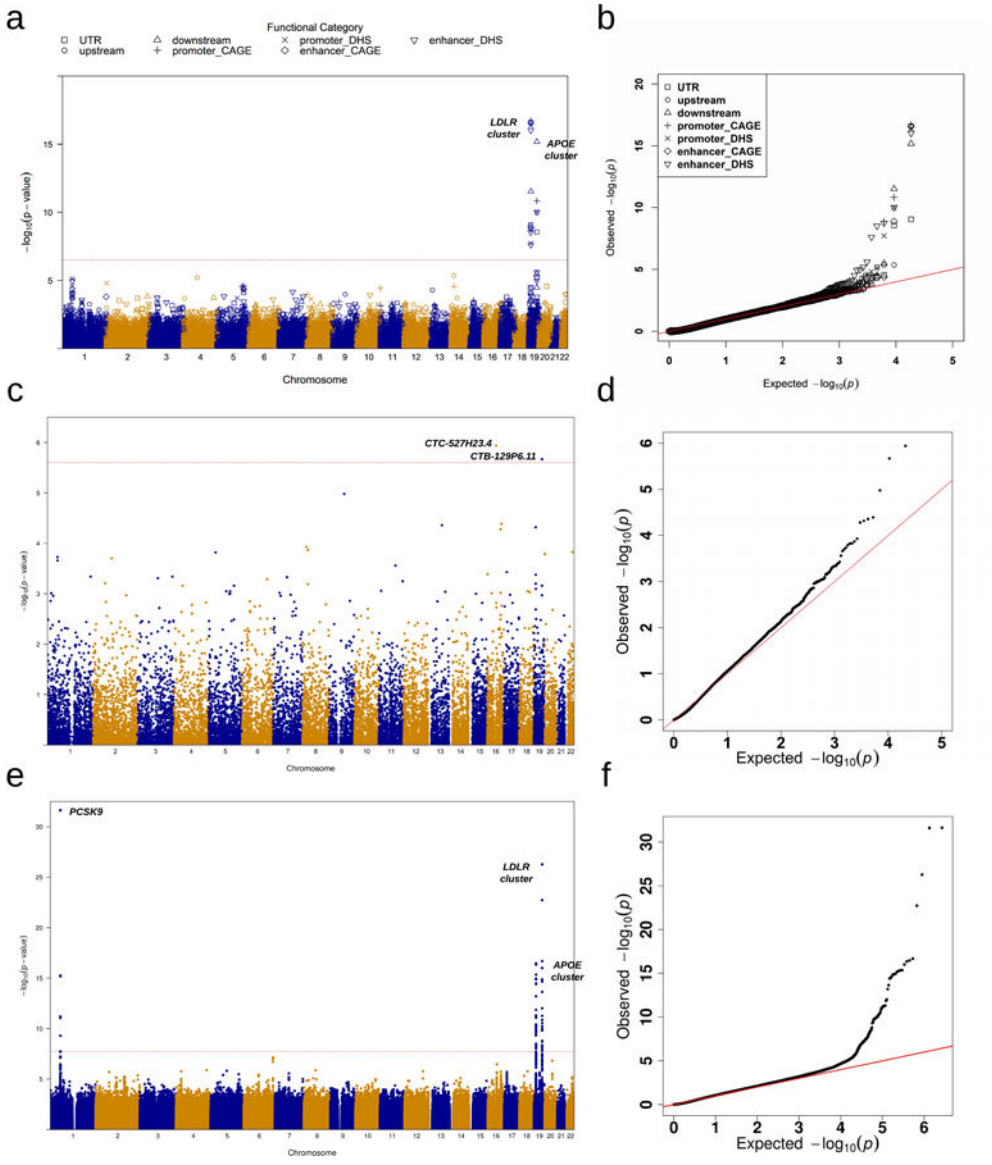
Variant categories are defined by GENCODE VEP categories.



Extended Data Fig. 2|. Manhattan plots and Q-Q plots for unconditional gene-centric noncoding analysis and sliding window analysis of high-density lipoprotein cholesterol (HDL-C) in the discovery phase (n=21,015).

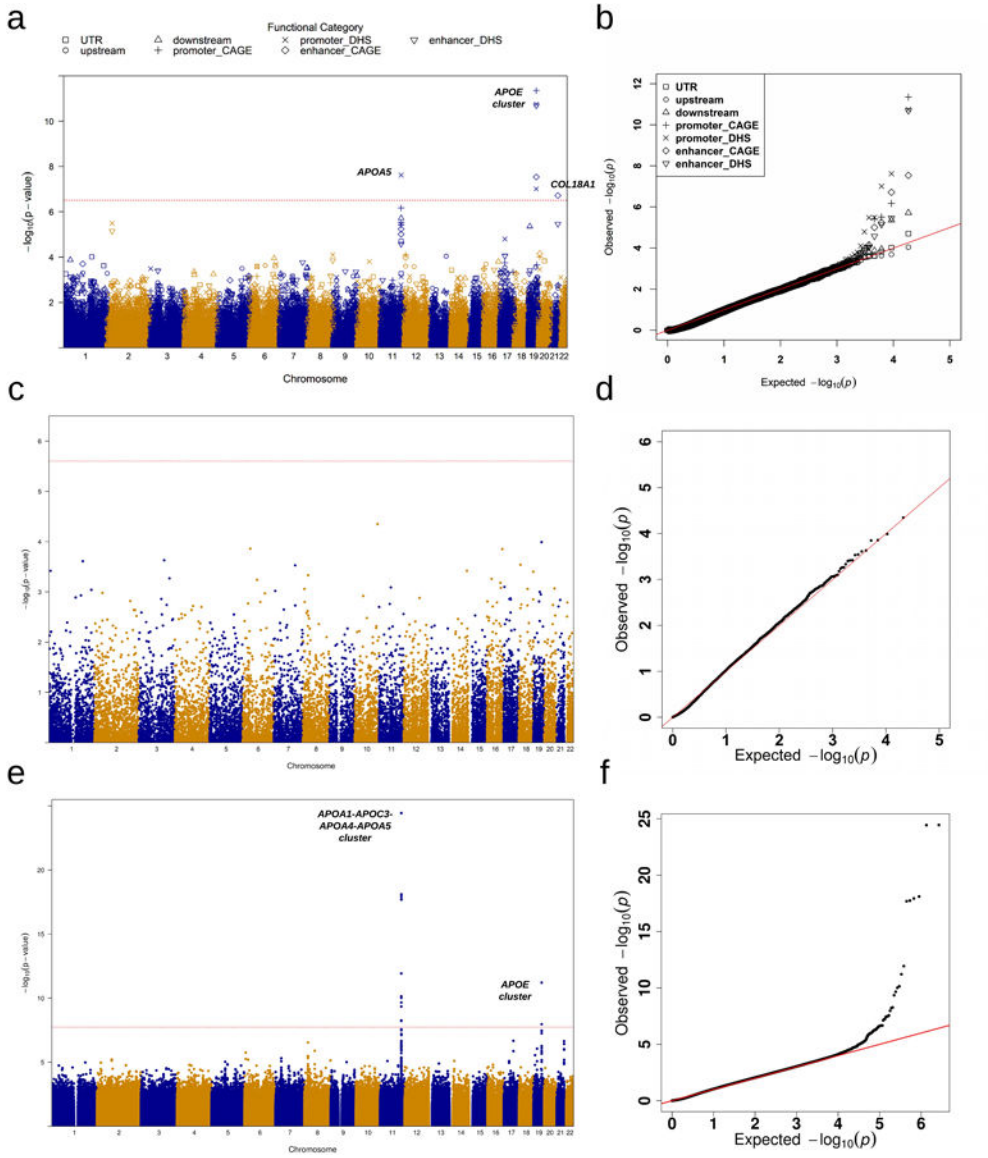
a, Manhattan plots for unconditional gene-centric noncoding analysis of protein-coding gene. The horizontal line indicates a genome-wide STAAR-O *P* value threshold of 3.57×10^{-7}. The significant threshold is defined by multiple comparisons using the Bonferroni correction $(0.05/(20,000 \times 7) = 3.57 \times 10^{-7})$. Different symbols represent the STAAR-O *P* value of the protein-coding gene using different functional categories (upstream, downstream, UTR, promoter_CAGE, promoter_DHS, enhancer_CAGE, enhancer_DHS). Promoter_CAGE and promoter_DHS are the promoters with overlap of Cap Analysis of Gene Expression (CAGE) sites and DNase hypersensitivity (DHS) sites for a given gene, respectively. Enhancer_CAGE and enhancer_DHS are the enhancers in GeneHancer predicted regions with the overlap of CAGE sites and DHS sites for a given gene, respectively. **b**, Quantile-quantile plots for unconditional gene-centric noncoding

analysis of protein-coding gene. Different symbols represent the STAAR-O *P*-value of the gene using different functional categories (upstream, downstream, UTR, promoter_CAGE, promoter_DHS, enhancer_CAGE, enhancer_DHS). **c,** Manhattan plots for unconditional gene-centric noncoding analysis of ncRNA gene. The horizontal line indicates a genome-wide STAAR-O *P* value threshold of 2.50×10^{-6}. The significant threshold is defined by multiple comparisons using the Bonferroni correction $(0.05/20,000 = 2.50 \times 10^{-6})$. **d,** Quantile-quantile plots for unconditional gene-centric noncoding analysis of ncRNA gene. **e,** Manhattan plot for 2-kb sliding windows. The horizontal line indicates a genome-wide *P* value threshold of 1.88×10^{-8}. The significant threshold is defined by multiple comparisons using the Bonferroni correction $(0.05/(2.66 \times 10^{6}) = 1.88 \times 10^{-8})$. **f**, Quantile-quantile plot for 2-kb sliding windows. In panels, **a**, **c** and **e**, the chromosome number are indicated by the colors of dots. In all panels, STAAR-O is a two-sided test.



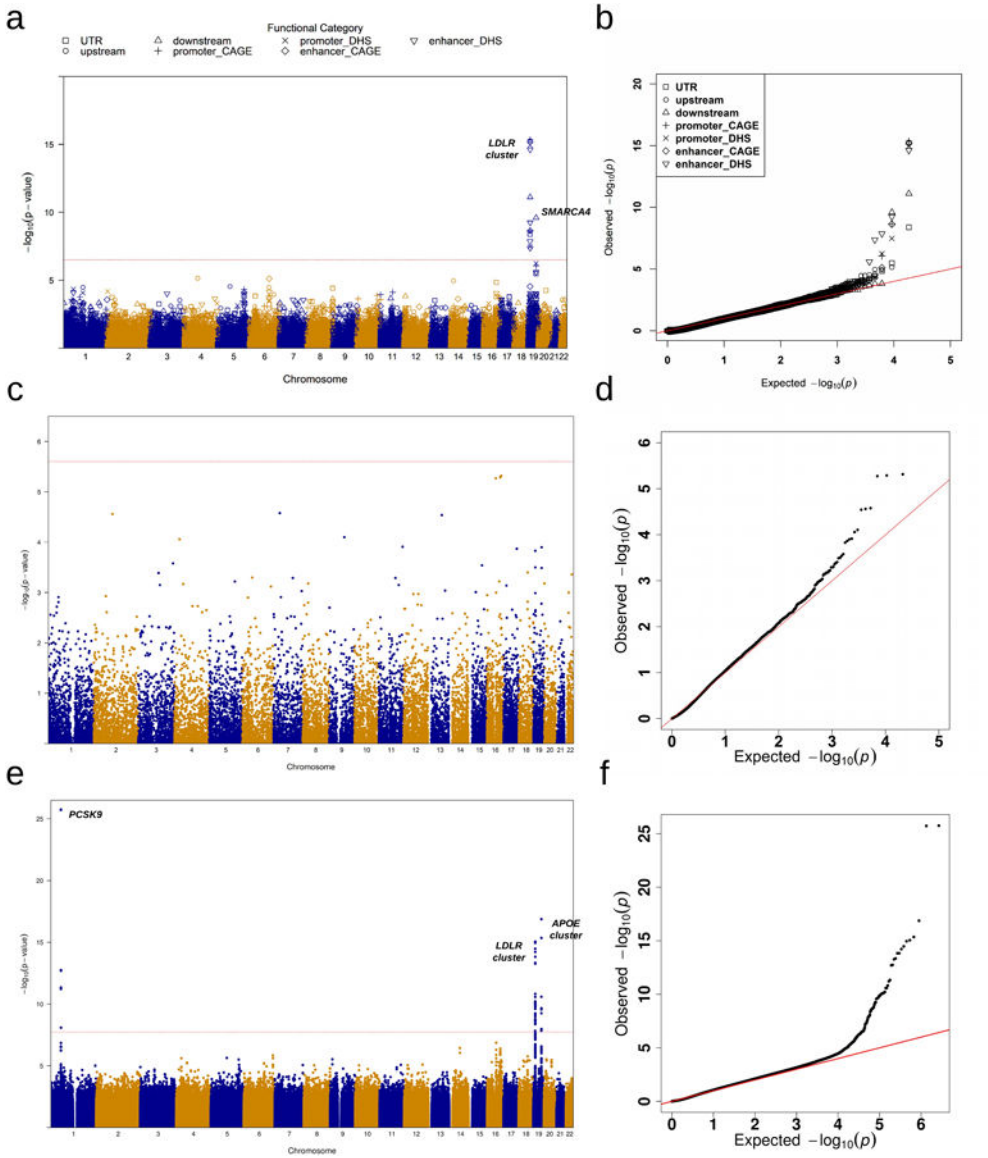
Extended Data Fig. 3|. Manhattan plots and Q-Q plots for unconditional gene-centric noncoding analysis and sliding window analysis of low-density lipoprotein cholesterol (LDL-C) in the discovery phase (n=21,015).

a, Manhattan plots for unconditional gene-centric noncoding analysis of protein-coding gene. The horizontal line indicates a genome-wide STAAR-O *P*-value threshold of 3.57×10^{-7}. The significant threshold is defined by multiple comparisons using the Bonferroni correction $(0.05/(20,000 \times 7) = 3.57 \times 10^{-7})$. Different symbols represent the STAAR-O *P*-value of the protein-coding gene using different functional categories (upstream, downstream, UTR, promoter_CAGE, promoter_DHS, enhancer_CAGE, enhancer_DHS). Promoter_CAGE and promoter_DHS are the promoters with overlap of Cap Analysis of Gene Expression (CAGE) sites and DNase hypersensitivity (DHS) sites for a given gene, respectively. Enhancer_CAGE and enhancer_DHS are the enhancers in GeneHancer predicted regions with the overlap of CAGE sites and DHS sites for a given gene, respectively. **b**, Quantile-quantile plots for unconditional gene-centric noncoding analysis of protein-coding gene. Different symbols represent the STAAR-O *P*-value of the gene using different functional categories (upstream, downstream, UTR, promoter_CAGE, promoter_DHS, enhancer_CAGE, enhancer_DHS). **c,** Manhattan plots for unconditional gene-centric noncoding analysis of ncRNA gene. The horizontal line indicates a genome-wide STAAR-O *P*-value threshold of 2.50×10^{-6}. The significant threshold is defined by multiple comparisons using the Bonferroni correction $(0.05/20,000 = 2.50 \times 10^{-6})$. **d,** Quantile-quantile plots for unconditional gene-centric noncoding analysis of ncRNA gene. **e,** Manhattan plot for 2-kb sliding windows. The horizontal line indicates a genome-wide *P*-value threshold of 1.88×10^{-8}. The significant threshold is defined by multiple comparisons using the Bonferroni correction $(0.05/(2.66 \times 10^{6}) = 1.88 \times 10^{-8})$. **f**, Quantile-quantile plot for 2-kb sliding windows. In panels, **a**, **c** and **e**, the chromosome number are indicated by the colors of dots. In all panels, STAAR-O is a two-sided test.



Extended Data Fig. 4|. Manhattan plots and Q-Q plots for unconditional gene-centric noncoding analysis and sliding window analysis of triglycerides (TG) in the discovery phase (n=21,015). **a**, Manhattan plots for unconditional gene-centric noncoding analysis of protein-coding gene. The horizontal line indicates a genome-wide STAAR-O *P*-value threshold of 3.57×10^{-7}. The significant threshold is defined by multiple comparisons using the Bonferroni correction $(0.05/(20,000 \times 7) = 3.57 \times 10^{-7})$. Different symbols represent the STAAR-O *P*-value of the protein-coding gene using different functional categories (upstream, downstream, UTR, promoter_CAGE, promoter_DHS, enhancer_CAGE, enhancer_DHS). Promoter_CAGE and promoter_DHS are the promoters with overlap of Cap Analysis of Gene Expression (CAGE) sites and DNase hypersensitivity (DHS) sites for a given gene, respectively. Enhancer_CAGE and enhancer_DHS are the enhancers in GeneHancer predicted regions with the overlap of CAGE sites and DHS sites for a given gene, respectively. **b**, Quantile-quantile plots for unconditional gene-centric noncoding

analysis of protein-coding gene. Different symbols represent the STAAR-O *P*-value of the gene using different functional categories (upstream, downstream, UTR, promoter_CAGE, promoter_DHS, enhancer_CAGE, enhancer_DHS). **c,** Manhattan plots for unconditional gene-centric noncoding analysis of ncRNA gene. The horizontal line indicates a genome-wide STAAR-O *P*-value threshold of 2.50×10^{-6}. The significant threshold is defined by multiple comparisons using the Bonferroni correction $(0.05/20,000 = 2.50 \times 10^{-6})$. **d,** Quantile-quantile plots for unconditional gene-centric noncoding analysis of ncRNA gene. **e,** Manhattan plot for 2-kb sliding windows. The horizontal line indicates a genome-wide *P*-value threshold of 1.88×10^{-8}. The significant threshold is defined by multiple comparisons using the Bonferroni correction $(0.05/(2.66 \times 10^{6}) = 1.88 \times 10^{-8})$. **f**, Quantile-quantile plot for 2-kb sliding windows. In panels, **a**, **c** and **e**, the chromosome number are indicated by the colors of dots. In all panels, STAAR-O is a two-sided test.



Extended Data Fig. 5|. Manhattan plots and Q-Q plots for unconditional gene-centric noncoding analysis and sliding window analysis of total cholesterol (TC) in the discovery phase (n=21,015). **a**, Manhattan plots for unconditional gene-centric noncoding analysis of protein-coding gene. The horizontal line indicates a genome-wide STAAR-O *P*-value threshold of 3.57×10^{-7}. The significant threshold is defined by multiple comparisons using the Bonferroni correction $(0.05/(20,000 \times 7) = 3.57 \times 10^{-7})$. Different symbols represent the STAAR-O *P*-value of the protein-coding gene using different functional categories (upstream, downstream, UTR, promoter_CAGE, promoter_DHS, enhancer_CAGE, enhancer_DHS). Promoter_CAGE and promoter_DHS are the promoters with overlap of Cap Analysis of Gene Expression (CAGE) sites and DNase hypersensitivity (DHS) sites for a given gene, respectively. Enhancer_CAGE and enhancer_DHS are the enhancers in GeneHancer predicted regions with the overlap of CAGE sites and DHS sites for a given gene, respectively. **b**, Quantile-quantile plots for unconditional gene-centric noncoding analysis of protein-coding gene. Different symbols represent the STAAR-O *P*-value of the gene using different functional categories (upstream, downstream, UTR, promoter_CAGE, promoter_DHS, enhancer_CAGE, enhancer_DHS). **c,** Manhattan plots for unconditional gene-centric noncoding analysis of ncRNA gene. The horizontal line indicates a genome-wide STAAR-O *P*-value threshold of 2.50×10^{-6}. The significant threshold is defined by multiple comparisons using the Bonferroni correction $(0.05/20,000 = 2.50 \times 10^{-6})$. **d,** Quantile-quantile plots for unconditional gene-centric noncoding analysis of ncRNA gene. **e,** Manhattan plot for 2-kb sliding windows. The horizontal line indicates a genome-wide *P*-value threshold of 1.88×10^{-8}. The significant threshold is defined by multiple comparisons using the Bonferroni correction $(0.05/(2.66 \times 10^{6}) = 1.88 \times 10^{-8})$. **f**, Quantile-quantile plot for 2-kb sliding windows. In panels, **a**, **c** and **e**, the chromosome number are indicated by the colors of dots. In all panels, STAAR-O is a two-sided test.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Authors

Zilin Li[1,2,54,*], Xihao Li[1,54], Hufeng Zhou[1], Sheila M. Gaynor[1], Margaret Sunitha Selvaraj[3,4,5], Theodore Arapoglou[1], Corbin Quick[1], Yaowu Liu[6], Han Chen[7,8], Ryan Sun[9], Rounak Dey[1], Donna K. Arnett[10], Paul L. Auer[11], Lawrence F. Bielak[12], Joshua C. Bis[13], Thomas W. Blackwell[14], John Blangero[15], Eric Boerwinkle[7,16], Donald W. Bowden[17], Jennifer A. Brody[13], Brian E. Cade[4,18,19], Matthew P. Conomos[20], Adolfo Correa[21], L. Adrienne Cupples[22,23], Joanne E. Curran[15], Paul S. de Vries[7], Ravindranath Duggirala[15], Nora Franceschini[24], Barry I. Freedman[25], Harald H. H. Göring[15], Xiuqing Guo[26], Rita R. Kalyani[27], Charles Kooperberg[28], Brian G. Kral[29], Leslie A. Lange[30], Bridget M. Lin[31], Ani Manichaikul[32], Alisa K. Manning[5,33,34], Lisa W. Martin[32], Rasika A. Mathias[27], James B. Meigs[4,5,35], Braxton D. Mitchell[36,37], May E. Montasser[38], Alanna C. Morrison[7], Take Naseri[39], Jeffrey R. O'Connell[36], Nicholette D. Palmer[17], Patricia A. Peyser[12], Bruce M. Psaty[13,40,41], Laura M. Raffield[42], Susan Redline[18,19,43], Alexander P. Reiner[28,40], Muagututi'a Sefuiva Reupena[44], Kenneth M. Rice[20], Stephen S. Rich[30], Jennifer

A. Smith[12,45], Kent D. Taylor[26], Margaret A. Taub[46], Ramachandran S. Vasan[23,47], Daniel E. Weeks[48], James G. Wilson[49], Lisa R. Yanek[27], Wei Zhao[12], NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium, TOPMed Lipids Working Group, Jerome I. Rotter[26], Cristen J. Willer[50,51,52], Pradeep Natarajan[3,4,5], Gina M. Peloso[22,23], Xihong Lin[1,4,53,*]

Affiliations

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

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

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

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

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

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

[7]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.

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

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

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

[11]Division of Biostatistics, Institute for Health & Equity and Cancer Center, Medical College of Wisconsin, Milwaukee, WI, USA

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

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

[14]Department of Biostatistics and Center for Statistical Genetics, University of Michigan, Ann Arbor, Michigan, USA.

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

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

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

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

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

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

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

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

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

[24]Department of Epidemiology, Gillings School of Global Public Health, University of North Carolina, Chapel Hill, NC, USA.

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

[26]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.

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

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

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

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

[31]Department of Biostatistics, University of North Carolina, Chapel Hill, NC, USA.

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

[33]Metabolism Program, The Broad Institute of MIT and Harvard, Cambridge, MA, USA.

[34]Clinical and Translational Epidemiology Unit, Mongan Institute, Massachusetts General Hospital, Boston, MA, USA.

[35]Division of General Internal Medicine, Massachusetts General Hospital, Boston, MA, USA.

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

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

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

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

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

[41]Departments of Health Systems and Population Health, University of Washington, Seattle, WA, USA.

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

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

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

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

[46]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA.

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

[48]Department of Human Genetics and Biostatistics, University of Pittsburgh, Pittsburgh, PA, USA.

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

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

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

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

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

[54]These authors contributed equally: Zilin Li, Xihao Li.

Acknowledgments

This work was supported by grants R35-CA197449, U19-CA203654, R01-HL113338, and U01-HG009088 (X. Lin), R01-HL142711 and R01-HL127564 (P.N. and G.M.P), R35-HL135824 (C.J.W.), 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.), 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.), HHSN268201800001I and U01-HL137162 (K.M.R., M.P.C.), R01-HL133040 (D.E.W.), 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.), R01-HL141944 (R.A.M.), R.A.M. receives support as the Sarah Miller Coulson Scholar in the Johns Hopkins Center for Innovative Medicine, HHSN268201600018C, HHSN268201600001C, HHSN268201600002C, HHSN268201600003C, and HHSN268201600004C (C.L.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.), R01-DK117445 and R01-MD012765 (N.F. and B.M.L.), U01-DK078616, UM1-DK0786 and R01-DK078616 (J.B.M.), 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. We gratefully acknowledge the support from The Samoan Obesity, Lifestyle and Genetic Adaptations Study (OLaGA) Group. The full study specific acknowledgements are detailed in Supplementary Note.

Data availability

This paper used the TOPMed Freeze 5 Whole Genome Sequencing data and phenotype data of lipids, C-reactive protein, estimated glomerular filtration rate, fasting glucose, fasting insulin and telomere length. The genotype and phenotype data are both available in dbGAP. The TOPMed data were from the following fourteen studies, where the accession numbers are provided in parenthesis: Framingham Heart Study (phs000974.v1.p1), Old Order Amish (phs000956.v1.p1), Jackson Heart Study (phs000964.v1.p1), Multi-Ethnic Study of Atherosclerosis (phs001416.v1.p1), Genome-wide Association Study of Adiposity in Samoans (phs000972) and Women's Health Initiative (phs001237), 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) and San Antonio Family Heart Study (phs001215).

The functional annotation data are publicly available and were downloaded from the following links: GRCh38 CADD v1.4 (https://cadd.gs.washington.edu/download), ANNOVAR dbNSFP v3.3a (https://annovar.openbioinformatics.org/en/latest/user-guide/download), LINSIGHT (https://github.com/CshlSiepelLab/LINSIGHT), FATHMM-XF (http://fathmm.biocompute.org.uk/fathmm-xf), CAGE (https://fantom.gsc.riken.jp/5/data), GeneHancer (https://www.genecards.org), and Umap/Bismap (https://bismap.hoffmanlab.org). In addition, recombination rate and nucleotide diversity were obtained from Gazal et al[55]. The tissue-specific functional annotations were downloaded

from ENCODE (https://www.encodeproject.org/report/?type=Experiment). The assembled functional annotation data from these sources are available at http://favor.genohub.org.

NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium

Namiko Abe[55], Gonçalo Abecasis[56], Francois Aguet[57], Christine Albert[58], Laura Almasy[59], Alvaro Alonso[60], Seth Ament[61], Peter Anderson[62], Pramod Anugu[63], Deborah Applebaum-Bowden[64], Kristin Ardlie[57], Dan Arking[65], Allison Ashley-Koch[66], Stella Aslibekyan[67], Tim Assimes[68], Dimitrios Avramopoulos[65], Najib Ayas[69], Adithya Balasubramanian[70], John Barnard[71], Kathleen Barnes[72], R. Graham Barr[73], Emily Barron-Casella[65], Lucas Barwick[74], Terri Beaty[65], Gerald Beck[75], Diane Becker[76], Lewis Becker[65], Rebecca Beer[77], Amber Beitelshees[61], Emelia Benjamin[78], Takis Benos[79], Marcos Bezerra[80], Nathan Blue[81], Russell Bowler[82], Ulrich Broeckel[83], Jai Broome[62], Deborah Brown[84], Karen Bunting[55], Esteban Burchard[85], Carlos Bustamante[86], Erin Buth[87], Jonathan Cardwell[88], Vincent Carey[89], Julie Carrier[90], April Carson[91], Cara Carty[92], Richard Casaburi[93], Juan P. Casas Romero[94], James Casella[65], Peter Castaldi[95], Mark Chaffin[57], Christy Chang[61], Yi-Cheng Chang[96], Daniel Chasman[97], Sameer Chavan[88], Bo-Juen Chen[55], Wei-Min Chen[98], Yii-Der Ida Chen[99], Michael Cho[89], Seung Hoan Choi[57], Lee-Ming Chuang[100], Mina Chung[101], Ren-Hua Chung[102], Clary Clish[103], Suzy Comhair[104], Elaine Cornell[105], Carolyn Crandall[93], James Crapo[106], Jeffrey Curtis[107], Brian Custer[108], Coleen Damcott[61], Dawood Darbar[109], Sean David[110], Colleen Davis[62], Michelle Daya[88], Mariza de Andrade[111], Lisa de las Fuentes[112], Michael DeBaun[113], Ranjan Deka[114], Dawn DeMeo[89], Scott Devine[61], Huyen Dinh[70], Harsha Doddapaneni[115], Qing Duan[116], Shannon Dugan-Perez[70], Jon Peter Durda[105], Susan K. Dutcher[117], Charles Eaton[118], Lynette Ekunwe[63], Adel El Boueiz[119], Patrick Ellinor[120], Leslie Emery[62], Serpil Erzurum[71], Charles Farber[98], Jesse Farek[70], Tasha Fingerlin[121], Matthew Flickinger[56], Myriam Fornage[122], Chris Frazar[62], Mao Fu[61], Stephanie M. Fullerton[62], Lucinda Fulton[123], Stacey Gabriel[57], Weiniu Gan[77], Shanshan Gao[88], Yan Gao[63], Margery Gass[124], Heather Geiger[125], Bruce Gelb[126], Mark Geraci[127], Soren Germer[55], Robert Gerszten[128], Auyon Ghosh[89], Richard Gibbs[70], Chris Gignoux[68], Mark Gladwin[79], David Glahn[129], Stephanie Gogarten[62], Da-Wei Gong[61], Sharon Graw[130], Kathryn J. Gray[131], Daniel Grine[88], Colin Gross[56], C. Charles Gu[123], Yue Guan[61], Namrata Gupta[57], Michael Hall[132], Yi Han[70], Patrick Hanly[133], Daniel Harris[134], Nicola L. Hawley[135], Jiang He[136], Ben Heavner[87], Susan Heckbert[137], Ryan Hernandez[85], David Herrington[138], Craig Hersh[139], Bertha Hidalgo[67], James Hixson[122], Brian Hobbs[89], John Hokanson[88], Elliott Hong[61], Karin Hoth[140], Chao (Agnes) Hsiung[141], Jianhong Hu[70], Yi-Jen Hung[142], Haley Huston[143], Chii Min Hwu[144], Marguerite Ryan Irvin[67], Rebecca Jackson[145], Deepti Jain[62], Cashell Jaquish[146], Jill Johnsen[147], Andrew Johnson[77], Craig Johnson[62], Rich Johnston[60], Kimberly Jones[65], Hyun Min Kang[148], Robert Kaplan[149], Sharon Kardia[56], Shannon Kelly[150], Eimear Kenny[126], Michael Kessler[61], Alyna Khan[62], Ziad Khan[70], Wonji Kim[151], John Kimoff[152], Greg Kinney[153], Barbara Konkle[154], Holly Kramer[155], Christoph Lange[156], Ethan Lange[88], Cathy Laurie[62], Cecelia Laurie[62], Meryl LeBoff[89], Jiwon Lee[89], Sandra Lee[70], Wen-Jane Lee[144], Jonathon LeFaive[56], David Levine[62], Dan Levy[77], Joshua Lewis[61], Xiaohui Li[99], Yun Li[116], Henry Lin[99], Honghuang Lin[157], Simin Liu[158], Yongmei Liu[159], Yu Liu[160], Ruth J.F. Loos[161], Steven Lubitz[120], Kathryn Lunetta[162], James Luo[77], Ulysses Magalang[163],

Michael Mahaney[164], Barry Make[65], JoAnn Manson[89], Melissa Marton[125], Susan Mathai[88], Susanne May[87], Patrick McArdle[61], Merry-Lynn McDonald[165], Sean McFarland[151], Daniel McGoldrick[166], Caitlin McHugh[87], Becky McNeil[167], Hao Mei[63], Vipin Menon[70], Luisa Mestroni[130], Ginger Metcalf[70], Deborah A. Meyers[168], Emmanuel Mignot[169], Julie Mikulla[77], Nancy Min[63], Mollie Minear[170], Ryan L. Minster[79], Matt Moll[95], Zeineen Momin[70], Courtney Montgomery[171], Donna Muzny[70], Josyf C. Mychaleckyj[98], Girish Nadkarni[126], Rakhi Naik[65], Sergei Nekhai[172], Sarah C. Nelson[87], Bonnie Neltner[88], Caitlin Nessner[70], Deborah Nickerson[173], Osuji Nkechinyere[70], Kari North[116], Tim O'Connor[61], Heather Ochs-Balcom[174], Geoffrey Okwuonu[70], Allan Pack[175], David T. Paik[176], James Pankow[177], George Papanicolaou[77], Cora Parker[178], Juan Manuel Peralta[179], Marco Perez[68], James Perry[61], Ulrike Peters[180], Lawrence S Phillips[60], Jacob Pleiness[56], Toni Pollin[61], Wendy Post[181], Julia Powers Becker[182], Meher Preethi Boorgula[88], Michael Preuss[126], Pankaj Qasba[77], Dandi Qiao[89], Zhaohui Qin[60], Nicholas Rafaels[183], Mahitha Rajendran[70], D.C. Rao[123], Laura Rasmussen-Torvik[184], Aakrosh Ratan[98], Robert Reed[61], Catherine Reeves[185], Elizabeth Regan[106], Rebecca Robillard[186], Nicolas Robine[125], Dan Roden[187], Carolina Roselli[57], Ingo Ruczinski[65], Alexi Runnels[125], Pamela Russell[88], Sarah Ruuska[143], Kathleen Ryan[61], Ester Cerdeira Sabino[188], Danish Saleheen[189], Shabnam Salimi[190], Sejal Salvi[70], Steven Salzberg[65], Kevin Sandow[191], Vijay G. Sankaran[192], Jireh Santibanez[70], Karen Schwander[123], David Schwartz[88], Frank Sciurba[79], Christine Seidman[193], Jonathan Seidman[194], Frédéric Sériès[195], Vivien Sheehan[196], Stephanie L. Sherman[197], Amol Shetty[61], Aniket Shetty[88], Wayne Hui-Heng Sheu[144], M. Benjamin Shoemaker[198], Brian Silver[199], Edwin Silverman[89], Robert Skomro[200], Albert Vernon Smith[201], Josh Smith[62], Nicholas Smith[137], Tanja Smith[55], Sylvia Smoller[149], Beverly Snively[202], Michael Snyder[68], Tamar Sofer[89], Nona Sotoodehnia[62], Adrienne M. Stilp[62], Garrett Storm[203], Elizabeth Streeten[61], Jessica Lasky Su[204], Yun Ju Sung[123], Jody Sylvia[89], Adam Szpiro[62], Daniel Taliun[56], Hua Tang[205], Margaret Taub[65], Matthew Taylor[130], Simeon Taylor[61], Marilyn Telen[66], Timothy A. Thornton[62], Machiko Threlkeld[206], Lesley Tinker[207], David Tirschwell[62], Sarah Tishkoff[208], Hemant Tiwari[209], Catherine Tong[210], Russell Tracy[211], Michael Tsai[177], Dhananjay Vaidya[65], David Van Den Berg[212], Peter VandeHaar[56], Scott Vrieze[177], Tarik Walker[88], Robert Wallace[140], Avram Walts[88], Fei Fei Wang[62], Heming Wang[213], Jiongming Wang[214], Karol Watson[93], Jennifer Watt[70], Joshua Weinstock[148], Bruce Weir[62], Scott T. Weiss[215], Lu-Chen Weng[120], Jennifer Wessel[216], Kayleen Williams[87], L. Keoki Williams[217], Carla Wilson[89], Lara Winterkorn[125], Quenna Wong[62], Joseph Wu[176], Huichun Xu[61], Ivana Yang[88], Ketian Yu[56], Seyedeh Maryam Zekavat[57], Yingze Zhang[218], Snow Xueyan Zhao[106], Xiaofeng Zhu[219], Elad Ziv[220], Michael Zody[55], Sebastian Zoellner[56]

55 - New York Genome Center, New York, New York, 10013, US; 56 - University of Michigan, Ann Arbor, Michigan, 48109, US; 57 - Broad Institute, Cambridge, Massachusetts, 2142, US; 58 - Cedars Sinai, Boston, Massachusetts, 2114, US; 59 - Children's Hospital of Philadelphia, University of Pennsylvania, Philadelphia, Pennsylvania, 19104, US; 60 - Emory University, Atlanta, Georgia, 30322, US; 61 - University of Maryland, Baltimore, Maryland, 21201, US; 62 - University of Washington, Seattle, Washington, 98195, US; 63 - University of Mississippi Medical Center, Jackson, Mississippi, 39216, US; 64 - National Institutes of Health, Bethesda, Maryland, 20892,

US; 65 - Johns Hopkins University, Baltimore, Maryland, 21218, US; 66 - Duke University, Durham, North Carolina, 27708, US; 67 - University of Alabama, Birmingham, Alabama, 35487, US; 68 - Stanford University, Stanford, California, 94305, US; 69 - Providence Health Care, Medicine, Vancouver, CA; 70 - Baylor College of Medicine Human Genome Sequencing Center, Houston, Texas, 77030, US; 71 - Cleveland Clinic, Cleveland, Ohio, 44195, US; 72 - Tempus, University of Colorado Anschutz Medical Campus, Aurora, Colorado, 80045, US; 73 - Columbia University, New York, New York, 10032, US; 74 - The Emmes Corporation, LTRC, Rockville, Maryland, 20850, US; 75 - Cleveland Clinic, Quantitative Health Sciences, Cleveland, Ohio, 44195, US; 76 - Johns Hopkins University, Medicine, Baltimore, Maryland, 21218, US; 77 - National Heart, Lung, and Blood Institute, National Institutes of Health, Bethesda, Maryland, 20892, US; 78 - Boston University, Massachusetts General Hospital, Boston University School of Medicine, Boston, Massachusetts, 2114, US; 79 - University of Pittsburgh, Pittsburgh, Pennsylvania, 15260, US; 80 - Fundação de Hematologia e Hemoterapia de Pernambuco - Hemope, Recife, 52011-000, BR; 81 - University of Utah, Obstetrics and Gynecology, Salt Lake City, Utah, 84132, US; 82 - National Jewish Health, National Jewish Health, Denver, Colorado, 80206, US; 83 - Medical College of Wisconsin, Pediatrics, Milwaukee, Wisconsin, 53226, US; 84 - University of Texas Health at Houston, Pediatrics, Houston, Texas, 77030, US; 85 - University of California, San Francisco, San Francisco, California, 94143, US; 86 - Stanford University, Biomedical Data Science, Stanford, California, 94305, US; 87 - University of Washington, Biostatistics, Seattle, Washington, 98195, US; 88 - University of Colorado at Denver, Denver, Colorado, 80204, US; 89 - Brigham & Women's Hospital, Boston, Massachusetts, 2115, US; 90 - University of Montreal, US; 91 - University of Mississippi, Medicine, Jackson, Mississippi, 39213, US; 92 - Washington State University, Pullman, Washington, 99164, US; 93 - University of California, Los Angeles, Los Angeles, California, 90095, US; 94 - Brigham & Women's Hospital, US; 95 - Brigham & Women's Hospital, Medicine, Boston, Massachusetts, 2115, US; 96 - National Taiwan University, Taipei, 10617, TW; 97 - Brigham & Women's Hospital, Division of Preventive Medicine, Boston, Massachusetts, 2215, US; 98 - University of Virginia, Charlottesville, Virginia, 22903, US; 99 - Lundquist Institute, Torrance, California, 90502, US; 100 - National Taiwan University, National Taiwan University Hospital, Taipei, 10617, TW; 101 - Cleveland Clinic, Cleveland Clinic, Cleveland, Ohio, 44195, US; 102 - National Health Research Institute Taiwan, Miaoli County, 350, TW; 103 - Broad Institute, Metabolomics Platform, Cambridge, Massachusetts, 2142, US; 104 - Cleveland Clinic, Immunity and Immunology, Cleveland, Ohio, 44195, US; 105 - University of Vermont, Burlington, Vermont, 5405, US; 106 - National Jewish Health, Denver, Colorado, 80206, US; 107 - University of Michigan, Internal Medicine, Ann Arbor, Michigan, 48109, US; 108 - Vitalant Research Institute, San Francisco, California, 94118, US; 109 - University of Illinois at Chicago, Chicago, Illinois, 60607, US; 110 - University of Chicago, Chicago, Illinois, 60637, US; 111 - Mayo Clinic, Health Quantitative Sciences Research, Rochester, Minnesota, 55905, US; 112 - Washington University in St Louis, Department of Medicine, Cardiovascular Division, St. Louis, Missouri, 63110, US; 113 - Vanderbilt University, Nashville, Tennessee, 37235, US; 114 - University of Cincinnati, Cincinnati, Ohio, 45220, US; 115 - Baylor College of Medicine Human Genome Sequencing Center, Houston, Texas, 77030; 116 - University of North Carolina, Chapel Hill, North Carolina, 27599, US; 117 - Washington

University in St Louis, Genetics, St Louis, Missouri, 63110, US; 118 - Brown University, Providence, Rhode Island, 2912, US; 119 - Harvard University, Channing Division of Network Medicine, Cambridge, Massachusetts, 2138, US; 120 - Massachusetts General Hospital, Boston, Massachusetts, 2114, US; 121 - National Jewish Health, Center for Genes, Environment and Health, Denver, Colorado, 80206, US; 122 - University of Texas Health at Houston, Houston, Texas, 77225, US; 123 - Washington University in St Louis, St Louis, Missouri, 63130, US; 124 - Fred Hutchinson Cancer Research Center, Seattle, Washington, 98109, US; 125 - New York Genome Center, New York City, New York, 10013, US; 126 - Icahn School of Medicine at Mount Sinai, New York, New York, 10029, US; 127 - University of Pittsburgh, Pittsburgh, Pennsylvania, US; 128 - Beth Israel Deaconess Medical Center, Boston, Massachusetts, 2215, US; 129 - Boston Children's Hospital, Harvard Medical School, Department of Psychiatry, Boston, Massachusetts, 2115, US; 130 - University of Colorado Anschutz Medical Campus, Aurora, Colorado, 80045, US; 131 - Mass General Brigham, Obstetrics and Gynecology, Boston, Massachusetts, 2115, US; 132 - University of Mississippi, Cardiology, Jackson, Mississippi, 39216, US; 133 - University of Calgary, Medicine, Calgary, CA; 134 - University of Maryland, Genetics, Philadelphia, Pennsylvania, 19104, US; 135 - Yale University, Department of Chronic Disease Epidemiology, New Haven, Connecticut, 6520, US; 136 - Tulane University, New Orleans, Louisiana, 70118, US; 137 - University of Washington, Epidemiology, Seattle, Washington, 98195, US; 138 - Wake Forest Baptist Health, Winston-Salem, North Carolina, 27157, US; 139 - Brigham & Women's Hospital, Channing Division of Network Medicine, Boston, Massachusetts, 2115, US; 140 - University of Iowa, Iowa City, Iowa, 52242, US; 141 - National Health Research Institute Taiwan, Institute of Population Health Sciences, NHRI, Miaoli County, 350, TW; 142 - Tri-Service General Hospital National Defense Medical Center, TW; 143 - Blood Works Northwest, Seattle, Washington, 98104, US; 144 - Taichung Veterans General Hospital Taiwan, Taichung City, 407, TW; 145 - Oklahoma State University Medical Center, Internal Medicine, DIvision of Endocrinology, Diabetes and Metabolism, Columbus, Ohio, 43210, US; 146 - National Heart, Lung, and Blood Institute, National Institutes of Health, NHLBI, Bethesda, Maryland, 20892, US; 147 - Blood Works Northwest, Research Institute, Seattle, Washington, 98104, US; 148 - University of Michigan, Biostatistics, Ann Arbor, Michigan, 48109, US; 149 - Albert Einstein College of Medicine, New York, New York, 10461, US; 150 - University of California, San Francisco, San Francisco, California, 94118, US; 151 - Harvard University, Cambridge, Massachusetts, 2138, US; 152 - McGill University, Montréal, QC H3A 0G4, CA; 153 - University of Colorado at Denver, Epidemiology, Aurora, Colorado, 80045, US; 154 - Blood Works Northwest, Medicine, Seattle, Washington, 98104, US; 155 - Loyola University, Public Health Sciences, Maywood, Illinois, 60153, US; 156 - Harvard School of Public Health, Biostats, Boston, Massachusetts, 2115, US; 157 - Boston University, University of Massachusetts Chan Medical School, Worcester, Massachusetts, 1655, US; 158 - Brown University, Epidemiology and Medicine, Providence, Rhode Island, 2912, US; 159 - Duke University, Cardiology, Durham, North Carolina, 27708, US; 160 - Stanford University, Cardiovascular Institute, Stanford, California, 94305, US; 161 - Icahn School of Medicine at Mount Sinai, The Charles Bronfman Institute for Personalized Medicine, New York, New York, 10029, US; 162 - Boston University, Boston, Massachusetts, 2215, US; 163 - Ohio State University, Division of Pulmonary, Critical Care and Sleep

Medicine, Columbus, Ohio, 43210, US; 164 - University of Texas Rio Grande Valley School of Medicine, Brownsville, Texas, 78520, US; 165 - University of Alabama, University of Alabama at Birmingham, Birmingham, Alabama, 35487, US; 166 – University of Washington, Genome Sciences, Seattle, Washington, 98195, US; 167 - RTI International, US; 168 - University of Arizona, Tucson, Arizona, 85721, US; 169 - Stanford University, Center For Sleep Sciences and Medicine, Palo Alto, California, 94304, US; 170 - National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, Maryland, 20892, US; 171 - Oklahoma Medical Research Foundation, Genes and Human Disease, Oklahoma City, Oklahoma, 73104, US; 172 - Howard University, Washington, District of Columbia, 20059, US; 173 - University of Washington, Department of Genome Sciences, Seattle, Washington, 98195, US; 174 - University at Buffalo, Buffalo, New York, 14260, US; 175 - University of Pennsylvania, Division of Sleep Medicine/Department of Medicine, Philadelphia, Pennsylvania, 19104-3403, US; 176 - Stanford University, Stanford Cardiovascular Institute, Stanford, California, 94305, US; 177 - University of Minnesota, Minneapolis, Minnesota, 55455, US; 178 - RTI International, Biostatistics and Epidemiology Division, Research Triangle Park, North Carolina, 27709-2194, US; 179 - University of Texas Rio Grande Valley School of Medicine, Edinburg, Texas, 78539, US; 180 - Fred Hutchinson Cancer Research Center, Fred Hutch and UW, Seattle, Washington, 98109, US; 181 - Johns Hopkins University, Cardiology/Medicine, Baltimore, Maryland, 21218, US; 182 - University of Colorado at Denver, Medicine, Denver, Colorado, 80204, US; 183 - University of Colorado at Denver, CCPM, Denver, Colorado, 80045, US; 184 - Northwestern University, Chicago, Illinois, 60208, US; 185 - New York Genome Center, New York Genome Center, New York City, New York, 10013, US; 186 - University of Ottawa, Sleep Research Unit, University of Ottawa Institute for Mental Health Research, Ottawa, ON K1Z 7K4, CA; 187 - Vanderbilt University, Medicine, Pharmacology, Biomedicla Informatics, Nashville, Tennessee, 37235, US; 188 - Universidade de Sao Paulo, Faculdade de Medicina, Sao Paulo, 1310000, BR; 189 - Columbia University, New York, New York, 10027, US; 190 - University of Maryland, Pathology, Seattle, Washington, 98195, US; 191 - Lundquist Institute, TGPS, Torrance, California, 90502, US; 192 - Harvard University, Division of Hematology/Oncology, Boston, Massachusetts, 2115, US; 193 - Harvard Medical School, Genetics, Boston, Massachusetts, 2115, US; 194 - Harvard Medical School, Boston, Massachusetts, 2115, US; 195 - Université Laval, Quebec City, G1V 0A6, CA; 196 - Emory University, Pediatrics, Atlanta, Georgia, 30307, US; 197 - Emory University, Human Genetics, Atlanta, Georgia, 30322, US; 198 - Vanderbilt University, Medicine/Cardiology, Nashville, Tennessee, 37235, US; 199 - UMass Memorial Medical Center, Worcester, Massachusetts, 1655, US; 200 - University of Saskatchewan, Saskatoon, SK S7N 5C9, CA; 201 - University of Michigan; 202 - Wake Forest Baptist Health, Biostatistical Sciences, Winston-Salem, North Carolina, 27157, US; 203 - University of Colorado at Denver, Genomic Cardiology, Aurora, Colorado, 80045, US; 204 - Brigham & Women's Hospital, Channing Department of Medicine, Boston, Massachusetts, 2115, US; 205 - Stanford University, Genetics, Stanford, California, 94305, US; 206 - University of Washington, University of Washington, Department of Genome Sciences, Seattle, Washington, 98195, US; 207 - Fred Hutchinson Cancer Research Center, Cancer Prevention Division of Public Health Sciences, Seattle, Washington, 98109, US; 208 - University of Pennsylvania, Genetics, Philadelphia, Pennsylvania, 19104, US; 209 -

University of Alabama, Biostatistics, Birmingham, Alabama, 35487, US; 210 - University of Washington, Department of Biostatistics, Seattle, Washington, 98195, US; 211 - University of Vermont, Pathology & Laboratory Medicine, Burlington, Vermont, 5405, US; 212 - University of Southern California, USC Methylation Characterization Center, University of Southern California, California, 90033, US; 213 - Brigham & Women's Hospital, Mass General Brigham, Boston, Massachusetts, 2115, US; 214 - University of Michigan, US; 215 - Brigham & Women's Hospital, Channing Division of Network Medicine, Department of Medicine, Boston, Massachusetts, 2115, US; 216 - Indiana University, Epidemiology, Indianapolis, Indiana, 46202, US; 217 - Henry Ford Health System, Detroit, Michigan, 48202, US; 218 - University of Pittsburgh, Medicine, Pittsburgh, Pennsylvania, 15260, US; 219 - Case Western Reserve University, Department of Population and Quantitative Health Sciences, Cleveland, Ohio, 44106, US; 220 - University of California, San Francisco, US

TOPMed Lipids Working Group

Gonçalo Abecasis[56], Donna K. Arnett[10], Stella Aslibekyan[67], Tim Assimes[68], Elizabeth Atkinson[57], Christie Ballantyne[70], Wei Bao[140], Amber Beitelshees[61], Romit Bhattacharya[57], Larry Bielak[12], Joshua Bis[13], Corneliu Bodea[120], Eric Boerwinkle[7,16], Donald W. Bowden[17], Jennifer Brody[13], Brian Cade[4,18,19], Sarah Calvo[57], Jenna Carlson[79], I-Shou Chang[102], Yii-Der Ida Chen[99], So Mi Cho[57], Seung Hoan Choi[57], Ren-Hua Chung[102], Adolfo Correa[21], L. Adrienne Cupples[22,23], Coleen Damcott[61], Paul de Vries[7], Ana F. Diallo[221], Ron Do[126], Jacqueline Dron[57], Amanda Elliott[57], Hilary Finucane[57], Caitlin Floyd[222], Mao Fu[61], Andrea Ganna[57], Dawei Gong[61], Sarah Graham[107], Mary Haas[57], Bernhard Haring[223], Jiang He[136], Scott Heemann[222], Blanca Himes[224], James Hixson[122], Marguerite Ryan Irvin[67], Gail Jarvik[62], Jicai Jiang[61], Roby Joehanes[194], Paule Valery Joseph[64], Goo Jun[122], Rita Kalyani[27], Masahiro Kanai[57], Sharon Kardia[56], Sekar Kathiresan[225], Amit Khera[57], Sumeet Khetarpal[57], Derek Klarin[57], Charles Kooperberg[28], Satoshi Koyama[57], Brian Kral[29], Leslie Lange[30], Cathy Laurie[62], Rozenn Lemaitre[13], Zilin Li[1,2], Xihao Li[1], Changwei Li[136], Xihong Lin[1,4,53], Yingchang Lu[113], Michael Mahaney[164], Ani Manichaikul[32], Lisa Martin[32], Rasika Mathias[27], Ravi Mathur[167], Stephen McGarvey[118], John McLenithan[61], Julie Mikulla[77], Amy Miller[222], Braxton D. Mitchell[36,37], May E. Montasser[38], Vamsi Mootha[57], Andrew Moran[73], Alanna C. Morrison[7], Tetsushi Nakao[57], Pradeep Natarajan[3,4,5], Kari North[116], Jeff O'Connell[36], Christopher O'Donnell[77], Nicholette Palmer[17], Kaavya Paruchuri[57], Aniruddh Patel[57], Gina Peloso[22,23], James Perry[61], Ulrike Peters[180], Mary Pettinger[124], Patricia Peyser[12], James Pirruccello[120], Toni Pollin[61], Michael Preuss[126], Bruce Psaty[13,40,41], Susan Redline[18,19,43], Robert Reed[61], Alex Reiner[28,40], Stephen Rich[30], Samantha Rosenthal[127], Jerome Rotter[26], Margaret Sunitha Selvaraj[3,4,5], Wayne Hui-Heng Sheu[144], Jennifer Smith[12,45], Tamar Sofer[89], Adrienne M. Stilp[62], Shamil R. Sunyaev[94], Ida Surakka[56], Carole Sztalryd[61], Hua Tang[205], Kent D. Taylor[26], Mark Trinder[57], Michael Tsai[177], Md Mesbah Uddin[57], Sarah Urbut[57], Eric Van Buren[1], Marie Verbanck[126], Ann Von Holle[116], Heming Wang[213], Yuxuan Wang[22], Kerri Wiggins[62], John Wilkins[184], Cristen Willer[50,51,52], James Wilson[49], Brooke Wolford[56], Huichun Xu[61], Lisa Yanek[27], Zhi Yu[57], Norann Zaghloul[61], Seyedeh Maryam Zekavat[57], Jingwen Zhang[1], Ying Zhou[124]

221 - Virginia Commonwealth University, Richmond, Virginia, US; 222 - Westat, Atlanta, Georgia, US; 223 - Saarland University Medical Center, Homburg, Germany; 224 - Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, Pennsylvania, US; 225 - Verve Therapeutics, Cambridge, Massachusetts, US

References

1. Manolio TA et al. Finding the missing heritability of complex diseases. Nature 461, 747–753 (2009). [PubMed: 19812666]
2. 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]
3. Hernandez RD et al. Ultrarare variants drive substantial cis heritability of human gene expression. Nature genetics 51, 1349–1355 (2019). [PubMed: 31477931]
4. Taliun D et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. Nature 590, 290–299 (2021). [PubMed: 33568819]
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. Van Hout CV et al. Exome sequencing and characterization of 49,960 individuals in the UK Biobank. Nature 586, 749–756 (2020). [PubMed: 33087929]
7. Zhang F & Lupski JR Non-coding genetic variants in human disease. Human molecular genetics 24, R102–R110 (2015). [PubMed: 26152199]
8. Khurana E et al. Role of non-coding sequence variants in cancer. Nature Reviews Genetics 17, 93–108 (2016).
9. Lee PH et al. Principles and methods of in-silico prioritization of non-coding regulatory variants. Human genetics 137, 15–30 (2018). [PubMed: 29288389]
10. ENCODE Project Consortium An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57 (2012). [PubMed: 22955616]
11. Moore JE et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. Nature 583, 699–710 (2020). [PubMed: 32728249]
12. Bansal V, Libiger O, Torkamani A & Schork NJ Statistical analysis strategies for association studies involving rare variants. Nature Reviews Genetics 11, 773 (2010).
13. Lee S, Abecasis GR, 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]
14. Kiezun A et al. Exome sequencing and the genetic basis of complex traits. Nature genetics 44, 623 (2012). [PubMed: 22641211]
15. 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]
16. Wu MC 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]
17. 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]
18. 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]
19. 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]
20. He Z, Xu B, Buxbaum J & Ionita-Laza I A genome-wide scan statistic framework for whole-genome sequence data analysis. Nature communications 10, 1–11 (2019).

21. Natarajan P et al. Deep-coverage whole genome sequences and blood lipids among 16,324 individuals. Nature communications 9, 1–12 (2018).
22. Li Z, Liu Y & Lin X Simultaneous Detection of Signal Regions Using Quadratic Scan Statistics With Applications to Whole Genome Association Studies. Journal of the American Statistical Association 117, 823–834 (2022). [PubMed: 35845434]
23. Bocher O & Génin E Rare variant association testing in the non-coding genome. Human Genetics 139, 1345–1362 (2020). [PubMed: 32500240]
24. Fishilevich S et al. GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. Database 2017 (2017).
25. Fantom Consortium A promoter-level mammalian expression atlas. Nature 507, 462 (2014). [PubMed: 24670764]
26. Andersson R et al. An atlas of active enhancers across human cell types and tissues. Nature 507, 455–461 (2014). [PubMed: 24670763]
27. Breslow NE & Clayton DG Approximate inference in generalized linear mixed models. Journal of the American statistical Association 88, 9–25 (1993).
28. 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]
29. 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 (2019).
30. Zhou H, Arapoglou T, Li X, Li Z & Lin X, Edn. V1 (Harvard Dataverse, 2022).
31. Harrow J et al. GENCODE: the reference human genome annotation for The ENCODE Project. Genome research 22, 1760–1774 (2012). [PubMed: 22955987]
32. Frankish A et al. GENCODE reference annotation for the human and mouse genomes. Nucleic acids research 47, D766–D773 (2019). [PubMed: 30357393]
33. Kinsella RJ et al. Ensembl BioMarts: a hub for data retrieval across taxonomic space. Database 2011 (2011).
34. Povysil G et al. Rare-variant collapsing analyses for complex traits: guidelines and applications. Nature Reviews Genetics 20, 747–759 (2019).
35. Kircher M et al. A general framework for estimating the relative pathogenicity of human genetic variants. Nature genetics 46, 310 (2014). [PubMed: 24487276]
36. 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]
37. Rogers MF et al. FATHMM-XF: accurate prediction of pathogenic point mutations via extended features. Bioinformatics 34, 511–513 (2017).
38. 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]
39. 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 (2019). [PubMed: 30445434]
40. 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).
41. Moutsianas L et al. The power of gene-based rare variant methods to detect disease-associated variation and test hypotheses about complex disease. PLoS genetics 11, e1005165 (2015). [PubMed: 25906071]
42. Raffield LM et al. Allelic heterogeneity at the CRP locus identified by whole-genome sequencing in multi-ancestry cohorts. The American Journal of Human Genetics 106, 112–120 (2020). [PubMed: 31883642]

43. Lin BM et al. Whole genome sequence analyses of eGFR in 23,732 people representing multiple ancestries in the NHLBI trans-omics for precision medicine (TOPMed) consortium. EBioMedicine 63, 103157 (2021). [PubMed: 33418499]
44. DiCorpo D et al. Whole Genome Sequence Association Analysis of Fasting Glucose and Fasting Insulin Levels in Diverse Cohorts from the NHLBI TOPMed Program. medRxiv, 2020.2012. 2031.20234310 (2021).
45. Taub MA et al. Genetic determinants of telomere length from 109,122 ancestrally diverse whole-genome sequences in TOPMed. Cell Genomics 2, 100084 (2022). [PubMed: 35530816]
46. Schaffner SF et al. Calibrating a coalescent simulation of human genome sequence variation. Genome research 15, 1576–1583 (2005). [PubMed: 16251467]
47. Lee S, Wu MC & Lin X Optimal tests for rare variant effects in sequencing association studies. Biostatistics 13, 762–775 (2012). [PubMed: 22699862]
48. Zaidi AA & Mathieson I Demographic history mediates the effect of stratification on polygenic scores. Elife 9, e61548 (2020). [PubMed: 33200985]
49. Gogarten SM et al. Genetic association testing using the GENESIS R/Bioconductor package. Bioinformatics 35, 5346–5348 (2019). [PubMed: 31329242]
50. Zheng X & Davis JW SAIGEgds—an efficient statistical tool for large-scale PheWAS with mixed models. Bioinformatics (2020).
51. Peloso GM et al. Association of low-frequency and rare coding-sequence variants with blood lipids and coronary heart disease in 56,000 whites and blacks. The American Journal of Human Genetics 94, 223–232 (2014). [PubMed: 24507774]
52. Moon S, Lee Y, Won S & Lee J Multiple genotype–phenotype association study reveals intronic variant pair on SIDT2 associated with metabolic syndrome in a Korean population. Human genomics 12, 1–10 (2018). [PubMed: 29335020]
53. 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 (2015). [PubMed: 25552646]

Methods-only references

54. 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]
55. 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]
56. Li X & Li Z xihaoli/STAARpipeline: STAARpipeline_v0.9.6 Version 0.9.6 10.5281/zenodo.6871504 (2022).
57. Li X & Li Z xihaoli/STAARpipelineSummary: STAARpipelineSummary_v0.9.6 Version 0.9.6 10.5281/zenodo.6871524 (2022).
58. Li X & Li Z xihaoli/STAARpipeline-Tutorial: v0.9.6 Version 0.9.6 10.5281/zenodo.6871408 (2022).

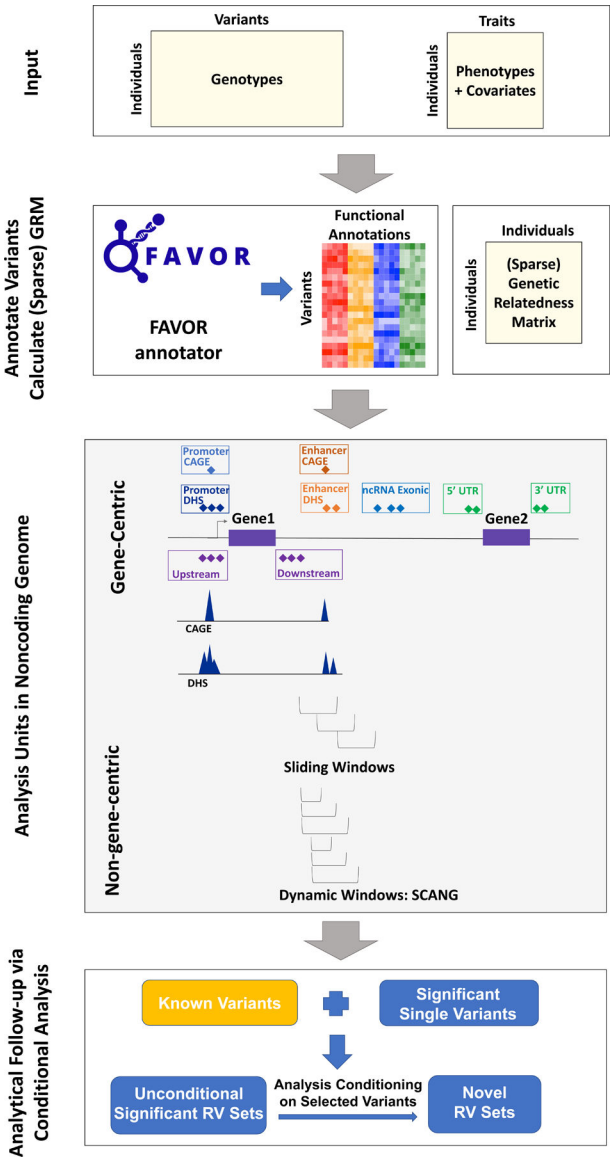


Fig. 1 |. Workflow of *STAARpipeline*.

(a) Prepare the input data of *STAARpipeline*, including genotypes, phenotypes and covariates. (b) Annotate all variants in the genome using FAVORannotator through FAVOR database and calculate the (sparse) genetic relatedness matrix. (c) Define analysis units in the noncoding genome: eight functional categories of regulatory regions, sliding windows and dynamic windows using SCANG. (d) Obtain genome-wide significant associations and perform analytical follow-up via conditional analysis.

Table 1 |

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

21,015 discovery samples and 9,123 replication samples from the NHLBI Trans-Omics for Precision Medicine (TOPMed) program are considered in the analysis. Results for the conditionally significant genes (unconditional STAAR-O $P < 3.57 \times 10^{-7}$ and conditional STAAR-O $P < 1.16 \times 10^{-3}$ for 7 different noncoding masks across protein-coding genes; unconditional STAAR-O $P < 2.50 \times 10^{-6}$ and conditional STAAR-O $P < 1.16 \times 10^{-3}$ for ncRNA genes) using discovery samples are presented in the table. The unconditional significant thresholds for both protein-coding genes of 7 different noncoding masks and ncRNA genes were defined by the multiple comparisons using the Bonferroni correction, that is, $0.05/(20{,}000 \times 7) = 3.57 \times 10^{-7}$ and $0.05/20{,}000 = 2.50 \times 10^{-6}$. STAAR-O is a two-sided test. Chr (Chromosome); Category (Functional category); #SNV (Number of rare variants (MAF < 1%) of the particular functional category in the gene); STAAR-O (STAAR-O *P* value); HDL-C (High-density lipoprotein cholesterol); LDL-C (Low-density lipoprotein cholesterol); TG (Triglycerides); TC (Total cholesterol); Variants Adjusted (Adjusted variants in conditional analysis); n/a, no variant adjusted in the conditional analysis.

Trait	Gene	Chr	Category	Discovery			Replication			Variants Adjusted
				#SNV	STAAR-O (Unconditional)	STAAR-O (Conditional)	#SNV	STAAR-O (Unconditional)	STAAR-O (Conditional)	
HDL-C	*APOA1*	11	enhancer_DHS	1862	2.19E-07	7.67E-07	1005	1.50E-03	3.17E-03	rs964184, rs12269901
LDL-C	*LDLR*	19	upstream	68	2.35E-17	4.24E-04	27	5.58E-01	6.31E-01	rs12151108, rs688, rs6511720
	LDLR	19	promoter_CAGE	131	1.88E-17	3.37E-04	56	2.51E-02	9.50E-02	rs12151108, rs688, rs6511720
	APOE	19	promoter_CAGE	91	1.45E-11	4.88E-12	35	1.86E-01	4.36E-02	rs7412, rs429358, rs35136575
	LDLR	19	promoter_DHS	257	4.03E-17	7.21E-04	113	5.74E-02	2.27E-01	rs12151108, rs688, rs6511720
	APOE	19	promoter_DHS	162	9.81E-11	3.41E-12	64	7.45E-02	3.42E-02	rs7412, rs429358, rs35136575
	LDLR	19	enhancer_CAGE	150	2.82E-17	5.01E-04	71	1.20E-02	4.05E-02	rs12151108, rs688, rs6511720
	APOE	19	enhancer_DHS	239	9.84E-11	2.03E-11	112	2.55E-01	1.34E-01	rs7412, rs429358, rs35136575
	CTC-527H23.4	16	ncRNA	32	1.15E-06	1.15E-06	17	9.12E-01	9.12E-01	n/a
TG	*APOE*	19	promoter_CAGE	92	4.45E-12	7.48E-06	36	9.45E-06	3.53E-05	rs12721054, rs5112, rs429358

Nat Methods. Author manuscript; available in PMC 2023 March 12.

Trait	Gene	Chr	Category	Discovery			Replication			Variants Adjusted
				#SNV	STAAR-O (Unconditional)	STAAR-O (Conditional)	#SNV	STAAR-O (Unconditional)	STAAR-O (Conditional)	
	APOA5	11	promoter_DHS	175	2.39E-08	3.47E-05	84	1.19E-04	8.78E-03	rs964184, rs9804646, rs3135506, rs2266788
	APOE	19	promoter_DHS	163	1.80E-11	9.56E-06	65	2.96E-06	1.13E-05	rs12721054, rs5112, rs429358
	COL18A1	21	enhancer_CAGE	256	1.92E-07	1.92E-07	147	4.57E-02	4.57E-02	n/a
	APOE	19	enhancer_DHS	241	2.02E-11	8.44E-05	116	1.12E-05	4.15E-05	rs12721054, rs5112, rs429358

Table 2 |

2-kb sliding window analysis results of unconditional analysis and analysis conditional on known common and low-frequency variants.

21,015 discovery samples and 9,123 replication samples from the NHLBI Trans-Omics for Precision Medicine (TOPMed) program are considered in the analysis. Results for the conditionally significant sliding windows (unconditional STAAR-O $P < 1.88 \times 10^{-8}$; conditional STAAR-O $P < 3.57 \times 10^{-4}$) using discovery samples are presented in the table. The unconditional significant threshold was defined by the multiple comparisons using the Bonferroni correction ($0.05/(2.66 \times 10^{6}) = 1.88 \times 10^{-8}$). STAAR-O is a two-sided test. Chr (Chromosome); Start Location (Start location of the 2kb sliding window); End Location (End location of the 2-kb sliding window); #SNV (Number of rare variants (MAF < 1%) in the 2-kb sliding window; STAAR-O (STAAR-O *P* value); HDL-C (High-density lipoprotein cholesterol); LDL-C (Low-density lipoprotein cholesterol); TG (Triglycerides); TC (Total cholesterol); Variants Adjusted (Adjusted variants in conditional analysis); n/a, no variant adjusted in the conditional analysis. Physical positions of each window are on build hg38.

Trait	Chr	Start Location	End Location	Gene	Discovery			Replication			Variants Adjusted
					#SNV	STAAR-O (Unconditional)	STAAR-O (Conditional)	#SNV	STAAR-O (Unconditional)	STAAR-O (Conditional)	
HDL-C	8	57,071,644	57,073,643	*Intergenic (IMPAD1)*	111	1.79E-08	1.79E-08	53	8.38E-01	8.38E-01	n/a
	11	116,802,930	116,804,929	*Intergenic (ZPR1)*	135	1.25E-08	4.31E-08	76	9.49E-05	2.02E-04	rs964184, rs12269901
	11	117,146,930	117,148,929	*Intronic (PAFAH1B2)*	165	5.98E-09	8.28E-08	98	6.02E-04	1.12E-03	rs964184, rs12269901
	11	117,147,930	117,149,929	*Intronic (PAFAH1B2)*	168	8.85E-09	1.22E-07	96	8.72E-04	1.64E-03	rs964184, rs12269901
	16	56,760,029	56,762,028	*Intronic (NUP93)*	132	1.38E-08	9.65E-06	68	2.45E-01	1.15E-01	rs247616, rs5883, rs7499892, rs17231520, rs5880
	16	56,761,029	56,763,028	*Intronic (NUP93)*	141	1.50E-08	1.09E-05	73	5.87E-01	2.26E-01	rs247616, rs5883, rs7499892, rs17231520, rs5880
LDL-C	1	55,333,498	55,335,497	*Intergenic (GOT2P1)*	171	6.66E-16	5.81E-07	95	1.27E-06	5.81E-07	rs11591147, rs28362263, rs505151, rs12117661, rs472495
	1	55,334,498	55,336,497	*Intergenic (GOT2P1)*	148	5.55E-16	5.49E-07	81	1.20E-06	5.49E-07	rs11591147, rs28362263, rs505151, rs12117661, rs472495

Trait	Chr	Start Location	End Location	Gene	Discovery			Replication			Variants Adjusted
					#SNV	STAAR-O (Unconditional)	STAAR-O (Conditional)	#SNV	STAAR-O (Unconditional)	STAAR-O (Conditional)	
TG	11	117,146,930	117,148,929	*Intronic (PAFAH1B2)*	164	7.81E-19	4.13E-18	93	2.17E-17	5.66E-17	rs964184, rs9804646, rs3135506, rs2266788
	11	117,147,930	117,149,929	*Intronic (PAFAH1B2)*	165	1.15E-18	6.11E-18	94	3.47E-17	9.13E-17	rs964184, rs9804646, rs3135506, rs2266788
	19	44,882,528	44,884,527	*Intronic (NECTIN2)*	145	1.06E-08	2.18E-07	88	2.71E-02	8.07E-01	rs12721054, rs5112, rs429358
TC	1	55,333,498	55,335,497	*Intergenic (GOT2P1)*	175	1.98E-13	3.83E-14	101	5.84E-07	1.88E-07	rs11591147, rs28362263, rs505151, rs12117661, rs2495477
	1	55,334,498	55,336,497	*Intergenic (GOT2P1)*	149	1.80E-13	3.49E-14	90	5.53E-07	1.78E-07	rs11591147, rs28362263, rs505151, rs12117661, rs2495477
	19	44,894,528	44,896,527	*Intronic (TOMM40)*	180	2.73E-10	8.95E-08	97	2.68E-03	4.22E-01	rs7412, rs429358, rs12721054

Table 3 |

Dynamic window analysis results of unconditional analysis and analysis conditional on known common and low-frequency variants.

21,015 discovery samples and 9,123 replication samples from the NHLBI Trans-Omics for Precision Medicine (TOPMed) program are considered in the analysis. Results for the conditionally significant sliding windows (unconditional genome-wide error rate $GWER < 0.05$; conditional STAAR-S $P < 5.56 \times 10^{-4}$) using discovery samples are presented in the table. STAAR-S is a two-sided test. Chr (Chromosome); Start Location (Start location of the dynamic window); End Location (End location of the dynamic window); #SNV (Number of rare variants (MAF < 1%) in the dynamic window; GWER (genome-wide error rate); STAAR-S (STAAR-S *P* value); HDL-C (High-density lipoprotein cholesterol); LDL-C (Low-density lipoprotein cholesterol); TG (Triglycerides); TC (Total cholesterol); Variants Adjusted (Adjusted variants in conditional analysis). Physical positions of each window are on build hg38.

Trait	Chr	Start Location	End Location	Gene	Discovery				Replication			Variants Adjusted
					#SNV	GWER	STAAR-S (Unconditional)	STAAR-S (Conditional)	#SNV	STAAR-S (Unconditional)	STAAR-S (Conditional)	
HDL-C	11	116,866,780	116,867,288	*Intronic (SIK3)*	40	0.0295	2.24E-09	8.45E-09	19	2.22E-05	5.46E-05	rs964184, rs12269901
	11	116,928,564	116,929,045	*Intronic (SIK3)*	40	0.0025	1.50E-10	4.43E-10	18	7.81E-04	1.06E-03	rs964184, rs12269901
LDL-C	1	55,335,150	55,335,701	*Intergenic (GOT2P1)*	40	<0.0005	8.58E-18	7.49E-19	21	9.29E-07	4.80E-07	rs11591147, rs28362263, rs505151, rs12117661, rs472495
	19	11,319,992	11,320,870	*Intronic (TSPAN16)*	60	0.02	1.44E-09	3.16E-05	41	5.04E-01	5.10E-01	rs12151108, rs688, rs6511720
TG	11	117,147,061	117,148,086	*Intronic (PAFAH1B2)*	80	<0.0005	5.10E-16	8.55E-15	41	9.48E-19	3.44E-18	rs964184, rs9804646, rs3135506, rs2266788
	11	117,182,856	117,183,310	*Intronic (SIDT2)*	40	<0.0005	3.96E-12	1.08E-11	15	3.77E-14	6.53E-14	rs964184, rs9804646, rs3135506, rs2266788
	11	117,349,560	117,350,171	*Intronic (CEP164)*	50	0.013	1.08E-09	1.26E-09	29	4.12E-11	6.39E-11	rs964184, rs9804646, rs3135506, rs2266788
TC	1	55,291,905	55,293,502	*Intergenic (GOT2P1)*	140	0.0055	3.17E-10	8.77E-05	68	4.76E-01	2.30E-01	rs11591147, rs28362263, rs505151,

Trait	Chr	Start Location	End Location	Gene	Discovery				Replication			Variants Adjusted
					#SNV	GWER	STAAR-S (Unconditional)	STAAR-S (Conditional)	#SNV	STAAR-S (Unconditional)	STAAR-S (Conditional)	
												rs12117661, rs2495477
	1	55,335,119	55,335,584	*Intergenic (GOT2P1)*	40	<0.0005	1.63E-15	4.44E-16	26	2.23E-07	7.03E-08	rs11591147, rs28362263, rs505151, rs12117661, rs2495477
	19	11,319,627	11,320,925	*Intronic (TSPAN16)*	110	<0.0005	2.95E-12	2.32E-05	75	3.40E-01	5.90E-01	rs73015024, rs688, rs2278426, rs6511720