Title

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

Journal

Nature Genetics, 52(9)

ISSN

1061-4036

Authors

Li, Xihao
Li, Zilin
Zhou, Hufeng
et al.

Publication Date

2020-09-01

DOI

10.1038/s41588-020-0676-4

Peer reviewed

HHS Public Access

Author manuscript

Nat Genet. Author manuscript; available in PMC 2021 February 24.

Published in final edited form as:

Nat Genet. 2020 September ; 52(9): 969–983. doi:10.1038/s41588-020-0676-4.

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

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

Abstract

Large-scale whole genome sequencing (WGS) studies have enabled the analysis of rare variants (RVs) associated with complex phenotypes. Commonly used RV association tests (RVATs) have limited scope to leverage variant functions. We propose STAAR (variant-Set Test for Association using Annotation infoRmation), a scalable and powerful RVAT method that effectively incorporates both variant categories and multiple complementary annotations using a dynamic weighting scheme. For the latter, we introduce "annotation Principal Components", multi-dimensional summaries of *in silico* variant annotations. STAAR accounts for population structure and relatedness, and is scalable for analyzing very large cohort and biobank WGS studies of continuous and dichotomous traits. We applied STAAR to identify RVs associated with four lipid traits in 12,316 discovery samples and 17,822 replication samples from the Trans-Omics for Precision Medicine program. We discovered and replicated novel RV associations, including disruptive missense RVs of *NPC1L1* and an intergenic region near *APOC1P1* associated with low-density lipoprotein cholesterol.

An increasing number of whole genome/exome sequencing (WGS/WES) studies are being conducted to investigate the genetic bases of human diseases and traits, including the Trans-Omics for Precision Medicine Program (TOPMed) of the National Heart, Lung and Blood Institute (NHLBI) and the Genome Sequencing Program (GSP) of the National Human Genome Research Institute (NHGRI). Such studies enable assessment of associations between complex traits and both coding and non-coding rare variants (RVs; minor allele frequency (MAF) < 1%) across the genome. However, single-variant analyses typically have

---

* xlin@hsph.harvard.edu.

Author contributions

X. Li, Z.L., H.Z., G.R.A., J.I.R., C.J.W., G.M.P., P.N. and X. Lin designed the experiments. X. Li, Z.L., H.Z., and X. Lin performed the experiments. X. Li, Z.L., H.Z., S.M.G., Y.L., H.C., R.S., R.D., D.K.A., S.A., C.M.B., L.F.B., J.B., E.B., D.W.B., J.G.B., M.P.C., A.C., L.A.C., J.E.C., B.I.F., X.G., G.H., M.R.I., S.L.R.K., S.K., A.T.K., C.L.K., C.C.L., X.S.L., M.C.M., A.W.M., L.W.M., R.A.M., S.T.M., B.D.M., M.E.M., J.E.M., A.C.M., J.R.O., N.D.P., A.P., J.M.P., P.A.P., B.M.P., S.R., K.M.R., S.S.R., J.A.S., H.K.T., M.Y.T., R.S.V., F.F.W., D.E.W., Z.W., J.G.W., L.R.Y., B.M.N., S.R.S., G.R.A., J.I.R., C.J.W., G.M.P., P.N., 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. X. Li, Z.L., S.M.G., J.I.R., G.M.P., P.N., and X. Lin drafted the manuscript and revised according to co-author suggestions. All authors critically reviewed the manuscript, suggested revisions as needed, and approved the final version.

URLs

STAAR (version 0.9.5), https://github.com/xihaoli/STAAR and https://content.sph.harvard.edu/xlin/software.html.

A full list of NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium members and TOPMed Lipids Working Group members appear at the end of the paper.

low power to identify associations with rare variants[1–3]. To improve power, variant-set tests have been proposed to jointly test the effects of given sets of multiple rare variants. These methods include the burden test[4–7], Sequence Kernel Association Test (SKAT)[8], and their various combinations[9–12]. In parallel, external biological information provided by functional annotations, such as conservation scores and predicted enhancer status, has been successfully used for prioritizing plausibly causal common variants in fine-mapping studies, partitioning heritability in GWAS, and predicting genetic risk[13–17]. It is of substantial interest to incorporate variant functional annotations effectively, to boost the power of RV analysis of WGS association studies[18,19].

Variant functional annotations take two forms: (i) qualitative functional groupings into genomic elements, such as Variant Effect Predictor (VEP) categories[20,21], and (ii) quantitative functional scores available for variants across the genome, including protein functional scores[22,23], evolutionary conservation scores[24,25], epigenetic measures[26], and integrative functional scores[27]. Different annotation scores capture diverse aspects of variant function[28,29]. Given the diversity of available annotations, efforts have been made to aggregate the evidence they provide on genomic function[30]. Simultaneous use of multiple, varied functional annotation scores in variant-set tests could improve rare variant association study (RVAS) power, for example, by optimally selecting and weighting plausibly-causal rare variants[31].

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

Recent methods[32–34] have incorporated functional annotations in genetic association studies. However, these methods are not scalable to analyze large-scale WGS studies while accounting for relatedness and population structure. Large scale WGS and WES studies, such as TOPMed and GSP, include a considerable fraction of related and ancestrally diverse samples. STAAR accounts for both relatedness and population structure, as well as longitudinal follow-up designs, for both quantitative and dichotomous traits, using a Generalized Linear Mixed Models (GLMM) framework[35] that includes linear and logistic mixed models[36,37]. Using sparse Genetic Relatedness Matrices (GRMs)[38], STAAR is computationally scalable for very large WGS studies and biobanks of hundreds of thousands of samples.

We perform herein extensive simulation studies to demonstrate that STAAR can achieve substantially greater power compared to conventional variant-set tests, while maintaining accurate type I error rates for both quantitative and dichotomous phenotypes. We then apply STAAR to perform WGS gene-centric and sliding window-based genetic region analysis of 12,316 discovery samples and 17,822 replication samples with four quantitative lipid traits: low-density lipoprotein cholesterol (LDL-C), high-density lipoprotein cholesterol (HDL-C),

triglycerides (TG), and total cholesterol (TC) from the NHLBI TOPMed program. We show that STAAR outperforms existing methods and identifies novel and replicated associations, including with LDL-C in disruptive missense RVs of *NPC1L1*, and in an intergenic region near *APOC1P1*.

Results

Overview of methods.

STAAR is a general framework for analyzing WGS RVAS at scale by using both qualitative functional categories as well as multiple *in silico* variant annotation scores within a variant-set, while accounting for population structure and relatedness by fitting linear and logistic mixed models for quantitative and dichotomous traits using fast and scalable algorithms. For each variant-set, there are two main components of the STAAR framework: (i) using annotation PCs to capture and prioritize multi-dimensional variant biological functions, and (ii) testing the association between each variant-set and phenotypes by incorporating these annotation PCs as well as other integrative functional scores and MAFs in the STAAR test statistics using an omnibus weighting scheme (Fig. 1).

Variants often influence genes and gene products through multiple mechanisms. We extract a broad set of variant functional annotations (Supplementary Table 1), including both individual and ensemble functional scores, from various databases, such as ENCODE[26], Roadmap Epigenomics[39], and other evolutionary and protein annotation databases[27,40,41]. A correlation heatmap across variants in the genome (Fig. 2) shows that the correlation structure among all individual annotations is approximately block-diagonal, with highly correlated blocks representing different classes of variant function, e.g., epigenetic function, evolutionary conservation, protein function, local nucleotide diversity. We introduce annotation Principal Components defined as the first PCs calculated from the set of individual functional annotation scores in each functional block (Supplementary Table 1 and Online Methods). Annotation PCs effectively reduce the dimensionality of the large number of individual annotations and summarize multiple aspects of variant function.

The STAAR framework first calculates a set of multiple candidate test statistics using different annotation weights under a particular testing approach (Fig. 1d). For each type of RV test, STAAR then uses ACAT (aggregated Cauchy association test) method to combine the resulting *P*-values calculated using different weights in order to effectively and powerfully aggregate the association strength from all annotations in a data-adaptive manner (Fig. 1d and Online Methods). The ACAT method for combining *P*-values is accurate and computationally efficient, while accounting for arbitrary correlation structure between tests[9,42]. To leverage the advantages of different types of tests, we propose an omnibus test in the STAAR framework (STAAR-O) by combining *P*-values across different types of multiple-annotation-weighted variant-set tests using the ACAT method (Fig. 1d and Online Methods).

Simulation studies.

To evaluate the type I error and power of STAAR compared to conventional variant-set tests, we performed simulation studies under a variety of configurations. We followed the steps described in the Data simulation section of the Online Methods to generate both continuous and dichotomous phenotypes. We generated genotypes by simulating 20,000 sequences for 100 different regions with each spanning 1 Mb. The data were generated to mimic the linkage disequilibrium (LD) structure of an African American population by using the calibration coalescent model (COSI)[43]. We randomly selected 5-kb regions from these 1-Mb regions and considered sample sizes of 2,500, 5,000, and 10,000 for each replicate. The simulation studies focused on aggregating uncommon variants with MAF < 5%.

Type I error simulations.

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

Empirical power simulations.

Next, we evaluated the power of STAAR empirically by incorporating MAF and 10 annotations into its analysis and comparing results with conventional variant-set tests in a variety of configurations. Power was estimated as the proportion of *P*-values less than $\alpha = 10^{-7}$ based on 10^4 replicates. Causality of variants was allowed to be dependent on different sets of annotations through a logistic model (Online Methods). We considered different proportions of causal variants (5%, 15%, 35% on average) in the signal region. For both continuous and dichotomous traits, STAAR-O incorporating all 10 annotations had higher power than the conventional variant-set tests in terms of signal region detection (Supplementary Figs. 1–4). Power simulation results of STAAR-O for different magnitudes of effect sizes and different proportions of effect size directions yielded the same conclusion (Supplementary Figs. 1, 5, and 6). Overall, our simulation studies showed that STAAR-O could provide considerably higher power than conventional variant-set tests.

Association analysis of lipid traits in the TOPMed WGS data.

We applied STAAR to identify RV-sets associated with four quantitative lipid traits (LDL-C, HDL-C, TG and TC) using TOPMed WGS data[44,45]. LDL-C and TC were adjusted for the presence of medications as before[44]. DNA samples were sequenced at >30X target coverage. The discovery phase consists of four study cohorts of TOPMed Freeze 3. The replication phase consists of ten different study cohorts in TOPMed Freeze 5 that were not in Freeze 3 (Supplementary Note and Supplementary Table 3).

Sample-level and variant-level quality control (QC) were performed[44,45]. There were 12,316 discovery samples, which had 155 million single nucleotide variants (SNVs), and 17,822 replication samples, which had 188 million SNVs. The TOPMed data consist of ancestrally diverse and multi-ethnic related samples. Race/ethnicity was defined using a combination of self-reported race/ethnicity and study recruitment information. The discovery cohorts consist of 4,580 (37.2%) Black or African American, 6,266 (50.9%) White, 543 (4.4%) Asian American, and 927 (7.5%) Hispanic/Latino American. Among all samples in discovery phase, 3,577 (29.0%) had first-degree relatedness, 491 (4.0%) had second-degree relatedness, and 273 (2.2%) had third-degree relatedness (Supplementary Fig. 7). Among all SNVs observed in the discovery samples, there were 6.5 million (4.2%) common variants (MAF > 5%), 5.3 million (3.4%) low frequency variants (1% ≤ MAF ≤ 5%), and 143.2 million (92.4%) rare variants (MAF < 1%). The race/ethnicity distribution, related sample distribution, and variant number distribution for replication phase and pooled samples (samples from both discovery phase and replication phase) are given in Supplementary Table 4.

Our study used the proposed STAAR-O method to perform (i) gene-centric analysis using RV-sets based on functional categories, and (ii) genetic region analysis using variant-sets defined by 2-kb sliding windows with 1-kb skip length across the genome. We adjusted for age, age^2, sex, race/ethnicity, study, and the first 10 ancestral PCs, while controlling for relatedness using linear mixed models, with inverse-rank normal transformation applied to phenotypes (Online Methods). Race/ethnicity was included as a covariate to adjust for sociocultural and environmental factors, while genetic ancestry differences were captured by the inclusion of the ancestral PCs. In addition to the two MAF weights[3], we incorporated 13 aggregated functional annotation scores in STAAR-O: 3 integrative scores (CADD[27], LINSIGHT[46], and FATHMM-XF[47]) and 10 aPCs. Figure 2 summarizes the correlation among all functional annotations, including 60 individual scores, 3 integrative scores, and 10 aPCs.

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

We performed gene-centric analysis to identify whether rare variants in coding, promoter, and enhancer regions of genes are associated with lipid traits using STAAR-O. For each of the four lipid traits, we analyzed five functional categories (masks) of coding and non-coding variants: (i) pLoF (stop gain, stop loss and splice) RVs, (ii) missense RVs, (iii) synonymous RVs, (iv) promoter RVs, and (v) enhancer RVs. The pLoF, missense, and synonymous RVs were defined by GENCODE VEP categories[20,21]. The promoter RVs were defined as RVs in the +/− 3-kb window of transcription starting site (TSS) with overlap of Cap Analysis of Gene Expression (CAGE) sites. The enhancer RVs were defined as RVs in GeneHancer predicted regions with overlap of CAGE sites[48–50]. Within each gene functional category, we tested for an association between rare variants (MAF < 1%) in the functional category and lipid traits using STAAR-O with the 13 aggregated functional annotations described above. For missense RVs, we incorporated an additional annotation functional category predicting functionally “disruptive” variants determined by MetaSVM[51], which measures the deleteriousness of missense mutations. The overall distributions of STAAR-O *P*-values were well calibrated for all four lipid phenotypes (Supplementary Fig. 8). We

considered in unconditional analysis a Bonferroni-corrected genome-wide significance threshold of $\alpha = 0.05/(20,000 \times 5) = 5.00 \times 10^{-7}$ accounting for five different masks across protein-coding genes.

STAAR-O identified 21 genome-wide significant associations with four lipid phenotypes using unconditional analysis of the discovery samples (Supplementary Table 5 and Supplementary Fig. 9). After conditioning on known lipids-associated variants[44,52–67], 11 out of the 21 associations remained significant at the Bonferroni correction level $0.05/21 = 2.38 \times 10^{-3}$ using the discovery samples. These included associations with LDL-C (pLoF RVs in *PCSK9* and *APOB*, missense RVs in *PCSK9*, *NPC1L1*, and *APOE*), association with HDL-C (pLoF RVs in *APOC3*), association with TG (pLoF RVs in *APOC3*), and associations with TC (pLoF RVs in *PCSK9* and *APOB*, missense RVs in *PCSK9* and *LIPG*) (Table 1). Of these 11 associations, 10 were replicated at the Bonferroni-corrected level $0.05/11 = 4.55 \times 10^{-3}$ after adjusting for known lipid-associated variants. The association between *APOC3* pLoF RVs and HDL-C was unreported in a previous study using the same TOPMed Freeze 3 data[44].

The association between missense RVs in *NPC1L1* and LDL-C was not detected by the conventional variant-set tests and has not been observed in previous studies[44,55,68,69]. In the discovery phase, its unconditional STAAR-O *P*-value was 1.29×10^{-7}, while the most significant conventional variant-set test was the burden test with $P = 7.04 \times 10^{-6}$. This association was not driven by any single RV (minimum single RV *P*-value $> 10^{-3}$) but was due to the aggregated effects of multiple missense RVs. The *P*-value of the burden test additionally weighted by MetaSVM was the smallest of all annotations ($P = 3.15 \times 10^{-9}$), highlighting the significant association between disruptive missense RVs in *NPC1L1* and LDL-C (Supplementary Fig. 10). Among all 174 missense RVs in *NPC1L1* from the discovery samples, the disruptive missense RVs as predicted by MetaSVM were enriched among variants with higher aPC-Conservation scores (Supplementary Table 6). This contributed to the test weighted by aPC-Conservation being the most significant across all quantitative annotation-weighted tests included in STAAR-O (burden $P = 3.12 \times 10^{-7}$). As aPC-Conservation summarizes variants' evolutionary conservation scores, it is informative in predicting whether or not variants are deleterious and thus functional[70,71]. Conditioning on the ten known common variants in *NPC1L1* associated with LDL-C (Supplementary Table 7)[57–61,65–67], the association between disruptive missense RVs in *NPC1L1* and LDL-C remained significant after Bonferroni correction with the conditional analysis $P = 9.27 \times 10^{-9}$ in discovery phase.

This association was validated in replication phase with $P = 2.59 \times 10^{-4}$ and with $P = 4.02 \times 10^{-11}$ in pooled samples in conditional analysis. This significant association was also validated using whole exome sequencing data from the UK Biobank[72] ($n = 40,519$) with $P = 2.49 \times 10^{-4}$ in conditional analysis.

Genetic region analysis of rare variants.

We performed genetic region analysis to determine whether RVs within sliding windows are associated with lipid traits. The sliding windows were defined to be 2 kb in length, start at position 0 bp for each chromosome, and have a skip length of 1 kb. Windows with a total minor allele count less than 10 were excluded from the analysis, resulting in a total of 2.66 million 2-kb overlapping windows, with a median of 104 RVs in each sliding window among discovery samples. For each 2-kb window, we tested for an association between the RVs in the window and each lipid trait using STAAR-O by incorporating 13 aggregated quantitative annotations. The overall distributions of STAAR-O *P*-values were well calibrated for all four lipid phenotypes (Fig. 3b and Supplementary Figs. 11b, 12b, and 13b). Using the Bonferroni correction, we set the genome-wide significance threshold at $\alpha = 0.05/(2.66 \times 10^6) = 1.88 \times 10^{-8}$ across sliding windows (Fig. 3a and Supplementary Figs. 11a, 12a, and 13a). Supplementary Table 8 summarizes the significant 2-kb sliding windows identified using STAAR-O. Overall, by dynamically incorporating multiple functional annotations capturing different aspects of variant function, STAAR-O was able to detect more significant sliding windows, and showed consistently smaller *P*-values for top sliding windows compared with conventional variant-set tests weighted using MAFs (Fig. 3c,d and Supplementary Figs. 11c–f, 12c, and 14). Burden tests were not able to detect any window that reached significance.

Among the 59 genome-wide significant sliding windows detected by STAAR-O in unconditional analysis, 17 remained significant at the Bonferroni correction level $0.05/59 = 8.47 \times 10^{-4}$ after conditioning on known lipids-associated variants using the discovery samples (Table 2). For LDL-C, the significant sliding windows were located in gene *PCSK9* or in a 50-kb region on chromosome 19 including the *APOE* cluster. For TC, all of the significant sliding windows were located in the same areas as for LDL-C. For TG, STAAR-O detected two consecutive significant sliding windows within *APOC3*, whereas no significant sliding windows were detected for HDL-C. Of these 17 associations, six were replicated at level $0.05\ /17 = 2.94 \times 10^{-3}$ after Bonferroni correction and another four were replicated at level $0.05/9 = 5.56 \times 10^{-3}$ after Bonferroni correction for nine non-overlapping sliding windows in conditional analysis of replication samples[17], including a sliding window located downstream of *APOC1P1* (Chr 19: 44,931,528 bp - 44,933,527 bp), which was significantly associated with LDL-C but undetected by the burden test, SKAT, and ACAT-V (Table 2 and Fig. 3c).

The top variant of the significant sliding window located downstream of *APOC1P1* was rs370625306 (MAF = 0.005, $P = 8.71 \times 10^{-8}$), which was not significant at a Bonferroni-corrected threshold ($\alpha = 0.05/(1.51 \times 10^7) = 3.31 \times 10^{-9}$) in individual variant analysis. This rare variant and the second top variant in these windows (rs9749443, MAF = 0.009, $P = 2.46 \times 10^{-5}$) were upweighted by aPC-Epigenetic in STAAR-O (Supplementary Fig. 15). Specifically, the aPC-Epigenetic scores of rs370625306 and rs9749443 ranked in the top 10% and top 30% among all RVs, respectively, in each sliding window. Conditioning on the two known common variants rs7412 and rs429358 in *APOE* associated with LDL-C[55], the strength of association of both sliding windows was reduced

but remained significant (Table 2). Similar results were found after further conditioning on *APOE* haplotypes using these two SNPs (Supplementary Table 8). This suggests that the effects of RVs in this sliding window are not fully captured by the two known common LDL-associated variants. STAAR-O also identified and replicated two highly significant windows in *APOC3* associated with TG in conditional analysis that were undetected by SKAT and burden test[73].

STAAR identifies more associations using relevant tissue functional annotations.

To evaluate the effect of tissue specificity, we compared the performance of STAAR-O in both gene-centric and genetic region analysis by incorporating liver (a central hub for lipid metabolism), heart, and brain annotations. For each tissue, we calculated a tissue-specific aPC from tissue-specific DNase, H3K4me3, H3K27ac and H3K27me3 from ENCODE (Supplementary Table 9)[26,74]. We used tissue-specific CAGE sites with overlap of RVs in the +/− 3-kb window of TSS and GeneHancers to define promoter and enhancer RV masks in gene-centric analysis. To make a fair comparison between tissues, we calculated STAAR-O *P*-values based solely on the tissue-specific aPC and without incorporating the MAF and other annotations.

Overall, the use of liver annotation resulted in more significant levels of association than heart and brain annotations, as would be expected for lipid traits, although no additional replicated conditionally significant association was detected by using tissue-specific annotations. STAAR-O identified 9 and 8 replicated conditionally significant associations by using liver annotation in gene-centric and genetic region analysis, respectively (Supplementary Tables 10 and 11). Among these 17 significant associations, two were not seen when heart annotation was used and two were not seen when brain annotation was used, and no additional associations were detected by using heart and brain annotations (Supplementary Tables 10 and 11). Furthermore, more suggestive significant associations were detected when using liver annotation than the other two tissues at various levels of unconditional *P*-value thresholds in the discovery phase (Supplementary Figs. 16 and 17).

Computation cost.

We developed an R package, STAAR, to perform scalable variant-set association tests incorporating multiple variant annotations for WGS RVAS. Using sparse GRMs[38], STAAR scales well both in terms of computation time and memory for very large-scale WGS association studies, such as sample sizes in TOPMed, GSP, and UK Biobank. The computation time for STAAR-O to perform WGS gene-centric and genetic region analysis on 30,000 related samples using the TOPMed data requires 15 hours for 100 2.10 GHz computing cores with 6 GB memory for each lipid trait. Analyzing 500,000 simulated related samples mimicking the UK Biobank sample size requires 26 hours for WGS analysis using the same approach and computational resources (Online Methods).

Discussion

We propose STAAR as a general, computationally scalable framework that effectively incorporates multiple qualitative and quantitative variant functional annotations to boost

power for variant-set tests for continuous and binary traits in WGS RVAS, while accounting for both population structure and relatedness using GLMMs.

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

In a WGS RV analysis of lipid traits using the TOPMed data, STAAR-O identified several conditionally significant functional categories associated with lipid traits in gene-centric analysis (including *NPC1L1* missense RVs and LDL-C; *APOC3* pLoF RVs and HDL-C; and *LIPG* missense RVs and TC) that were missed by the previous study using the same TOPMed data[44]. Earlier studies reported marginal association between inactivating mutations (pLoF RVs and frameshift indels) in *NPC1L1* and LDL-C with $P = 0.04$[69], which was replicated using the pooled TOPMed samples ($P = 0.02$), no significant association between pLoF RVs and LDL-C was found ($P = 0.15$). STAAR-O identified much more significant novel association, which replicated, between missense RVs in *NPC1L1* and LDL-C, which was driven by disruptive missense RVs (conditional $P = 4.02 \times 10^{-11}$ in pooled samples). None of these disruptive missense RVs was reported in ClinVar[75], suggesting that the findings from emerging WGS studies can help guide the expansion of the ClinVar database. *NPC1L1* is the direct molecular target of the lipid-lowering drug ezetimibe, which reduces the absorption of cholesterol by binding to *NPC1L1*[76]. STAAR-O also suggested several conditional associations in the discovery phase that were validated in our replication phase and achieved significance in pooled samples (Supplementary Table 12).

In agnostic sliding-window based genetic region analysis, STAAR-O detected and replicated 10 sliding windows after conditioning on known variants, including association between an intergenic region located downstream of *APOC1P1* and LDL-C, that were not detected using conventional tests. This detected *APOC1P1* region is located in the hepatic control region 2 (HCR-2) that regulates hepatic expression of apolipoproteins. By further conditioning on the *APOE* haplotypes and rs35136575, a common variant previously found in the downstream HCR-2 associated with LDL-C[77], the strength of association was reduced but remained significant (Supplementary Table 8). This discovery is due to upweighting several plausibly causal rare variants that have regulatory functions using aPC-Epigenetic scores in STAAR-O (Supplementary Fig. 15 and Supplementary Table 13). These results highlight that incorporating multiple functional annotations using STAAR can effectively boost power for WGS RVAS.

To capture multiple aspects of variant functionality, we introduced annotation PCs by performing dimension reduction of a large number of diverse individual annotations from various external databases. See Online Methods for an example demonstrating that aPCs explain diverse and complementary functionality of known LDL-associated functional rare

variants, and STAAR provides greater power for RV association tests by upweighting these variants using aPCs.

In practice, STAAR is very flexible and users can determine the set of individual annotations to calculate aPCs and the number of aPCs and integrative functional scores and other qualitative scores to be used, as well as tissue, cell-type and phenotype-specific variant annotations[78–80]. In this paper, we group the individual annotations based on biological knowledge; users can also apply data-driven approaches, such as clustering, to group annotations for aPC calculation. We also demonstrate that STAAR detects more associations using relevant tissue functional annotations. It will be of interest, in future research, to incorporate improved rare variant effect size models in the weights to further improve power for RVAS[81,82].

The STAAR procedure is fast and scalable for very large WGS studies and biobanks of hundreds of thousands to millions of samples for both quantitative and dichotomous phenotypes as it uses estimated sparse GRMs[38] to fit the null GLMM and to scan the genome. Besides using sliding windows of a pre-specified fixed window length, STAAR could be extended to flexibly detect the sizes and locations of coding and non-coding rare variant association regions using the dynamic window analysis method SCANG[83]. In addition, STAAR could be extended to settings with survival, unbalanced case-control, and multiple phenotypes, and hence could provide a comprehensive framework for WGS RVAS. Thus, STAAR provides a powerful and flexible tool for variant association discovery in many settings to explore the molecular basis of common diseases.

Online Methods

Notations and model.

Suppose there are n subjects with M total variants sequenced across the whole genome. Given a genetic set of p variants, for subject i, let Y_i denote a continuous or dichotomous trait with mean μ_i; $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{iq})^T$ 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 variant-set.

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

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

Where $g(\mu) = \mu$ for a continuous normally distributed trait, $g(\mu) = \text{logit}(\mu)$ for a dichotomous trait, α_0 is an intercept, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_q)^T$ 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 $\boldsymbol{G}_i$.

When the data consist of related samples, we consider the following Generalized Linear Mixed Model (GLMM)[35–37]

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

where the random effects b_i account for remaining population structure unaccounted by ancestral PCs, relatedness, and other between-observation correlation. We assume that $\boldsymbol{b} = (b_1, \ldots, b_n)^T \sim N\left(\boldsymbol{0}, \sum_{l=1}^{L} \theta_l \boldsymbol{\Phi}_l\right)$ with variance components θ_l and known covariance matrices $\boldsymbol{\Phi}_l$. The random effects $\boldsymbol{b}$ can be decomposed into a sum of multiple random effects to account for different sources of relatedness and correlation as $\boldsymbol{b} = \sum_{l=1}^{L} \boldsymbol{b}_l$ with $\boldsymbol{b}_l \sim N(\boldsymbol{0}, \theta_l \boldsymbol{\Phi}_l)$. For example, $\boldsymbol{b}_1$ accounts for population structure and family relatedness by using the Genetic Relatedness Matrices (GRMs) as its covariance matrix $\boldsymbol{\Phi}_1$[84,85]. A sparse GRM can be used to scale up computation[38]. Additional random effects $\boldsymbol{b}_2, \cdots, \boldsymbol{b}_L$ can be used to account for complex sampling designs, such as correlation between repeated measures from longitudinal studies using subject-specific random intercepts and slopes and hierarchical designs. The remaining variables are defined in the same way as those in the GLM (1). Under both the GLM and the GLMM, we are interested in testing the null hypothesis of whether the variant-set is associated with the phenotype, adjusting for covariates and relatedness, which corresponds to $H_0: \boldsymbol{\beta} = \boldsymbol{0}$, that is, $\beta_1 = \beta_2 = \cdots = \beta_p = 0$.

Conventional variant-set tests.

Conventional score-based aggregation methods allow for jointly testing the association between variants in the genetic set and phenotype. In particular, burden tests[4–7] assume that $\beta_j = w_j \beta$, where β is a constant for all variants, such that the corresponding burden test statistic to test $H_0: \boldsymbol{\beta} = \boldsymbol{0} \Leftrightarrow H_0: \beta = 0$ is given by

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

where $S_j = \sum_{i=1}^{n} G_{ij}(Y_i - \hat{\mu}_i)$ is the score statistic of the marginal model for variant j and $\hat{\mu}_i$ is the estimated mean of Y_i under the null GLM $g(\mu_i) = \alpha_0 + \boldsymbol{X_i^T}\boldsymbol{\alpha}$ or the null GLMM $g(\mu_i) = \alpha_0 + \boldsymbol{X_i^T}\boldsymbol{\alpha} + b_i$. Q_{Burden} asymptotically follows a chi-square distribution with 1 degree of freedom under the null hypothesis, and its P-value can be obtained analytically while accounting for linkage disequilibrium (LD) between variants[3,37].

For SKAT[8], the β_j's are assumed to be independent and identically distributed (i.i.d.) following an arbitrary distribution, with $E(\beta_j) = 0$ and $Var(\beta_j) = w_j^2 \tau$. The null hypothesis of no variant-set effect $H_0: \boldsymbol{\beta} = 0$ is equivalent to $H_0: \tau = 0$, and the corresponding SKAT test statistic is given by

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

Q_{SKAT} asymptotically follows a mixture of chi-square distributions under the null hypothesis, and its *P*-value can be obtained analytically while accounting for LD between variants[3,37].

Further, the recently proposed ACAT-V test uses a combination of transformed variant *P*-values rather than operating on the test statistics directly[9]. The ACAT-V test statistic is given by

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

where p' is the number of variants with minor allele count (MAC) greater than 10 and p_j is the association *P*-value of individual variant j corresponding the individual variant score statistics S_j for those variants with MAC > 10. p_0 is the burden test *P*-value of extremely rare variants with MAC ≤ 10 and $\overline{w^2\mathrm{MAF}(1-\mathrm{MAF})}$ is the average of the weights $w_j^2\mathrm{MAF}_j(1-\mathrm{MAF}_j)$ among the extremely rare variants with MAC ≤ 10. Q_{ACAT-V} can be well approximated by a Cauchy distribution under the null hypothesis, and its *P*-value can be obtained analytically while accounting for LD between variants[9]. For binary traits in highly unbalanced designs, one can improve individual *P*-value calculations using Saddlepoint approximation[86,87].

These conventional approaches consider a weight w_j defined as a threshold indicator or a function of minor allele frequency (MAF) for variant j, i.e. $w_j = \mathrm{Beta}(\mathrm{MAF}_j; a_1, a_2)^3$. Common choices of the parameters are $a_1 = 1$ and $a_2 = 25$ which upweights rarer variants, or $a_1 = 1$ and $a_2 = 1$, which corresponds to equal weights for all variants. In WGS studies, the vast majority of rare variants across the genome are not causal. Thus, choosing their weights according to MAF will incorrectly upweight many such “noise” variants in a variant-set and result in a loss of statistical power. Weighting using multiple variant functional annotations will help overcome this deficiency.

Calculation of annotation principal components using individual functional annotations.

To effectively capture the multi-faceted biological impact of a variant while reducing dimensionality, we propose variant annotation Principal Components (aPCs) as the PC summary of the functional annotation data by incorporating individual scores extracted from various functional databases[26,27,39–41,88]. We first group the individual scores into 10 major functional categories based on a priori knowledge, each capturing a specific aspect of variant biological function, including epigenetics, conservation, protein function, local nucleotide diversity, distance to coding, mutation density, transcription factors, mappability, distance to TSS/TES, and micro RNA (Fig. 2). For each category, we then center and standardize all individual scores within the category, such that higher value of each individual score indicates increased functionality of that annotation, and calculate aPC as the first PC from the standardized individual scores (Supplementary Table 1). To facilitate better

interpretation, these aPCs are then transformed into the PHRED-scaled scores for each variant across the genome, defined as $-10 \times \log_{10}(rank(-score)/M)$, where M is total number of variants sequenced across the whole genome.

Unlike ancestral PCs that are subject-specific and are calculated using genotypes across the genome to control for population structure, annotation PCs are variant-specific and are calculated using functional annotations for individual variants and are used to summarize multi-facet functions of individual variants. Complementary to other existing single-dimension integrative functional scores, annotation PCs summarize multiple aspects of variant function, with different blocks captured by different annotation PCs in the heatmap (Fig. 2).

STAAR incorporating multiple functional annotations.

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

$$\beta_j = c_j\gamma_j,$$

where c_j is the latent binary indicator of whether variant j is causal, and γ_j is the effect size of variant j if it is causal. The burden test, SKAT, and ACAT-V make direct assumptions on the variance of β_j using MAF information. This newly proposed variant effect model is expected to increase association power since a variant's causal status can be prioritized using its functional annotations[13,14]. Let $\pi_j = E(c_j)$ denote the probability of variant j being causal, then the effect of variant j given above is equivalent to

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

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

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

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

where A_{jk} is the kth annotation for the jth variant. For $k = 0$, we set $A_{j0} = 1$ as the intercept, which gives $\hat{\pi}_{j0} = 1$. For a quantitative annotation, A_{jk} represents its numeric value, e.g., the

kth annotation PC. The quantitative A_{jk} we consider in this paper include 10 aPCs (Supplementary Table 1) and existing integrative scores, including CADD[27], LINSIGHT[46], and FATHMM-XF[47]. For a qualitative annotation, we define $A_{jk} = 1$ for variants in the functional group (yes) and $A_{jk} = 0$ for variants otherwise (no). For example, A_{jk} denotes whether a variant is a disruptive missense variant using MetaSVM[51]. Hence, $\hat{\pi}_{jk} = 1$ for variants in the functional group and $\hat{\pi}_{jk} = 0$ otherwise, e.g., disruptive missense variants (yes/no). This corresponds to the RV tests using variants of this functional group.

In the STAAR framework, we model the effect sizes of causal variants γ_j in the same way as that used in conventional variant-set tests. Specifically, we assume $|\gamma_j| \propto w_j$, where w_j is assumed as a function of MAFs. For simplicity, we model w_j using $Beta(\mathrm{MAF}_j; a_1, a_2)$ and set (a_1, a_2) to be (1, 1) or (1,25). Then, the burden test statistic using kth variant functional annotation as the weight, e.g., aPC-Epigenetic, is given by $Q_{Burden,k} = \left(\sum_{j=1}^{p} \hat{\pi}_{jk} w_j S_j\right)^2$, whose P-value is denoted by $p_{Burden,k}(k = 0, \cdots, K)$. Under the assumption of SKAT, by estimating the probability of jth variant being causal using the kth annotation $(k = 0, \cdots, K)$, we have $E(\beta_j) = 0$ and $Var(\beta_j) = Var(c_j\gamma_j) = \pi_{jk} w_j^2 \tau_k$. Hence, the SKAT test statistic using kth variant functional annotation as the weight is given by

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

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

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

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

We denote by $p_{Burden,k}, p_{SKAT,k}, p_{ACAT-V,k}$ the P-values of burden, SKAT, and ACAT-V tests, respectively calculated using the kth annotation as the weight. For each type of RV tests, to robustly aggregate information from multiple annotations to boost power RV association tests in a data-adaptive manner, we propose to use the STAAR framework to combine individual annotation weighted tests using the ACAT P-value combination method[9,42]. Specifically, we define STAAR-Burden (STAAR-B), STAAR-SKAT (STAAR-S), and STAAR-ACAT-V (STAAR-A) as

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

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

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

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

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

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

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

To further aggregate information from different types tests and different weights, we propose an omnibus test in the STAAR framework (STAAR-O) by combining STAAR-B, STAAR-S and STAAR-A using the ACAT method[9,42]. We define the STAAR-O test statistic as

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

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

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

By combining different types of tests into an omnibus test, STAAR-O has a robust power with respect to the sparsity of causal variants and the directionality of effects of causal

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

Data simulation.

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

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

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

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

where X_{1i} and X_{2i} were defined the same as continuous traits and α_0 was determined to set the prevalence to 1%. In this setting, we used a balanced case-control design. We generated genotypes by simulating 20,000 sequences for 100 different regions each spanning 1 Mb. The data were generated to mimic the LD structure of an African American population by using the calibration coalescent model (COSI)[43]. In each simulation replicate, 10 annotations were generated as $A_1, \ldots, A_{10}$ i.i.d. $N(0,1)$ for each variant, and we randomly selected 5-kb regions from these 1-Mb regions for type I error simulations. We applied STAAR-B, STAAR-S, STAAR-A, and STAAR-O by incorporating MAFs and the 10 annotations and repeated the procedure with 10^9 replicates to examine the type I error rate at $\alpha = 10^{-5}, 10^{-6}, 10^{-7}$ levels. Total sample sizes considered were 2,500, 5,000, and 10,000.

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

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

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

We generated continuous traits from a linear model given by

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

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

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

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

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

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

Statistical analysis of lipid traits in the TOPMed data.

The TOPMed WGS data consist of ancestrally diverse and multi-ethnic related samples[45]. Race/ethnicity was defined using a combination of self-reported race/ethnicity and study recruitment information. The discovery cohorts consist of 4,580 (37.2%) Black or African

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

We applied STAAR-O to identify RV-sets associated with four quantitative lipid traits (LDL, HDL, TG and TC) using the TOPMed WGS data. LDL-C and TC were adjusted for the presence of medications as before[44]. Linear regression model adjusting for age, age^2, sex was first fit for each study-race/ethnicity-specific group. In addition, for Old Order Amish (OOA), we also adjusted for *APOB* p.R3527Q in LDL-C and TC analyses and adjusted for *APOC3* p.R19Ter in TG and HDL-C analyses[44]. The residuals were rank-based inverse normal transformed and rescaled by the standard deviation of the original phenotype within each group. We then fit a heteroscedastic linear mixed model (HLMM) for the rank normalized residuals, adjusting for 10 ancestral PCs, study-ethnicity group indicators, and a variance component for empirically derived kinship matrix plus separate group-specific residual variance components to account for population structure and relatedness. The output of HLMM was then used to perform following variant set analyses for rare variants (MAF < 1%) by scanning the genome, including gene-centric analysis using five variant categories (pLoF RVs, missense RVs, synonymous RVs, promoter RVs, and enhancer RVs) for each protein coded gene, and agnostic genetic region analysis using 2-kb sliding windows across the genome with a 1-kb skip length. The WGS RVAS analysis was performed using the R package STAAR (version 0.9.5).

The aPCs provide diverse and complementary information on variant functionality, and are incorporated in rare variant association tests using an omnibus weighting scheme via the proposed STAAR method. We demonstrate using the following example that STAAR boosts the rare variant association test power by properly upweighting known LDL-associated functional rare variants. For example, the association between a 2-kb sliding window located at 55,038,498 bp - 55,040,497 bp on chromosome 1 and LDL-C using STAAR-O is more significant than conventional tests in unconditional analysis (Supplementary Table 14). This power gain of STAAR-O is due to upweighting functional variants, e.g., the known tolerated missense variant rs11591147 within the sliding window through incorporating multiple aPCs[59]. Specifically, the aPC-Epigenetic, aPC-Protein, and aPC-Mappability PHRED scores are greater than 20 (top 1% across the genome), and the aPC-MutationDensity, aPC-TF, and CADD PHRED scores are greater than 10 (top 10% across the genome) for this variant, highlighting the multi-dimensional functionality of this variant. The aPC-Protein and aPC-Mappability weighted SKAT *P*-values are 6.69×10^{-13} and 3.78×10^{-12}, which are more significant than SKAT ($P = 1.12 \times 10^{-9}$) and burden test ($P = 4.68 \times 10^{-4}$).

Statistical analysis of LDL-C in the UK Biobank data.

We used UK Biobank whole exome sequences (WES) from the functionally equivalent (FE) pipeline. Sample and variant quality control measures were previously described[72,89]. In brief, samples with mismatch between genetically inferred and reported sex, high rates of heterozygosity or contamination (D-stat > 0.4), low sequence coverage (less than 85% of

targeted bases achieving 20X coverage), duplicates, and WES variants discordant with genotyping chip were removed. A total of 43,243 individuals with genetically inferred European ancestry were included; 40,519 of those had data on LDL cholesterol. Total cholesterol was adjusted by dividing the value by 0.8 among individuals reporting lipid lowering medication use after 1994 or statin use at any time point. LDL cholesterol was calculated from adjusted total cholesterol levels by the Friedewald equation for individuals with triglyceride levels < 400 mg/dl. If LDL cholesterol levels were directly measured, then their values were divided by 0.7 among reporting lipid lowering medication use after 1994 or statin use at any time point. Residuals were created after adjustment for age, age^2, sex, and the first 10 ancestral principal components. Residuals were then rank-based inverse-normal transformed and multiplied by the standard deviation. Analyses were restricted to missense variants in the *NPC1L1* gene predicted to be damaging according to the MetaSVM prediction algorithm and conditioned on ten known common variants in *NPC1L1* associated with LDL-C (rs10234070, rs73107473, rs2072183, rs41279633, rs17725246, rs2073547, rs10260606, rs217386, rs7791240, rs2300414) obtained from the UK Biobank imputed genotype data. We performed a burden test for the association between disruptive missense RVs in *NPC1L1* and LDL-C.

Reporting summary.

Further information on research design is available in the Nature Research Reporting Summary linked to this technical report.

Genome build.

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

Code availability.

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

Data availability.

This paper used the TOPMed Freeze 5 Whole Genome Sequencing data and lipids phenotype data. The genotype and phenotype data are both available in dbGAP. The discovery phase used the data from the following four study cohorts, where the accession numbers are provided in parenthesis: Framingham Heart Study (phs000974.v1.p1), Old Order Amish (phs000956.v1.p1), Jackson Heart Study (phs000964.v1.p1), and Multi-Ethnic Study of Atherosclerosis (phs001416.v1.p1). The replication phase used the data from the following ten study cohorts: Atherosclerosis Risk in Communities Study (phs001211), Cleveland Family Study (phs000954), Cardiovascular Health Study (phs001368), Diabetes Heart Study (phs001412), Genetic Study of Atherosclerosis Risk (phs001218), Genetic Epidemiology Network of Arteriopathy (phs001345), Genetics of Lipid Lowering Drugs and Diet Network (phs001359), San Antonio Family Heart Study (phs001215), Genome-wide Association Study of Adiposity in Samoans (phs000972) and Women's Health Initiative (phs001237). The sample sizes, ethnicity and phenotype summary statistics of these cohorts are given in Supplementary Table 3.

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[90]. The tissue-specific functional annotations were downloaded from ENCODE (https://www.encodeproject.org/report/?type=Experiment).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Authors

Xihao Li[1,58], Zilin Li[1,58], Hufeng Zhou[1], Sheila M. Gaynor[1], Yaowu Liu[2], Han Chen[3,4], Ryan Sun[5], Rounak Dey[1], Donna K. Arnett[6], Stella Aslibekyan[7], Christie M. Ballantyne[8], Lawrence F. Bielak[9], John Blangero[10], Eric Boerwinkle[3,11], Donald W. Bowden[12], Jai G. Broome[13], Matthew P. Conomos[14], Adolfo Correa[15], L. Adrienne Cupples[16,17], Joanne E. Curran[10], Barry I. Freedman[18], Xiuqing Guo[19], George Hindy[20], Marguerite R. Irvin[7], Sharon L. R. Kardia[9], Sekar Kathiresan[21,22,23], Alyna T. Khan[14], Charles L. Kooperberg[24], Cathy C. Laurie[14], X. Shirley Liu[25,26], Michael C. Mahaney[10], Ani W. Manichaikul[27], Lisa W. Martin[28], Rasika A. Mathias[29], Stephen T. McGarvey[30], Braxton D. Mitchell[31,32], May E. Montasser[33], Jill E. Moore[34], Alanna C. Morrison[3], Jeffrey R. O'Connell[31], Nicholette D. Palmer[12], Akhil Pampana[35,36], Juan M. Peralta[10], Patricia A. Peyser[9], Bruce M. Psaty[37,38], Susan Redline[39,40,41], Kenneth M. Rice[14], Stephen S. Rich[27], Jennifer A. Smith[9,42], Hemant K. Tiwari[43], Michael Y. Tsai[44], Ramachandran S. Vasan[17,45], Fei Fei Wang[14], Daniel E. Weeks[46], Zhiping Weng[34], James G. Wilson[47,48], Lisa R. Yanek[29], NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium, TOPMed Lipids Working Group, Benjamin M. Neale[35,49,50], Shamil R. Sunyaev[35,51,52], Gonçalo R. Abecasis[53,54], Jerome I. Rotter[19], Cristen J. Willer[55,56,57], Gina M. Peloso[16], Pradeep Natarajan[23,35,36], Xihong Lin[1,26,35,*]

Affiliations

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

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

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

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

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

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

[7] Department of Epidemiology, University of Alabama at Birmingham, Birmingham, AL, USA.

[8] Department of Medicine, Baylor College of Medicine, Houston, TX, USA.

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

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

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

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

[13] Division of Medical Genetics, University of Washington, Seattle, WA, USA.

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

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

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

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

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

[19] 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.

[20] Department of Population Medicine, Qatar University College of Medicine, QU Health, Doha, Qatar.

[21] Verve Therapeutics, Cambridge, MA, USA.

[22] Cardiology Division, Massachusetts General Hospital, Boston, MA, USA.

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

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

[25] Department of Data Sciences, Dana-Farber Cancer Institute and Harvard T.H. Chan School of Public Health, Boston, MA, USA.

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

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

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

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

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

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

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

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

[34] Program in Bioinformatics and Integrative Biology, University of Massachusetts Medical School, Worcester, MA, USA.

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

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

[37] Cardiovascular Health Research Unit, Departments of Medicine, Epidemiology, and Health Services, University of Washington, Seattle, WA, USA.

[38] Kaiser Permanente Washington Health Research Institute, Seattle, WA, USA.

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

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

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

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

[43] Department of Biostatistics, School of Public Health, University of Alabama at Birmingham, Birmingham, AL, USA.

[44] Department of Laboratory Medicine and Pathology, University of Minnesota, Minneapolis, MN, USA.

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

[46] Departments of Human Genetics and Biostatistics, University of Pittsburgh, Pittsburgh, PA, USA.

[47] Department of Physiology and Biophysics, University of Mississippi Medical Center, Jackson, MS, USA.

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

[49] Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA.

[50] Stanley Center for Psychiatric Research, Broad Institute of Harvard and MIT, Cambridge, MA, USA.

[51] Division of Genetics, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA.

[52] Department of Biomedical Informatics, Harvard Medical School, Boston, MA, USA.

[53] Regeneron Pharmaceuticals, Tarrytown, NY, USA.

[54] Department of Biostatistics, University of Michigan, Ann Arbor, MI, USA.

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

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

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

[58] These authors contributed equally to this work.

Acknowledgments

This work was supported by grants R35-CA197449, P01-CA134294, U19-CA203654, and R01-HL113338 (X. Lin), U01-HG009088 (X. Lin, S.R.S., and B.M.N.), R01-HL142711 (P.N. and G.M.P), K01-HL125751 and R03-HL141439 (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, UL1TR001881, and DK063491 (J.I.R. and X.G.), HHSN268201800002I (G.R.A.), R35-GM127131 and R01-MH101244 (S.R.S.), U01-HL72518, HL087698, HL49762, HL59684, HL58625, HL071025, HL112064, NR0224103, and M01-RR000052 (to the Johns Hopkins General Clinical Research Center), R01-HL093093, R01-HL133040 (D.E.W.), 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 (K.M.R., A.T.K., M.P.C., and J.G.B.), U01-HL137162 (K.M.R. and M.P.C.), R35-HL135818 and R01-HL113338 (S.R.), R01-HL113323, U01-DK085524, R01-HL045522, R01-MH078143, R01-MH078111, and R01-MH083824 (J.M.P., M.C.M., J.E.C., and J.B.), 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-HL093093 (S.T.M.), 1U24CA237617 and 5U24HG009446 (X.S.L.), HHSN268201600018C, HHSN268201600001C, HHSN268201600002C, HHSN268201600003C, and HHSN268201600004C (C.L.K.), U01-HL072524, R01-HL104135-04S1, U01-HL054472, U01-HL054473, U01-HL054495, U01-HL054509, and R01-HL055673-18S1 (M.R.I., S.A., and D.K.A.), Swedish Research Council 201606830 (G.H.), HHSN2682018000010I, HHSN2682018000011I, HHSN2682018000012I, HHSN2682018000013I, HHSN2682018000014I, and HHSN2682018000015I (A.C.), HHSN2682017000001I, HHSN2682017000002I, HHSN2682017000003I, HHSN2682017000005I, and HHSN2682017000004I (E.B.), and R01-HL134320 (C.M.B.). Whole genome sequencing (WGS) for the Trans-Omics in Precision Medicine (TOPMed) program was supported by the National Heart, Lung and Blood Institute (NHLBI). Centralized read mapping and genotype calling, along with variant quality metrics and filtering were provided by the TOPMed Informatics Research Center (3R01HL-117626-02S1; contract HHSN2682018000002I). Phenotype harmonization, data management, sample-identity QC, and general study coordination, were provided by the TOPMed Data Coordinating Center (3R01HL-120393-02S1; contract HHSN2682018000001I). We gratefully acknowledge the studies and participants who provided biological samples and data for TOPMed. The full study specific acknowledgements are detailed in Supplementary Note.

Competing interests

S.A. reports equity and employment by 23andMe, Inc. L.A.C. spends part of her time consulting for Dyslipidemia Foundation, which is a non-profit company, as a statistical consultant. X.S.L. is co-founder, board member, and Scientific Advisory Board of GV20 Oncotherapy; Scientific Advisory Board of 3DMedCare; consultant of Genentech; research grants from Sanofi and Takeda; all unrelated to the present work. 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. For S.R.: Jazz Pharma, Eisai Pharm, Respircardia, unrelated to the present work. Z.W. cofounded Rgenta Therapeutics and directs its Scientific Advisory Board. B.M.N. is on the Scientific Advisory Board of Deep Genomics, a consultant for Camp4 Therapeutics, Takeda Pharmaceutical and Biogen. S.R.S. is consultant to NGM Biopharmaceuticals and Inari agriculture. He is also on Scientific Advisory Board of Veritas Genetics. G.R.A. is an employee of Regeneron Pharmaceuticals and owns stock and stock options for Regeneron Pharmaceuticals. The spouse of C.J.W. works at Regeneron Pharmaceuticals. P.N. reports grants from Amgen, Apple, and Boston Scientific, and consulting income from Apple and Blackstone Life Sciences, all unrelated to the present work. X. Lin is a consultant of AbbVie Pharmaceuticals.

Appendix

NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium

Namiko Abe[59], Gonçalo R. Abecasis[53,54], Francois Aguet[60], Christine Albert[61], Laura Almasy[62], Alvaro Alonso[63], Seth Ament[64], Peter Anderson[65], Pramod Anugu[66], Deborah Applebaum-Bowden[67], Kristin Ardlie[60], Dan Arking[68], Donna K. Arnett[6], Allison Ashley-Koch[69], Stella Aslibekyan[7], Tim Assimes[70], Paul Auer[71], Dimitrios Avramopoulos[68], John Barnard[72], Kathleen Barnes[73], R. Graham Barr[74], Emily Barron-Casella[68], Lucas Barwick[75], Terri Beaty[68], Gerald Beck[76], Diane Becker[77], Lewis Becker[68], Rebecca Beer[78], Amber Beitelshees[64], Emelia Benjamin[79], Takis Benos[80], Marcos Bezerra[81], Lawrence F. Bielak[9], Joshua Bis[82], Thomas Blackwell[83], John Blangero[10], Eric Boerwinkle[3,11], Donald W. Bowden[12], Russell Bowler[84], Jennifer Brody[65], Ulrich Broeckel[85], Jai G. Broome[13], Karen Bunting[59], Esteban Burchard[86], Carlos Bustamante[87], Erin Buth[88], Brian Cade[89], Jonathan Cardwell[90], Vincent Carey[91], Cara Carty[92], Richard Casaburi[93], James Casella[68], Peter Castaldi[94], Mark Chaffin[95], Christy Chang[64], Yi-Cheng Chang[96], Daniel Chasman[97], Sameer Chavan[90], Bo-Juen Chen[59], Wei-Min Chen[98], Yii-Der Ida Chen[99], Michael Cho[100], Seung Hoan Choi[95], Lee-Ming Chuang[101], Mina Chung[102], Ren-Hua Chung[103], Clary Clish[104], Suzy Comhair[105], Matthew P. Conomos[14], Elaine Cornell[106], Adolfo Correa[15], Carolyn Crandall[93], James Crapo[107], L. Adrienne Cupples[16,17], Joanne E. Curran[10], Jeffrey Curtis[83], Brian Custer[108], Coleen Damcott[64], Dawood Darbar[109], Sayantan Das[83], Sean David[110], Colleen Davis[65], Michelle Daya[90], Mariza de Andrade[111], Lisa de las Fuentes[112], Michael DeBaun[113], Ranjan Deka[114], Dawn DeMeo[100], Scott Devine[64], Qing Duan[115], Ravi

Duggirala[116], Jon Peter Durda[106], Susan Dutcher[117], Charles Eaton[118], Lynette Ekunwe[66], Adel El Boueiz[119], Patrick Ellinor[120], Leslie Emery[65], Serpil Erzurum[121], Charles Farber[98], Tasha Fingerlin[122], Matthew Flickinger[83], Myriam Fornage[123], Nora Franceschini[124], Chris Frazar[125], Mao Fu[64], Stephanie M. Fullerton[65], Lucinda Fulton[117], Stacey Gabriel[95], Weiniu Gan[78], Shanshan Gao[126], Yan Gao[66], Margery Gass[127], Bruce Gelb[128], Xiaoqi (Priscilla) Geng[83], Mark Geraci[129], Soren Germer[59], Robert Gerszten[130], Auyon Ghosh[131], Richard Gibbs[132], Chris Gignoux[70], Mark Gladwin[133], David Glahn[134], Stephanie Gogarten[65], Da-Wei Gong[64], Harald Goring[135], Sharon Graw[136], Daniel Grine[126], C. Charles Gu[117], Yue Guan[64], Xiuqing Guo[19], Namrata Gupta[60], Jeff Haessler[127], Michael Hall[66], Daniel Harris[64], Nicola L. Hawley[137], Jiang He[138], Ben Heavner[88], Susan Heckbert[65], Ryan Hernandez[139], David Herrington[140], Craig Hersh[141], Bertha Hidalgo[142], James Hixson[123], Brian Hobbs[91], John Hokanson[90], Elliott Hong[64], Karin Hoth[143], Chao (Agnes) Hsiung[144], Yi-Jen Hung[145], Haley Huston[146], Chii Min Hwu[147], Marguerite R. Irvin[7], Rebecca Jackson[148], Deepti Jain[65], Cashell Jaquish[78], Min A Jhun[83], Jill Johnsen[149], Andrew Johnson[78], Craig Johnson[65], Rich Johnston[63], Kimberly Jones[68], Hyun Min Kang[150], Robert Kaplan[151], Sharon L.R. Kardia[9], Sekar Kathiresan[21,22,23], Shannon Kelly[108], Eimear Kenny[128], Michael Kessler[64], Alyna T. Khan[14], Wonji Kim[152], Greg Kinney[90], Barbara Konkle[153], Charles L. Kooperberg[24], Holly Kramer[154], Christoph Lange[155], Ethan Lange[90], Leslie Lange[90], Cathy C. Laurie[14], Cecelia Laurie[65], Meryl LeBoff[100], Jiwon Lee[91], Seunggeun Shawn Lee[83], Wen-Jane Lee[147], Jonathon LeFaive[83], David Levine[65], Dan Levy[78], Joshua Lewis[64], Xiaohui Li[156], Yun Li[115], Henry Lin[156], Honghuang Lin[157], Keng Han Lin[83], Xihong Lin[1,26,35], Simin Liu[158], Yongmei Liu[159], Yu Liu[160], Ruth J.F. Loos[161], Steven Lubitz[120], Kathryn Lunetta[157], James Luo[78], Michael C. Mahaney[10], Barry Make[68], Ani W. Manichaikul[27], JoAnn Manson[100], Lauren Margolin[95], Lisa W. Martin[28], Susan Mathai[90], Rasika A. Mathias[29], Susanne May[162], Patrick McArdle[64], Merry-Lynn McDonald[142], Sean McFarland[163], Stephen T. McGarvey[30], Daniel McGoldrick[125], Caitlin McHugh[164], Hao Mei[66], Luisa Mestroni[136], Deborah A Meyers[165], Julie Mikulla[78], Nancy Min[66], Mollie Minear[78], Ryan L Minster[133], Braxton D. Mitchell[31,32], Matt Moll[166], May E. Montasser[33], Courtney Montgomery[167], Arden Moscati[168], Solomon Musani[169], Stanford Mwasongwe[66], Josyf C Mychaleckyj[98], Girish Nadkarni[128], Rakhi Naik[68], Take Naseri[170], Pradeep Natarajan[23,35,36], Sergei Nekhai[171], Sarah C. Nelson[88], Bonnie Neltner[126], Deborah Nickerson[65], Kari North[115], Jeffrey R. O'Connell[31], Tim O'Connor[64], Heather Ochs-Balcom[172], David Paik[173], Nicholette D. Palmer[12], James Pankow[174], George Papanicolaou[78], Afshin Parsa[64], Juan M. Peralta[10], Marco Perez[70], James Perry[64], Ulrike Peters[175], Patricia A. Peyser[9], Lawrence S Phillips[63], Toni Pollin[64], Wendy Post[176], Julia Powers Becker[177], Meher Preethi Boorgula[90], Michael Preuss[128], Bruce M. Psaty[37,38], Pankaj Qasba[78], Dandi Qiao[100], Zhaohui Qin[63], Nicholas Rafaels[178], Laura Raffield[179], Ramachandran S. Vasan[17,45], D.C. Rao[117], Laura Rasmussen-Torvik[180], Aakrosh Ratan[98], Susan Redline[39,40,41], Robert Reed[64], Elizabeth Regan[107], Alex Reiner[175], Muagututi'a Sefuiva Reupena[181], Kenneth M. Rice[14], Stephen S. Rich[27], Dan Roden[182], Carolina Roselli[95], Jerome I. Rotter[19], Ingo Ruczinski[68], Pamela Russell[90], Sarah Ruuska[183], Kathleen Ryan[64], Ester Cerdeira Sabino[184], Danish Saleheen[185], Shabnam Salimi[64], Steven Salzberg[68], Kevin Sandow[186], Vijay G. Sankaran[187], Christopher Scheller[83], Ellen Schmidt[83], Karen Schwander[117], David Schwartz[90], Frank Sciurba[133], Christine Seidman[188], Jonathan Seidman[189], Vivien Sheehan[190], Stephanie L. Sherman[191], Amol

Shetty[64], Aniket Shetty[90], Wayne Hui-Heng Sheu[147], M. Benjamin Shoemaker[192], Brian Silver[193], Edwin Silverman[100], Jennifer A. Smith[9,42], Josh Smith[65], Nicholas Smith[194], Tanja Smith[59], Sylvia Smoller[151], Beverly Snively[195], Michael Snyder[196], Tamar Sofer[100], Nona Sotoodehnia[65], Adrienne M. Stilp[65], Garrett Storm[126], Elizabeth Streeten[64], Jessica Lasky Su[91], Yun Ju Sung[117], Jody Sylvia[100], Adam Szpiro[65], Carole Sztalryd[64], Daniel Taliun[83], Hua Tang[197], Margaret Taub[68], Kent D. Taylor[198], Matthew Taylor[199], Simeon Taylor[64], Marilyn Telen[69], Timothy A. Thornton[65], Machiko Threlkeld[200], Lesley Tinker[127], David Tirschwell[65], Sarah Tishkoff[201], Hemant K. Tiwari[43], Catherine Tong[202], Russell Tracy[203], Michael Y. Tsai[44], Dhananjay Vaidya[68], David Van Den Berg[204], Peter VandeHaar[83], Scott Vrieze[205], Tarik Walker[90], Robert Wallace[143], Avram Walts[90], Fei Fei Wang[14], Heming Wang[206], Karol Watson[93], Daniel E. Weeks[46], Bruce Weir[65], Scott Weiss[100], Lu-Chen Weng[120], Jennifer Wessel[207], Cristen J. Willer[55,56,57], Kayleen Williams[65], L. Keoki Williams[208], Carla Wilson[209], James G. Wilson[47,48], Quenna Wong[65], Joseph Wu[210], Huichun Xu[64], Lisa R. Yanek[29], Ivana Yang[90], Rongze Yang[64], Norann Zaghloul[64], Maryam Zekavat[95], Yingze Zhang[211], Snow Xueyan Zhao[107], Wei Zhao[212], Degui Zhi[123], Xiang Zhou[83], Xiaofeng Zhu[213], Michael Zody[59], Sebastian Zoellner[83]

[59]New York Genome Center, New York, NY, USA. [60]Broad Institute, Cambridge, MA, USA. [61]Brigham and Women's Hospital, Cedars Sinai, Boston, MA, USA. [62]Children's Hospital of Philadelphia, University of Pennsylvania, Philadelphia, PA, USA. [63]Emory University, Atlanta, GA, USA. [64]University of Maryland, Baltimore, MD, USA. [65]University of Washington, Seattle, WA, USA. [66]University of Mississippi, Jackson, MS, USA. [67]National Institutes of Health, Bethesda, MD, USA. [68]Johns Hopkins University, Baltimore, MD, USA. [69]Duke University, Durham, NC, USA. [70]Stanford University, Stanford, CA, USA. [71]University of Wisconsin Milwaukee, Milwaukee, WI, USA. [72]Cleveland Clinic, Cleveland, OH, USA. [73]University of Colorado Anschutz Medical Campus, Aurora, CO, USA. [74]Columbia University, New York, NY, USA. [75]The Emmes Corporation, LTRC, Rockville, MD, USA. [76]Cleveland Clinic, Quantitative Health Sciences, Cleveland, OH, USA. [77]Johns Hopkins University, Medicine, Baltimore, MD, USA. [78]National Heart, Lung, and Blood Institute, National Institutes of Health, Bethesda, MD, USA. [79]Boston University, Massachusetts General Hospital, Boston University School of Medicine, Boston, MA, USA. [80]University of Pittsburgh, Pittsburgh, PA, USA. [81]Fundação de Hematologia e Hemoterapia de Pernambuco - Hemope, Recife, Brazil. [82]University of Washington, Cardiovascular Health Research Unit, Department of Medicine, Seattle, WA, USA. [83]University of Michigan, Ann Arbor, MI, USA. [84]National Jewish Health, National Jewish Health, Denver, CO, USA. [85]Medical College of Wisconsin, Milwaukee, WI, USA. [86]University of California, San Francisco, San Francisco, CA, USA. [87]Stanford University, Stanford, CA, USA. [88]University of Washington, Biostatistics, Seattle, WA, USA. [89]Brigham and Women's Hospital, Boston, MA, USA. [90]University of Colorado at Denver, Denver, CO, USA. [91]Brigham and Women's Hospital, Boston, MA, USA. [92]Washington State University, Seattle, WA, USA. [93]University of California, Los Angeles, Los Angeles, CA, USA. [94]Brigham and Women's Hospital, Medicine, Boston, MA, USA. [95]Broad Institute, Cambridge, MA, USA. [96]National Taiwan University, Taipei, Taiwan. [97]Brigham and Women's Hospital, Division of Preventive Medicine, Boston, MA, USA. [98]University of Virginia, Charlottesville, VA, USA. [99]Lundquist Institute, Charlottesville, VA, USA.

[100]Brigham and Women's Hospital, Boston, MA, USA. [101]National Taiwan University, National Taiwan University Hospital, Taipei, Taiwan. [102]Cleveland Clinic, Cleveland, OH, USA. [103]National Health Research Institute, Taiwan. [104]Broad Institute, Metabolomics Platform, Cambridge, MA, USA. [105]Cleveland Clinic, Immunity and Immunology, Cleveland, OH, USA. [106]University of Vermont, Burlington, VT, USA. [107]National Jewish Health, Denver, CO, USA. [108]Vitalant Research Institute, San Francisco, CA, USA. [109]University of Illinois at Chicago, Chicago, IL, USA. [110]University of Chicago, Chicago, IL, USA. [111]Mayo Clinic, Health Sciences Research, Rochester, MN, USA. [112]Washington University in St Louis, Department of Medicine, Cardiovascular Division, St. Louis, MO, USA. [113]Vanderbilt University, Nashville, TN, USA. [114]University of Cincinnati, Cincinnati, OH, USA. [115]University of North Carolina, Chapel Hill, NC, USA. [116]University of Texas Rio Grande Valley School of Medicine, Edinburg, TX, USA. [117]Washington University in St Louis, St Louis, MO, USA. [118]Brown University, Providence, RI, USA. [119]Harvard University, Channing Division of Network Medicine, Boston, MA, USA. [120]Massachusetts General Hospital, Boston, MA, USA. [121]Cleveland Clinic, Cleveland, OH, USA. [122]National Jewish Health, Center for Genes, Environment and Health, Denver, CO, USA. [123]University of Texas Health at Houston, Houston, TX, USA. [124]University of North Carolina, Epidemiology, Chapel Hill, NC, USA. [125]University of Washington, Seattle, WA, USA. [126]University of Colorado at Denver, Denver, CO, USA. [127]Fred Hutchinson Cancer Research Center, Seattle, WA, USA. [128]Icahn School of Medicine at Mount Sinai, New York, NY, USA. [129]Indiana University, Medicine, Indianapolis, IN, USA. [130]Beth Israel Deaconess Medical Center, Boston, MA, USA. [131]Brigham and Women's Hospital, Boston, MA, USA. [132]Baylor College of Medicine Human Genome Sequencing Center, Houston, TX, USA. [133]University of Pittsburgh, Pittsburgh, PA, USA. [134]Yale University, New Haven, CT, USA. [135]University of Texas Rio Grande Valley School of Medicine, San Antonio, TX, USA. [136]University of Colorado Anschutz Medical Campus, Aurora, CO, USA. [137]Yale University, Department of Chronic Disease Epidemiology, New Haven, CT, USA. [138]Tulane University, New Orleans, LA, USA. [139]McGill University, University of California, San Francisco, San Francisco, CA, USA. [140]Wake Forest Baptist Health, Winston-Salem, NC, USA. [141]Brigham and Women's Hospital, Channing Division of Network Medicine, Boston, MA, USA. [142]University of Alabama, Birmingham, AL, USA. [143]University of Iowa, Iowa City, IA, USA. [144]National Health Research Institute Taiwan, Institute of Population Health Sciences, NHRI, Miaoli County, Taiwan. [145]Tri-Service General Hospital National Defense Medical Center, Taipei, Taiwan. [146]Blood Works Northwest, Seattle, WA, USA. [147]Taichung Veterans General Hospital Taiwan, Taichung City, Taiwan. [148]Ohio State University Wexner Medical Center, Internal Medicine, Division of Endocrinology, Diabetes and Metabolism, Columbus, OH, USA. [149]Blood Works Northwest, University of Washington, Seattle, WA, USA. [150]University of Michigan, Biostatistics, Ann Arbor, MI, USA. [151]Albert Einstein College of Medicine, New York, NY, USA. [152]Harvard University, Cambridge, USA. [153]Blood Works Northwest, Seattle, WA, USA. [154]Loyola University, Public Health Sciences, Maywood, IL, USA. [155]Harvard School of Public Health, Biostats, Boston, MA, USA. [156]Lundquist Institute, Torrance, CA, USA. [157]Boston University, Boston, MA, USA. [158]Brown University, Epidemiology and Medicine, Providence, RI, USA. [159]Duke University, Cardiology, Durham, NC, USA. [160]Stanford University, Cardiovascular Institute, Palo Alto, CA, USA. [161]Icahn School of Medicine at Mount Sinai, The Charles Bronfman

Institute for Personalized Medicine, New York, NY, USA. [162]University of Washington, Biostatistics, Seattle, WA, USA. [163]Harvard University, Cambridge, MA, USA. [164]University of Washington, Biostatistics, Seattle, WA, USA. [165]University of Arizona, Tucson, AZ, USA. [166]Brigham and Women's Hospital, Medicine, Boston, MA, USA. [167]Oklahoma Medical Research Foundation, Genes and Human Disease, Oklahoma City, OK, USA. [168]Icahn School of Medicine at Mount Sinai, New York, NY, USA. [169]University of Mississippi, Medicine, Jackson, MS, USA. [170]Ministry of Health, Government of Samoa, Apia, Samoa. [171]Howard University, Washington, DC, USA. [172]University at Buffalo, Buffalo, NY, USA. [173]Stanford University, Stanford Cardiovascular Institute, Stanford, CA, USA. [174]University of Minnesota, Minneapolis, MN, USA. [175]Fred Hutchinson Cancer Research Center, University of Washington, Seattle, WA, USA. [176]Johns Hopkins University, Cardiology/Medicine, Baltimore, MD, USA. [177]University of Colorado at Denver, Medicine, Denver, CO, USA. [178]University of Colorado at Denver, Denver, CO, USA. [179]University of North Carolina, Genetics, Chapel Hill, NC, USA. [180]Northwestern University, Chicago, IL, USA. [181]Lutia I Puava Ae Mapu I Fagalele, Apia, Samoa. [182]Vanderbilt University, Medicine, Pharmacology, Biomedicla Informatics, Nashville, TN, USA. [183]Blood Works Northwest, Seattle, WA, USA. [184]Universidade de Sao Paulo, Faculdade de Medicina, Sao Paulo, Brazil. [185]Columbia University, New York, NY, USA. [186]Lundquist Institute, TGPS, Torrance, CA, USA. [187]Broad Institute, Harvard University, Division of Hematology/Oncology, Boston, MA, USA. [188]Harvard Medical School, Genetics, Boston, MA, USA. [189]Harvard Medical School, Boston, MA, USA. [190]Baylor College of Medicine, Pediatrics, Houston, TX, USA. [191]Emory University, Human Genetics, Atlanta, GA, USA. [192]Vanderbilt University, Medicine/Cardiology, Nashville, TN, USA. [193]UMass Memorial Medical Center, Worcester, MA, USA. [194]University of Washington, Epidemiology, Seattle, WA, USA. [195]Wake Forest Baptist Health, Biostatistical Sciences, Winston-Salem, NC, USA. [196]Stanford University, Stanford, CA, USA. [197]Stanford University, Genetics, Stanford, CA, USA. [198]Lundquist Institute, Institute for Translational Genomics and Populations Sciences, Torrance, CA, USA. [199]University of Colorado at Denver, Denver, CO, USA. [200]University of Washington, Department of Genome Sciences, Seattle, WA, USA. [201]University of Pennsylvania, Genetics, Philadelphia, PA, USA. [202]University of Washington, Department of Biostatistics, Seattle, WA, USA. [203]University of Vermont, Pathology & Laboratory Medicine, Burlington, VT, USA. [204]University of Southern California, USC Methylation Characterization Center, University of Southern California, Los Angeles, CA, USA. [205]University of Colorado at Boulder, University of Minnesota, Boulder, CO, USA. [206]Brigham and Women's Hospital, Partners, Boston, MA, USA. [207]Indiana University, Epidemiology, Indianapolis, IN, USA. [208]Henry Ford Health System, Detroit, MI, USA. [209]Brigham and Women's Hospital, Boston, MA, USA. [210]Stanford University, Stanford Cardiovascular Institute, Stanford, CA, USA. [211]University of Pittsburgh, Medicine, Pittsburgh, PA, USA. [212]University of Michigan, Department of Epidemiology, Ann Arbor, MI, USA. [213]Case Western Reserve University, Department of Population and Quantitative Health Sciences, Cleveland, OH, USA.

TOPMed Lipids Working Group

Moustafa Abdalla[95], Gonçalo R. Abecasis[53,54], Donna K. Arnett[6], Stella Aslibekyan[7], Tim Assimes[70], Elizabeth Atkinson[50], Christie M. Ballantyne[8], Amber Beitelshees[64], Lawrence F. Bielak[9], Joshua Bis[82], Corneliu Bodea[100], Eric Boerwinkle[3,11], Donald W. Bowden[12], Jennifer Brody[65], Brian Cade[89], Jenna Carlson[80], I-Shou Chang[103], Yii-Der Ida Chen[99], Sung Chun[100], Ren-Hua Chung[103], Matthew P. Conomos[14], Adolfo Correa[15], L. Adrienne Cupples[16,17], Coleen Damcott[64], Paul de Vries[123], Ron Do[128], Amanda Elliott[95], Mao Fu[64], Andrea Ganna[95], Da-Wei Gong[64], Sarah Graham[55], Mary Haas[95], Bernhard Haring[214], Jiang He[138], Susan Heckbert[65], Blanca Himes[215], James Hixson[123], Marguerite R. Irvin[7], Deepti Jain[65], Gail Jarvik[65], Min A Jhun[83], Jicai Jiang[64], Goo Jun[123], Rita Kalyani[29], Sharon L.R. Kardia[9], Sekar Kathiresan[21,22,23], Amit Khera[95], Derek Klarin[95], Charles L. Kooperberg[24], Brian Kral[29], Leslie Lange[90], Cathy C. Laurie[14], Cecelia Laurie[65], Rozenn Lemaitre[82], Zilin Li[1], Xihao Li[1], Xihong Lin[1,26,35], Michael C. Mahaney[10], Ani W. Manichaikul[27], Lisa W. Martin[28], Rasika A. Mathias[29], Ravi Mathur[216], Stephen T. McGarvey[30], Caitlin McHugh[164], John McLenithan[64], Julie Mikulla[78], Braxton D. Mitchell[31,32], May E. Montasser[33], Andrew Moran[185], Alanna C. Morrison[3], Tetsushi Nakao[95], Pradeep Natarajan[23,35,36], Deborah Nickerson[65], Kari North[115], Jeffrey R. O'Connell[31], Christopher O'Donnell[217], Nicholette D. Palmer[12], Akhil Pampana[35,36], Aniruddh Patel[95], Gina M. Peloso[16], James Perry[64], Ulrike Peters[175], Patricia A. Peyser[9], James Pirruccello[95], Toni Pollin[64], Michael Preuss[128], Bruce M. Psaty[37,38], D. C. Rao[117], Susan Redline[39,40,41], Robert Reed[64], Alex Reiner[175], Stephen S. Rich[27], Samantha Rosenthal[80], Jerome I. Rotter[19], Jenny Schoenberg[127], Margaret Sunitha Selvaraj[35,36], Wayne Hui-Heng Sheu[147], Jennifer A. Smith[9,42], Tamar Sofer[100], Adrienne M. Stilp[65], Shamil R. Sunyaev[35,51,52], Ida Surakka[55], Carole Sztalryd[64], Hua Tang[197], Kent D. Taylor[198], Michael Y. Tsai[44], Md Mesbah Uddin[95], Sarah Urbut[95], Marie Verbanck[128], Ann Von Holle[115], Heming Wang[206], Fei Fei Wang[14], Kerri Wiggins[65], Cristen J. Willer[55,56,57], James G. Wilson[47,48], Brooke Wolford[56], Huichun Xu[64], Lisa R. Yanek[29], Norann Zaghloul[64], Maryam Zekavat[95], Jingwen Zhang[1]

[214]University of Würzburg, Department of Medicine I / Cardiology, Wuerzburg, Germany. [215]University of Pennsylvania, Department of Biostatistics, Epidemiology and Informatics, Philadelphia, PA, USA. [216]RTI International, Biostatistics and Epidemiology, Research Triangle Park, NC, USA. [217]VA Boston Healthcare System, Department of Medicine, Boston, MA, USA.

References

1. Bansal V, Libiger O, Torkamani A & Schork NJ Statistical analysis strategies for association studies involving rare variants. Nat. Rev. Genet 11, 773–785 (2010). [PubMed: 20940738]
2. Kiezun A et al. Exome sequencing and the genetic basis of complex traits. Nat. Genet 44, 623–630 (2012). [PubMed: 22641211]
3. Lee S, Abecasis GR, Boehnke M & Lin X Rare-variant association analysis: study designs and statistical tests. Am. J. Hum. Genet 95, 5–23 (2014). [PubMed: 24995866]
4. Morgenthaler S & Thilly WG A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis 615, 28–56 (2007). [PubMed: 17101154]

5. Li B & Leal SM Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. Am. J. Hum. Genet 83, 311–321 (2008). [PubMed: 18691683]
6. Madsen BE & Browning SR A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet. 5, e1000384 (2009). [PubMed: 19214210]
7. Morris AP & Zeggini E An evaluation of statistical approaches to rare variant analysis in genetic association studies. Genet. Epidemiol 34, 188–193 (2010). [PubMed: 19810025]
8. Wu MC et al. Rare-variant association testing for sequencing data with the sequence kernel association test. Am. J. Hum. Genet 89, 82–93 (2011). [PubMed: 21737059]
9. Liu Y et al. ACAT: a fast and powerful p value combination method for rare-variant analysis in sequencing studies. Am. J. Hum. Genet 104, 410–421 (2019). [PubMed: 30849328]
10. Lee S, Wu MC & Lin X Optimal tests for rare variant effects in sequencing association studies. Biostatistics 13, 762–775 (2012). [PubMed: 22699862]
11. Sun J, Zheng Y & Hsu L A unified mixed-effects model for rare-variant association in sequencing studies. Genet. Epidemiol 37, 334–344 (2013). [PubMed: 23483651]
12. Pan W, Kim J, Zhang Y, Shen X & Wei P A powerful and adaptive association test for rare variants. Genetics 197, 1081–1095 (2014). [PubMed: 24831820]
13. Kichaev G et al. Integrating functional data to prioritize causal variants in statistical fine-mapping studies. PLoS Genet. 10, e1004722 (2014). [PubMed: 25357204]
14. Kichaev G et al. Improved methods for multi-trait fine mapping of pleiotropic risk loci. Bioinformatics 33, 248–255 (2017). [PubMed: 27663501]
15. Finucane HK et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat. Genet 47, 1228–1235 (2015). [PubMed: 26414678]
16. Hu Y et al. Leveraging functional annotations in genetic risk prediction for human complex diseases. PLoS Comp. Biol 13, e1005589 (2017).
17. Morrison AC et al. Practical approaches for whole-genome sequence analysis of heart-and blood-related traits. Am. J. Hum. Genet 100, 205–215 (2017). [PubMed: 28089252]
18. Schaid DJ, Chen W & Larson NB From genome-wide associations to candidate causal variants by statistical fine-mapping. Nat. Rev. Genet 19, 491–504 (2018). [PubMed: 29844615]
19. Claussnitzer M et al. A brief history of human disease genetics. Nature 577, 179–189 (2020). [PubMed: 31915397]
20. Harrow J et al. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. 22, 1760–1774 (2012). [PubMed: 22955987]
21. Frankish A et al. GENCODE reference annotation for the human and mouse genomes. Nucleic Acids Res. 47, D766–D773 (2018).
22. Ng PC & Henikoff S SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res. 31, 3812–3814 (2003). [PubMed: 12824425]
23. Adzhubei IA et al. A method and server for predicting damaging missense mutations. Nat. Methods 7, 248–249 (2010). [PubMed: 20354512]
24. Siepel A et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res. 15, 1034–1050 (2005). [PubMed: 16024819]
25. Pollard KS, Hubisz MJ, Rosenbloom KR & Siepel A Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. 20, 110–121 (2010). [PubMed: 19858363]
26. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature 489, 57–74 (2012). [PubMed: 22955616]
27. Kircher M et al. A general framework for estimating the relative pathogenicity of human genetic variants. Nat. Genet 46, 310–315 (2014). [PubMed: 24487276]
28. Tang H & Thomas PD Tools for predicting the functional impact of nonsynonymous genetic variation. Genetics 203, 635–647 (2016). [PubMed: 27270698]
29. Lee PH et al. Principles and methods of in-silico prioritization of non-coding regulatory variants. Hum. Genet 137, 15–30 (2018). [PubMed: 29288389]
30. Kellis M et al. Defining functional DNA elements in the human genome. Proc. Natl. Acad. Sci. USA 111, 6131–6138 (2014). [PubMed: 24753594]

31. Zuk O et al. Searching for missing heritability: designing rare variant association studies. Proc. Natl. Acad. Sci. USA 111, E455–E464 (2014). [PubMed: 24443550]
32. Hao X, Zeng P, Zhang S & Zhou X Identifying and exploiting trait-relevant tissues with multiple functional annotations in genome-wide association studies. PLoS Genet. 14, e1007186 (2018). [PubMed: 29377896]
33. He Z, Xu B, Lee S & Ionita-Laza I Unified sequence-based association tests allowing for multiple functional annotations and meta-analysis of noncoding variation in Metabochip data. Am. J. Hum. Genet 101, 340–352 (2017). [PubMed: 28844485]
34. Ma Y & Wei P FunSPU: a versatile and adaptive multiple functional annotation-based association test of whole-genome sequencing data. PLoS Genet. 15, e1008081 (2019). [PubMed: 31034468]
35. Breslow NE & Clayton DG Approximate inference in generalized linear mixed models. J. Am. Stat. Assoc 88, 9–25 (1993).
36. Chen H et al. Control for population structure and relatedness for binary traits in genetic association studies via logistic mixed models. Am. J. Hum. Genet 98, 653–666 (2016). [PubMed: 27018471]
37. Chen H et al. Efficient variant set mixed model association tests for continuous and binary traits in large-scale whole-genome sequencing studies. Am. J. Hum. Genet 104, 260–274 (2019). [PubMed: 30639324]
38. Gogarten SM et al. Genetic association testing using the GENESIS R/Bioconductor package. Bioinformatics 35, 5346–5348 (2019). [PubMed: 31329242]
39. Kundaje A et al. Integrative analysis of 111 reference human epigenomes. Nature 518, 317–330 (2015). [PubMed: 25693563]
40. Rentzsch P, Witten D, Cooper GM, Shendure J & Kircher M CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Res. 47, D886–D894 (2018).
41. Liu X, Wu C, Li C & Boerwinkle E dbNSFP v3. 0: A one-stop database of functional predictions and annotations for human nonsynonymous and splice-site SNVs. Hum. Mutat 37, 235–241 (2016). [PubMed: 26555599]
42. Liu Y & Xie J Cauchy combination test: a powerful test with analytic p-value calculation under arbitrary dependency structures. J. Am. Stat. Assoc 115, 393–402 (2018).
43. Schaffner SF et al. Calibrating a coalescent simulation of human genome sequence variation. Genome Res. 15, 1576–1583 (2005). [PubMed: 16251467]
44. Natarajan P et al. Deep-coverage whole genome sequences and blood lipids among 16,324 individuals. Nat. Commun 9, 3391 (2018). [PubMed: 30140000]
45. Taliun D et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. BioRxiv, 563866 (2019).
46. Huang Y-F, Gulko B & Siepel A Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. Nat. Genet 49, 618–624 (2017). [PubMed: 28288115]
47. Rogers MF et al. FATHMM-XF: accurate prediction of pathogenic point mutations via extended features. Bioinformatics 34, 511–513 (2018). [PubMed: 28968714]
48. Forrest AR et al. A promoter-level mammalian expression atlas. Nature 507, 462–470 (2014). [PubMed: 24670764]
49. Andersson R et al. An atlas of active enhancers across human cell types and tissues. Nature 507, 455–461 (2014). [PubMed: 24670763]
50. Fishilevich S et al. GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. Database 2017, bax028 (2017).
51. Dong C et al. Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. Hum. Mol. Genet 24, 2125–2137 (2014). [PubMed: 25552646]
52. Sabatti C et al. Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. Nat. Genet 41, 35–46 (2009). [PubMed: 19060910]
53. Kathiresan S et al. Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. Nat. Genet 40, 189–197 (2008). [PubMed: 18193044]

54. Huang C-C et al. Longitudinal association of PCSK9 sequence variations with low-density lipoprotein cholesterol levels: the Coronary Artery Risk Development in Young Adults Study. Circ.Cardiovasc. Genet 2, 354–361 (2009). [PubMed: 20031607]
55. Lange LA et al. Whole-exome sequencing identifies rare and low-frequency coding variants associated with LDL cholesterol. Am. J. Hum. Genet 94, 233–245 (2014). [PubMed: 24507775]
56. Bomba L, Walter K & Soranzo N The impact of rare and low-frequency genetic variants in common disease. Genome Biol. 18, 77 (2017). [PubMed: 28449691]
57. Ference BA, Majeed F, Penumetcha R, Flack JM & Brook RD Effect of naturally random allocation to lower low-density lipoprotein cholesterol on the risk of coronary heart disease mediated by polymorphisms in NPC1L1, HMGCR, or both: a 2 × 2 factorial Mendelian randomization study. J. Am. Coll. Cardiol 65, 1552–1561 (2015). [PubMed: 25770315]
58. Teslovich TM et al. Biological, clinical and population relevance of 95 loci for blood lipids. Nature 466, 707–713 (2010). [PubMed: 20686565]
59. Surakka I et al. The impact of low-frequency and rare variants on lipid levels. Nat. Genet 47, 589–597 (2015). [PubMed: 25961943]
60. Kathiresan S et al. Common variants at 30 loci contribute to polygenic dyslipidemia. Nat. Genet 41, 56–65 (2009). [PubMed: 19060906]
61. Kamatani Y et al. Genome-wide association study of hematological and biochemical traits in a Japanese population. Nat. Genet 42, 210–215 (2010). [PubMed: 20139978]
62. Nagy R et al. Exploration of haplotype research consortium imputation for genome-wide association studies in 20,032 Generation Scotland participants. Genome Med. 9, 23 (2017). [PubMed: 28270201]
63. Aulchenko YS et al. Loci influencing lipid levels and coronary heart disease risk in 16 European population cohorts. Nat. Genet 41, 47–55 (2009). [PubMed: 19060911]
64. Deelen J et al. Genome-wide association study identifies a single major locus contributing to survival into old age; the APOE locus revisited. Aging Cell 10, 686–698 (2011). [PubMed: 21418511]
65. Klarin D et al. Genetics of blood lipids among~ 300,000 multi-ethnic participants of the Million Veteran Program. Nat. Genet 50, 1514–1523 (2018). [PubMed: 30275531]
66. Hoffmann TJ et al. A large electronic-health-record-based genome-wide study of serum lipids. Nat. Genet 50, 401–413 (2018). [PubMed: 29507422]
67. Willer CJ et al. Discovery and refinement of loci associated with lipid levels. Nat. Genet 45, 1274–1283 (2013). [PubMed: 24097068]
68. Cohen JC et al. Multiple rare variants in NPC1L1 associated with reduced sterol absorption and plasma low-density lipoprotein levels. Proc. Natl. Acad. Sci. USA 103, 1810–1815 (2006). [PubMed: 16449388]
69. Myocardial Infarction Genetics Consortium Investigators. Inactivating mutations in NPC1L1 and protection from coronary heart disease. N. Engl. J. Med 371, 2072–2082 (2014). [PubMed: 25390462]
70. Cooper GM et al. Single-nucleotide evolutionary constraint scores highlight disease-causing mutations. Nat. Methods 7, 250–251 (2010). [PubMed: 20354513]
71. Cooper GM & Shendure J Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. Nat. Rev. Genet 12, 628–640 (2011). [PubMed: 21850043]
72. Van Hout CV et al. Whole exome sequencing and characterization of coding variation in 49,960 individuals in the UK Biobank. bioRxiv, 572347 (2019).
73. TG HDL Working Group of the Exome Sequencing Project, Lung NH, & Institute, B. Loss-of-function mutations in APOC3, triglycerides, and coronary disease. N. Engl. J. Med 371, 22–31 (2014). [PubMed: 24941081]
74. ENCODE Project Consortium. A user's guide to the encyclopedia of DNA elements (ENCODE). PLoS Biol. 9, e1001046 (2011). [PubMed: 21526222]
75. Landrum MJ et al. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res. 46, D1062–D1067 (2018). [PubMed: 29165669]

76. Davis HR & Veltri EP Zetia: inhibition of Niemann-Pick C1 Like 1 (NPC1L1) to reduce intestinal cholesterol absorption and treat hyperlipidemia. J. Atheroscler. Thromb 14, 99–108 (2007). [PubMed: 17587760]

77. Klos K et al. APOE/C1/C4/C2 hepatic control region polymorphism influences plasma apoE and LDL cholesterol levels. Hum. Mol. Genet 17, 2039–2046 (2008). [PubMed: 18378515]

78. Lu Q, Powles RL, Wang Q, He BJ & Zhao H Integrative tissue-specific functional annotations in the human genome provide novel insights on many complex traits and improve signal prioritization in genome wide association studies. PLoS Genet. 12, e1005947 (2016). [PubMed: 27058395]

79. Backenroth D et al. FUN-LDA: A latent Dirichlet allocation model for predicting tissue-specific functional effects of noncoding variation: methods and applications. Am. J. Hum. Genet 102, 920–942 (2018). [PubMed: 29727691]

80. Bodea CA et al. PINES: phenotype-informed tissue weighting improves prediction of pathogenic noncoding variants. Genome Biol. 19, 173 (2018). [PubMed: 30359302]

81. Park J-H et al. Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. Nat. Genet 42, 570–575 (2010). [PubMed: 20562874]

82. Derkach A, Zhang H & Chatterjee N Power Analysis for Genetic Association Test (PAGEANT) provides insights to challenges for rare variant association studies. Bioinformatics 34, 1506–1513 (2017).

83. Li Z et al. Dynamic scan procedure for detecting rare-variant association regions in whole-genome sequencing studies. Am. J. Hum. Genet 104, 802–814 (2019). [PubMed: 30982610]

Methods-only references

84. Yang J, Lee SH, Goddard ME & Visscher PM GCTA: a tool for genome-wide complex trait analysis. Am. J.Hum. Genet 88, 76–82 (2011). [PubMed: 21167468]

85. Conomos MP, Reiner AP, Weir BS & Thornton TA Model-free estimation of recent genetic relatedness. Am. J. Hum. Genet 98, 127–148 (2016). [PubMed: 26748516]

86. Dey R, Schmidt EM, Abecasis GR & Lee S A fast and accurate algorithm to test for binary phenotypes and its application to PheWAS. Am. J. Hum. Genet 101, 37–49 (2017). [PubMed: 28602423]

87. Zhou W et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. Nat. Genet 50, 1335–1341 (2018). [PubMed: 30104761]

88. Karimzadeh M, Ernst C, Kundaje A & Hoffman MM Umap and Bismap: quantifying genome and methylome mappability. Nucleic Acids Res. 46, e120–e120 (2018). [PubMed: 30169659]

89. Regier AA et al. Functional equivalence of genome sequencing analysis pipelines enables harmonized variant calling across human genetics projects. Nat. Commun 9, 4038 (2018). [PubMed: 30279509]

90. Gazal S et al. Linkage disequilibrium–dependent architecture of human complex traits shows action of negative selection. Nat. Genet 49, 1421–1427 (2017). [PubMed: 28892061]

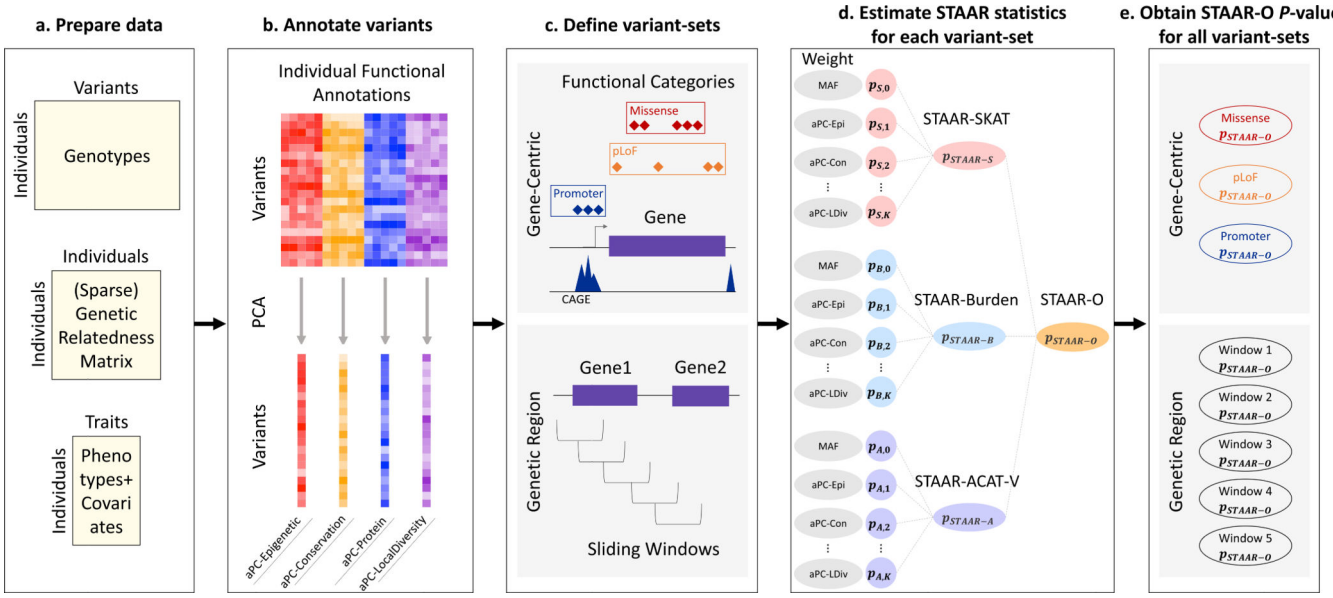

Figure 1 |. STAAR workflow.

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

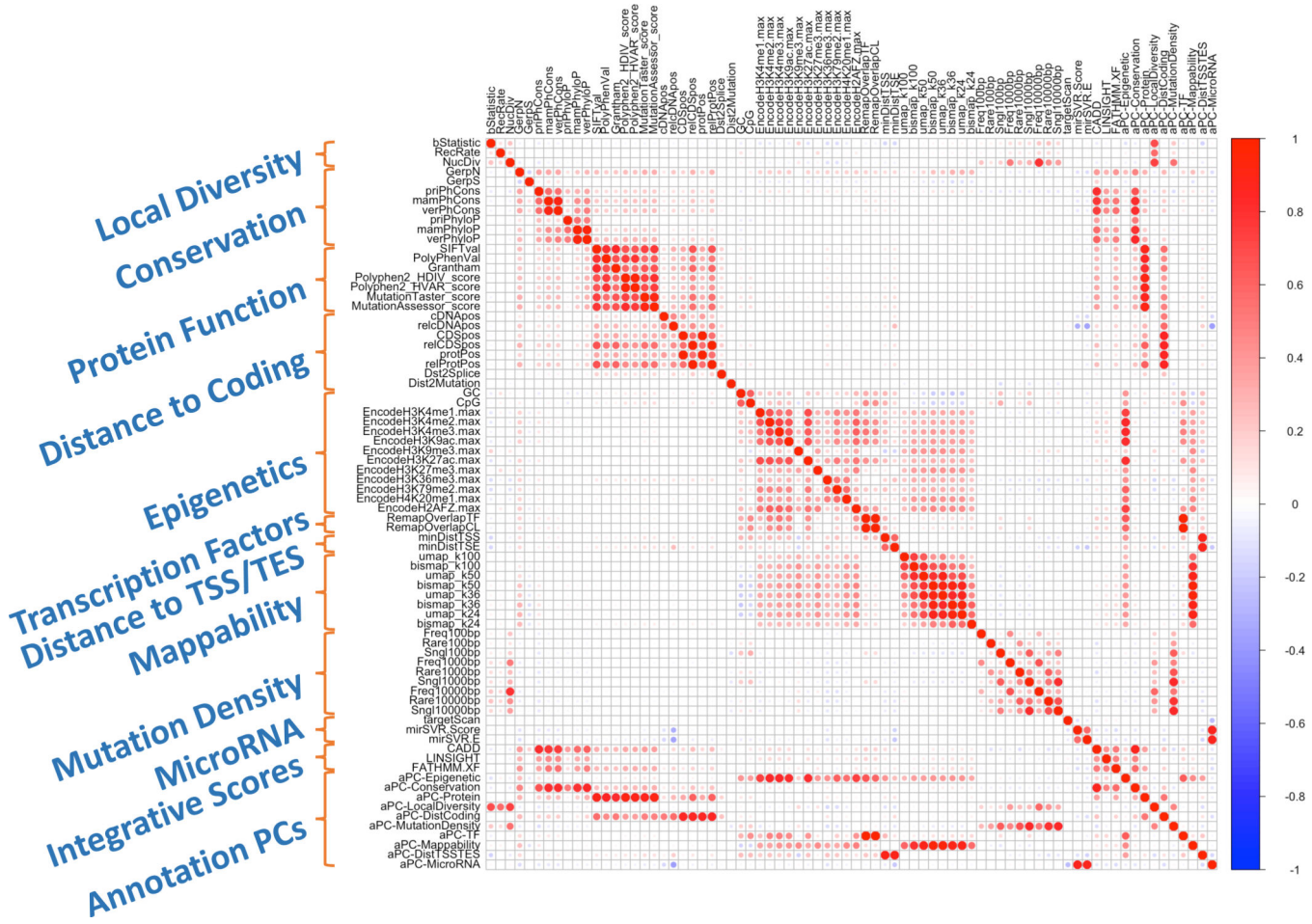

Figure 2 |. Correlation heatmap of functional annotation scores.

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

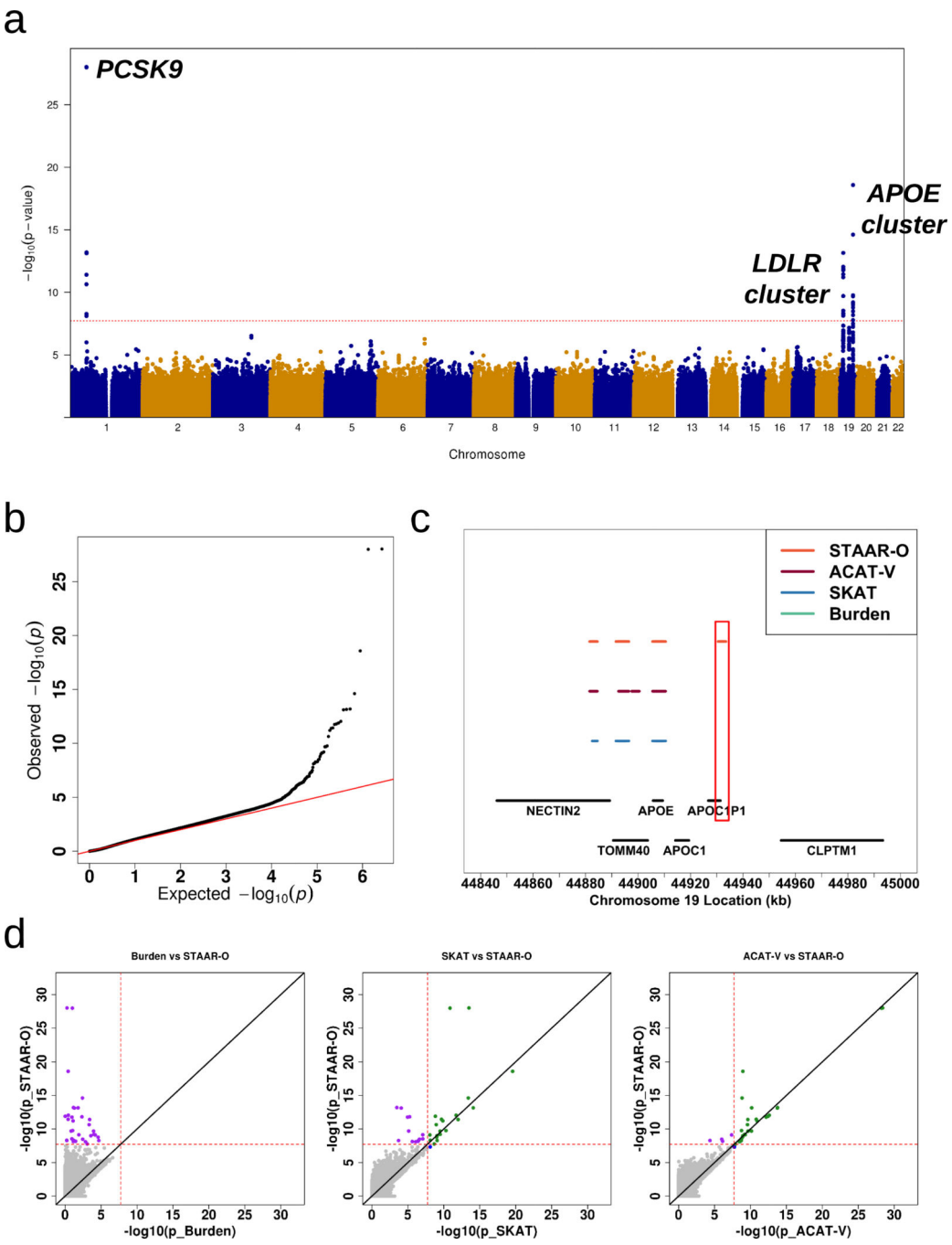

Figure 3 |. Genetic region (2-kb sliding window) unconditional analysis results of LDL-C in discovery phase using the TOPMed cohort.

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

indicates that the sliding window at this location is significant using the statistical test that the color of the dot represents (n = 12,316). **d**, Scatterplot of P-values for the 2-kb sliding windows comparing STAAR-O with Burden, SKAT and ACAT-V tests. Each dot represents a sliding window with x-axis label being the $-\log_{10}(P\ value)$ of the conventional test and y-axis label being the $-\log_{10}(P\ value)$ of STAAR-O (n = 12,316).

Table 1 |

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

12,316 discovery samples, 17,822 replication samples and 30,138 pooled samples from TOPMed program were considered in the analysis. Results for the conditionally significant genes (unconditional STAAR-O $P < 5.00 \times 10^{-7}$; conditional STAAR-O $P < 2.38 \times 10^{-3}$) using discovery samples are presented in the table. Chr (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); LDL-C (low-density lipoprotein cholesterol); HDL-C (high-density lipoprotein cholesterol); TG (triglycerides); TC (total cholesterol); Variants Adjusted (adjusted variants in conditional analysis).

Trait	Gene	Chr	Category	Discovery			Replication			Pooled			Variants Adjusted
				#SNV	STAAR-O (Unconditional)	STAAR-O (Conditional)	#SNV	STAAR-O (Unconditional)	STAAR-O (Conditional)	#SNV	STAAR-O (Unconditional)	STAAR-O (Conditional)	
LDL-C	*PCSK9*	1	pLoF	5	3.09E-38	1.94E-07	8	6.97E-27	5.29E-10	9	4.59E-65	7.52E-17	rs28362286, rs28362263, rs11591147, rs12117661
	APOB	2	pLoF	11	1.91E-14	2.38E-14	5	1.97E-09	1.76E-09	16	3.91E-21	4.08E-21	rs934197
	PCSK9	1	missense	92	1.09E-16	2.65E-08	129	1.90E-06	1.15E-06	167	2.11E-15	1.14E-14	rs28362286, rs28362263, rs11591147, rs12117661
	NPC1L1	7	missense	174	1.29E-07	3.83E-07	219	2.19E-03	3.28E-03	293	3.25E-10	1.58E-09	rs10234070, rs73107473, rs2072183, rs41279633, rs17725246, rs2073547, rs10260606, rs217386, rs7791240, rs2300414
	NPC1L1	7	disruptive missense	94	3.15E-09*	9.27E-09*	129	1.46E-04*	2.59E-04*	173	8.05E-12*	4.02E-11*	rs10234070, rs73107473, rs2072183, rs41279633, rs17725246, rs2073547, rs10260606, rs217386, rs7791240, rs2300414
	APOE	19	missense	54	3.11E-10	9.88E-11	58	6.61E-05	3.47E-04	88	1.07E-13	2.02E-12	rs7412, rs429358

Nat Genet. Author manuscript; available in PMC 2021 February 24.

Trait	Gene	Chr	Category	Discovery			Replication			Pooled			Variants Adjusted
				#SNV	STAAR-O (Unconditional)	STAAR-O (Conditional)	#SNV	STAAR-O (Unconditional)	STAAR-O (Conditional)	#SNV	STAAR-O (Unconditional)	STAAR-O (Conditional)	
HDL-C	*APOC3*	11	pLoF	5	2.20E-07	6.82E-07	6	5.73E-18	2.89E-17	7	3.18E-23	4.51E-22	rs66505542
TG	*APOC3*	11	pLoF	5	1.10E-14	5.53E-14	6	2.67E-49	2.73E-46	7	3.98E-56	1.04E-52	rs66505542, rs964184, rs7350481
TC	*PCSK9*	1	pLoF	5	4.60E-33	2.04E-10	8	1.83E-25	9.74E-11	9	9.83E-58	4.23E-20	rs28362286, rs11591147, rs191448952
	APOB	2	pLoF	11	7.29E-13	8.78E-13	5	2.62E-09	2.30E-09	16	9.76E-20	1.01E-19	rs934197
	PCSK9	1	missense	92	6.00E-15	1.11E-06	131	2.14E-05	1.13E-05	169	5.18E-12	3.16E-12	rs28362286, rs11591147, rs191448952
	LIPG	18	missense	62	9.61E-08	4.34E-06	68	3.45E-04	1.47E-01	101	2.04E-09	5.62E-04	rs4939883, rs7241918, rs149615216

*Burden test *P*-value.

Table 2 |

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

12,316 discovery samples, 17,822 replication samples and 30,138 pooled samples from the TOPMed program were considered in the analysis. Results for the conditionally significant sliding windows (unconditional STAAR-O $P < 1.88 \times 10^{-8}$; conditional STAAR-O $P < 8.47 \times 10^{-4}$) using discovery samples are presented in the table. Chr (chromosome); Start Location (start location of the 2-kb 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); 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			Pooled			Variants Adjusted
					#SNV	STAAR-O (Unconditional)	STAAR-O (Conditional)	#SNV	STAAR-O (Unconditional)	STAAR-O (Conditional)	#SNV	STAAR-O (Unconditional)	STAAR-O (Conditional)	
LDL-C	1	55045498	55047497	*PCSK9*	114	7.83E-09	1.06E-04	124	3.33E-06	4.10E-04	186	1.89E-15	2.90E-09	rs28362286, rs28362263, rs11591147, rs12117661
	1	55046498	55048497	*PCSK9*	124	5.32E-09	2.13E-05	130	1.79E-06	8.79E-05	191	1.33E-15	1.15E-09	rs28362286, rs28362263, rs11591147, rs12117661
	19	44881528	44883527	*NECTIN2*	118	7.31E-10	1.81E-08	155	5.16E-04	2.42E-01	202	8.15E-08	5.26E-06	rs7412, rs429358
	19	44882528	44884527	*NECTIN2*	104	2.08E-10	3.90E-09	133	1.23E-01	3.59E-01	176	1.38E-08	7.47E-07	rs7412, rs429358
	19	44893528	44895527	*TOMM40*	110	2.64E-19	2.33E-11	136	4.54E-09	2.60E-02	187	7.29E-29	7.62E-13	rs7412, rs429358
	19	44894528	44896527	*TOMM40*	120	2.44E-15	4.31E-11	153	7.62E-05	1.74E-02	205	6.73E-20	5.28E-13	rs7412, rs429358
	19	44905528	44907527	*APOE*	91	1.73E-10	1.64E-10	115	1.22E-02	4.91E-03	169	7.68E-12	9.00E-12	rs7412, rs429358
	19	44906528	44908527	*APOE*	84	1.67E-09	1.90E-10	115	8.65E-03	3.24E-03	165	8.34E-11	6.25E-12	rs7412, rs429358
	19	44907528	44909527	*APOE*	113	1.01E-09	1.97E-10	143	5.92E-03	3.58E-03	205	4.88E-11	8.71E-12	rs7412, rs429358
	19	44908528	44910527	*APOE*	140	6.30E-10	1.32E-10	152	4.14E-03	6.10E-03	228	2.40E-11	5.21E-12	rs7412, rs429358
	19	44931528	44933527	*APOC1P1*	114	6.63E-09	7.60E-04	123	5.78E-11	5.40E-03	181	1.34E-19	4.15E-06	rs7412, rs429358

Trait	Chr	Start Location	End Location	Gene	Discovery			Replication			Pooled			Variants Adjusted
					#SNV	STAAR-O (Unconditional)	STAAR-O (Conditional)	#SNV	STAAR-O (Unconditional)	STAAR-O (Conditional)	#SNV	STAAR-O (Unconditional)	STAAR-O (Conditional)	
TG	11	116828930	116830929	*APOC3*	125	4.63E-10	2.80E-09	155	1.35E-36	3.94E-34	207	7.32E-45	2.73E-41	rs66505542, rs964184, rs7350481
	11	116829930	116831929	*APOC3*	109	3.61E-10	5.99E-10	140	2.85E-36	4.25E-34	187	5.75E-45	2.17E-41	rs66505542, rs964184, rs7350481
TC	1	55045498	55047497	*PCSK9*	114	3.05E-09	2.86E-07	130	3.12E-06	1.92E-06	189	2.22E-15	9.21E-14	rs28362286, rs11591147, rs191448952
	1	55046498	55048497	*PCSK9*	124	2.24E-09	2.06E-07	138	2.19E-06	1.34E-06	195	1.78E-15	7.04E-14	rs28362286, rs11591147, rs191448952
	19	44893528	44895527	*TOMM40*	111	9.35E-13	4.37E-07	146	1.12E-07	4.02E-01	196	7.57E-21	7.91E-08	rs7412, rs429358
	19	44894528	44896527	*TOMM40*	120	1.80E-09	1.99E-06	164	1.08E-04	8.31E-01	213	8.40E-14	2.19E-07	rs7412, rs429358