nature genetics

TECHNICAL REPORT

https://doi.org/10.1038/s41588-020-0676-4

Check for updates

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

Xihao Li[1,218], Zilin Li[1,218], 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] and Xihong Lin[1,26,35] ✉

Large-scale whole-genome sequencing studies have enabled the analysis of rare variants (RVs) associated with complex phenotypes. Commonly used RV association tests have limited scope to leverage variant functions. We propose STAAR (variant-set test for association using annotation information), a scalable and powerful RV association test method that effectively incorporates both variant categories and multiple complementary annotations using a dynamic weighting scheme. For the latter, we introduce 'annotation principal components', multidimensional summaries of in silico variant annotations. STAAR accounts for population structure and relatedness and is scalable for analyzing very large cohort and biobank whole-genome sequencing studies of continuous and dichotomous traits. We applied STAAR to identify RVs associated with four lipid traits in 12,316 discovery and 17,822 replication samples from the Trans-Omics for Precision Medicine Program. We discovered and replicated new 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 and the Genome Sequencing Program (GSP) of the National Human Genome Research Institute. Such studies enable the assessment of associations between complex traits and both coding and noncoding RVs (minor allele frequency (MAF) < 1%) across the genome. However, single-variant analyses typically have low power to identify associations with RVs[1–3]. To improve power, variant set tests have been proposed to jointly test the effects of given sets of multiple RVs. 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 to prioritize 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].

A full list of affiliations appears at the end of the paper.

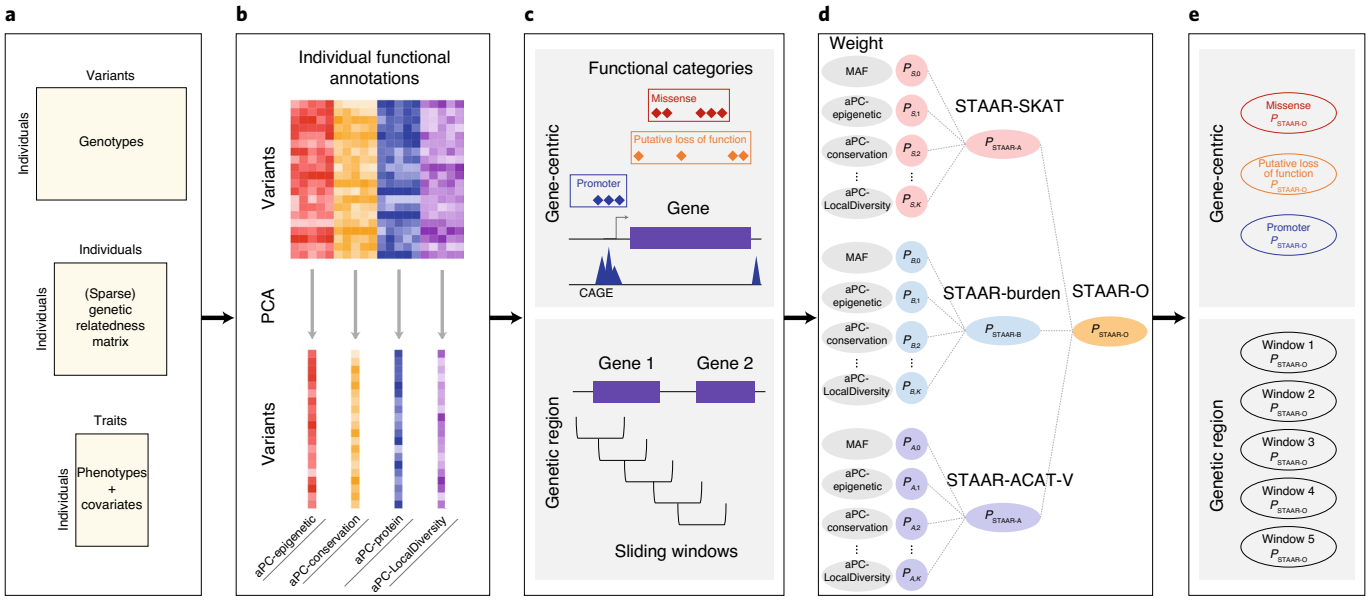

Fig. 1 | STAAR workflow. a, Input data of STAAR, including genotypes, phenotypes, covariates and (sparse) genetic relatedness matrix is prepared. **b**, All variants in the genome are annotated and the annotation principal components for different classes of variant function are calculated. PCA, principal component analysis. **c**, Two types of variant sets are defined: gene-centric analysis by grouping variants into functional genomic elements for each protein-coding gene; and genetic region analysis using agnostic sliding windows. **d**, The STAAR statistics for each variant set is calculated. aPC, annotation principal component. **e**, The STAAR-O *P* values for all variant sets defined in **c** are obtained and significant findings are reported.

Variant functional annotations take two forms: (1) qualitative functional groupings into genomic elements, such as variant effect predictor categories[20,21]; and (2) 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 RV association study power, for example, by optimally selecting and weighting plausibly causal RVs[31].

To boost power for variant set tests in the WGS RV association study, 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 multidimensional weighting scheme. For the latter, to effectively capture the multifaceted biological impact of a variant, we introduce annotation principal components, multidimensional 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 cannot be scaled 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 model (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.

In the present study, we performed 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 applied STAAR to perform WGS gene-centric and sliding window-based genetic region analysis of 12,316 discovery and 17,822 replication samples with 4 quantitative lipid traits: low-density lipoprotein cholesterol (LDL-C); high-density lipoprotein cholesterol (HDL-C); triglycerides (TG) and total cholesterol (TC) from the National Heart, Lung, and Blood Institute (NHLBI) TOPMed program. We show that STAAR outperforms existing methods and identifies new 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 RV association study at scale by using both qualitative functional categories and 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: (1) using annotation principal components to capture and prioritize multidimensional variant biological functions; and (2) testing the association between each variant set and phenotypes by incorporating these annotation principal components 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 extracted a broad set of variant functional annotations (Supplementary Table 1), including individual and ensemble functional scores, from various databases, such as ENCODE (v.2)[26] and Roadmap Epigenomics (Human Epigenome Atlas release 9)[39], as well as 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, for example, epigenetic function, evolutionary conservation, protein function and local nucleotide diversity. We introduce annotation principal components defined as the first principal components calculated from the set of individual functional annotation scores in each functional block (Supplementary Table 1 and Methods). Annotation principal components 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 the aggregated Cauchy association test (ACAT) method to combine the resulting *P* values calculated using different weights to effectively and powerfully aggregate the association strength from all annotations in a data-adaptive manner (Fig. 1d and 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 Methods).

Simulation studies. To evaluate type I errors and the power of STAAR compared to conventional variant set tests, we performed simulation studies under a variety of configurations. We followed the steps described in Data simulation (Methods) to generate both continuous and dichotomous phenotypes. We generated genotypes by simulating 20,000 sequences for 100 different regions with each spanning 1 megabase (Mb). The data were generated to mimic the linkage disequilibrium structure of an African-American population by using the calibration coalescent model[43]. We randomly selected 5-kilobase (kb) regions from these 1-Mb regions and considered sample sizes of 2,500, 5,000 and 10,000 for each replicate. The simulation studies focused on aggregating uncommon variants with an 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 was 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 (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 ten annotations into its analysis and comparing the 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. The causality of variants was allowed to be dependent on different sets of annotations through a logistic model (Methods). We considered different proportions of causal variants (5, 15 and 35% on average) in the signal region. For both continuous and dichotomous traits, STAAR-O incorporating all ten annotations had higher power than conventional variant set tests in terms of signal region detection (Supplementary Figs. 1–4). The 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 described elsewhere[44]. DNA samples were sequenced at >30× target coverage. The discovery phase consisted of four study cohorts of TOPMed Freeze 3. The replication phase consisted of ten different study cohorts in TOPMed Freeze 5 that were not in Freeze 3 (Supplementary Note and Supplementary Table 3).

We performed sample- and variant-level quality control[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-ancestry-related samples. Race/ethnicity was defined using a combination of self-reported race/ethnicity and study recruitment information. The discovery cohorts consisted of 4,580 (37.2%) Black or African-American individuals, 6,266 (50.9%) White, 543 (4.4%) Asian-American and 927 (7.5%) Hispanic/Latino American. Among all samples in the 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%) RVs (MAF < 1%). The race/ethnicity, related sample and variant number distributions for the replication phase and pooled samples (samples from both discovery and replication phases) are given in Supplementary Table 4.

Our study used the proposed STAAR-O method to perform (1) gene-centric analysis using RV sets based on functional categories and (2) 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 ten ancestral principal components, while controlling for relatedness using linear mixed models, with inverse rank normal transformation applied to phenotypes (Methods). Race/ethnicity was included as a covariate to adjust for sociocultural and environmental factors, while genetic ancestry differences were captured by the inclusion of the ancestral principal components. In addition to the 2 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 annotation principal components. Figure 2 summarizes the correlation among all functional annotations, including 63 individual scores, 3 integrative scores and 10 annotation principal components.

Gene-centric association analysis of coding and noncoding RVs. We performed gene-centric analysis to identify whether RVs 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 noncoding variants of each gene: (1) putative loss of function (stop gain, stop loss and splice) RVs; (2) missense RVs; (3) synonymous RVs; (4) promoter RVs; and (5) enhancer RVs. The putative loss of function, missense and synonymous RVs were defined by GENCODE variant effect predictor categories[20,21]. Promoter RVs were defined as RVs in the ±3-kb window of the transcription starting site (TSS) with overlap of cap analysis of gene expression (CAGE) sites. 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 RVs (MAF < 1%) in the functional category and lipid traits using STAAR-O with the 13 aggregated functional annotations described earlier. For missense RVs, we incorporated an additional annotation functional category predicting functionally 'disruptive' variants determined by meta-analytic support vector machine (MetaSVM)[51], which measures the

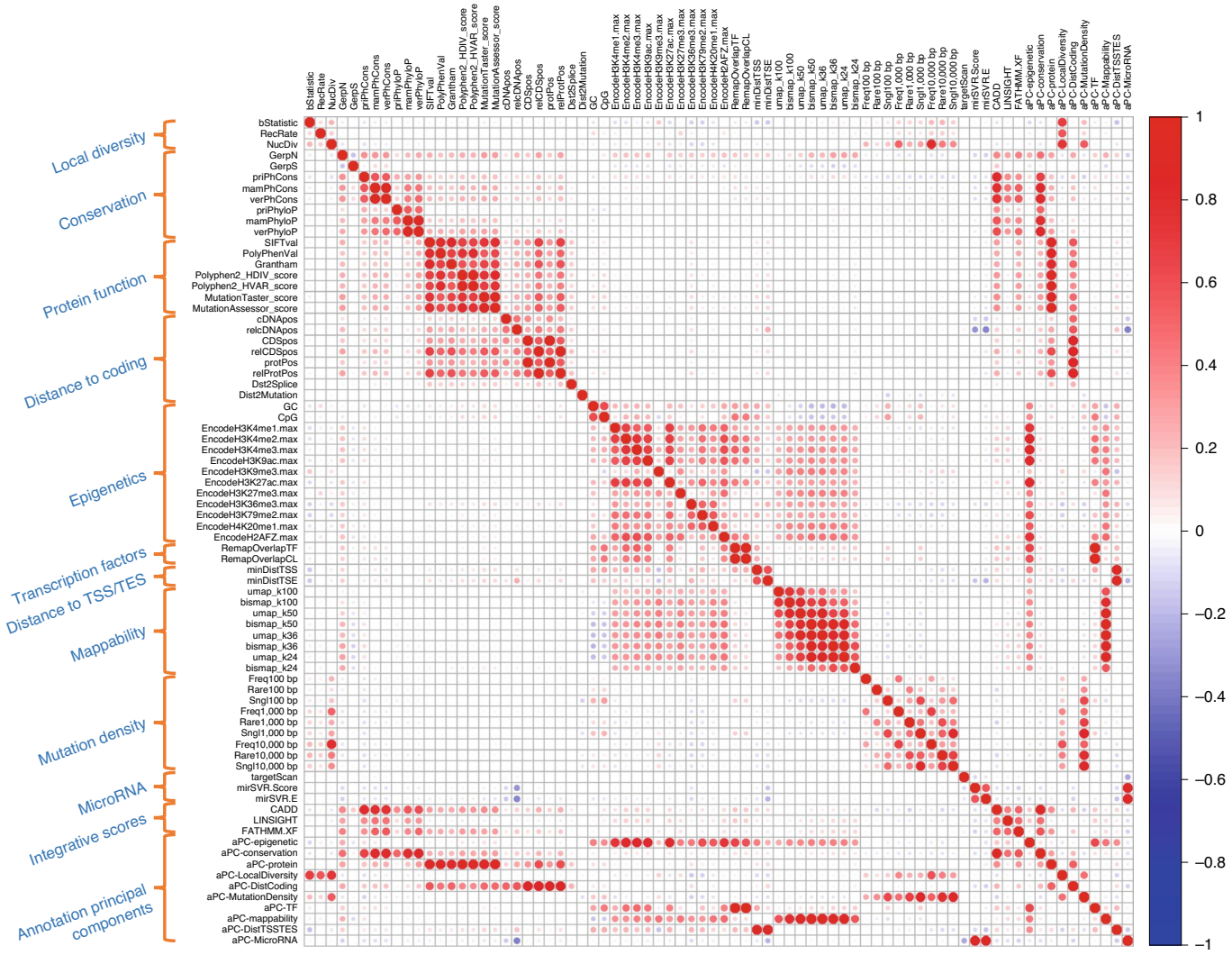

Fig. 2 | Correlation heatmap of functional annotation scores. Pairwise correlations between 76 individual and integrative functional annotations using variants from the pooled samples of lipid traits in the TOPMed data. The cells in the visualization are colored by Pearson's correlation coefficient values with the deeper colors indicating higher positive (red) or negative (blue) correlations. Each annotation principal component is the first principal component calculated from the set of individual functional annotations that measure similar biological function. These annotation principal components are then transformed into the Phred-scaled scores for each variant across the genome (Methods).

deleteriousness of missense mutations. The overall distributions of STAAR-O *P* values were well calibrated for all four lipid phenotypes (Supplementary Fig. 8). For unconditional analysis, we considered a Bonferroni-corrected genome-wide significance threshold of $\alpha = 0.05/(20{,}000 \times 5) = 5.00 \times 10^{-7}$ accounting for 5 different masks across protein-coding genes.

STAAR-O identified 21 genome-wide significant associations with 4 lipid phenotypes using unconditional analysis of the discovery samples (Supplementary Table 5 and Supplementary Fig. 9). After conditioning on known lipid-associated variants[44,52–67], 11 out of the 21 associations were significant at the Bonferroni correction level of $0.05/21 = 2.38 \times 10^{-3}$ using the discovery samples. These included associations with LDL-C (putative loss-of-function RVs in *PCSK9* and *APOB*, missense RVs in *PCSK9*, *NPC1L1* and *APOE*), association with HDL-C (putative loss-of-function RVs in *APOC3*), association with TG (putative loss-of-function RVs in *APOC3*) and associations with TC (putative loss-of-function 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 of $0.05/11 = 4.55 \times 10^{-3}$ after adjusting for known lipid-associated variants. The association between *APOC3* putative loss-of-function 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 annotation principal component-conservation scores (Supplementary Table 6). This contributed to the test weighted by annotation principal component-conservation being the most

Table 1 | Gene-centric analysis results of both unconditional analysis and analysis conditional on known common and low-frequency variants

Trait	Gene	Chromosome no.	Category	Discovery			Replication			Pooled		Variants (adjusted)
				No. of SNVs	STAAR-O (unconditional)	STAAR-O (conditional)	No. of SNVs	STAAR-O (unconditional)	STAAR-O (conditional)	No. of SNVs	STAAR-O (unconditional)	STAAR-O (conditional)	
LDL-C	*PCSK9*	1	Putative loss of function	5	3.09×10^{-38}	1.94×10^{-7}	8	6.9×10^{-27}	5.29×10^{-10}	9	4.59×10^{-65}	7.52×10^{-17}	rs28362286, rs28362263, rs11591147, rs12117661
	APOB	2	Putative loss of function	11	1.91×10^{-14}	2.38×10^{-14}	5	1.97×10^{-9}	1.76×10^{-9}	16	3.91×10^{-21}	4.08×10^{-21}	rs934197
	PCSK9	1	Missense	92	1.09×10^{-16}	2.65×10^{-8}	129	1.90×10^{-6}	1.15×10^{-6}	167	2.11×10^{-15}	1.14×10^{-14}	rs28362286, rs28362263, rs11591147, rs12117661
	NPC1L1	7	Missense	174	1.29×10^{-7}	3.83×10^{-7}	219	2.19×10^{-3}	3.28×10^{-3}	293	3.25×10^{-10}	1.58×10^{-9}	rs10234070, rs73107473, rs2072183, rs41279633, rs17725246, rs2073547, rs10260606, rs217386, rs7791240, rs2300414
	NPC1L1	7	Disruptive missense	94	3.15×10^{-9}[a]	9.27×10^{-9}[a]	129	1.46×10^{-4}[a]	2.59×10^{-4}[a]	173	8.05×10^{-12}[a]	4.02×10^{-11}[a]	rs10234070, rs73107473, rs2072183, rs41279633, rs17725246, rs2073547, rs10260606, rs217386, rs7791240, rs2300414
	APOE	19	Missense	54	3.11×10^{-10}	9.88×10^{-11}	58	6.61×10^{-5}	3.47×10^{-4}	88	1.07×10^{-13}	2.02×10^{-12}	rs7412, rs429358
HDL-C	*APOC3*	11	Putative loss of function	5	2.20×10^{-7}	6.82×10^{-7}	6	5.73×10^{-18}	2.89×10^{-17}	7	3.18×10^{-23}	4.51×10^{-22}	rs66505542
TG	*APOC3*	11	Putative loss of function	5	1.10×10^{-14}	5.53×10^{-14}	6	2.67×10^{-49}	2.73×10^{-46}	7	3.98×10^{-56}	1.04×10^{-52}	rs66505542, rs964184, rs7350481
TC	*PCSK9*	1	Putative loss of function	5	4.60×10^{-33}	2.04×10^{-10}	8	1.83×10^{-25}	9.74×10^{-11}	9	9.83×10^{-58}	4.23×10^{-20}	rs28362286, rs11591147, rs191448952
	APOB	2	Putative loss of function	11	7.29×10^{-13}	8.78×10^{-13}	5	2.62×10^{-9}	2.30×10^{-9}	16	9.76×10^{-20}	1.01×10^{-19}	rs934197
	PCSK9	1	Missense	92	6.00×10^{-15}	1.11×10^{-6}	131	2.14×10^{-5}	1.13×10^{-5}	169	5.18×10^{-12}	3.16×10^{-12}	rs28362286, rs11591147, rs191448952
	LIPG	18	Missense	62	9.61×10^{-8}	4.34×10^{-6}	68	3.45×10^{-4}	1.47×10^{-1}	101	2.04×10^{-9}	5.62×10^{-4}	rs4939883, rs7241918, rs149615216

[a]Burden test *P* value. A total of 12,316 discovery samples, 17,822 replication samples and 30,138 pooled samples from the TOPMed Program were considered in the analysis. Results for the conditionally significant genes (unconditional STAAR-O, $P<5.00\times10^{-7}$; conditional STAAR-O, $P<2.38\times10^{-3}$) using the discovery samples are presented in the table. Category, functional category; number of SNVs, number of RVs with a MAF <1% of the particular functional category in the gene; STAAR-O, STAAR-O *P* value; variants (adjusted), adjusted variants in the conditional analysis.

significant across all quantitative annotation-weighted tests included in STAAR-O (burden $P=3.12\times10^{-7}$). Since annotation principal component-conservation summarizes the evolutionary conservation scores of variants, 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; refs. [57–61,65–67]) resulted in the association between disruptive missense RVs in *NPC1L1* and LDL-C remaining significant after Bonferroni correction with the conditional analysis $P=9.27\times10^{-9}$ in the discovery phase. This association was validated in the replication phase with $P=2.59\times10^{-4}$ and $P=4.02\times10^{-11}$ in pooled samples in the conditional analysis. This significant association was also validated using WES data from the UK Biobank[72] ($n=40{,}519$) with $P=2.49\times10^{-4}$ in the conditional analysis.

Genetic region analysis of RVs. We performed genetic region analysis to determine whether RVs within sliding windows were associated with lipid traits. The sliding windows were 2 kb in length, started at position 0 base pairs (bp) for each chromosome and had a skip length of 1 kb. Windows with a total minor allele count (MAC) < 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 the 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 $\alpha=0.05/(2.66\times10^{6})=1.88\times10^{-8}$ 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 detected 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 did not detect any window that reached significance.

Among the 59 genome-wide significant sliding windows detected by STAAR-O in the unconditional analysis, 17 were significant at the Bonferroni correction level of $0.05/59=8.47\times10^{-4}$ after conditioning on known lipid-associated variants using the discovery samples (Table 2). For LDL-C, the significant sliding windows were located in the *PCSK9* gene or in a 50-kb region on chromosome 19 including the *APOE* cluster. For TC, all 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, 6 were replicated at $0.05/17=2.94\times10^{-3}$ after Bonferroni correction and another 4 were replicated at $0.05/9=5.56\times10^{-3}$ after Bonferroni correction for 9 nonoverlapping sliding windows in the conditional analysis of replication samples[17], including a sliding window located downstream of *APOC1P1* (chromosome 19: 44,931,528–44,933,527 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, $P=8.71\times10^{-8}$), which was not significant at a Bonferroni-corrected threshold ($\alpha=0.05/(1.51\times10^{7})=3.31\times10^{-9}$) in individual variant analysis. This RV and the second top variant in these windows (rs9749443, MAF = 0.009, $P=2.46\times10^{-5}$) were upweighted by annotation principal component-epigenetic in STAAR-O (Supplementary Fig. 15). Specifically, the annotation principal component-epigenetic scores of rs370625306 and rs9749443 ranked in the top 10 and 30% among all RVs, respectively, in the 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 SNVs (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 the conditional analysis that were undetected by SKAT and the 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 annotation principal component from tissue-specific DNase, H3K4me3, H3K27ac and H3K27me3 from ENCODE (Supplementary Table 9) (refs. [26,74]). We used tissue-specific CAGE sites with overlap of RVs in the ±3-kb window of the 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 annotation principal component and without incorporating the MAF and other annotations.

Overall, the use of liver annotation resulted in more significant levels of association than the 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 nine and eight replicated conditionally significant associations by using liver annotation in gene-centric and genetic region analyses, respectively (Supplementary Tables 10 and 11). Among these 17 significant associations, 2 were not seen when the heart annotation was used and 2 were not seen when the brain annotation was used; no additional associations were detected by using the heart and brain annotations (Supplementary Tables 10 and 11). Furthermore, more suggestive significant associations were detected when using the 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

Fig. 3 | Genetic region (2-kb sliding window) unconditional analysis results of LDL-C in the discovery phase using the TOPMed cohort. a, Manhattan plot showing the associations of 2.66 million 2-kb sliding windows for LDL-C versus $-\log_{10}(P)$ of STAAR-O. The horizontal line indicates a genome-wide *P* value threshold of 1.88×10^{-8} ($n=12{,}316$). **b**, Quantile–quantile plot of 2-kb sliding window STAAR-O *P* values for LDL-C ($n=12{,}316$). **c**, Genetic landscape of the windows significantly associated with LDL-C that are located in the 150-kb region on chromosome 19. Four statistical tests were compared: burden; SKAT; ACAT-V; and STAAR-O. A dash indicates that the sliding window at this location was significant using the statistical test that the color of the dash represents ($n=12{,}316$). **d**, Scatterplot of *P* values for the 2-kb sliding windows comparing STAAR-O with the burden, SKAT and ACAT-V tests. Each dot represents a sliding window with the *x* axis label being the $-\log_{10}(P)$ of the conventional test and the *y* axis label being the $-\log_{10}(P)$ of STAAR-O ($n=12{,}316$).

variant annotations for WGS RV association studies. Using sparse GRMs[38], STAAR scaled 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 required

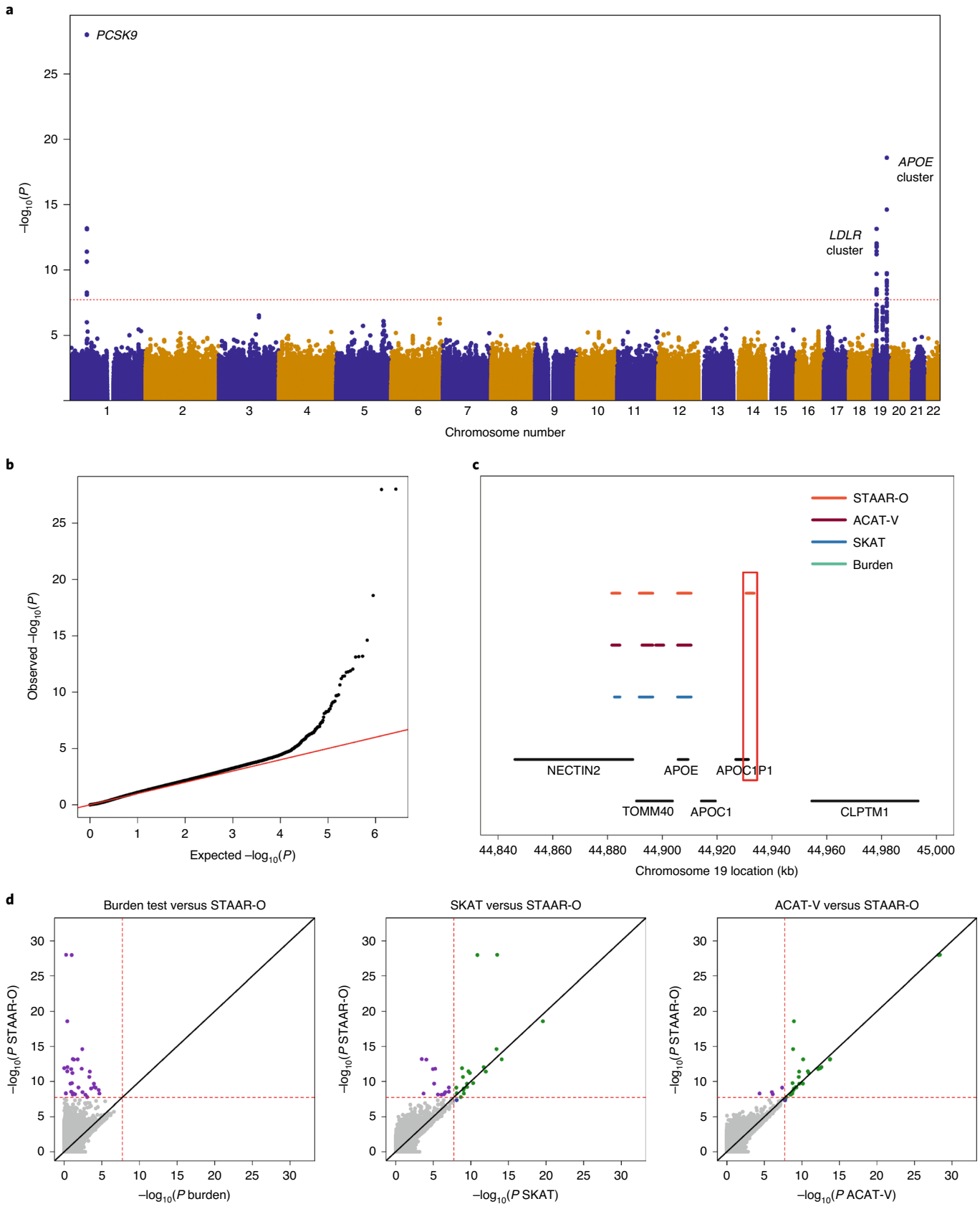

Table 2 | Genetic region (2-kb sliding window) analysis results of both unconditional analysis and analysis conditional on known common and low-frequency variants

Trait	Chromosome no.	Start location	End location	Gene	Discovery			Replication			Pooled			Variants (adjusted)
					No. of SNVs	STAAR-O (unconditional)	STAAR-O (conditional)	No. of SNVs	STAAR-O (unconditional)	STAAR-O (conditional)	No. of SNVs	STAAR-O (unconditional)	STAAR-O (conditional)	
LDL-C	1	55,045,498	55,047,497	*PCSK9*	114	7.83×10^{-9}	1.06×10^{-4}	124	3.33×10^{-6}	4.10×10^{-4}	186	1.89×10^{-15}	2.90×10^{-9}	rs28362286, rs28362263, rs11591147, rs12117661
	1	55,046,498	55,048,497	*PCSK9*	124	5.32×10^{-9}	2.13×10^{-5}	130	1.79×10^{-6}	8.79×10^{-5}	191	1.33×10^{-15}	1.15×10^{-9}	rs28362286, rs28362263, rs11591147, rs12117661
	19	44,881,528	44,883,527	*NECTIN2*	118	7.31×10^{-10}	1.81×10^{-8}	155	5.16×10^{-4}	2.42×10^{-1}	202	8.15×10^{-8}	5.26×10^{-6}	rs7412, rs429358
	19	44,882,528	44,884,527	*NECTIN2*	104	2.08×10^{-10}	3.90×10^{-9}	133	1.23×10^{-1}	3.59×10^{-1}	176	1.38×10^{-8}	7.47×10^{-7}	rs7412, rs429358
	19	44,893,528	44,895,527	*TOMM40*	110	2.64×10^{-19}	2.33×10^{-11}	136	4.54×10^{-9}	2.60×10^{-2}	187	7.29×10^{-29}	7.62×10^{-13}	rs7412, rs429358
	19	44,894,528	44,896,527	*TOMM40*	120	2.44×10^{-15}	4.31×10^{-11}	153	7.62×10^{-5}	1.74×10^{-2}	205	6.73×10^{-20}	5.28×10^{-13}	rs7412, rs429358
	19	44,905,528	44,907,527	*APOE*	91	1.73×10^{-10}	1.64×10^{-10}	115	1.22×10^{-2}	4.91×10^{-3}	169	7.68×10^{-12}	9.00×10^{-12}	rs7412, rs429358
	19	44,906,528	44,908,527	*APOE*	84	1.67×10^{-9}	1.90×10^{-10}	115	8.65×10^{-3}	3.24×10^{-3}	165	8.34×10^{-11}	6.25×10^{-12}	rs7412, rs429358
	19	44,907,528	44,909,527	*APOE*	113	1.01×10^{-9}	1.97×10^{-10}	143	5.92×10^{-3}	3.58×10^{-3}	205	4.88×10^{-11}	8.71×10^{-12}	rs7412, rs429358
	19	44,908,528	44,910,527	*APOE*	140	6.30×10^{-10}	1.32×10^{-10}	152	4.14×10^{-3}	6.10×10^{-3}	228	2.40×10^{-11}	5.21×10^{-12}	rs7412, rs429358
	19	44,931,528	44,933,527	*APOC1P1*	114	6.63×10^{-9}	7.60×10^{-4}	123	5.78×10^{-11}	5.40×10^{-3}	181	1.34×10^{-19}	4.15×10^{-6}	rs7412, rs429358
TG	11	116,828,930	116,830,929	*APOC3*	125	4.63×10^{-10}	2.80×10^{-9}	155	1.35×10^{-36}	3.94×10^{-34}	207	7.32×10^{-45}	2.73×10^{-41}	rs66505542, rs964184, rs7350481
	11	116,829,930	116,831,929	*APOC3*	109	3.61×10^{-10}	5.99×10^{-10}	140	2.85×10^{-36}	4.25×10^{-34}	187	5.75×10^{-45}	2.17×10^{-41}	rs66505542, rs964184, rs7350481
TC	1	55,045,498	55,047,497	*PCSK9*	114	3.05×10^{-9}	2.86×10^{-7}	130	3.12×10^{-6}	1.92×10^{-6}	189	2.22×10^{-15}	9.21×10^{-14}	rs28362286, rs11591147, rs191448952
	1	55,046,498	55,048,497	*PCSK9*	124	2.24×10^{-9}	2.06×10^{-7}	138	2.19×10^{-6}	1.34×10^{-6}	195	1.78×10^{-15}	7.04×10^{-14}	rs28362286, rs11591147, rs191448952
	19	44,893,528	44,895,527	*TOMM40*	111	9.35×10^{-13}	4.37×10^{-7}	146	1.12×10^{-7}	4.02×10^{-1}	196	7.57×10^{-21}	7.91×10^{-8}	rs7412, rs429358
	19	44,894,528	44,896,527	*TOMM40*	120	1.80×10^{-9}	1.99×10^{-6}	164	1.08×10^{-4}	8.31×10^{-1}	213	8.40×10^{-14}	2.19×10^{-7}	rs7412, rs429358

A total of 12,316 discovery samples, 17,822 replication samples and 30,138 pooled samples from the TOPMed Program were considered in the analysis. Results for the conditionally significant sliding windows (unconditional STAAR-O $P<1.88\times10^{-8}$; conditional STAAR-O $P<8.47\times10^{-4}$) using the discovery samples are presented in the table. Start location, start location of the 2-kb sliding window; end location, end location of the 2-kb sliding window; no. of SNVs, number of RVs (MAF < 1%) in the 2-kb sliding window; STAAR-O, STAAR-O *P* value; variants (adjusted), adjusted variants in the conditional analysis. The physical positions of each window are on build hg38.

15 h for 100 2.10 GHz computing cores with 6 gigabyte memory for each lipid trait. Analyzing 500,000 simulated related samples mimicking the UK Biobank sample size required 26 h for WGS analysis using the same approach and computational resources (Methods).

Discussion

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

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

In a WGS RV analysis of lipid traits using the TOPMed data, STAAR-O identified several conditionally significant functional categories associated with lipid traits in gene-centric analysis (including *NPC1L1* missense RVs and LDL-C; *APOC3* putative loss-of-function 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 (putative loss-of-function RVs and frameshift indels) in *NPC1L1* and LDL-C with $P=0.04$ (ref. [69]), which was replicated using the pooled TOPMed samples ($P=0.02$), although no significant association between putative loss-of-function RVs and LDL-C was found ($P=0.15$). STAAR-O identified a much more significant previously unknown association, which was replicated, between missense RVs in *NPC1L1* and LDL-C; this was driven by disruptive missense RVs (conditional $P=4.02\times10^{-11}$ in pooled samples). None of these disruptive missense RVs was reported in ClinVar (5 September 2017; 20170905)[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* (ref. [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 ten sliding windows after conditioning on known variants, including an association between an intergenic region located downstream of *APOC1P1* and LDL-C, which were not detected using conventional tests. The detected *APOC1P1* region is located in the hepatic control region 2 that regulates hepatic expression of apolipoproteins. By further conditioning on the *APOE* haplotypes and rs35136575, a common variant previously found in the downstream hepatic control region 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 RVs that have regulatory functions using annotation principal component-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 RV association studies.

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

In practice, STAAR is very flexible and users can determine the set of individual annotations to calculate annotation principal components and the number of annotation principal components 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 study, we grouped individual annotations based on biological knowledge; users can also apply data-driven approaches, such as clustering, to group annotations for annotation principal component calculation. We also demonstrated that STAAR detects more associations using relevant tissue functional annotations. It will be of interest, in future research, to incorporate improved RV effect size models in the weights to further improve the power for RV association studies[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 since it uses estimated sparse GRMs[38] to fit the null GLMM and scan the genome. Besides using sliding windows of a prespecified fixed window length, STAAR could be extended to flexibly detect the sizes and locations of coding and noncoding RV 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; hence, it could provide a comprehensive framework for WGS RV association studies. Thus, STAAR provides a powerful and flexible tool for variant association discovery in many settings to explore the molecular basis of common diseases. STAAR v.0.9.5 can be downloaded from https://github.com/xihaoli/STAAR and https://content.sph.harvard.edu/xlin/software.html.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-020-0676-4.

Received: 17 July 2019; Accepted: 2 July 2020;
Published online: 24 August 2020

References

1. Bansal, V., Libiger, O., Torkamani, A. & Schork, N. J. Statistical analysis strategies for association studies involving rare variants. *Nat. Rev. Genet.* **11**, 773–785 (2010).
2. Kiezun, A. et al. Exome sequencing and the genetic basis of complex traits. *Nat. Genet.* **44**, 623–630 (2012).
3. Lee, S., Abecasis, G. R., Boehnke, M. & Lin, X. Rare-variant association analysis: study designs and statistical tests. *Am. J. Hum. Genet.* **95**, 5–23 (2014).
4. Morgenthaler, S. & Thilly, W. G. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). *Mutat. Res.* **615**, 28–56 (2007).
5. Li, B. & Leal, S. M. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am. J. Hum. Genet.* **83**, 311–321 (2008).
6. Madsen, B. E. & Browning, S. R. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet.* **5**, e1000384 (2009).
7. Morris, A. P. & Zeggini, E. An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genet. Epidemiol.* **34**, 188–193 (2010).
8. Wu, M. C. et al. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.* **89**, 82–93 (2011).

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).
10. Lee, S., Wu, M. C. & Lin, X. Optimal tests for rare variant effects in sequencing association studies. *Biostatistics* **13**, 762–775 (2012).
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).
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).
13. Kichaev, G. et al. Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS Genet.* **10**, e1004722 (2014).
14. Kichaev, G. et al. Improved methods for multi-trait fine mapping of pleiotropic risk loci. *Bioinformatics* **33**, 248–255 (2017).
15. Finucane, H. K. et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
16. Hu, Y. et al. Leveraging functional annotations in genetic risk prediction for human complex diseases. *PLoS Comp. Biol.* **13**, e1005589 (2017).
17. Morrison, A. C. et al. Practical approaches for whole-genome sequence analysis of heart-and blood-related traits. *Am. J. Hum. Genet.* **100**, 205–215 (2017).
18. Schaid, D. J., Chen, W. & Larson, N. B. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat. Rev. Genet.* **19**, 491–504 (2018).
19. Claussnitzer, M. et al. A brief history of human disease genetics. *Nature* **577**, 179–189 (2020).
20. Harrow, J. et al. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* **22**, 1760–1774 (2012).
21. Frankish, A. et al. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* **47**, D766–D773 (2019).
22. Ng, P. C. & Henikoff, S. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res.* **31**, 3812–3814 (2003).
23. Adzhubei, I. A. et al. A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010).
24. Siepel, A. et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15**, 1034–1050 (2005).
25. Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res.* **20**, 110–121 (2010).
26. Dunham, I. et al. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
27. Kircher, M. et al. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
28. Tang, H. & Thomas, P. D. Tools for predicting the functional impact of nonsynonymous genetic variation. *Genetics* **203**, 635–647 (2016).
29. Lee, P. H. et al. Principles and methods of in-silico prioritization of non-coding regulatory variants. *Hum. Genet.* **137**, 15–30 (2018).
30. Kellis, M. et al. Defining functional DNA elements in the human genome. *Proc. Natl Acad. Sci. USA* **111**, 6131–6138 (2014).
31. Zuk, O. et al. Searching for missing heritability: designing rare variant association studies. *Proc. Natl Acad. Sci. USA* **111**, E455–E464 (2014).
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).
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).
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).
35. Breslow, N. E. & Clayton, D. G. 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).
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).
38. Gogarten, S. M. et al. Genetic association testing using the GENESIS R/Bioconductor package. *Bioinformatics* **35**, 5346–5348 (2019).
39. Kundaje, A. et al. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
40. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* **47**, D886–D894 (2019).
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).
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 (2020).
43. Schaffner, S. F. et al. Calibrating a coalescent simulation of human genome sequence variation. *Genome Res.* **15**, 1576–1583 (2005).
44. Natarajan, P. et al. Deep-coverage whole genome sequences and blood lipids among 16,324 individuals. *Nat. Commun.* **9**, 3391 (2018).
45. Taliun, D. et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. Preprint at *bioRxiv* https://doi.org/10.1101/563866 (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).
47. Rogers, M. F. et al. FATHMM-XF: accurate prediction of pathogenic point mutations via extended features. *Bioinformatics* **34**, 511–513 (2018).
48. Forrest, A. R. R. et al. A promoter-level mammalian expression atlas. *Nature* **507**, 462–470 (2014).
49. Andersson, R. et al. An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455–461 (2014).
50. Fishilevich, S. et al. GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database (Oxford)* **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 (2015).
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).
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).
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).
55. Lange, L. A. et al. Whole-exome sequencing identifies rare and low-frequency coding variants associated with LDL cholesterol. *Am. J. Hum. Genet.* **94**, 233–245 (2014).
56. Bomba, L., Walter, K. & Soranzo, N. The impact of rare and low-frequency genetic variants in common disease. *Genome Biol.* **18**, 77 (2017).
57. Ference, B. A., Majeed, F., Penumetcha, R., Flack, J. M. & Brook, R. D. 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).
58. Teslovich, T. M. et al. Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707–713 (2010).
59. Surakka, I. et al. The impact of low-frequency and rare variants on lipid levels. *Nat. Genet.* **47**, 589–597 (2015).
60. Kathiresan, S. et al. Common variants at 30 loci contribute to polygenic dyslipidemia. *Nat. Genet.* **41**, 56–65 (2009).
61. Kamatani, Y. et al. Genome-wide association study of hematological and biochemical traits in a Japanese population. *Nat. Genet.* **42**, 210–215 (2010).
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).
63. Aulchenko, Y. S. et al. Loci influencing lipid levels and coronary heart disease risk in 16 European population cohorts. *Nat. Genet.* **41**, 47–55 (2009).
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).
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).
66. Hoffmann, T. J. et al. A large electronic-health-record-based genome-wide study of serum lipids. *Nat. Genet.* **50**, 401–413 (2018).
67. Willer, C. J. et al. Discovery and refinement of loci associated with lipid levels. *Nat. Genet.* **45**, 1274–1283 (2013).
68. Cohen, J. C. 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).
69. Stitziel, N. O. et al. Inactivating mutations in *NPC1L1* and protection from coronary heart disease. *N. Engl. J. Med.* **371**, 2072–2082 (2014).
70. Cooper, G. M. et al. Single-nucleotide evolutionary constraint scores highlight disease-causing mutations. *Nat. Methods* **7**, 250–251 (2010).
71. Cooper, G. M. & Shendure, J. Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data. *Nat. Rev. Genet.* **12**, 628–640 (2011).
72. Van Hout, C. V. et al. Whole exome sequencing and characterization of coding variation in 49,960 individuals in the UK Biobank. Preprint at *bioRxiv* https://doi.org/10.1101/572347 (2019).
73. Crosby, J. et al. Loss-of-function mutations in *APOC3*, triglycerides, and coronary disease. *N. Engl. J. Med.* **371**, 22–31 (2014).

74. Myers, R. M. et al. A user's guide to the Encyclopedia of DNA Elements (ENCODE). *PLoS Biol.* **9**, e1001046 (2011).
75. Landrum, M. J. et al. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* **46**, D1062–D1067 (2018).
76. Davis, H. R. & Veltri, E. P. Zetia: inhibition of Niemann-Pick C1 Like 1 (NPC1L1) to reduce intestinal cholesterol absorption and treat hyperlipidemia. *J. Atheroscler. Thromb.* **14**, 99–108 (2007).
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).
78. Lu, Q., Powles, R. L., Wang, Q., He, B. J. & 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).
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).
80. Bodea, C. A. et al. PINES: phenotype-informed tissue weighting improves prediction of pathogenic noncoding variants. *Genome Biol.* **19**, 173 (2018).
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).
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 (2018).
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).

[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, 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]College of Public Health, University of Kentucky, 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]Department 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]Stanley Center for Psychiatric Research, Broad Institute of Harvard and MIT, Cambridge, MA, USA. [50]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, 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. [218]These authors contributed equally: Xihao Li, Zilin Li. *Lists of authors and their affiliations appear at the end of the paper. ✉e-mail: xlin@hsph.harvard.edu

NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium

Namiko Abe[58], Gonçalo R. Abecasis[53,54], Francois Aguet[59], Christine Albert[60], Laura Almasy[61], Alvaro Alonso[62], Seth Ament[63], Peter Anderson[64], Pramod Anugu[65], Deborah Applebaum-Bowden[66], Kristin Ardlie[59], Dan Arking[67], Donna K. Arnett[6], Allison Ashley-Koch[68], Stella Aslibekyan[7], Tim Assimes[69], Paul Auer[70], Dimitrios Avramopoulos[67], John Barnard[71], Kathleen Barnes[72],

R. Graham Barr[73], Emily Barron-Casella[67], Lucas Barwick[74], Terri Beaty[67], Gerald Beck[75], Diane Becker[76], Lewis Becker[67], Rebecca Beer[77], Amber Beitelshees[63], Emelia Benjamin[78], Takis Benos[79], Marcos Bezerra[80], Lawrence F. Bielak[9], Joshua Bis[81], Thomas Blackwell[82], John Blangero[10], Eric Boerwinkle[3,11], Donald W. Bowden[12], Russell Bowler[83], Jennifer Brody[64], Ulrich Broeckel[84], Jai G. Broome[13], Karen Bunting[58], Esteban Burchard[85], Carlos Bustamante[86], Erin Buth[87], Brian Cade[88], Jonathan Cardwell[89], Vincent Carey[90], Cara Carty[91], Richard Casaburi[92], James Casella[67], Peter Castaldi[93], Mark Chaffin[94], Christy Chang[63], Yi-Cheng Chang[95], Daniel Chasman[96], Sameer Chavan[89], Bo-Juen Chen[58], Wei-Min Chen[97], Yii-Der Ida Chen[98], Michael Cho[99], Seung Hoan Choi[94], Lee-Ming Chuang[100], Mina Chung[101], Ren-Hua Chung[102], Clary Clish[103], Suzy Comhair[104], Matthew P. Conomos[14], Elaine Cornell[105], Adolfo Correa[15], Carolyn Crandall[92], James Crapo[106], L. Adrienne Cupples[16,17], Joanne E. Curran[10], Jeffrey Curtis[82], Brian Custer[107], Coleen Damcott[63], Dawood Darbar[108], Sayantan Das[82], Sean David[109], Colleen Davis[64], Michelle Daya[89], Mariza de Andrade[110], Lisa de las Fuentes[111], Michael DeBaun[112], Ranjan Deka[113], Dawn DeMeo[99], Scott Devine[63], Qing Duan[114], Ravi Duggirala[115], Jon Peter Durda[105], Susan Dutcher[116], Charles Eaton[117], Lynette Ekunwe[65], Adel El Boueiz[118], Patrick Ellinor[119], Leslie Emery[64], Serpil Erzurum[120], Charles Farber[97], Tasha Fingerlin[121], Matthew Flickinger[82], Myriam Fornage[122], Nora Franceschini[123], Chris Frazar[124], Mao Fu[63], Stephanie M. Fullerton[64], Lucinda Fulton[116], Stacey Gabriel[94], Weiniu Gan[77], Shanshan Gao[125], Yan Gao[65], Margery Gass[126], Bruce Gelb[127], Xiaoqi (Priscilla) Geng[82], Mark Geraci[128], Soren Germer[58], Robert Gerszten[129], Auyon Ghosh[130], Richard Gibbs[131], Chris Gignoux[69], Mark Gladwin[132], David Glahn[133], Stephanie Gogarten[64], Da-Wei Gong[63], Harald Goring[134], Sharon Graw[135], Daniel Grine[125], C. Charles Gu[116], Yue Guan[63], Xiuqing Guo[19], Namrata Gupta[59], Jeff Haessler[126], Michael Hall[65], Daniel Harris[63], Nicola L. Hawley[136], Jiang He[137], Susan Heckbert[64], Ryan Hernandez[138], David Herrington[139], Craig Hersh[140], Bertha Hidalgo[141], James Hixson[122], Brian Hobbs[90], John Hokanson[89], Elliott Hong[63], Karin Hoth[142], Chao (Agnes) Hsiung[143], Yi-Jen Hung[144], Haley Huston[145], Chii Min Hwu[146], Marguerite R. Irvin[7], Rebecca Jackson[147], Deepti Jain[64], Cashell Jaquish[77], Min A. Jhun[82], Jill Johnsen[148], Andrew Johnson[77], Craig Johnson[64], Rich Johnston[62], Kimberly Jones[67], Hyun Min Kang[149], Robert Kaplan[150], Sharon L. R. Kardia[9], Sekar Kathiresan[21,22,23], Shannon Kelly[107], Eimear Kenny[127], Michael Kessler[63], Alyna T. Khan[14], Wonji Kim[151], Greg Kinney[89], Barbara Konkle[152], Charles L. Kooperberg[24], Holly Kramer[153], Christoph Lange[154], Ethan Lange[89], Leslie Lange[89], Cathy C. Laurie[14], Cecelia Laurie[64], Meryl LeBoff[99], Jiwon Lee[90], Seunggeun Shawn Lee[82], Wen-Jane Lee[146], Jonathon LeFaive[82], David Levine[64], Dan Levy[77], Joshua Lewis[63], Xiaohui Li[155], Yun Li[114], Henry Lin[155], Honghuang Lin[156], Keng Han Lin[82], Xihong Lin[1,26,35], Simin Liu[157], Yongmei Liu[158], Yu Liu[159], Ruth J. F. Loos[160], Steven Lubitz[119], Kathryn Lunetta[156], James Luo[77], Michael C. Mahaney[10], Barry Make[67], Ani W. Manichaikul[27], JoAnn Manson[99], Lauren Margolin[94], Lisa W. Martin[28], Susan Mathai[89], Rasika A. Mathias[29], Susanne May[161], Patrick McArdle[63], Merry-Lynn McDonald[141], Sean McFarland[162], Stephen T. McGarvey[30], Daniel McGoldrick[124], Caitlin McHugh[163], Hao Mei[65], Luisa Mestroni[135], Deborah A. Meyers[164], Julie Mikulla[77], Nancy Min[65], Mollie Minear[77], Ryan L. Minster[132], Braxton D. Mitchell[31,32], Matt Moll[165], May E. Montasser[33], Courtney Montgomery[166], Arden Moscati[167], Solomon Musani[168], Stanford Mwasongwe[65], Josyf C. Mychaleckyj[97], Girish Nadkarni[127], Rakhi Naik[67], Take Naseri[169], Pradeep Natarajan[23,35,36], Sergei Nekhai[170], Sarah C. Nelson[87], Bonnie Neltner[125], Deborah Nickerson[64], Kari North[114], Jeffrey R. O'Connell[31], Tim O'Connor[63], Heather Ochs-Balcom[171], David Paik[172], Nicholette D. Palmer[12], James Pankow[173], George Papanicolaou[77], Afshin Parsa[63],

Juan M. Peralta[10], Marco Perez[69], James Perry[63], Ulrike Peters[174], Patricia A. Peyser[9], Lawrence S. Phillips[62], Toni Pollin[63], Wendy Post[175], Julia Powers Becker[176], Meher Preethi Boorgula[89], Michael Preuss[127], Bruce M. Psaty[37,38], Pankaj Qasba[77], Dandi Qiao[99], Zhaohui Qin[62], Nicholas Rafaels[177], Laura Raffield[178], Ramachandran S. Vasan[17,45], D. C. Rao[116], Laura Rasmussen-Torvik[179], Aakrosh Ratan[97], Susan Redline[39,40,41], Robert Reed[63], Elizabeth Regan[106], Alex Reiner[174], Muagututi'a Sefuiva Reupena[180], Kenneth M. Rice[14], Stephen S. Rich[27], Dan Roden[181], Carolina Roselli[94], Jerome I. Rotter[19], Ingo Ruczinski[67], Pamela Russell[89], Sarah Ruuska[182], Kathleen Ryan[63], Ester Cerdeira Sabino[183], Danish Saleheen[184], Shabnam Salimi[63], Steven Salzberg[67], Kevin Sandow[185], Vijay G. Sankaran[186], Christopher Scheller[82], Ellen Schmidt[82], Karen Schwander[116], David Schwartz[89], Frank Sciurba[132], Christine Seidman[187], Jonathan Seidman[188], Vivien Sheehan[189], Stephanie L. Sherman[190], Amol Shetty[63], Aniket Shetty[89], Wayne Hui-Heng Sheu[146], M. Benjamin Shoemaker[191], Brian Silver[192], Edwin Silverman[99], Jennifer A. Smith[9,42], Josh Smith[64], Nicholas Smith[193], Tanja Smith[58], Sylvia Smoller[150], Beverly Snively[194], Michael Snyder[195], Tamar Sofer[99], Nona Sotoodehnia[64], Adrienne M. Stilp[64], Garrett Storm[125], Elizabeth Streeten[63], Jessica Lasky Su[90], Yun Ju Sung[116], Jody Sylvia[99], Adam Szpiro[64], Carole Sztalryd[63], Daniel Taliun[82], Hua Tang[196], Margaret Taub[67], Kent D. Taylor[197], Matthew Taylor[198], Simeon Taylor[63], Marilyn Telen[68], Timothy A. Thornton[64], Machiko Threlkeld[199], Lesley Tinker[126], David Tirschwell[64], Sarah Tishkoff[200], Hemant K. Tiwari[43], Catherine Tong[201], Russell Tracy[202], Michael Y. Tsai[44], Dhananjay Vaidya[67], David Van Den Berg[203], Peter VandeHaar[82], Scott Vrieze[204,205], Tarik Walker[89], Robert Wallace[142], Avram Walts[89], Fei Fei Wang[14], Heming Wang[206], Karol Watson[92], Daniel E. Weeks[46], Bruce Weir[64], Scott Weiss[99], Lu-Chen Weng[119], Jennifer Wessel[207], Cristen J. Willer[55,56,57], Kayleen Williams[64], L. Keoki Williams[208], Carla Wilson[209], James G. Wilson[47,48], Quenna Wong[64], Joseph Wu[210], Huichun Xu[63], Lisa R. Yanek[29], Ivana Yang[89], Rongze Yang[63], Norann Zaghloul[63], Maryam Zekavat[94], Yingze Zhang[211], Snow Xueyan Zhao[106], Wei Zhao[212], Degui Zhi[122], Xiang Zhou[82], Xiaofeng Zhu[213], Michael Zody[58] and Sebastian Zoellner[82]

[58]New York Genome Center, New York, NY, USA. [59]Broad Institute, Cambridge, MA, USA. [60]Brigham and Women's Hospital, Cedars-Sinai, Boston, MA, USA. [61]Children's Hospital of Philadelphia, University of Pennsylvania, Philadelphia, PA, USA. [62]Emory University, Atlanta, GA, USA. [63]University of Maryland, Baltimore, MD, USA. [64]University of Washington, Seattle, WA, USA. [65]University of Mississippi, Jackson, MS, USA. [66]National Institutes of Health, Bethesda, MD, USA. [67]Johns Hopkins University, Baltimore, MD, USA. [68]Duke University, Durham, NC, USA. [69]Stanford University, Stanford, CA, USA. [70]University of Wisconsin Milwaukee, Milwaukee, WI, USA. [71]Cleveland Clinic, Cleveland, OH, USA. [72]University of Colorado Anschutz Medical Campus, Aurora, CO, USA. [73]Columbia University, New York, NY, USA. [74]The Emmes Corporation, Lung Tissue Research Consortium, Rockville, MD, USA. [75]Cleveland Clinic, Quantitative Health Sciences, Cleveland, OH, USA. [76]Johns Hopkins University, School of Medicine, Baltimore, MD, USA. [77]National Heart, Lung, and Blood Institute, National Institutes of Health, Bethesda, MD, USA. [78]Boston University, Massachusetts General Hospital, Boston University School of Medicine, Boston, MA, USA. [79]University of Pittsburgh, Pittsburgh, PA, USA. [80]Fundação de Hematologia e Hemoterapia de Pernambuco - Hemope, Recife, Brazil. [81]Cardiovascular Health Research Unit, Department of Medicine, University of Washington, Seattle, WA, USA. [82]University of Michigan, Ann Arbor, MI, USA. [83]National Jewish Health, Denver, CO, USA. [84]Medical College of Wisconsin, Milwaukee, WI, USA. [85]University of California, San Francisco, San Francisco, CA, USA. [86]Stanford University, Stanford, CA, USA. [87]Biostatistics, University of Washington, Seattle, WA, USA. [88]Brigham and Women's Hospital, Boston, MA, USA. [89]University of Colorado at Denver, Denver, CO, USA. [90]Brigham and Women's Hospital, Boston, MA, USA. [91]Washington State University, Seattle, WA, USA. [92]University of California, Los Angeles, Los Angeles, CA, USA. [93]Medicine, Brigham and Women's Hospital, Boston, MA, USA. [94]Broad Institute, Cambridge, MA, USA. [95]National Taiwan University, Taipei, Taiwan. [96]Division of Preventive Medicine, Brigham and Women's Hospital, Boston, MA, USA. [97]University of Virginia, Charlottesville, VA, USA. [98]Lundquist Institute, Torrance, CA, USA. [99]Brigham and Women's Hospital, Boston, MA, USA. [100]National Taiwan University, National Taiwan University Hospital, Taipei, Taiwan. [101]Cleveland Clinic, Cleveland, OH, USA. [102]National Health Research Institutes, Zhunan, Taiwan. [103]Broad Institute, Metabolomics Platform, Cambridge, MA, USA. [104]Cleveland Clinic, Immunity and Immunology, Cleveland, OH, USA. [105]University of Vermont, Burlington, VT, USA. [106]National Jewish Health, Denver, CO, USA. [107]Vitalant Research Institute, San Francisco, CA, USA. [108]University of Illinois at Chicago, Chicago, IL, USA. [109]University of Chicago, Chicago, IL, USA. [110]Mayo Clinic, Health Sciences Research, Rochester, MN, USA. [111]Department of Medicine, Cardiovascular Division Washington University in St Louis, St. Louis, MO, USA. [112]Vanderbilt University, Nashville, TN, USA. [113]University of Cincinnati, Cincinnati, OH, USA. [114]University of North Carolina, Chapel Hill, NC, USA. [115]University of Texas Rio Grande Valley School of Medicine, Edinburg, TX, USA. [116]Washington University in St Louis, St Louis, MO, USA. [117]Brown University, Providence, RI, USA. [118]Channing Division of Network Medicine, Harvard University, Boston, MA, USA. [119]Massachusetts General Hospital, Boston, MA, USA. [120]Cleveland Clinic, Cleveland, OH, USA. [121]National Jewish Health, Center for Genes, Environment and Health, Denver, CO, USA. [122]University of Texas Health at Houston, Houston, TX, USA. [123]Epidemiology, University of North Carolina, Chapel Hill, NC, USA. [124]University of Washington, Seattle, WA, USA. [125]University of Colorado at Denver, Denver, CO, USA. [126]Fred Hutchinson Cancer Research Center,

Seattle, WA, USA. [127]Icahn School of Medicine at Mount Sinai, New York, NY, USA. [128]Medicine, Indiana University, Indianapolis, IN, USA. [129]Beth Israel Deaconess Medical Center, Boston, MA, USA. [130]Brigham and Women's Hospital, Boston, MA, USA. [131]Baylor College of Medicine Human Genome Sequencing Center, Houston, TX, USA. [132]University of Pittsburgh, Pittsburgh, PA, USA. [133]Yale University, New Haven, CT, USA. [134]University of Texas Rio Grande Valley School of Medicine, San Antonio, TX, USA. [135]University of Colorado Anschutz Medical Campus, Aurora, CO, USA. [136]Department of Chronic Disease Epidemiology, Yale University, New Haven, CT, USA. [137]Tulane University, New Orleans, LA, USA. [138]University of California, San Francisco, San Francisco, CA, USA. [139]Wake Forest Baptist Health, Winston-Salem, NC, USA. [140]Channing Division of Network Medicine, Brigham and Women's Hospital, Boston, MA, USA. [141]University of Alabama, Birmingham, AL, USA. [142]University of Iowa, Iowa City, IA, USA. [143]Institute of Population Health Sciences, National Health Research Institutes, Zhunan, Taiwan. [144]Tri-Service General Hospital National Defense Medical Center, Taipei, Taiwan. [145]Bloodworks Northwest, Seattle, WA, USA. [146]Taichung Veterans General Hospital Taiwan, Taichung City, Taiwan. [147]Ohio State University Wexner Medical Center, Internal Medicine, Division of Endocrinology, Diabetes and Metabolism, Columbus, OH, USA. [148]Bloodworks Northwest, Seattle, WA, USA. [149]Biostatistics, University of Michigan, Ann Arbor, MI, USA. [150]Albert Einstein College of Medicine, New York, NY, USA. [151]Harvard University, Cambridge, MA, USA. [152]Bloodworks Northwest, Seattle, WA, USA. [153]Public Health Sciences, Loyola University, Maywood, IL, USA. [154]Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA. [155]Lundquist Institute, Torrance, CA, USA. [156]Boston University, Boston, MA, USA. [157]Epidemiology and Medicine, Brown University, Providence, RI, USA. [158]Cardiology, Duke University, Durham, NC, USA. [159]Cardiovascular Institute, Stanford University, Palo Alto, CA, USA. [160]Icahn School of Medicine at Mount Sinai, The Charles Bronfman Institute for Personalized Medicine, New York, NY, USA. [161]Biostatistics, University of Washington, Seattle, WA, USA. [162]Harvard University, Cambridge, MA, USA. [163]Biostatistics, University of Washington, Seattle, WA, USA. [164]University of Arizona, Tucson, AZ, USA. [165]Medicine, Brigham and Women's Hospital, Boston, MA, USA. [166]Oklahoma Medical Research Foundation, Genes and Human Disease Research Program, Oklahoma City, OK, USA. [167]Icahn School of Medicine at Mount Sinai, New York, NY, USA. [168]Medicine, University of Mississippi, Jackson, MS, USA. [169]Ministry of Health, Government of Samoa, Apia, Samoa. [170]Howard University, Washington, DC, USA. [171]University at Buffalo, Buffalo, NY, USA. [172]Stanford Cardiovascular Institute, Stanford University, Stanford, CA, USA. [173]University of Minnesota, Minneapolis, MN, USA. [174]Fred Hutchinson Cancer Research Center, University of Washington, Seattle, WA, USA. [175]Cardiology/Medicine, Johns Hopkins University, Baltimore, MD, USA. [176]Medicine, University of Colorado at Denver, Denver, CO, USA. [177]University of Colorado at Denver, Denver, CO, USA. [178]Genetics, University of North Carolina, Chapel Hill, NC, USA. [179]Northwestern University, Chicago, IL, USA. [180]Lutia I Puava Ae Mapu I Fagalele, Apia, Samoa. [181]Medicine, Pharmacology, Biomedical Informatics, Vanderbilt University, Nashville, TN, USA. [182]Bloodworks Northwest, Seattle, WA, USA. [183]Faculdade de Medicina, Universidade de São Paulo, São Paulo, Brazil. [184]Columbia University, New York, NY, USA. [185]Lundquist Institute, Torrance, CA, USA. [186]Division of Hematology/Oncology, Broad Institute, Harvard University, Boston, MA, USA. [187]Genetics, Harvard Medical School, Boston, MA, USA. [188]Harvard Medical School, Boston, MA, USA. [189]Pediatrics, Baylor College of Medicine, Houston, TX, USA. [190]Human Genetics, Emory University, Atlanta, GA, USA. [191]Medicine/Cardiology, Vanderbilt University, Nashville, TN, USA. [192]University of Massachusetts Memorial Medical Center, Worcester, MA, USA. [193]Epidemiology, University of Washington, Seattle, WA, USA. [194]Biostatistical Sciences, Wake Forest Baptist Health, Winston-Salem, NC, USA. [195]Stanford University, Stanford, CA, USA. [196]Genetics, Stanford University, Stanford, CA, USA. [197]Lundquist Institute, Institute for Translational Genomics and Populations Sciences, Torrance, CA, USA. [198]University of Colorado at Denver, Denver, CO, USA. [199]Department of Genome Sciences, University of Washington, Seattle, WA, USA. [200]Genetics, University of Pennsylvania, Philadelphia, PA, USA. [201]Department of Biostatistics, University of Washington, Seattle, WA, USA. [202]Pathology & Laboratory Medicine, University of Vermont, Burlington, VT, USA. [203]Methylation Characterization Center, University of Southern California, Los Angeles, CA, USA. [204]University of Colorado at Boulder, Boulder, CO, USA. [205]University of Minnesota, Minneapolis, MN, USA. [206]Brigham and Women's Hospital, Partners, Boston, MA, USA. [207]Epidemiology, Indiana University, Indianapolis, IN, USA. [208]Henry Ford Health System, Detroit, MI, USA. [209]Brigham and Women's Hospital, Boston, MA, USA. [210]Stanford Cardiovascular Institute, Stanford University, Stanford, CA, USA. [211]Medicine, University of Pittsburgh, Pittsburgh, PA, USA. [212]Department of Epidemiology, University of Michigan, Ann Arbor, MI, USA. [213]Department of Population and Quantitative Health Sciences, Case Western Reserve University, Cleveland, OH, USA.

TOPMed Lipids Working Group

Moustafa Abdalla[94], Gonçalo R. Abecasis[53,54], Donna K. Arnett[6], Stella Aslibekyan[7], Tim Assimes[69], Elizabeth Atkinson[49], Christie M. Ballantyne[8], Amber Beitelshees[63], Lawrence F. Bielak[9], Joshua Bis[81], Corneliu Bodea[99], Eric Boerwinkle[3,11], Donald W. Bowden[12], Jennifer Brody[64], Brian Cade[88], Jenna Carlson[79], I-Shou Chang[102], Yii-Der Ida Chen[98], Sung Chun[99], Ren-Hua Chung[102], Matthew P. Conomos[14], Adolfo Correa[15], L. Adrienne Cupples[16,17], Coleen Damcott[63], Paul de Vries[122], Ron Do[127], Amanda Elliott[94], Mao Fu[63], Andrea Ganna[94], Da-Wei Gong[63], Sarah Graham[55], Mary Haas[94], Bernhard Haring[214], Jiang He[137], Susan Heckbert[64], Blanca Himes[215], James Hixson[122], Marguerite R. Irvin[7], Deepti Jain[64], Gail Jarvik[64], Min A. Jhun[82], Jicai Jiang[63], Goo Jun[122], Rita Kalyani[29], Sharon L. R. Kardia[9], Sekar Kathiresan[21,22,23], Amit Khera[94], Derek Klarin[94], Charles L. Kooperberg[24], Brian Kral[29], Leslie Lange[89], Cathy C. Laurie[14], Cecelia Laurie[64], Rozenn Lemaitre[81], 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[163], John McLenithan[63], Julie Mikulla[77], Braxton D. Mitchell[31,32], May E. Montasser[33], Andrew Moran[184], Alanna C. Morrison[3], Tetsushi Nakao[94], Pradeep Natarajan[23,35,36], Deborah Nickerson[64], Kari North[114], Jeffrey R. O'Connell[31], Christopher O'Donnell[217], Nicholette D. Palmer[12], Akhil Pampana[35,36], Aniruddh Patel[94], Gina M. Peloso[16], James Perry[63], Ulrike Peters[174], Patricia A. Peyser[9], James Pirruccello[94],

Toni Pollin[63], Michael Preuss[127], Bruce M. Psaty[37,38], D. C. Rao[116], Susan Redline[39,40,41], Robert Reed[63], Alex Reiner[174], Stephen S. Rich[27], Samantha Rosenthal[79], Jerome I. Rotter[19], Jenny Schoenberg[126], Margaret Sunitha Selvaraj[35,36], Wayne Hui-Heng Sheu[146], Jennifer A. Smith[9,42], Tamar Sofer[99], Adrienne M. Stilp[64], Shamil R. Sunyaev[35,51,52], Ida Surakka[55], Carole Sztalryd[63], Hua Tang[196], Kent D. Taylor[197], Michael Y. Tsai[44], Md Mesbah Uddin[94], Sarah Urbut[94], Marie Verbanck[127], Ann Von Holle[114], Heming Wang[206], Fei Fei Wang[14], Kerri Wiggins[64], Cristen J. Willer[55,56,57], James G. Wilson[47,48], Brooke Wolford[56], Huichun Xu[63], Lisa R. Yanek[29], Norann Zaghloul[63], Maryam Zekavat[94] and Jingwen Zhang[1]

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

Methods

Notation and model. Suppose there are n individuals 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; $\mathbf{X}_i = (X_{i1}, \ldots, X_{iq})^{\mathrm{T}}$ denotes q covariates, such as age, age^2, sex and ancestral principal components; $\mathbf{G}_i = (G_{i1}, \ldots, G_{ip})^{\mathrm{T}}$ denotes 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):

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

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

When the data consist of related samples, we consider the following GLMM[35–37]: $g(\mu_i) = \alpha_0 + \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\alpha} + \mathbf{G}_i^{\mathrm{T}}\boldsymbol{\beta} + b_i$, where the random effects b_i account for the remaining population structure unaccounted by ancestral principal components, relatedness and other between-observation correlation. We assume that $\mathbf{b} = (b_1, \ldots, b_n)^{\mathrm{T}} \approx N(\mathbf{0}, \sum_{l=1}^{L} \theta_l \boldsymbol{\Phi}_l)$ with variance components θ_l and known covariance matrices $\boldsymbol{\Phi}_l$. The random effects $\mathbf{b}$ can be decomposed into a sum of multiple random effects to account for different sources of relatedness and correlation as $\mathbf{b} = \sum_{l=1}^{L} \mathbf{b}_l$ with $\mathbf{b}_l \approx N(\mathbf{0}, \theta_l\boldsymbol{\Phi}_l)$. For example, $\mathbf{b}_1$ accounts for unaccounted population structure and family relatedness by using the adjusted GRM as its covariance matrix $\boldsymbol{\Phi}_1$ (refs. [84,85]). A sparse GRM can be used to scale up computation[38]. Additional random effects $\mathbf{b}_2$,...,$\mathbf{b}_L$ can be used to account for complex sampling designs, such as correlation between repeated measures from longitudinal studies using individual-specific random intercepts and slopes and hierarchical designs. The remaining variables are defined in the same way as those in the GLM (equation (1)). Under both GLM and GLMM, we are interested in testing the null hypothesis of whether the variant set is associated with the phenotype, adjusting for covariates and relatedness, which corresponds to H_0:$\boldsymbol{\beta} = \mathbf{0}$, that is, $\beta_1 = \beta_2 = \ldots = \beta_p = 0$.

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 $H_0 : \boldsymbol{\beta} = \mathbf{0} \Leftrightarrow \mathbf{H_0} : \beta = \mathbf{0}$ is given by:

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

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 $g(\mu_i) = \alpha_0 + \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\alpha}$ or the null GLMM $g(\mu_i) = \alpha_0 + \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\alpha} + b_i$. Q_{Burden} asymptotically follows a chi-square distribution with 1 d.f. under the null hypothesis; its P value can be obtained analytically while accounting for linkage disequilibrium between variants[3,37].

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

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

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

Furthermore, 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:

$$Q_{\mathrm{ACAT-V}} = \overline{w^2\mathrm{MAF}(1-\mathrm{MAF})}\tan((0.5 - P_0)\pi) + \sum_{j=1}^{p'} w_j^2\mathrm{MAF}_j(1-\mathrm{MAF}_j)\tan((0.5 - P_j)\pi)$$

where p' is the number of variants with an MAC > 10 and P_j is the association P value of the individual variant j corresponding to 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 an 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 extremely rare variants with an MAC ≤ 10. $Q_{\mathrm{ACAT-V}}$ can be well approximated using a Cauchy distribution under the null hypothesis and its P value can be obtained analytically while accounting for linkage disequilibrium 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 MAF for variant j, that is, $w_j = Beta(\mathrm{MAF}_j;a_1,a_2)$ (ref. [3]), where Beta(;a_1,a_2) is the Beta density function with parameters a_1 and a_2. Common choices of the parameters are $a_1 = 1$ and $a_2 = 25$, which upweight rarer variants, or $a_1 = 1$ and $a_2 = 1$, which correspond to equal weights for all variants. In WGS studies, most RVs 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 multifaceted biological impact of a variant while reducing dimensionality, we propose variant annotation principal components as the principal component summary of the functional annotation data by incorporating individual scores extracted from various functional databases[26,27,39–41,88]. We first grouped the individual scores into ten 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 factor (TF), mappability, distance to the TSS/transcription end site (TES) and microRNA (Fig. 2). For each category, we then centered and standardized all individual scores within the category, such that a higher value of each individual score indicates increased functionality of that annotation, and calculated the annotation principal component as the first principal component from the standardized individual scores (Supplementary Table 1). To facilitate better interpretation, these annotation principal components were then transformed into the Phred-scaled scores for each variant across the genome, defined as $-10\times\log_{10}(\mathrm{rank}(-\mathrm{score})/M)$, where M is total number of variants sequenced across the whole genome.

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

STAAR incorporates multiple functional annotations. STAAR constructs the weights by modeling the probability of a variant being causal using its functional annotation information via qualitative annotations (for example, functional categories) and quantitative annotations (for example, annotation principal components and integrative annotations), as well as modeling the effect sizes of causal variants. Specifically, 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 the 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.

Defining $\hat{\pi}_{jk}$ as the estimated probability of jth variant being causal using the kth annotation (k=0,...,K), for example, $\hat{\pi}_{j1}$ measures the estimated probability that the jth variant is causal using epigenetic annotation, annotation principal component-epigenetic. We estimated $\hat{\pi}_{jk}$ using the empirical cumulative distribution function of the kth annotation for variant j using its rank among all variants as:

$$\hat{\pi}_{jk} = \mathrm{ECDF}_k(A_{jk}) = \frac{\mathrm{rank}(A_{jk})}{M}$$

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 numerical value, for example, the kth annotation principal component. The quantitative A_{jk} we consider in this study includes ten annotation principal components (Supplementary Table 1) and existing integrative scores, including CADD[27], LINSIGHT[46] and FATHMM-XF[47]. For a qualitative annotation, we defined $A_{jk} = 1$ 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, for example, 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 MAF. 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, for example, annotation principal component-epigenetic, is given by $Q_{\text{Burden},k} = \left(\sum_{j=1}^{p} \hat{\pi}_{jk} w_j S_j\right)^2$, whose P value is denoted by $P_{\text{Burden},k}$ ($k=0,...,K$). Under the assumption of SKAT, by estimating the probability of jth variant being causal using the kth annotation ($k=0,..., K$), we have $E(\beta_j)=0$ and $\text{Var}\left(\beta_j\right) = \text{Var}\left(c_j\gamma_j\right) = \pi_{jk} w_j^2 \tau_k$. Hence, the SKAT test statistic using the kth variant functional annotation as the weight is given by:

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

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

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

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

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

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

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

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

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

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

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

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

To further aggregate information from different types of tests and different weights, we proposed 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 defined the STAAR-O test statistic as:

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

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

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

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 multifacet functions and MAFs. Specifically, by including the burden test, STAAR-O is powerful when most 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 linkage disequilibrium; 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 principal components 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: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 linkage disequilibrium structure of an African-American population by using the calibration coalescent model[43]. In each simulation replicate, 10 annotations were generated as $A_1,\ldots,A_{10}$ independent and identically distributed $N(0,1)$ for each variant; 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 the $\alpha = 10^{-5}, 10^{-6}$ and 10^{-7} levels. The total sample sizes considered were 2,500, 5,000 and 10,000.

Empirical power simulations. Next, we carried out a 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 $\{k_1, \ldots, k_5\} \subset \{1, \ldots, 10\}$ were randomly sampled for each region. For different regions, the 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 $\delta_0 = \text{logit}(0.0015)$, logit(0.015) and 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:

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

where X_{1i}, X_{2i} and ϵ_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:

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

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 odds ratios (ORs) of the s causal variants.

Under both settings, we modeled the effect sizes of causal variants using $\beta_j = \gamma_j = c_0 |\log_{10} \text{MAF}_j|$. Therefore, the effect size of causal variant was 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 OR = 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 $\alpha = 10^{-7}$. The 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 RV association studies 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-ancestry-related samples[45]. Race/ethnicity was defined using a combination of self-reported race/ethnicity and study recruitment information. The discovery cohorts consisted of 4,580 (37.2%) Black or African-American individuals, 6,266 (50.9%) White, 543 (4.4%) Asian-American and 927 (7.5%) Hispanic/Latino American. The replication cohorts consisted of 3,534 (19.8%) Black or African-American individuals, 11,662 (65.4%) White, 132 (0.7%) Asian-American and 2,494 (14.0%) others. The 'others' category in the replication cohort included many Hispanic/Latino American individuals as well as a cohort of Samoans.

We applied STAAR-O to identify RV sets associated with four quantitative lipid traits (LDL-C, HDL-C, TG and TC) using the TOPMed WGS data. LDL-C and TC were adjusted for the presence of medications as outlined elsewhere[44]. A linear regression model adjusting for age, age^2 and sex was first fitted for each study-race/ethnicity-specific group. In addition, for Old Order Amish, we also adjusted for *APOB* p.R3527Q in the LDL-C and TC analyses and adjusted for *APOC3* p.R19Ter in the 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 fitted a heteroscedastic linear mixed model for the rank-normalized residuals, adjusting for ten ancestral principal components, study-ancestry group indicators and a variance component for an empirically derived kinship matrix plus separate group-specific residual variance components to account for population structure and relatedness. The output of the heteroscedastic linear mixed model was then used to perform the following variant set analyses for RVs (MAF < 1%) by scanning the genome, including gene-centric analysis using five variant categories (putative loss-of-function 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 RV association study analysis was performed using the R package STAAR v.0.9.5.

The annotation principal components provided diverse and complementary information on variant functionality and were incorporated in RV association tests using an omnibus weighting scheme via the proposed STAAR method. We demonstrated using the following example that STAAR boosts the RV association test power by properly upweighting known LDL-associated functional RVs. For example, the association between a 2-kb sliding window located at 55,038,498–55,040,497 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, for example, the known tolerated missense variant rs11591147, within the sliding window through incorporating multiple annotation principal components[59]. Specifically, the annotation principal component-epigenetic, annotation principal component-protein and annotation principal component-mappability Phred scores are greater than 20 (top 1% across the genome) and the annotation principal component-MutationDensity, annotation principal component-TF and CADD Phred scores are greater than 10 (top 10% across the genome) for this variant, highlighting the multidimensional functionality of this variant. The annotation principal component-protein and annotation principal component-mappability-weighted SKAT P values are 6.69×10^{-13} and 3.78×10^{-12}, which are more significant than SKAT ($P=1.12\times10^{-9}$) and the burden test ($P=4.68\times10^{-4}$).

Statistical analysis of LDL-C in the UK Biobank data. We used the UK Biobank WES from the functionally equivalent pipeline. Sample and variant quality control measures were described previously[72,89]. Briefly, samples with mismatch between genetically inferred and reported sex, high rates of heterozygosity or contamination (D statistic > 0.4), low sequence coverage (less than 85% of targeted bases achieving 20× coverage), duplicates and WES variants discordant with genotyping chip were removed. A total of 43,243 individuals with genetically inferred European ancestry were included; 40,519 of them had data on LDL cholesterol. TC 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 TC levels by the Friedewald equation for individuals with TG levels <400 mg dl^{-1}. If LDL cholesterol levels were directly measured, then their values were divided by 0.7 among those reporting lipid-lowering medication use after 1994 or statin use at any time point. Residuals were created after adjusting for age, age^2, sex and the first ten 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 and 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.

Genome build. All genome coordinates are given in the NCBI GRCh38/UCSC hg38 version of the human genome.

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

Data availability

This paper used the TOPMed Freeze 5 WGS data and lipids phenotype data. Genotype and phenotype data are both available in database of Genotypes and Phenotypes. The discovery phase used data from the following four study cohorts (accession numbers provided in parentheses): Framingham Heart Study (phs000974.v1.p1); Genetics of Cardiometabolic Health in the Amish (phs000956.v1.p1); Jackson Heart Study (phs000964.v1.p1); and Multi-Ethnic Study of Atherosclerosis (phs001416.v1.p1). The replication phase used 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, ancestry 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 v.1.4 (https://cadd.gs.washington.edu/download); ANNOVAR dbNSFP v.3.3a (https://annovar.openbioinformatics.org/en/latest/user-guide/download); LINSIGHT (https://github.com/CshlSiepelLab/LINSIGHT); FATHMM-XF (http://fathmm.biocompute.org.uk/fathmm-xf); FANTOM5 CAGE (https://fantom.gsc.riken.jp/5/data); GeneCards (https://www.genecards.org; v.4.7 for hg38); and Umap/Bismap (https://bismap.hoffmanlab.org; 'before March 2020' version). In addition, recombination rate and nucleotide diversity were obtained from Gazal et al[90]. The whole-genome individual functional annotation data assembled from a variety of sources and the computed annotation principal components are available at the Functional Annotation of Variant–Online Resource (FAVOR) site (http://favor.genohub.org). The tissue-specific functional annotations were downloaded from ENCODE (https://www.encodeproject.org/report/?type=Experiment).

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.

References

84. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
85. Conomos, M. P., Reiner, A. P., Weir, B. S. & Thornton, T. A. Model-free estimation of recent genetic relatedness. *Am. J. Hum. Genet.* **98**, 127–148 (2016).
86. Dey, R., Schmidt, E. M., Abecasis, G. R. & 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).
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).
88. Karimzadeh, M., Ernst, C., Kundaje, A. & Hoffman, M. M. Umap and Bismap: quantifying genome and methylome mappability. *Nucleic Acids Res.* **46**, e120 (2018).
89. Regier, A. A. et al. Functional equivalence of genome sequencing analysis pipelines enables harmonized variant calling across human genetics projects. *Nat. Commun.* **9**, 4038 (2018).
90. Gazal, S. et al. Linkage disequilibrium-dependent architecture of human complex traits shows action of negative selection. *Nat. Genet.* **49**, 1421–1427 (2017).

Acknowledgements

This work was supported by grant nos. R35-CA197449, P01-CA134294, U19-CA203654 and R01-HL113338 (to X. Lin), U01-HG009088 (to X. Lin, S.R.S. and B.M.N.), R01-HL142711 (to P.N. and G.M.P.), K01-HL125751 and R03-HL141439 (to G.M.P.), R35-HL135824 (to 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 (to J.I.R. and X.G.), HHSN268201800002I (to G.R.A.), R35-GM127131 and R01-MH101244 (to 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 (to D.E.W.), NO1-HC-25195, HHSN268201500001I, 75N92019D00031 and R01-HL092577-06S1 (to 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 (to R.S.V.), HHSN268201800001I (to K.M.R., A.T.K., M.P.C. and J.G.B.), U01-HL137162 (to K.M.R. and M.P.C.), R35-HL135818 and R01-HL113338 (to S.R.), R01-HL113323, U01-DK085524, R01-HL045522, R01-MH078143, R01-MH078111 and R01-MH083824 (to J.M.P., M.C.M., J.E.C. and J.B.), R01-HL92301, R01-HL67348, R01-NS058700, R01-AR48797 and R01-AG058921 (to N.D.P. and D.W.B.), R01-DK071891 (to 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 (to J.R.O.), R01-HL093093 (to S.T.M.), 1U24CA237617 and 5U24HG009446 (to X.S.L.), HHSN268201600018C, HHSN268201600001C, HHSN268201600002C, HHSN268201600003C and HHSN268201600004C (to C.L.K.), U01-HL072524, R01-HL104135-04S1, U01-HL054472, U01-HL054473, U01-HL054495, U01-HL054509

and R01-HL055673-18S1 (to M.R.I., S.A. and D.K.A.), Swedish Research Council grant no. 201606830 (to G.H.), grant nos. HHSN268201800010I, HHSN268201800011I, HHSN268201800012I, HHSN268201800013I, HHSN268201800014I and HHSN268201800015I (to A.C.), HHSN268201700001I, HHSN268201700002I, HHSN268201700003I, HHSN268201700005I and HHSN268201700004I (to E.B.), and R01-HL134320 (to C.M.B.). WGS for the TOPMed program was supported by the NHLBI. Centralized read mapping and genotype calling, along with variant quality metrics and filtering were provided by the TOPMed Informatics Research Center (no. 3R01HL-117626-02S1; contract no. HHSN268201800002I). Phenotype harmonization, data management, sample identity quality control and general study coordination were provided by the TOPMed Data Coordinating Center (no. 3R01HL-120393-02S1; contract no. HHSN268201800001I). We gratefully acknowledge the studies and participants who provided biological samples and data for TOPMed. The full study-specific acknowledgements are detailed in the Supplementary Note.

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 the data. G.M.P., P.N. and the 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 it according to suggestions by the coauthors. All authors critically reviewed the manuscript, suggested revisions as needed and approved the final version.

Competing interests

S.A. reports equity and employment by 23andMe. L.A.C. spends part of her time consulting for the Dyslipidemia Foundation, a nonprofit company, as a statistical consultant. X.S.L. is cofounder, board member and scientific advisory board of GV20 Oncotherapy, board member of the scientific advisory board of 3DMedCare, consultant of Genentech and is a recipient of research grants from Sanofi and Takeda, all unrelated to the present work. For The Amish Research Program receives partial support from Regeneron Pharmaceuticals for B.D.M. M.E.M reports a grant from Regeneron Pharmaceuticals that is unrelated to the present work. B.M.P. serves on the steering committee of the Yale Open Data Access Project funded by Johnson & Johnson. S.R. reports interests in Jazz Pharmaceuticals, Eisai and Respircardia, all 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, and is a consultant for CAMP4 Therapeutics, Takeda and Biogen. S.R.S. is a consultant to NGM Biopharmaceuticals and Inari Agriculture. He is also on the 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 to AbbVie Pharmaceuticals.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-020-0676-4.

Correspondence and requests for materials should be addressed to X.L.

nature research

Corresponding author(s): Xihong Lin

Last updated by author(s): May 14, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Software was used for downloading the data as follows: Wget (https://www.gnu.org/software/wget/wget.html) and ANNOVAR (annotate_variation.pl http://annovar.openbioinformatics.org).
Data analysis	Data analysis was performed in R (3.5.1). STAAR v0.9.5 was used in both simulation and real data analysis and is implemented as an open-source R package available at https://github.com/xihaoli/STAAR.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on 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 studies, where the accession numbers are provided in parenthesis: Framingham Heart Study (phs000974.v1.p1), Old Order Amish (phs000956.v1.p1), Jackson Heart Study (phs000964.v1.p1), Multi-Ethnic Study of Atherosclerosis (phs001416.v1.p1). The replication phase used the data from the following ten studies: 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 are given in Supplementary Table 3.

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

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, S. et al. Linkage disequilibrium–dependent architecture of human complex traits shows action of negative selection. Nature genetics 49, 1421 (2017). The tissue-specific functional annotations were downloaded from ENCODE (https://www.encodeproject.org/report/?type=Experiment).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	The discovery phase consists of all four study cohorts of TOPMed Freeze 3 and had 12,316 samples with lipid traits. The replication phase consists of all ten independent study cohorts in TOPMed Freeze 5 that were not in Freeze 3 and had 17,822 samples with lipid traits. The UK Biobank whole exome sequencing data had 40,519 samples with lipid traits.
Data exclusions	For TOPMed data, failed variants and variants with sequencing depth <= 10 were excluded in the QC procedure. For UK Biobank data, 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 whole exome sequencing variants discordant with genotyping chip were removed. All data exclusions criteria were pre-established.
Replication	The significant associations between lipids and rare variants identified in gene-centric functional category and sliding window analyses using samples in the discovery phase were tested using the replication samples. The significant association between disruptive missense rare variants in NPC1L1 and LDL-C was also tested using UK Biobank data. All attempts at replication were successful. Experimental replication was not attempted.
Randomization	This is a method and data analysis manuscript, where the methods were applied to analyze large whole genome sequencing genetic epidemiological studies of the TOPMed program. No randomization was used in the study design.
Blinding	Not relevant, as this is not a clinical trial. De-identified coded data were used for analysis.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	The TOPMed data consist of ancestrally diverse and multi-ethnic related samples. 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. Race/ethnicity was defined using a combination of self-reported race/ethnicity from study recruitment information.

Recruitment	The TOPMed Freeze 5 lipids data included whole genome sequencing data of 30,138 samples from multiple existing NHLBI deep phenotyped study cohorts. The study participants of the TOPMed data have diverse ethnicities. The sample sizes, ethnicity and phenotype summary statistics can be found in Supplementary Table 3. Detailed information of participant recruitment of each study cohort can be found in Supplementary Note. More details can be found at https://www.nhlbiwgs.org.
Ethics oversight	The study protocol was approved by the NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium.

Note that full information on the approval of the study protocol must also be provided in the manuscript.