nature computational science

Article

https://doi.org/10.1038/s43588-024-00764-8

A statistical framework for multi-trait rare variant analysis in large-scale whole-genome sequencing studies

Received: 12 November 2023

Accepted: 20 December 2024

Published online: 7 February 2025

Check for updates

A list of authors and their affiliations appears at the end of the paper

Large-scale whole-genome sequencing (WGS) studies have improved our understanding of the contributions of coding and noncoding rare variants to complex human traits. Leveraging association effect sizes across multiple traits in WGS rare variant association analysis can improve statistical power over single-trait analysis, and also detect pleiotropic genes and regions. Existing multi-trait methods have limited ability to perform rare variant analysis of large-scale WGS data. We propose MultiSTAAR, a statistical framework and computationally scalable analytical pipeline for functionally informed multi-trait rare variant analysis in large-scale WGS studies. MultiSTAAR accounts for relatedness, population structure and correlation among phenotypes by jointly analyzing multiple traits, and further empowers rare variant association analysis by incorporating multiple functional annotations. We applied MultiSTAAR to jointly analyze three lipid traits in 61,838 multi-ethnic samples from the Trans-Omics for Precision Medicine (TOPMed) Program. We discovered and replicated new associations with lipid traits missed by single-trait analysis.

Advances in next-generation sequencing technologies and the decreasing cost of whole-exome/whole-genome sequencing (WES/WGS) have made it possible to study the genetic underpinnings of rare variants (that is, minor allele frequency (MAF) < 1%) in complex human traits. Large nationwide consortia and biobanks, such as the Trans-Omics for Precision Medicine (TOPMed) Program[1] of the National Heart, Lung and Blood Institute (NHLBI), the National Human Genome Research Institute's Genome Sequencing Program (GSP), the National Institute of Health's All of Us Research Program[2] and the UK's Biobank WGS Program[3], are expected to sequence more than a million individuals in total, at more than one billion genetic variants in both coding and noncoding regions of the human genome, while also recording thousands of phenotypes. To mitigate the lack of power of single-variant analyses to identify rare variant associations[4], variant set tests have been proposed to analyze the joint effects of multiple rare variants[5–9], with most of the work focusing on single trait analysis.

Pleiotropy occurs when genetic variants influence multiple traits[10]. There is growing empirical evidence from genome-wide association studies (GWASs) that many variants have pleiotropic effects[11,12]. Identifying these effects can provide valuable insights into the genetic architecture of complex traits[13]. As such, it is of increasing interest to identify pleiotropic rare variants by jointly analyzing multiple traits in WGS rare variant association studies (RVASs).

Several existing methods for multi-trait rare variant association analysis, including MSKAT[14], Multi-SKAT[15] and MTAR[16], have shown that leveraging the cross-phenotype correlation structure can improve the power of multi-trait analyses compared to single-trait analyses when analyzing pleiotropic genes[14–17]. However, existing methods do not scale well, and are not feasible when analyzing large-scale WGS studies with hundreds of millions of rare variants in samples exhibiting relatedness and population structure. Furthermore, none of the existing multi-trait rare variant analysis methods leverage functional annotations that predict the biological functionality of variants, resulting in limited interpretability and power loss. Although the STAAR method[18] dynamically incorporates multiple variant functional annotations to maximize the power of rare variant

e-mail: li@hsph.harvard.edu; zl2509@cumc.columbia.edu; xlin@hsph.harvard.edu

https://doi.org/10.1038/s43588-024-00764-8

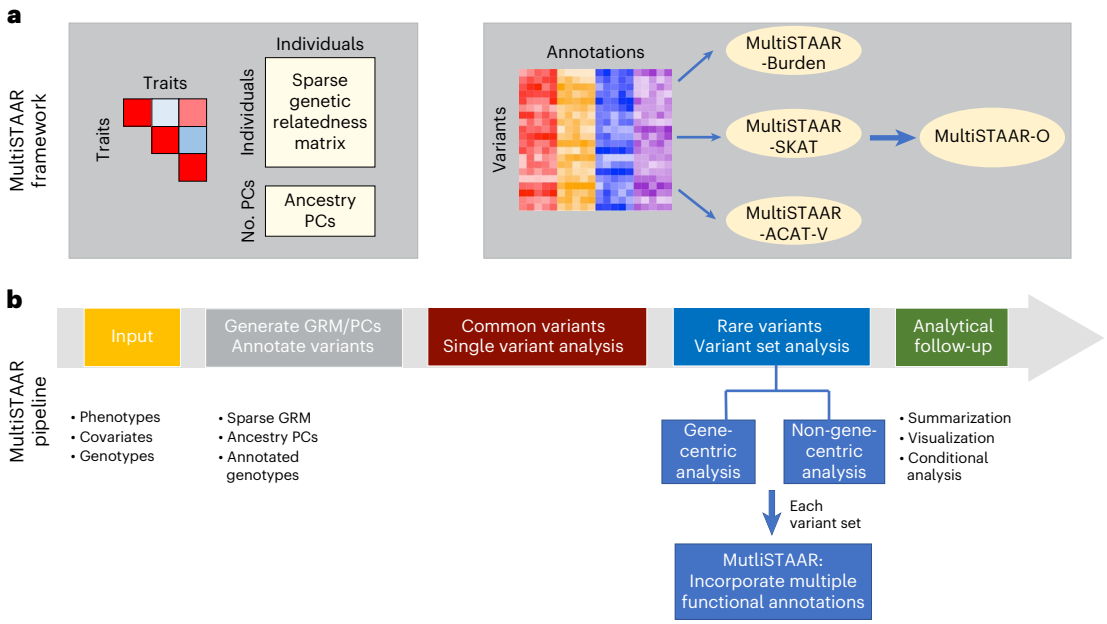


Fig. 1 | MultiSTAAR framework and pipeline. a, MultiSTAAR framework: (i) fit null MLMMs using sparse GRM and ancestry PCs to account for population structure, relatedness and the correlation between phenotypes; (ii) test for associations between each variant set and multiple traits by dynamically incorporating multiple variant functional annotations. **b**, MultiSTAAR pipeline: (i) prepare the input data of MultiSTAAR, including genotypes, multiple phenotypes and covariates; (ii) calculate sparse GRM, ancestry PCs and annotate all variants in the genome; (iii) perform single-variant analysis for common and low-frequency variants; (iv) define the rare variant analysis units, including gene-centric analysis of five coding functional categories and eight noncoding functional categories and non-gene-centric analysis of sliding windows; (v) provide result summarization and perform analytical follow-up via conditional analysis.

association tests, it is designed for single-trait analysis and cannot be directly applied to multiple traits.

To overcome these limitations, we propose the 'Multi-trait variant-Set Test for Association using Annotation infoRmation' (MultiSTAAR), a statistical framework for multi-trait rare variant analyses of large-scale WGS studies and biobanks. It has several features. First, by fitting a null multivariate linear mixed model (MLMM)[19] for multiple quantitative traits simultaneously, adjusting for ancestry principal components (PCs)[20] and using a sparse genetic relatedness matrix (GRM)[21,22], MultiSTAAR scales well but also accounts for relatedness and population structure, as well as correlations among the multiple traits. Second, MultiSTAAR enables the incorporation of multiple variant functional annotations as weights to improve the power of RVASs. Furthermore, we provide MultiSTAAR via a comprehensive pipeline for large-scale WGS studies that facilitates functionally informed multi-trait analysis of both coding and noncoding rare variants. Third, MultiSTAAR enables conditional multi-trait analysis to assess rare variant association signals beyond known common and low-frequency variants.

In the current study we conducted extensive simulation studies to demonstrate the validity of MultiSTAAR and to assess the power gain of MultiSTAAR by incorporating multiple relevant variant functional annotations, and its ability to preserve type I error rates. We then applied MultiSTAAR to perform WGS RVAS of 61,838 ancestrally diverse participants from NHLBI's TOPMed consortium by jointly analyzing three circulating lipid traits.

Results

Overview of the methods

MultiSTAAR is a statistical framework and an analytic pipeline for jointly analyzing multiple traits in large-scale WGS RVASs. There are two main components in the MultiSTAAR framework: (1) fitting null MLMMs using ancestry PCs and sparse GRMs to account for population structure, relatedness and the correlation between phenotypes and (2) testing for associations between each aggregated variant set and multiple traits by dynamically incorporating multiple variant functional annotations[18] (Fig. 1a). Specifically, MultiSTAAR utilizes annotation PCs to capture and prioritize the multidimensional biological functions of variants. MultiSTAAR then integrates these annotation PCs with other integrative functional scores and minor allele frequencies within the MultiSTAAR test statistics using an omnibus weighting scheme.

In WGS RVASs, an important but often underemphasized challenge is selecting biologically meaningful and functionally interpretable analysis units, especially for the noncoding genome[23,24]. In gene-centric analyses of multiple traits, MultiSTAAR provides five functional categories (masks) to aggregate coding rare variants of each protein-coding gene, as well as an additional eight masks of regulatory regions to aggregate noncoding rare variants. In non-gene-centric analyses of multiple traits, MultiSTAAR performs agnostic genetic-region analyses using sliding windows[18,25] (Fig. 1b).

For each rare variant set analyzed, MultiSTAAR first constructs the multi-trait burden, SKAT and ACAT-V test statistics (Methods). For each type of rare variant test, MultiSTAAR calculates multiple candidate *P* values using different variant functional annotations as weights, following the STAAR framework[18]. MultiSTAAR then aggregates the association strength by combining the *P* values from all annotations using the ACAT method, which provides robustness to correlation between tests[9], to obtain the functionally informed multi-trait burden (MultiSTAAR-B), SKAT (MultiSTAAR-S) and ACAT-V (MultiSTAAR-A) tests, and proposes an omnibus test, MultiSTAAR-O, which leverages the advantages of the different types of test using the ACAT method (Fig. 1a and Methods). Furthermore, MultiSTAAR can test multi-trait rare variants' associations conditional on a set of known associations (Fig. 1b).

Simulation studies

To evaluate the type I error rates and the power of MultiSTAAR, we performed simulation studies under several configurations. Following the steps described in Data Simulation (Methods), we generated three quantitative traits with a correlation matrix similar to the empirical correlation in the three lipid traits[26–28]. We then generated genotypes

by simulating 20,000 sequences for 100 different 1-megabase (Mb) regions, each of which was generated to mimic the linkage disequilibrium structure of an African American population by using the calibration coalescent model[29]. Throughout the simulation studies, we randomly and uniformly selected 5-kilobase (kb) regions from these 1-Mb regions and considered sample sizes of 10,000 for each replicate. The simulation studies focused on aggregating uncommon variants with MAF < 5%.

Type I error rate evaluations

We performed 10^8 simulations to evaluate the type I error rates of the multi-trait burden, SKAT, ACAT-V and MultiSTAAR-O tests at $\alpha = 10^{-4}$, 10^{-5} and 10^{-6} (Supplementary Table 1). The results show that, for multi-trait rare variant analysis, all four MultiSTAAR tests controlled the type I error rates at very close to nominal α levels.

Empirical power simulations

We next assessed the power of MultiSTAAR-O for the analysis of multiple phenotypes under different genetic architectures, while also comparing its power with existing methods. Specifically, we considered four models, in which variants in the signal region (variant–phenotype association regions) were associated with (1) one phenotype only, (2) two positively correlated phenotypes, (3) two negatively correlated phenotypes and (4) all three phenotypes. In addition, we considered different proportions (5%, 15% and 35% on average) of causal variants in the signal region, where the causality of variants depended on different sets of annotations, and the effect size directions of causal variants were allowed to vary (Methods). Power was evaluated as the proportions of P values less than $\alpha = 10^{-7}$ based on 10^4 simulations. Overall, MultiSTAAR-O consistently delivered higher power to detect signal regions compared to multi-trait burden, SKAT and ACAT-V tests, through its incorporation of multiple annotations (Supplementary Figs. 1–32). This power advantage was also robust to the existence of non-informative annotations.

Application to the TOPMed lipids WGS data

We applied MultiSTAAR to identify rare variant associations with three quantitative lipid traits (low-density lipoprotein cholesterol (LDL-C), high-density lipoprotein cholesterol (HDL-C) and trigylcerides (TG)) through a multi-trait analysis using TOPMed Freeze 8 WGS data, comprising 61,838 individuals from 20 multi-ethnic studies (Supplementary Note). LDL-C values were adjusted for the usage of lipid-lowering medication[26,30] (Methods), and DNA samples were sequenced at more than 30× target coverage. Sample- and variant-level quality control (QC) steps were performed for each participating study[1,26,30].

Race/ethnicity was measured using a combination of self-reported race/ethnicity and study recruitment information[31] (Supplementary Note). Of the 61,838 samples, 15,636 (25.3%) were Black or African American, 27,439 (44.4%) were White, 4,461 (7.2%) were Asian or Asian American, 13,138 (21.2%) were Hispanic/Latino American and 1,164 (1.9%) were Samoans. There were 414 million single-nucleotide variants (SNVs) observed overall, with 6.5 million (1.6%) common variants (MAF > 5%), 5.2 million (1.2%) low-frequency variants (1% ≤ MAF ≤ 5%) and 402 million (97.2%) rare variants (MAF < 1%). The study-specific demographics and baseline characteristics are provided in Supplementary Table 2.

Gene-centric multi-trait analysis of rare variants

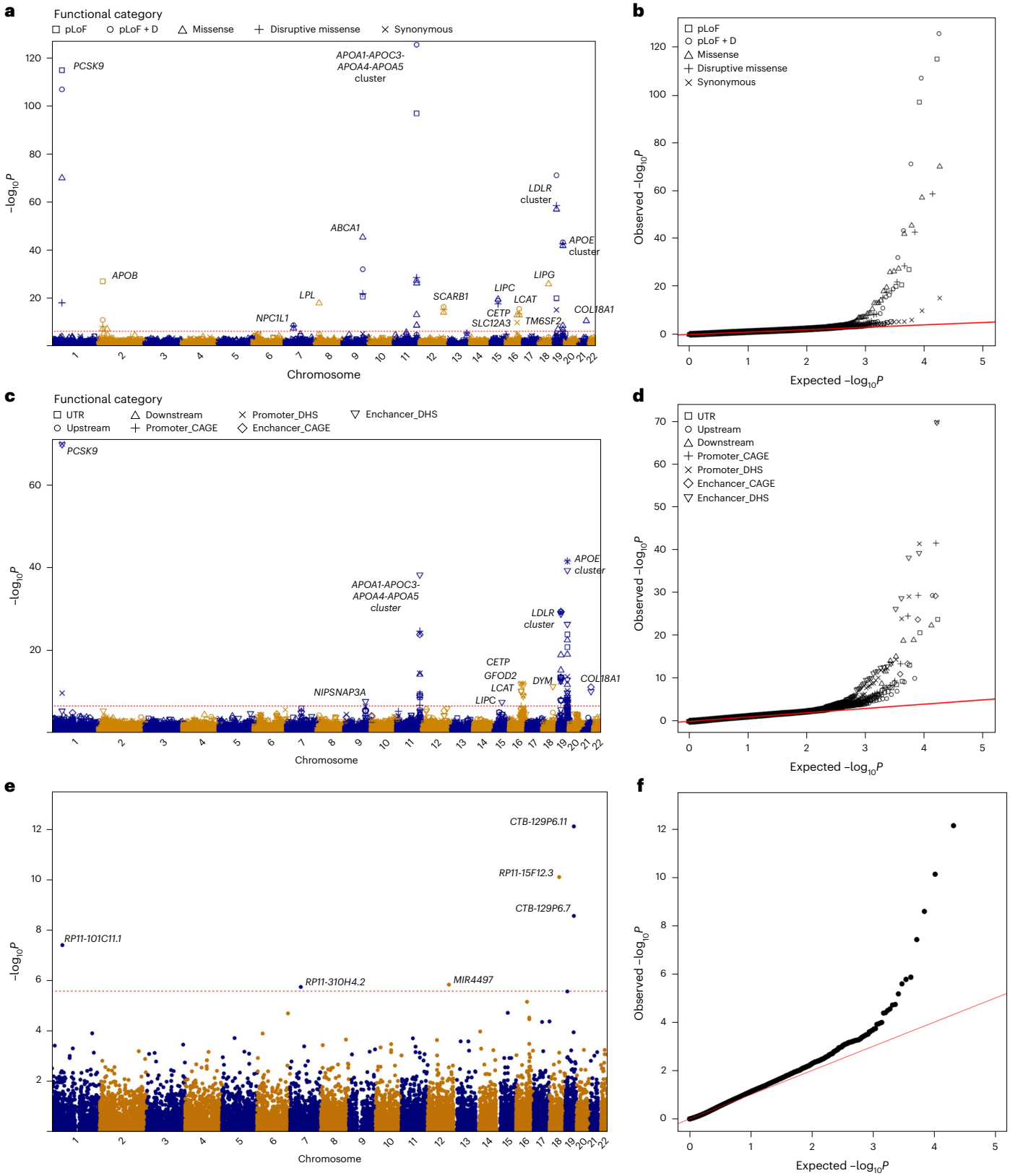
We applied MultiSTAAR-O on the gene-centric multi-trait analysis of coding and noncoding rare variants of genes with lipid traits in TOPMed. For coding variants, rare variants (MAF < 1%) from five coding functional categories (masks) were aggregated, separately, and analyzed using a joint model for LDL-C, HDL-C and TG, including (1) putative loss-of-function (stop gain, stop loss and splice) rare variants, (2) missense rare variants, (3) disruptive missense rare variants, (4) putative loss-of-function and disruptive missense rare variants and (5) synonymous rare variants of each protein-coding gene. The putative loss-of-function, missense and synonymous rare variants were defined by GENCODE variant effect predictor (VEP) categories[32]. The disruptive missense variants were further defined by MetaSVM[33], which measures the deleteriousness of missense mutations. We incorporated nine annotation principal components (aPCs)[18,24,26], CADD[34], LINSIGHT[35], FATHMM-XF[36] and MetaSVM[33] (for missense rare variants only) along with the two MAF-based weights[4] in MultiSTAAR-O (Supplementary Table 3). The overall distribution of MultiSTAAR-O P values was well calibrated for the multi-trait analysis of coding rare variants (Fig. 2b). At a Bonferroni-corrected significance threshold of $\alpha = 0.05/(20{,}000 \times 5) = 5.00 \times 10^{-7}$, accounting for five different coding masks across protein-coding genes, MultiSTAAR-O identified 51 genome-wide significant associations using unconditional multi-trait analysis (Fig. 2a and Supplementary Table 4). After conditioning on previously reported variants associated with LDL-C, HDL-C or TG located within a 1-Mb broader region of each coding mask in the GWAS Catalog and Million Veteran Program (MVP)[26,37,38], 34 out of the 51 associations remained significant at the Bonferroni-corrected threshold of $\alpha = 0.05/51 = 9.80 \times 10^{-4}$ (Supplementary Table 5). We then performed replication analyses of these 34 conditionally significant associations using the UK Biobank WGS data of 170,104 individuals (Methods), and 32 were replicated with a conditional $P < 9.80 \times 10^{-4}$ in UK Biobank (Supplementary Table 5).

For noncoding variants, rare variants from eight noncoding masks were analyzed in a similar fashion: (1) promoter rare variants overlaid with cap analysis of gene expression (CAGE) sites[39], (2) promoter rare variants overlaid with DNase hypersensitivity (DHS) sites[40], (3) enhancer rare variants overlaid with CAGE sites[41,42], (4) enhancer rare variants overlaid with DHS sites[40,42], (5) untranslated region (UTR) rare

Fig. 2 | Manhattan plots and Q–Q plots for unconditional gene-centric coding, noncoding and ncRNA multi-trait analysis of LDL-C, HDL-C and TG using TOPMed data (n = 61,838). a, Manhattan plots for unconditional gene-centric coding analysis of protein-coding genes. The horizontal red dotted line indicates a genome-wide MultiSTAAR-O P value threshold of 5.00×10^{-7}. The significant threshold is defined by multiple comparisons using the Bonferroni correction ($0.05/(20{,}000 \times 5) = 5.00 \times 10^{-7}$). Different symbols represent the MultiSTAAR-O P value of the protein-coding gene using different functional categories (putative loss-of-function (pLoF), putative loss-of-function and disruptive missense (pLoF + D), missense, disruptive missense, synonymous). **b**, Q–Q plots for unconditional gene-centric coding analysis of protein-coding genes. Different symbols represent the MultiSTAAR-O P value of the gene using different functional categories. The red solid line is a 45° reference line. **c**, Manhattan plots for unconditional gene-centric noncoding analysis of protein-coding genes. The horizontal red dotted line indicates a genome-wide MultiSTAAR-O P value threshold of 3.57×10^{-7}. The significant threshold is defined by multiple comparisons using the Bonferroni correction ($0.05/20{,}000 \times 7) = 3.57 \times 10^{-7}$). Different symbols represent the MultiSTAAR-O P value of the protein-coding gene using different functional categories (upstream, downstream, UTR, promoter_CAGE, promoter_DHS, enhancer_CAGE, enhancer_DHS). Promoter_CAGE and promoter_DHS are the promoters with overlap of CAGE sites and DHS sites for a given gene, respectively. Enhancer_CAGE and enhancer_DHS are the enhancers in GeneHancer-predicted regions with the overlap of CAGE sites and DHS sites for a given gene, respectively. **d**, Q–Q plots for unconditional gene-centric noncoding analysis of protein-coding genes. Different symbols represent the MultiSTAAR-O P value of the gene using different functional categories. **e**, Manhattan plot for unconditional gene-centric noncoding analysis of ncRNA genes. The horizontal line indicates a genome-wide MultiSTAAR-O P value threshold of 2.50×10^{-6}. The significant threshold is defined by multiple comparisons using the Bonferroni correction ($0.05/20{,}000 = 2.50 \times 10^{-6}$). **f**, Q–Q plot for unconditional gene-centric noncoding analysis of ncRNA genes. In **a**, **c** and **e**, the chromosome numbers are indicated by the colors of dots. In all panels, MultiSTAAR-O is a two-sided test.

https://doi.org/10.1038/s43588-024-00764-8

variants, (6) upstream region rare variants, (7) downstream region rare variants of each protein-coding gene and (8) rare variants in noncoding RNA (ncRNA) genes[24]. The promoter rare variants were defined as rare variants in the ±3-kb window of transcription start sites with the overlap of CAGE sites or DHS sites. The enhancer rare variants were defined as rare variants in GeneHancer-predicted regions with the overlap of CAGE sites or DHS sites. The UTR, upstream, downstream and ncRNA rare variants were defined by GENCODE VEP categories[32]. With a well-calibrated overall distribution of MultiSTAAR-O *P* values (Fig. 2d) and at a Bonferroni-corrected significance threshold of $\alpha = 0.05/(20{,}000 \times 7) = 3.57 \times 10^{-7}$, accounting for seven different noncoding masks across protein-coding genes, MultiSTAAR-O



https://doi.org/10.1038/s43588-024-00764-8

Table 1 | TOPMed gene-centric noncoding multi-trait analysis results of both unconditional analysis and analysis conditional on known lipids-associated variants

Gene	Chr.[a]	Category[b]	Discovery (TOPMed)			Replication (UK Biobank)			Variants[e] (adjusted)
			No. of SNVs[c]	MultiSTAAR-O[d] (unconditional)	MultiSTAAR-O[d] (conditional)	No. of SNVs[c]	MultiSTAAR-O[d] (unconditional)	MultiSTAAR-O[d] (conditional)	
APOA1	11	Promoter (CAGE)	230	2.33×10^{-7}	9.45×10^{-7}	316	8.86×10^{-23}	1.81×10^{-20}	rs509728, rs61905072, rs66505542, rs7102314, rs964184, rs75198898, rs142958146, rs2075291, rs3135506, rs651821, rs45611741, rs662799, rs10750097, rs9804646, rs978880643, rs2070669, rs76353203, rs138326449, rs147210663, rs140621530, rs525028, rs141469619, rs188287950, rs202207736
CETP	16	Promoter (DHS)	411	1.21×10^{-12}	5.75×10^{-4}	533	6.65×10^{-32}	2.24×10^{-4}	rs35571500, rs247617, rs17231506, rs34498052, rs34119551, rs34065661, rs1597000001[f], rs7499892, rs5883, rs289719, rs11860407, rs189866004, rs5880
APOA1	11	Enhancer (CAGE)	642	1.88×10^{-24}	6.23×10^{-4}	872	6.77×10^{-21}	1.21×10^{-18}	rs509728, rs61905072, rs66505542, rs7102314, rs964184, rs75198898, rs142958146, rs2075291, rs3135506, rs651821, rs45611741, rs662799, rs10750097, rs9804646, rs978880643, rs2070669, rs76353203, rs138326449, rs147210663, rs140621530, rs525028, rs141469619, rs188287950, rs202207736
SPC24	19	Enhancer (CAGE)	366	1.33×10^{-8}	4.88×10^{-4}	536	6.73×10^{-13}	2.61×10^{-16}	rs140753491, rs138294113, rs17242353, rs17242843, rs10422256, rs72658860, rs11669576, rs2738447, rs72658867, rs2738464, rs6511728, rs3760782, rs59168178, rs2278426, rs112942459
NIPSNAP3A	9	Enhancer (DHS)	767	2.63×10^{-8}	8.46×10^{-6}	1,031	1.70×10^{-4}	7.13×10^{-6}	rs2150867, rs33918808, rs112853430, rs4149307, rs9282541, rs1883025, rs1800978
LIPC	15	Enhancer (DHS)	3,714	4.26×10^{-8}	1.25×10^{-4}	5,073	1.48×10^{-8}	9.04×10^{-6}	rs1973688, rs1601935, rs2043082, rs10468017, rs1532085, rs436965, rs35980001, rs1800588, rs2070895, rs113298164
RP11-310H4.2	7	ncRNA	154	1.69×10^{-6}	1.69×10^{-6}	NA	NA	NA	NA[g]
MIR4497	12	ncRNA	23	1.37×10^{-6}	1.42×10^{-6}	37	8.48×10^{-1}	8.49×10^{-1}	rs5800864
RP11-15F12.3	18	ncRNA	64	7.53×10^{-11}	7.50×10^{-3}	NA	NA	NA	rs77960347, rs117623631, rs9958734, rs7229562, rs8086351, rs10048323, rs8084172

A total of 61,838 samples from the TOPMed Program were considered in the analysis. Results for the conditionally significant genes (unconditional MultiSTAAR-O $P<3.57\times10^{-7}$ and conditional MultiSTAAR-O $P<6.58\times10^{-4}$ for seven different noncoding masks across protein-coding genes; unconditional MultiSTAAR-O $P<2.50\times10^{-6}$ and conditional MultiSTAAR-O $P<8.33\times10^{-3}$ for ncRNA genes) are presented. MultiSTAAR-O is a two-sided test. NA, not available. [a]Chromosome number. [b]Functional category. [c]Number of rare variants (MAF < 1%) of the particular noncoding functional category in the gene. [d]*P* value. [e]Adjusted variants in the conditional analysis. [f]Samoan-specific missense variant[62]. [g]No variant adjusted in the conditional analysis.

identified 76 genome-wide significant associations using unconditional multi-trait analysis (Fig. 2c and Supplementary Table 6). After conditioning on known lipids-associated variants[26,37,38], six of the 76 associations remained significant at the Bonferroni-corrected threshold of $\alpha = 0.05/76 = 6.58\times10^{-4}$ (Table 1). These included promoter CAGE and enhancer CAGE rare variants in *APOA1*, promoter DHS rare variants in *CETP*, enhancer CAGE rare variants in *SPC24*, and enhancer DHS rare variants in *NIPSNAP3A* and *LIPC*. All of these six conditionally significant associations were replicated with a conditional $P < 6.58\times10^{-4}$ using the UK Biobank WGS data (Table 1 and Methods).

MultiSTAAR-O further identified six genome-wide significant associations using unconditional multi-trait analysis at

https://doi.org/10.1038/s43588-024-00764-8

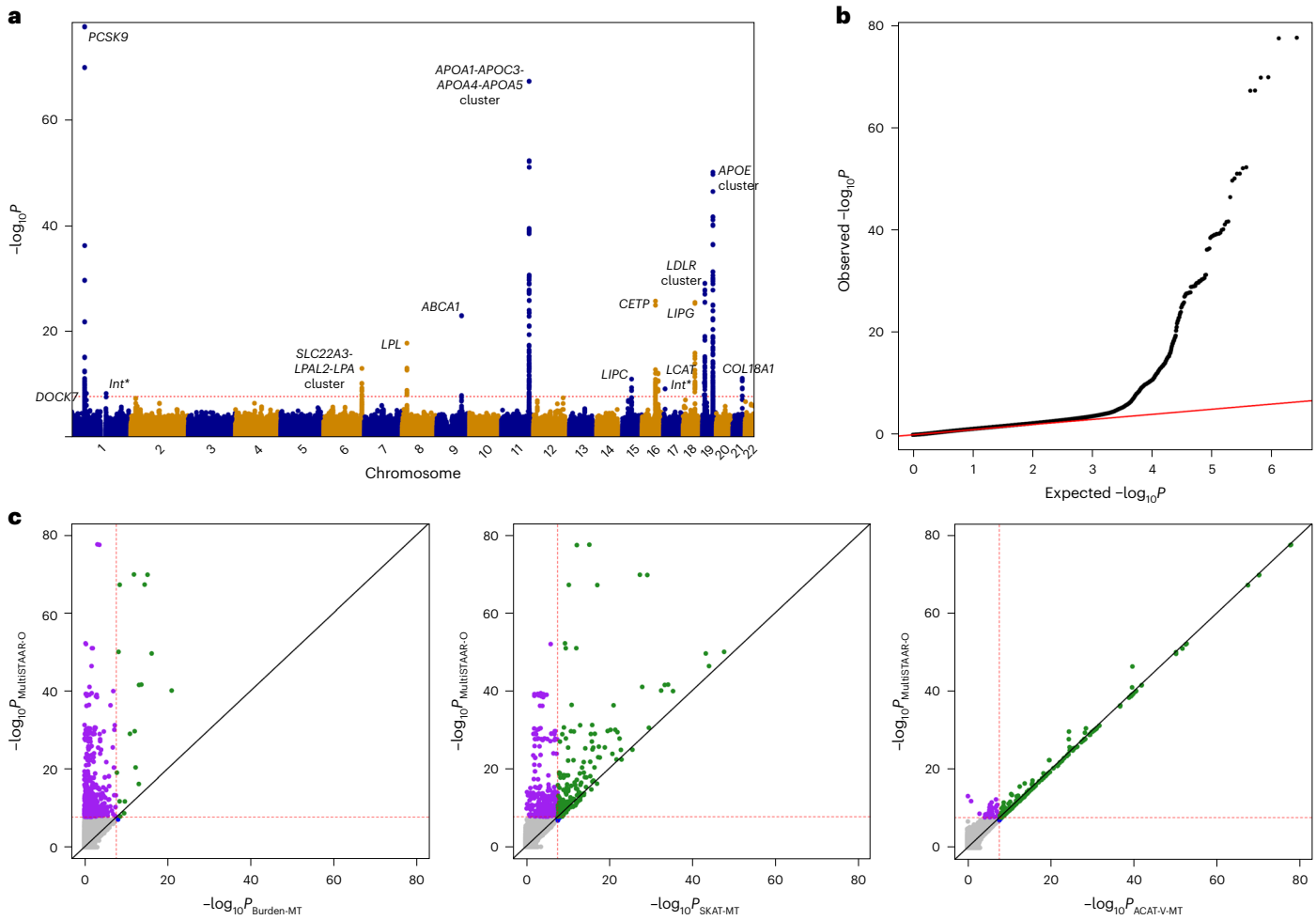


Fig. 3 | TOPMed genetic-region (2-kb sliding window) unconditional multi-trait analysis results for LDL-C, HDL-C and TG using TOPMed data ($n = 61{,}838$). **a**, Manhattan plot showing the associations of 2.65 million 2-kb sliding windows versus $-\log_{10}P$ of MultiSTAAR-O. The horizontal red dotted line indicates a genome-wide P value threshold of 1.89×10^{-8}. **b**, Q–Q plot of 2-kb sliding window MultiSTAAR-O P values. **c**, Scatterplot of P values for the 2-kb sliding windows comparing MultiSTAAR-O with burden-MT, SKAT-MT and ACAT-V-MT tests (MT, multi-trait). Each dot represents a sliding window, with the x-axis label being the $-\log_{10}P$ of the conventional multi-trait test and the y-axis label being the $-\log_{10}(P)$ of MultiSTAAR-O. Burden-MT, SKAT-MT, ACAT-V-MT and MultiSTAAR-O are two-sided tests. Int*, intergenic sliding window.

$\alpha = 0.05/20{,}000 = 2.50 \times 10^{-6}$ accounting for ncRNA genes (Fig. 2e and Supplementary Table 6), with three rare variant associations in *RP11-15F12.3*, *RP11-310H4.2* and *MIR4497* remaining significant at $\alpha = 0.05/6 = 8.33 \times 10^{-3}$ after conditioning on known lipids-associated variants[26,37,38] (Table 1). Among these three conditionally significant associations, none was replicated with a conditional $P < 8.33 \times 10^{-3}$ using the UK Biobank WGS data (Table 1 and Methods).

Notably, among the nine conditionally significant noncoding rare variants associations with lipid traits, four were not detected by any of the three single-trait analyses (LDL-C, HDL-C or TG) using unconditional analysis of STAAR-O, including the associations of enhancer DHS rare variants in *NIPSNAP3A* and *LIPC* as well as ncRNA rare variants in *RP11-310H4.2* and *MIR4497* (Supplementary Table 6). These results demonstrate that MultiSTAAR-O has increased power over existing methods, and identifies additional trait-associated signals by leveraging cross-phenotype correlations between multiple traits.

Genetic-region multi-trait analysis of rare variants

We next applied MultiSTAAR-O to perform genetic-region multi-trait analysis to identify rare variants associated with lipid traits in TOPMed. Rare variants residing in 2-kb sliding windows with a 1-kb skip length were aggregated and analyzed using a joint model for LDL-C, HDL-C and TG. We incorporated 12 quantitative annotations, including nine aPCs, CADD, LINSIGHT and FATHMM-XF, along with the two MAF weights in MultiSTAAR-O (Methods). The overall distribution of MultiSTAAR-O P values was well-calibrated for the multi-trait analysis (Fig. 3b). At a Bonferroni-corrected significance threshold of $\alpha = 0.05/(2.65 \times 10^6) = 1.89 \times 10^{-8}$ accounting for 2.65 million 2-kb sliding windows across the genome, MultiSTAAR-O identified 502 genome-wide significant associations using unconditional multi-trait analysis (Fig. 3a and Supplementary Table 7). By dynamically incorporating multiple functional annotations capturing different aspects of variant function, MultiSTAAR-O detected more significant sliding windows and showed consistently smaller P values for the top sliding windows compared with multi-trait analysis using only MAFs as the weight (Fig. 3c). After conditioning on known lipids-associated variants[26,37,38], seven of the 502 associations remained significant at the Bonferroni-corrected threshold of $\alpha = 0.05/502 = 9.96 \times 10^{-5}$ (Table 2). Among these seven conditionally significant associations, six were replicated with a conditional $P < 9.96 \times 10^{-5}$ using the UK Biobank WGS data (Table 2 and Methods), including two sliding windows in *DOCK7* (chromosome 1, 62,651,447–62,653,446 bp; chromosome 1, 62,652,447–62,654,446 bp) and an intergenic sliding window (chromosome 1, 145,530,447–145,532,446 bp) that were not detected by any of the three single-trait analyses (LDL-C, HDL-C or TG) using STAAR-O (Supplementary Table 7). Notably, all known lipids-associated variants

https://doi.org/10.1038/s43588-024-00764-8

Table 2 | TOPMed genetic-region (2-kb sliding window) multi-trait analysis results of both unconditional analysis and analysis conditional on known lipid-associated variants

Chr.[a]	Start location[b]	End location[c]	Gene	Discovery (TOPMed)			Replication (UK Biobank)			Variants[f] (adjusted)
				No. of SNVs[d]	MultiSTAAR-O[e] (unconditional)	MultiSTAAR-O[e] (conditional)	No. of SNVs[d]	MultiSTAAR-O[e] (unconditional)	MultiSTAAR-O[e] (conditional)	
1	55,051,447	55,053,446	*PCSK9*	327	7.11×10^{-11}	6.60×10^{-8}	458	1.90×10^{-35}	3.75×10^{-41}	rs12117661, rs2495491, rs11591147, rs67608943, rs72646508, rs693668, rs28362261, rs28362263, rs141502002, rs505151, rs28362286
1	55,052,447	55,054,446	*PCSK9*	320	9.37×10^{-9}	9.07×10^{-6}	442	5.28×10^{-37}	4.01×10^{-41}	rs12117661, rs2495491, rs11591147, rs67608943, rs72646508, rs693668, rs28362261, rs28362263, rs141502002, rs505151, rs28362286
1	62,651,447	62,653,446	*DOCK7*	277	5.08×10^{-9}	7.56×10^{-10}	396	1.51×10^{-41}	7.74×10^{-45}	rs67461605
1	62,652,447	62,654,446	*DOCK7*	257	4.87×10^{-9}	7.24×10^{-10}	357	9.59×10^{-42}	4.93×10^{-45}	rs67461605
1	145,530,447	145,532,446	*intergenic*	233	5.12×10^{-9}	5.12×10^{-9}	386	4.54×10^{-28}	4.54×10^{-28}	NA[g]
19	11,104,367	11,106,366	*LDLR*	336	1.15×10^{-12}	8.33×10^{-13}	437	9.84×10^{-5}	2.11×10^{-5}	rs140753491, rs138294113, rs17242353, rs17242843, rs10422256, rs72658860, rs11669576, rs2738447, rs72658867, rs2738464, rs6511728, rs3760782, rs59168178, rs2278426, rs112942459
19	11,105,367	11,107,366	*LDLR*	338	5.97×10^{-14}	5.55×10^{-15}	480	4.51×10^{-4}	2.54×10^{-3}	rs140753491, rs138294113, rs17242353, rs17242843, rs10422256, rs72658860, rs11669576, rs2738447, rs72658867, rs2738464, rs6511728, rs3760782, rs59168178, rs2278426, rs112942459

A total of 61,838 samples from the TOPMed Program were considered in the analysis. Results for the conditionally significant sliding windows (unconditional MultiSTAAR-O $P<1.89\times10^{-8}$ and conditional MultiSTAAR-O $P<9.96\times10^{-5}$) are presented. MultiSTAAR-O is a two-sided test. [a]Chromosome number. [b]Start location of the 2-kb sliding window. [c]End location of the 2-kb sliding window. [d]Number of rare variants (MAF < 1%) in the 2-kb sliding window. [e]*P* value. [f]Adjusted variants in the conditional analysis. [g]No variant adjusted in the conditional analysis. Physical positions of each window are on build NCBI GRCh38/UCSC hg38.

indexed in the previous literature were at least 1 Mb away from the intergenic sliding window.

Comparison of MultiSTAAR-O with existing methods

Using TOPMed Freeze 8 WGS data, our gene-centric multi-trait analysis of coding rare variants identified 34 conditionally significant associations with lipid traits (Supplementary Table 5), including *NPC1L1* and *SCARB1* missense rare variants that were missed by multi-trait burden, SKAT and ACAT-V tests (Supplementary Table 4). Among the nine and seven conditionally significant associations detected in gene-centric multi-trait analyses of noncoding rare variants and genetic-region multi-trait analysis, MultiSTAAR-O identified one and two associations, respectively, that were missed by the multi-trait burden, SKAT and ACAT-V tests (Supplementary Tables 6 and 7). These associations

included enhancer CAGE rare variants in *SPC24* and two sliding windows in *LDLR* (chromosome 19, 11,104,367–11,106,366 bp; chromosome 19, 11,105,367–11,107,366 bp).

Analysis of non-lipid phenotypes in the TOPMed WGS data

We further applied MultiSTAAR to analyzing a broader spectrum of phenotypes in the TOPMed WGS data, including (1) multi-trait analysis of fasting glucose (FG) and fasting insulin (FI) (n = 21,731)[43] and (2) multi-trait analysis of four inflammation biomarkers (C-reactive protein (CRP), interleukin-6 (IL-6), lipoprotein-associated phospholipase A2 (Lp-PLA2) activity and lipoprotein-associated phospholipase A2 (Lp-PLA2) mass (n = 9,380)[44]). Similar to the lipids analysis, for each multi-trait analysis we performed gene-centric coding and noncoding analysis and genetic-region analysis to detect rare variant associations (Methods).

In gene-centric coding unconditional analysis, MultiSTAAR identified seven and seven genome-wide significant associations for glycemic and inflammation biomarker analyses, respectively. Seven and four associations remained significant at the Bonferroni-corrected level $\alpha = 0.05/7 = 7.14 \times 10^{-3}$ after conditioning on known phenotype-specific variants[24,43,44] (Supplementary Tables 8 and 9 and Extended Data Figs. 1a,b and 2a,b). In gene-centric noncoding unconditional analysis, MultiSTAAR identified six genome-wide significant associations for inflammation biomarker analysis, but no association remained significant at the Bonferroni-corrected level $\alpha = 0.05/6 = 8.33 \times 10^{-3}$ after conditioning on known phenotype-specific variants[24,43,44] (Supplementary Table 10 and Extended Data Fig. 2c,d).

In genetic-region unconditional analysis using 2-kb sliding windows, MultiSTAAR identified 41 genome-wide significant associations for inflammation biomarker analysis, and two associations remained significant at the Bonferroni-corrected level $\alpha = 0.05/41 = 1.22 \times 10^{-3}$ after conditioning on known phenotype-specific variants[24,43,44] (Supplementary Table 11 and Extended Data Fig. 2e,f). No genome-wide significant associations were identified in gene-centric noncoding and genetic-region analyses for glycemic analysis (Extended Data Fig. 1c–f).

Computation cost

The computational cost for MultiSTAAR-O to perform WGS multi-trait rare variant analysis of n = 61,838 related TOPMed lipids samples was 2 h using 250 2.10-GHz computing cores with 12 GB of memory for gene-centric coding analysis; 20 h using 250 2.10-GHz computing cores with 24 GB of memory for gene-centric noncoding analysis; 2 h using 250 2.10-GHz computing cores with 12 GB of memory of ncRNA analysis; and 20 h using 500 2.10-GHz computing cores with 24 GB of memory for sliding-window analysis. The runtime for all analyses scaled linearly with sample size[24].

Discussion

In this Article we have introduced MultiSTAAR as a general statistical framework and a flexible analytical pipeline for performing functionally informed multi-trait RVAS in large-scale WGS studies. By jointly analyzing multiple quantitative traits using an MLMM in the first step, MultiSTAAR explicitly leverages the correlation among multiple phenotypes to enhance the power for detecting additional association signals, outperforming single-trait analyses of the individual phenotypes. MultiSTAAR also enables conditional multi-trait analysis to identify putatively novel rare variant associations independent of a set of known variants. Using TOPMed Freeze 8 WGS data, our gene-centric multi-trait analysis of noncoding rare variants identified nine conditionally significant associations with lipid traits (Table 1), including four noncoding associations that were missed by single-trait analysis using STAAR (Supplementary Table 6). Our genetic-region multi-trait analysis of rare variants identified seven conditionally significant 2-kb sliding windows associated with lipid traits (Table 2), including three associations that were missed by single-trait analysis using STAAR (Supplementary Table 7).

Among the seven associations that were conditionally significant in multi-trait analysis but missed by single-trait analysis, five of them were replicated using the UK Biobank WGS data of 170,104 samples (Tables 1 and 2), including the associations of enhancer DHS rare variants in *NIPSNAP3A* and *LIPC*, and the associations of two sliding windows in *DOCK7* (chromosome 1, 62,651,447–62,653,446 bp; chromosome 1, 62,652,447–62,654,446 bp) and an intergenic sliding window (chromosome 1, 145,530,447–145,532,446 bp). Previous research has demonstrated that both common and rare coding variants in these genes are associated with lipid levels[45–47]. Our findings extend this understanding by suggesting that rare noncoding variants may also contribute to alterations in lipid levels. These results demonstrate the robustness of the MultiSTAAR method.

We additionally performed three double-trait analyses of these seven results. The association of ncRNA rare variants in *RP11-310H4.2* was also missed by double-trait analysis using MultiSTAAR. The remaining six results were detected by at least one of the double-trait analyses, though not consistently by the same analysis (Supplementary Tables 6 and 7). This observation highlights the complexity of pleiotropic effects in multi-trait analyses. We also applied MultiSTAAR to non-lipid phenotypes in the TOPMed WGS data by conducting multi-trait analyses for two glycemic traits and four inflammation biomarker traits. The quantile–quantile (Q–Q) plots were well-calibrated for all analyses, demonstrating the validity of MultiSTAAR for use at genome-wide significance levels (Extended Data Figs. 1b,d,f and 2b,d,f).

Our gene-centric analysis primarily focuses on detecting associations within coding and regulatory regions of protein-coding and ncRNA genes. Complementing this, our agnostic genetic-region analysis employs sliding windows, and focuses on detecting associations in intergenic regions, a supplementary approach to gene-centric analysis. The sliding-window approach covers all variants across the genome, including those that are not used in gene-centric analysis, such as the noncoding variants that are not in the estimated promoters and enhancers of protein-coding genes. Although there is some overlap in that both analyses include coding and regulatory regions, the gene-centric method incorporates categorical functional annotations of protein-coding genes that define analysis units for the analysis, and the non-gene-centric method defines analysis units using sliding windows.

Among the five new lipid trait associations identified using MultiSTAAR-O, which were conditionally significant in our three-trait analyses but undetected in single-trait analysis using TOPMed data and replicated using UK Biobank data, two were detected by gene-centric analysis and three by sliding-window analysis. Notably, the associations identified by these two approaches do not overlap, underscoring the distinct yet complementary nature of the approaches.

By dynamically incorporating multiple annotations capturing diverse aspects of variant biological function in the second step, MultiSTAAR further improves the power over existing multi-trait rare variant analysis methods. Our simulation studies demonstrated that MultiSTAAR-O maintained accurate type I error rates at genome-wide significance levels while achieving considerable power gains over multi-trait burden, SKAT and ACAT-V tests, which do not incorporate functional annotation information (Supplementary Table 1 and Supplementary Figs. 1–32). Notably, the existing ACAT-V method[9] does not support multi-trait analysis. We extended it to accommodate multi-trait settings and incorporated the multi-trait ACAT-V test into the MultiSTAAR framework (Methods).

Implemented as a flexible analytical pipeline, MultiSTAAR allows for customized input phenotype selection, variant set definition and user-specified annotation weights to facilitate functionally informed multi-trait analyses. In practice, we recommend utilizing a biological knowledge-based approach to define trait groups, as it ensures a biologically meaningful interpretation. Alternately, users could adopt a data-driven approach where traits are clustered based on their correlation matrix and subsequently grouped using clustering or similar methods.

In addition to rare variant association analysis of coding and non-coding regions, MultiSTAAR also provides multi-trait single-variant analysis for common and low-frequency variants under a given MAF or minor allele count (MAC) cutoff (for example, MAC ≥ 20). We performed single-variant analysis of three lipid traits of 61,838 TOPMed samples using MultiSTAAR (Supplementary Fig. 33 and Supplementary Table 12). It took 8 h using 250 2.10-GHz computing cores with 12 GB of memory for multi-trait single-variant analysis of all genetic variants with MAC ≥ 20 (72,762,611 in total), which is scalable for large WGS/WES datasets.

There are several limitations to this study. First, MultiSTAAR performs sliding-window analysis with fixed sizes and could be further developed to allow for dynamic windows with data-adaptive sizes in genetic-region analysis using SCANG-STAAR[24,48]. Second, MultiSTAAR could be improved to properly leverage synthetic surrogates in the presence of partially missing phenotypes[49]. Third, MultiSTAAR is currently designed for analyzing individual-level genotype and phenotype data, which could be extended to incorporate summary statistics for meta-analyses of multiple WGS/WES studies[50].

Despite these limitations, MultiSTAAR provides a powerful statistical framework and a computationally scalable analytical pipeline for large-scale WGS multi-trait analysis with complex study samples. As the sample sizes and number of available phenotypes increase in biobank-scale sequencing studies, our proposed method may contribute to a better understanding of the genetic architecture of complex traits by elucidating the role of rare variants with pleiotropic effects.

Methods

Ethics statement

This study relied on analyses of genetic data from TOPMed cohorts. The study has been approved by the TOPMed Publications Committee, TOPMed Lipids Working Group and all participating cohorts, including Old Order Amish (phs000956.v1.p1), Atherosclerosis Risk in Communities Study (phs001211), Mt Sinai BioMe Biobank (phs001644), Coronary Artery Risk Development in Young Adults (phs001612), Cleveland Family Study (phs000954), Cardiovascular Health Study (phs001368), Diabetes Heart Study (phs001412), Framingham Heart Study (phs000974), Genetic Study of Atherosclerosis Risk (phs001218), Genetic Epidemiology Network of Arteriopathy (phs001345), Genetic Epidemiology Network of Salt Sensitivity (phs001217), Genetics of Lipid Lowering Drugs and Diet Network (phs001359), Hispanic Community Health Study–Study of Latinos (phs001395), Hypertension Genetic Epidemiology Network and Genetic Epidemiology Network of Arteriopathy (phs001293), Jackson Heart Study (phs000964), Multi-Ethnic Study of Atherosclerosis (phs001416), San Antonio Family Heart Study (phs001215), Genome-wide Association Study of Adiposity in Samoans (phs000972), Taiwan Study of Hypertension using Rare Variants (phs001387) and Women's Health Initiative (phs001237) (accession numbers are provided in parentheses). The use of human genetics data from TOPMed cohorts was approved by the Harvard T.H. Chan School of Public Health IRB (IRB13-0353).

Notation and model

Suppose there are n subjects with a total of M variants sequenced across the whole genome. For the ith subject, let $\mathbf{Y}_i = (y_{i1}, y_{i2}, \ldots, y_{iK})^{\mathrm{T}}$ denotes a vector of K quantitative phenotypes; $\mathbf{X}_i = (x_{i1}, x_{i2}, \ldots, x_{iq})^{\mathrm{T}}$ denotes q covariates, such as age, gender and ancestral PCs; $\mathbf{G}_i = (G_{i1}, G_{i2}, \ldots, G_{ip})^{\mathrm{T}}$ denotes the genotype matrix of the p genetic variants in a variant set. Because these K phenotypes may be defined on different measurement scales, we assume that each phenotype has been rescaled to have zero mean and unit variance.

When the data consist of unrelated samples, we consider the following multivariate linear model (MLM):

$$\mathbf{Y}_i = \begin{bmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{iK} \end{bmatrix} = \begin{bmatrix} \alpha_{0,1} + \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\alpha}_1 + \mathbf{G}_i^{\mathrm{T}}\boldsymbol{\beta}_1 \\ \alpha_{0,2} + \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\alpha}_2 + \mathbf{G}_i^{\mathrm{T}}\boldsymbol{\beta}_2 \\ \vdots \\ \alpha_{0,K} + \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\alpha}_K + \mathbf{G}_i^{\mathrm{T}}\boldsymbol{\beta}_K \end{bmatrix} + \begin{bmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \vdots \\ \varepsilon_{iK} \end{bmatrix} \quad (1)$$

where $\alpha_{0,k}$ is an intercept, $\boldsymbol{\alpha}_k = (\alpha_{1,k}, \alpha_{2,k}, \ldots, \alpha_{q,k})^{\mathrm{T}}$ and $\boldsymbol{\beta}_k = (\beta_{1,k}, \beta_{2,k}, \ldots, \beta_{q,k})^{\mathrm{T}}$ are column vectors of regression coefficients for covariates $\mathbf{X}_i$ and genotype $\mathbf{G}_i$ in phenotype k, respectively. The error terms $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \varepsilon_{i2}, \ldots, \varepsilon_{iK})^{\mathrm{T}}$ are independent and identically distributed and follow a multivariate normal distribution with the mean a vector of zeros and variance-covariance matrix $\boldsymbol{\Sigma}_{K\times K}$, assumed identical for all subjects. For all n subjects, using matrix notation we can write model (1) as

$$\mathbf{Y}_{n\times K} = \mathbf{1}_n\boldsymbol{\alpha}_0^{\mathrm{T}} + \mathbf{X}_{n\times q}\boldsymbol{\alpha}_{q\times K} + \mathbf{G}_{n\times p}\boldsymbol{\beta}_{p\times K} + \boldsymbol{\varepsilon}_{n\times K} \quad (2)$$

where $\mathbf{1}_n$ is a column vector of 1s of length n, $\boldsymbol{\alpha}_0 = (\alpha_{0,1}, \alpha_{0,2}, \ldots, \alpha_{0,K})^{\mathrm{T}}$ is a column vector of regression intercepts, the kth columns of $\boldsymbol{\alpha}_{q\times K}$ and $\boldsymbol{\beta}_{p\times K}$ are $\boldsymbol{\alpha}_k$ and $\boldsymbol{\beta}_k$, respectively, and $\boldsymbol{\varepsilon}_{n\times K} = (\boldsymbol{\varepsilon}_1, \boldsymbol{\varepsilon}_2, \ldots, \boldsymbol{\varepsilon}_n)^{\mathrm{T}} \sim \mathrm{MatrixNormal}_{n,K}(\mathbf{0}_{n\times K}, \mathbf{I}_{n\times n}, \boldsymbol{\Sigma}_{K\times K})$ follows a matrix normal distribution. We calculate the scaled residual for each subject on each phenotype, defined as $\hat{\mathbf{e}}_{n\times K} = (\mathbf{Y}_{n\times K} - \hat{\boldsymbol{\mu}}_{n\times K})\hat{\boldsymbol{\Sigma}}_{K\times K}^{-1}$, where $\hat{\boldsymbol{\mu}}_{n\times K}$ (a matrix of fitted values) and $\hat{\boldsymbol{\Sigma}}_{K\times K}$ are estimated under the null MLM $\mathbf{Y}_{n\times K} = \mathbf{1}_n\boldsymbol{\alpha}_0^{\mathrm{T}} + \mathbf{X}_{n\times q}\boldsymbol{\alpha}_{q\times K} + \boldsymbol{\varepsilon}_{n\times K}$, where no variant has any effect on any phenotype.

When the data consist of related samples, we consider the following MLMM[19,51,52]:

$$\mathbf{Y}_i = \begin{bmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{iK} \end{bmatrix} = \begin{bmatrix} \alpha_{0,1} + \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\alpha}_1 + \mathbf{G}_i^{\mathrm{T}}\boldsymbol{\beta}_1 \\ \alpha_{0,2} + \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\alpha}_2 + \mathbf{G}_i^{\mathrm{T}}\boldsymbol{\beta}_2 \\ \vdots \\ \alpha_{0,K} + \mathbf{X}_i^{\mathrm{T}}\boldsymbol{\alpha}_K + \mathbf{G}_i^{\mathrm{T}}\boldsymbol{\beta}_K \end{bmatrix} + \begin{bmatrix} b_{i1} \\ b_{i2} \\ \vdots \\ b_{iK} \end{bmatrix} + \begin{bmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \vdots \\ \varepsilon_{iK} \end{bmatrix} \quad (3)$$

where the random effects b_{iK} account for relatedness and remaining population structure unaccounted by ancestral PCs[20]. We assume that $\mathbf{b}_{n\times K} = (b_{iK})_{n\times K} \sim \mathrm{MatrixNormal}_{n,K}(\mathbf{0}_{n\times K}, \boldsymbol{\Phi}_{n\times n}, \boldsymbol{\Theta}_{K\times K})$ with a variance component matrix $\boldsymbol{\Theta}_{K\times K}$ and a sparse genetic relatedness matrix $\boldsymbol{\Phi}_{n\times n}$ (refs. 21,22). For all n subjects, using matrix notation we can rewrite equation (3) as

$$\mathbf{Y}_{n\times K} = \mathbf{1}_n\boldsymbol{\alpha}_0^{\mathrm{T}} + \mathbf{X}_{n\times q}\boldsymbol{\alpha}_{q\times K} + \mathbf{G}_{n\times p}\boldsymbol{\beta}_{p\times K} + \mathbf{b}_{n\times K} + \boldsymbol{\varepsilon}_{n\times K} \quad (4)$$

We calculate the scaled residual for each subject on each phenotype, defined as $\hat{\mathbf{e}}_{n\times K} = (\mathbf{Y}_{n\times K} - \hat{\boldsymbol{\mu}}_{n\times K})\hat{\boldsymbol{\Sigma}}_{K\times K}^{-1}$, where $\hat{\boldsymbol{\mu}}_{n\times K}$ and $\hat{\boldsymbol{\Sigma}}_{K\times K}$ are estimated under the null MLMM $\mathbf{Y}_{n\times K} = \mathbf{1}_n\boldsymbol{\alpha}_0^{\mathrm{T}} + \mathbf{X}_{n\times q}\boldsymbol{\alpha}_{q\times K} + \mathbf{b}_{n\times K} + \boldsymbol{\varepsilon}_{n\times K}$. Under both MLM and MLMM, our goal is to test for an association between a set of p genetic variants and K quantitative phenotypes, adjusting for covariates and relatedness. This corresponds to testing $H_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \cdots = \boldsymbol{\beta}_K = \mathbf{0}$.

Multi-trait rare variant association tests using MultiSTAAR

Single-trait score-based aggregation methods[5–9] can be extended to allow for jointly testing the association between rare variants in a variant set and multiple quantitative phenotypes. For a given variant set, let $\mathbf{S}_{p\times K} = (S_{jk})_{p\times K} = (\mathbf{G}_{n\times p})^{\mathrm{T}}\hat{\mathbf{e}}_{n\times K}$ denote the matrix of score statistics, where S_{jk} is the score statistic for the jth variant on the kth phenotype. For the multi-trait burden test using MultiSTAAR (Burden-MT), we consider the test statistic

$$Q_{\text{Burden-MT}} = \left(\sum_{j=1}^{p} w_j\mathbf{S}_{j\cdot}\right)\hat{\mathbf{V}}^{-1}\left(\sum_{j=1}^{p} w_j\mathbf{S}_{j\cdot}\right)^{\mathrm{T}}$$

where w_j is the weight defined as a function of the MAF for the jth variant[4,18], $\mathbf{S}_{j\cdot} = (S_{j1}, S_{j2}, \ldots, S_{jK})$ is the jth row of $\mathbf{S}$ and $\hat{\mathbf{V}}$ is the estimated variance-covariance matrix of $\sum_{j=1}^{p} w_j \mathbf{S}_{j\cdot} = \boldsymbol{w}^{\mathrm{T}}\mathbf{S}$. $Q_{\text{Burden-MT}}$ asymptotically follows a standard χ^2 distribution with K degrees of freedom under the null hypothesis, and its P value can be obtained analytically while accounting for the linkage disequilibrium (LD) between variants and the correlation between phenotypes.

For multi-trait SKAT using MultiSTAAR (SKAT-MT), we consider the test statistic

$$Q_{\text{SKAT-MT}} = \sum_{k=1}^{K}\sum_{j=1}^{p} w_j^2 S_{jk}^2$$

$Q_{\text{SKAT-MT}}$ asymptotically follows a mixture of χ^2 distributions under the null hypothesis, and its P value can be obtained analytically while accounting for the LD between variants and the correlation between phenotypes[14,15].

For multi-trait ACAT-V using MultiSTAAR (ACAT-V-MT), we propose the test statistic

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

where p' is the number of variants with MAC > 10, and p_j is the multi-trait association P value of individual variant j for those variants with MAC > 10 whose test statistic is given by the K degrees of freedom multivariate score test

$$Q_j = \mathbf{S}_{j\cdot}\hat{\mathbf{V}}_{\mathbf{S}_{j\cdot}}^{-1}\mathbf{S}_{j\cdot}^{\mathrm{T}}$$

where $\hat{\mathbf{V}}_{\mathbf{S}_{j\cdot}}$ is the estimated variance-covariance matrix of $\mathbf{S}_{j\cdot}$; p_0 is the multi-trait burden test P value of extremely rare variants with MAC ≤ 10 as described above, and $\overline{w^2\text{MAF}\,(1-\text{MAF})}$ is the average of the weights $w_j^2\text{MAF}_j\,(1-\text{MAF}_j)$ among the extremely rare variants with MAC ≤ 10. $Q_{\text{ACAT-V-MT}}$ is approximated well by a scaled Cauchy distribution under the null hypothesis, and its P value can be obtained analytically while accounting for the LD between variants and the correlation between phenotypes[9,53]. Note that when $K = 1$, the multi-trait burden, SKAT and ACAT-V tests reduce to the original single-trait burden, SKAT and ACAT-V tests.

Suppose we have a collection of L annotations, then let A_{jl} denote the lth annotation for the jth variant in the variant set. We define the functionally informed multi-trait burden, SKAT and ACAT-V test statistics weighted by the lth annotation as follows:

$$Q_{\text{Burden-MT},l,(a_1,a_2)} = \left(\sum_{j=1}^{p}\hat{\pi}_{jl}w_{j,(a_1,a_2)}\mathbf{S}_{j\cdot}\right)\hat{\mathbf{V}}_{l,(a_1,a_2)}^{-1}\left(\sum_{j=1}^{p}\hat{\pi}_{jl}w_{j,(a_1,a_2)}\mathbf{S}_{j\cdot}\right)^{\mathrm{T}}$$

$$Q_{\text{SKAT-MT},l,(a_1,a_2)} = \sum_{k=1}^{K}\sum_{j=1}^{p}\hat{\pi}_{jl}w_{j,(a_1,a_2)}^2 S_{jk}^2$$

$$Q_{\text{ACAT-V-MT},l,(a_1,a_2)} = \overline{\hat{\pi}_{\cdot l}w_{(a_1,a_2)}^2\text{MAF}\,(1-\text{MAF})}\tan\left((0.5-p_{0,l})\pi\right) + \sum_{j=1}^{M'}\hat{\pi}_{jl}w_{j,(a_1,a_2)}^2\text{MAF}_j\,(1-\text{MAF}_j)\tan\left((0.5-p_j)\pi\right)$$

where $\hat{\pi}_{jl} = \frac{\text{rank}(A_{jl})}{M}$, $w_{j,(a_1,a_2)} = \text{Beta}\,(\text{MAF}_j; a_1, a_2)$ with $(a_1, a_2) \in \mathcal{A} = \{(1.25), (1,1)\}$, $\hat{\mathbf{V}}_{l,(a_1,a_2)}$ is the estimated variance-covariance matrix of $\sum_{j=1}^{p}\hat{\pi}_{jl}w_{j,(a_1,a_2)}\mathbf{S}_{j\cdot}$, and $\overline{\hat{\pi}_{\cdot l}w_{(a_1,a_2)}^2\text{MAF}\,(1-\text{MAF})}$ is the average of the weights $\hat{\pi}_{jl}w_{j,(a_1,a_2)}^2\text{MAF}_j\,(1-\text{MAF}_j)$ among the extremely rare variants with MAC ≤ 10.

For each type of rare variant test, we define MultiSTAAR-B, MultiSTAAR-S and MultiSTAAR-A to incorporate multiple functional annotations through the STAAR framework for multi-trait burden, SKAT and ACAT-V as

$$T_{\text{MultiSTAAR}-B(a_1,a_2)} = \sum_{l=0}^{L}\frac{\tan\left\{(0.5-p_{\text{Burden}-\text{MT},l,(a_1,a_2)})\pi\right\}}{L+1}$$

$$T_{\text{MultiSTAAR}-S(a_1,a_2)} = \sum_{l=0}^{L}\frac{\tan\left\{(0.5-p_{\text{SKAT-MT},l,(a_1,a_2)})\pi\right\}}{L+1}$$

$$T_{\text{MultiSTAAR}-A(a_1,a_2)} = \sum_{l=0}^{L}\frac{\tan\left\{(0.5-p_{\text{ACAT-V-MT},l,(a_1,a_2)})\pi\right\}}{L+1}$$

where $T_{\text{MultiSTAAR-B}(1,1)}$, $T_{\text{MultiSTAAR-S}(1,25)}$ and $T_{\text{MultiSTAAR-A}(1,25)}$ are the test statistics of MultiSTAAR-B, MultiSTAAR-S and MultiSTAAR-A, respectively. The P values of $T_{\text{MultiSTAAR-B}(a_1,a_2)}$, $T_{\text{MultiSTAAR-S}(a_1,a_2)}$ and $T_{\text{MultiSTAAR-A}(a_1,a_2)}$ can be calculated by

$$P_{\text{MultiSTAAR-B}(a_1,a_2)} = \frac{1}{2} - \frac{\{\arctan\left(T_{\text{MultiSTAAR-B}(a_1,a_2)}\right)\}}{\pi}$$

$$P_{\text{MultiSTAAR-S}(a_1,a_2)} = \frac{1}{2} - \frac{\{\arctan\left(T_{\text{MultiSTAAR-S}(a_1,a_2)}\right)\}}{\pi}$$

$$P_{\text{MultiSTAAR-A}(a_1,a_2)} = \frac{1}{2} - \frac{\{\arctan\left(T_{\text{MultiSTAAR-A}(a_1,a_2)}\right)\}}{\pi}$$

Finally, we define the omnibus MultiSTAAR-O test statistic as

$$\begin{aligned}T_{\text{MultiSTAAR-O}} &= \frac{1}{3|\mathcal{A}|}\sum_{(a_1,a_2)\in\mathcal{A}}\left[T_{\text{MultiSTAAR-B}(a_1,a_2)} + T_{\text{MultiSTAAR-S}(a_1,a_2)} + T_{\text{MultiSTAAR-A}(a_1,a_2)}\right]\\ &= \frac{1}{3|\mathcal{A}|}\sum_{(a_1,a_2)\in\mathcal{A}}\sum_{l=0}^{L}\left[\frac{\tan\left\{(0.5-p_{\text{Burden-MT},l,(a_1,a_2)})\pi\right\}}{L+1} + \frac{\tan\left\{(0.5-p_{\text{SKAT-MT},l,(a_1,a_2)})\pi\right\}}{L+1} + \frac{\tan\left\{(0.5-p_{\text{ACAT-V-MT},l,(a_1,a_2)})\pi\right\}}{L+1}\right],\end{aligned}$$

and the P value of $T_{\text{MultiSTAAR-O}}$ can be calculated by

$$P_{\text{MultiSTAAR-O}} = \frac{1}{2} - \frac{\{\arctan\left(T_{\text{MultiSTAAR-O}}\right)\}}{\pi}$$

MultiSTAAR-O integrates different types of test into an omnibus approach to achieve robust power with respect to the sparsity of causal variants and the directionality of effects of causal variants in a variant set. Specifically, by including Burden-MT, MultiSTAAR-O is powerful when most variants in a variant set are causal and have effects in the same direction; by including SKAT-MT, MultiSTAAR-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-MT, MultiSTAAR-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 incorporating multiple functional annotations, MultiSTAAR-O is powerful when any of these functional annotations can pinpoint causal variants.

Data simulation

Type I error rate simulations. We performed simulation studies to evaluate how accurately MultiSTAAR controls the type I error rate. We generated three quantitative traits from a MLM, conditional on two covariates:

$$\mathbf{Y}_i = \begin{bmatrix} Y_{i1} \\ Y_{i2} \\ Y_{i3} \end{bmatrix} = \begin{bmatrix} 0.5X_{i1} + 0.5X_{i2} \\ 0.5X_{i1} + 0.5X_{i2} \\ 0.5X_{i1} + 0.5X_{i2} \end{bmatrix} + \begin{bmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \varepsilon_{i3} \end{bmatrix}$$

where $X_{i1} \sim N(0, 1)$, $X_{i2} \sim$ Bernoulli (0.5) and

$$\begin{bmatrix} \epsilon_{i1} \\ \epsilon_{i2} \\ \epsilon_{i3} \end{bmatrix} \sim \text{MVN}\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1.0 & -0.1 & 0.2 \\ -0.1 & 1.0 & -0.4 \\ 0.2 & -0.4 & 1.0 \end{bmatrix} \right)$$

where MVN denotes a multivariate normal distribution. The correlation matrix of error terms $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \varepsilon_{i2}, \varepsilon_{i3})^{\mathrm{T}}$ was chosen to mimic the correlations between three lipid traits, LDL-C, HDL-C and TG, estimated from the TOPMed data[26]. We considered a sample size of 10,000 and generated genotypes by simulating 20,000 sequences for 100 different regions each spanning 1 Mb. The data generation used the calibration coalescent model (COSI)[29] with parameters set to mimic the LD structure of African Americans. In each simulation replicate, ten annotations were generated as $A_1, \ldots, A_{10}$, all independently and identically distributed as $N(0, 1)$ for each variant, and we randomly selected 5-kb regions from these 1-Mb regions for type I error rate simulations. We applied MultiSTAAR-B, MultiSTAAR-S, MultiSTAAR-A and MultiSTAAR-O by incorporating MAFs and the ten annotations together with the Burden-MT, SKAT-MT and ACAT-V-MT tests. We repeated the procedure with 10^8 replicates to examine the type I error rate at levels $\alpha = 10^{-4}$, 10^{-5} and 10^{-6}.

Empirical power simulations. We next carried out simulation studies under a variety of configurations to assess the power of MultiSTAAR-O and how its incorporation of multiple functional annotations affects the power compared to the multi-trait burden, SKAT and ACAT-V tests implemented in MultiSTAAR. In each simulation replicate, we randomly selected 5-kb regions from a 1-Mb region for power evaluations. For each selected 5-kb region, we generated three quantitative traits from an MLM:

$$\mathbf{Y}_i = \begin{bmatrix} Y_{i1} \\ Y_{i2} \\ Y_{i3} \end{bmatrix} = \begin{bmatrix} 0.5X_{i1} + 0.5X_{i2} + \mathbf{G}_i^{\mathrm{T}}\boldsymbol{\beta}_1 \\ 0.5X_{i1} + 0.5X_{i2} + \mathbf{G}_i^{\mathrm{T}}\boldsymbol{\beta}_2 \\ 0.5X_{i1} + 0.5X_{i2} + \mathbf{G}_i^{\mathrm{T}}\boldsymbol{\beta}_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \varepsilon_{i3} \end{bmatrix}$$

where X_{1i}, X_{2i} and $\boldsymbol{\varepsilon}_i$ are defined as in the type I error rate simulations, $\mathbf{G}_i = (G_{i1}, G_{i2}, \ldots, G_{ip})^{\mathrm{T}}$ and $\boldsymbol{\beta}_k = (\beta_{1,k}, \beta_{2,k}, \ldots, \beta_{p,k})^{\mathrm{T}}$ are the genotypes and effect sizes of the p genetic variants in the signal region.

The genetic effect of variant j on phenotype k was defined as $\beta_{j,k} = c_j d_k \gamma_j$ to allow for heterogeneous effect sizes among variants and phenotypes. Specifically, we generated the causal variant indicator c_j according to a logistic model:

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

where $\{l_1, \cdots, l_5\} \subset \{1, \cdots, 10\}$ were randomly sampled for each region. For different regions, the causality of variants depended on different sets of annotations. We set $\delta_{l_\cdot} = \log(5)$ for all annotations and varied the proportions of causal variants in the signal region by setting $\delta_0 = \text{logit}(0.0015)$, $\text{logit}(0.015)$ and $\text{logit}(0.18)$, which correspond to averaging 5%, 15% and 35% causal variants in the signal region, respectively. We considered four scenarios of phenotypic indicator d_k that reflect different underlying genetic architectures across phenotypes: $(d_1, d_2, d_3) = (1, 0, 0)$, $(1, 0, 1)$, $(1, 1, 0)$ and $(1, 1, 1)$. These correspond to causal variants in the signal region being associated with (1) one phenotype only, (2) two positively correlated phenotypes, (3) two negatively correlated phenotypes and (4) all three phenotypes. We modeled the absolute effect sizes of causal variants using $|\gamma_j| = c_0 |\log_{10}\text{MAF}_j|$, such that it was a decreasing function of MAF. c_0 was set to be 0.13, 0.1, 0.1 and 0.07, respectively, to ensure a decent power of tests under each scenario. We additionally varied the proportions of causal variant effect size directions (signs of γ_j) by randomly generating 100%, 80% and 50% variants on average to have positive effects. We applied MultiSTAAR-B, MultiSTAAR-S, MultiSTAAR-A and MultiSTAAR-O using MAFs and all ten annotations together with the Burden-MT, SKAT-MT and ACAT-V-MT tests. We repeated the procedure with 10^4 replicates to examine the power at level $\alpha = 10^{-7}$. The sample size was 10,000 across all scenarios.

To assess how different correlation structures between phenotypes influence the enhancement of statistical power, we conducted additional power simulation studies, including (1) independent, by considering $\begin{bmatrix} \epsilon_{i1} \\ \epsilon_{i2} \\ \epsilon_{i3} \end{bmatrix} \sim \text{MVN}\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \right)$; (2) low phenotypic correlation, by considering $\begin{bmatrix} \epsilon_{i1} \\ \epsilon_{i2} \\ \epsilon_{i3} \end{bmatrix} \sim \text{MVN}\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1.0 & -0.05 & 0.1 \\ -0.05 & 1.0 & -0.2 \\ 0.1 & -0.2 & 1.0 \end{bmatrix} \right)$; and (3) high phenotypic correlation, by considering $\begin{bmatrix} \epsilon_{i1} \\ \epsilon_{i2} \\ \epsilon_{i3} \end{bmatrix} \sim \text{MVN}\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1.0 & -0.2 & 0.4 \\ -0.2 & 1.0 & -0.8 \\ 0.4 & -0.8 & 1.0 \end{bmatrix} \right)$. For each correlation structure, the causal variant proportions and causal variant effect sizes were considered the same as previous power simulation studies. Our simulation results demonstrate that MultiSTAAR achieves robust and considerable power gain in identifying pleiotropic loci across all correlation structures compared with existing multi-trait analysis methods (Supplementary Figs. 9–32).

Computational cost benchmarking. We benchmarked the computational cost of MultiSTAAR along with (1) the number of traits and (2) the sample size using simulation studies. Specifically, for (1), we varied the number traits among 2, 3, 4 and 5 while considering the sample size at 10,000 and randomly selecting 5-kb regions. For (2), we varied the sample sizes among 10,000, 20,000, 30,000, 40,000 and 50,000 while considering three traits and randomly selecting regions with 150 variants. Computational time was benchmarked by averaging over 10,000 simulation replicates. Our benchmarking results show that for both the null model fitting step and the MultiSTAAR testing step, the computational time increased approximately quadratically with the number of traits, and the computational time increased approximately linearly with the sample size (Supplementary Figs. 34 and 35). All analyses were completed with less than 2 GB of memory.

Lipid traits

Conventionally measured plasma lipids, including LDL-C, HDL-C and TGs, were included for analysis. LDL-C was either calculated by the Friedewald equation when TG levels were <400 mg dl^{-1} or directly measured. Given the average effect of statins, when statins were present, LDL-C was adjusted by dividing by 0.7. Triglycerides were natural-log-transformed for analysis. Phenotypes were harmonized by each cohort and deposited into the dbGaP TOPMed Exchange Area.

Multi-trait analysis of lipids in the TOPMed WGS data

The TOPMed WGS data consist of multi-ethnic related samples[1]. Race/ethnicity was defined using a combination of self-reported race/ethnicity from participant questionnaires and study recruitment information (Supplementary Note)[31]. A plot of ancestry PCs was presented to illustrate the genetic diversity among the populations studied (Supplementary Fig. 36). In this study, we applied MultiSTAAR to perform multi-trait rare variant analysis of three quantitative lipid traits (LDL-C, HDL-C and TG) using 20 study cohorts from the TOPMed Freeze 8 WGS data. LDL-C was adjusted for the presence of medications as before[30]. For each study, we first fit a linear regression model adjusting for age, age^2 and sex for each race/ethnicity-specific group. In addition, for

Old Order Amish (OOA), we also adjusted for *APOB* p.R3527Q in LDL-C analysis and adjusted for *APOC3* p.R19Ter in TG and HDL-C analyses[30]. The covariate distributions of samples with and without missing lipid traits are similar (Supplementary Fig. 37), indicating that data are plausibly missing at random.

Total cholesterol (TC) was not included in the multi-trait analysis based on the Friedewald equation TC ≈ LDL-C + HDL-C + TG/5. Given that TC is a linear combination of LDL-C, HDL-C and TG, it does not provide additional biological information when LDL-C, HDL-C and TG are already included in the model.

We performed rank-based inverse-normal transformation of the residuals of LDL-C, HDL-C and TG within each race/ethnicity-specific group. We then fit a MLMM for the rank-normalized residuals, adjusting for 11 ancestral PCs, ethnicity group indicators and a variance component for empirically derived sparse kinship matrix to account for population structure, relatedness and correlation between phenotypes.

We next applied MultiSTAAR-O to perform multi-trait variant set analyses for rare variants (MAF < 1%) by scanning the genome, including gene-centric analysis of each protein-coding gene using five coding variant functional categories (putative loss-of-function rare variants, missense rare variants, disruptive missense rare variants, putative loss-of-function and disruptive missense rare variants and synonymous rare variants); seven noncoding variant functional categories (promoter rare variants overlaid with CAGE sites, promoter rare variants overlaid with DHS sites, enhancer rare variants overlaid with CAGE sites, enhancer rare variants overlaid with DHS sites, UTR rare variants, upstream region rare variants, downstream region rare variants) and rare variants in ncRNA genes; and genetic-region analysis using 2-kb sliding windows across the genome with a 1-kb skip length.

Our analysis revealed that MultiSTAAR-O detected 325 significant associations that were missed by both existing multi-trait-based methods Burden-MT and SKAT-MT. Conversely, Burden-MT identified four and SKAT-MT identified 11 significant associations not detected by MultiSTAAR-O (Supplementary Fig. 38). This demonstrates the robust power of MultiSTAAR-O, particularly in handling the sparsity and directionality of causal variant effects through an integrated omnibus approach.

The WGS multi-trait rare variant analysis was performed using the R packages MultiSTAAR (version 0.9.7, https://github.com/xihaoli/MultiSTAAR)[54] and STAARpipeline (version 0.9.7, https://github.com/xihaoli/STAARpipeline)[55]. The WGS rare variant single-trait analysis of LDL-C, HDL-C and TG was performed using the R packages STAAR (version 0.9.7, http://github.com/xihaoli/STAAR)[56] and STAARpipeline (version 0.9.7)[55]. Both multi-trait and single-trait analyses results were summarized and visualized using the R package STAARpipelineSummary (version 0.9.7, https://github.com/xihaoli/STAARpipelineSummary)[57].

Multi-trait analysis of lipids in the UK Biobank WGS data

We used pVCF format files for the WGS data of 200,004 UK Biobank participants (UK Biobank Field #24304) and followed the same QC procedure as in a previous study of UK Biobank WGS data[3]. We kept all variants with a pass indicated by QC label and AAScore greater than 0.5, where AAScore was generated by GraphTyper, the software used by the UK Biobank to perform genotype calling. We harmonized three lipid traits (LDL-C, HDL-C and TG) of the UK Biobank WGS data. For LDL-C, we excluded individuals with LDL-C < 10 mg dl^{-1} or TG > 400 mg ml^{-1}. LDL-C was then adjusted by dividing the value by 0.7 among individuals reporting lipid-lowering medication use or statin use at any time point. TG levels were natural-logarithm-transformed. A total of 170,104 individuals had data on LDL-C, HDL-C and TG.

We fit a linear regression model adjusting for age, age^2, sex and the first ten ancestral PCs. Residuals were then rank-based inverse-normal transformed and multiplied by the standard deviation. We next fit an MLMM for the rank-normalized residuals of LDL-C, HDL-C and TG, adjusting for age, age^2, sex and the ten ancestral PCs, and a variance component for an empirically derived sparse kinship matrix to account for population structure, relatedness and correlation between phenotypes.

We next applied MultiSTAAR-O to perform multi-trait variant set analyses for rare variants (MAF < 1%), including gene-centric analysis of protein-coding genes using five coding variant functional categories; seven noncoding variant functional categories and rare variants in ncRNA genes; and genetic-region analysis using 2-kb sliding windows. For each analysis, the same set of annotations were incorporated as weights in MultiSTAAR-O (Supplementary Table 3). Our analysis was performed on the UK Biobank Research Analysis Platform (RAP). Specifically, the gene-centric coding analysis of five different masks for protein-coding genes across the genome required 1,183 central processing unit (CPU) hours with 16 GB of memory on average. The gene-centric noncoding analysis of seven different masks for protein-coding genes across the genome would require 17,173 CPU hours with 32 GB of memory on average. The gene-centric noncoding analysis of seven different masks for ncRNA genes across the genome required 2,453 CPU hours with 21 GB of memory on average.

Analysis of non-lipid phenotypes in the TOPMed WGS data

We applied MultiSTAAR to identify rare variants (MAF < 1%) associated with non-lipid phenotypes in the TOPMed WGS data, including (1) multi-trait analysis of FG and FI (n = 21,731) and (2) multi-trait analysis of four inflammation biomarkers, including CRP, IL-6, Lp-PLA2 activity and Lp-PLA2 mass (n = 9,380). The definitions and phenotype harmonization for these two glycemic traits and four inflammation biomarker traits were the same as those used in previous studies[43,44].

For each trait, we first fit a linear regression model adjusting for age and sex for each study-race/ethnicity group, with additional adjustment of age^2 and body mass index for FG and FI, ten ancestral PCs for FG and FI and 11 ancestral PCs for CRP, IL-6, Lp-PLA2 activity and Lp-PLA2 mass. The residuals were transformed using the rank-based inverse-normal transformation. We then fit an MLMM for the rank-normalized residuals, adjusting for 10 and 11 ancestral PCs for glycemic and inflammation biomarker analysis, respectively, study-ethnicity group indicators and a variance component for the empirically derived kinship matrix to account for population structure, relatedness and correlation between phenotypes. The outputs of two corresponding MLMM null models were then used in the multi-trait rare variant analysis of glycemic and inflammation biomarker traits by applying MultiSTAAR-O integrated in the STAARpipeline, including gene-centric analysis of protein-coding genes using five coding variant functional categories; seven noncoding variant functional categories and rare variants in ncRNA genes; and genetic-region analysis using 2-kb sliding windows. For each analysis, the same set of annotations were incorporated as weights in MultiSTAAR-O (Supplementary Table 3).

Genome build

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

Statistics and reproducibility

Sample size was not predetermined. The multi-trait analysis consisted of 20 study cohorts of TOPMed Freeze 8 and had 61,838 samples with lipid traits. We did not use any study design that required randomization or blinding.

Reporting summary

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

Data availability

This Article used TOPMed Freeze 8 WGS data and lipids phenotype data. Genotype and phenotype data are both available in the database of Genotypes and Phenotypes. The TOPMed WGS data were from the

following 20 study cohorts (accession numbers provided in parentheses): Old Order Amish (phs000956.v1.p1), Atherosclerosis Risk in Communities Study (phs001211), Mt Sinai BioMe Biobank (phs001644), Coronary Artery Risk Development in Young Adults (phs001612), Cleveland Family Study (phs000954), Cardiovascular Health Study (phs001368), Diabetes Heart Study (phs001412), Framingham Heart Study (phs000974), Genetic Study of Atherosclerosis Risk (phs001218), Genetic Epidemiology Network of Arteriopathy (phs001345), Genetic Epidemiology Network of Salt Sensitivity (phs001217), Genetics of Lipid Lowering Drugs and Diet Network (phs001359), Hispanic Community Health Study—Study of Latinos (phs001395), Hypertension Genetic Epidemiology Network and Genetic Epidemiology Network of Arteriopathy (phs001293), Jackson Heart Study (phs000964), Multi-Ethnic Study of Atherosclerosis (phs001416), San Antonio Family Heart Study (phs001215), Genome-Wide Association Study of Adiposity in Samoans (phs000972), Taiwan Study of Hypertension using Rare Variants (phs001387) and Women's Health Initiative (phs001237). The sample sizes, ancestry and phenotype summary statistics of these cohorts are provided in Supplementary Table 2. Source data for Figs. 2 and 3 and Extended Data Figs. 1 and 2 are available via Zenodo (https://doi.org/10.5281/zenodo.14213842)[58]. The UK Biobank analyses were conducted using the UK Biobank resource under application 52008. The functional annotation data are publicly available and can be downloaded from the following links: GRCh38 CADD v1.4 (https://cadd.gs.washington.edu/download); ANNOVAR dbNSFP v3.3a (https://annovar.openbioinformatics.org/en/latest/user-guide/download); LINSIGHT (https://github.com/CshlSiepelLab/LINSIGHT); FATHMM-XF (http://fathmm.biocompute.org.uk/fathmm-xf); FANTOM5 CAGE (https://fantom.gsc.riken.jp/5/data); GeneCards (https://www.genecards.org; v4.7 for hg38); and Umap/Bismap (https://bismap.hoffmanlab.org; 'before March 2020' version). In addition, recombination rate and nucleotide diversity were obtained from ref. 59. The whole-genome individual functional annotation data were assembled from a variety of sources, and the computed annotation PCs are available at the Functional Annotation of Variant-Online Resource (FAVOR) site (https://favor.genohub.org)[60] and the FAVOR database (https://doi.org/10.7910/DVN/1VGTJI)[61].

Code availability

MultiSTAAR is implemented as an open-source R package available at https://github.com/xihaoli/MultiSTAAR and https://hsph.harvard.edu/research/lin-lab/software. Data analysis was performed in R (4.1.0). STAAR v0.9.7 and MultiSTAAR v0.9.7 were used in simulation and real data analysis and implemented as open-source R packages available at https://github.com/xihaoli/STAAR (ref. 56) and https://github.com/xihaoli/MultiSTAAR (ref. 54). The assembled functional annotation data were downloaded from FAVOR using Wget (https://www.gnu.org/software/wget/wget.html).

References

1. Taliun, D. et al. Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021).
2. The All of Us Research Program Investigators. The 'All of Us' research program. *New Engl. J. Med.* **381**, 668–676 (2019).
3. Halldorsson, B. V. et al. The sequences of 150,119 genomes in the UK Biobank. *Nature* **607**, 732–740 (2022).
4. Lee, S., Abecasis, G. R., Boehnke, M. & Lin, X. Rare-variant association analysis: study designs and statistical tests. *Am. J. Human Genet.* **95**, 5–23 (2014).
5. Li, B. & Leal, S. M. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am. J. Human 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. Human 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. Human Genet.* **104**, 410–421 (2019).
10. Solovieff, N., Cotsapas, C., Lee, P. H., Purcell, S. M. & Smoller, J. W. Pleiotropy in complex traits: challenges and strategies. *Nat. Rev. Genet.* **14**, 483–495 (2013).
11. Sivakumaran, S. et al. Abundant pleiotropy in human complex diseases and traits. *Am. J. Human Genet.* **89**, 607–618 (2011).
12. Abdellaoui, A., Yengo, L., Verweij, K. J. H. & Visscher, P. M. 15 years of GWAS discovery: realizing the promise. *Am. J. Human Genet* **110**, 179–194 (2023).
13. Watanabe, K. et al. A global overview of pleiotropy and genetic architecture in complex traits. *Nat. Genet.* **51**, 1339–1348 (2019).
14. Wu, B. & Pankow, J. S. Sequence kernel association test of multiple continuous phenotypes. *Genet. Epidemiol.* **40**, 91–100 (2016).
15. Dutta, D., Scott, L., Boehnke, M. & Lee, S. Multi-SKAT: general framework to test for rare-variant association with multiple phenotypes. *Genet. Epidemiol.* **43**, 4–23 (2019).
16. Luo, L. et al. Multi-trait analysis of rare-variant association summary statistics using MTAR. *Nat. Commun.* **11**, 2850 (2020).
17. Broadaway, K. A. et al. A statistical approach for testing cross-phenotype effects of rare variants. *Am. J. Human Genet.* **98**, 525–540 (2016).
18. Li, X. et al. Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale. *Nat. Genet.* **52**, 969–983 (2020).
19. Sammel, M., Lin, X. & Ryan, L. Multivariate linear mixed models for multiple outcomes. *Stat. Med.* **18**, 2479–2492 (1999).
20. Conomos, M. P., Miller, M. B. & Thornton, T. A. Robust inference of population structure for ancestry prediction and correction of stratification in the presence of relatedness. *Genet. Epidemiol.* **39**, 276–293 (2015).
21. Conomos, M. P., Reiner, A. P., Weir, B. S. & Thornton, T. A. Model-free estimation of recent genetic relatedness. *Am. J. Human Genet.* **98**, 127–148 (2016).
22. Gogarten, S. M. et al. Genetic association testing using the GENESIS R/Bioconductor package. *Bioinformatics* **35**, 5346–5348 (2019).
23. Lee, P. H. et al. Principles and methods of in silico prioritization of noncoding regulatory variants. *Human Genet.* **137**, 15–30 (2018).
24. Li, Z. et al. A framework for detecting noncoding rare-variant associations of large-scale whole-genome sequencing studies. *Nat. Methods* **19**, 1599–1611 (2022).
25. Morrison, A. C. et al. Practical approaches for whole-genome sequence analysis of heart-and blood-related traits. *Am. J. Human Genet.* **100**, 205–215 (2017).
26. Selvaraj, M. S. et al. Whole genome sequence analysis of blood lipid levels in >66,000 individuals. *Nat. Commun.* **13**, 5995 (2022).
27. Liu, Z. & Lin, X. Multiple phenotype association tests using summary statistics in genome-wide association studies. *Biometrics* **74**, 165–175 (2018).
28. Teslovich, T. M. et al. Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707–713 (2010).
29. Schaffner, S. F. et al. Calibrating a coalescent simulation of human genome sequence variation. *Genome Res.* **15**, 1576–1583 (2005).
30. Natarajan, P. et al. Deep-coverage whole genome sequences and blood lipids among 16,324 individuals. *Nat. Commun.* **9**, 3391 (2018).

31. Stilp, A. M. et al. A system for phenotype harmonization in the National Heart, Lung and Blood Institute Trans-Omics for Precision Medicine (TOPMed) program. *Am. J. Epidemiol.* **190**, 1977–1992 (2021).
32. Frankish, A. et al. GENCODE reference annotation for the human and mouse genomes. *Nucleic Acids Res.* **47**, D766–D773 (2019).
33. Dong, C. et al. Comparison and integration of deleteriousness prediction methods for non-synonymous SNVs in whole exome sequencing studies. *Human Mol. Genet.* **24**, 2125–2137 (2014).
34. Kircher, M. et al. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
35. 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).
36. Rogers, M. F. et al. FATHMM-XF: accurate prediction of pathogenic point mutations via extended features. *Bioinformatics* **34**, 511–513 (2017).
37. Buniello, A. et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).
38. 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).
39. Forrest, A. R. et al. A promoter-level mammalian expression atlas. *Nature* **507**, 462–470 (2014).
40. Abascal, F. et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699–710 (2020).
41. Andersson, R. et al. An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455–461 (2014).
42. Fishilevich, S. et al. GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database* **2017**, bax028 (2017).
43. DiCorpo, D. et al. Whole genome sequence association analysis of fasting glucose and fasting insulin levels in diverse cohorts from the NHLBI TOPMed program. *Commun. Biol.* **5**, 756 (2022).
44. Jiang, M.-Z. et al. Whole genome sequencing based analysis of inflammation biomarkers in the Trans-Omics for Precision Medicine (TOPMed) consortium. *Human Mol. Genet* **33**, 1429–1441 (2024).
45. Dijk, W. et al. Identification of a gain-of-function LIPC variant as a novel cause of familial combined hypocholesterolemia. *Circulation* **146**, 724–739 (2022).
46. Ottensmann, L. et al. Genome-wide association analysis of plasma lipidome identifies 495 genetic associations. *Nat. Commun.* **14**, 6934 (2023).
47. Guo, T. et al. Association between the DOCK7, PCSK9 and GALNT2 gene polymorphisms and serum lipid levels. *Sci. Rep.* **6**, 19079 (2016).
48. Li, Z. et al. Dynamic scan procedure for detecting rare-variant association regions in whole-genome sequencing studies. *Am. J. Human Genet.* **104**, 802–814 (2019).
49. McCaw, Z. R., Gao, J., Lin, X. & Gronsbell, J. Synthetic surrogates improve power for genome-wide association studies of partially missing phenotypes in population biobanks. *Nat. Genet.* **56**, 1527–1536 (2024).
50. Li, X. et al. Powerful, scalable and resource-efficient meta-analysis of rare variant associations in large whole genome sequencing studies. *Nat. Genet.* **55**, 154–164 (2023).
51. Chen, H. et al. Control for population structure and relatedness for binary traits in genetic association studies via logistic mixed models. *Am. J. Human Genet.* **98**, 653–666 (2016).
52. 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. Human Genet.* **104**, 260–274 (2019).
53. 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).
54. Li, X. xihaoli/MultiSTAAR: MultiSTAAR_v0.9.7 (v0.9.7). Zenodo. https://doi.org/10.5281/zenodo.13955413 (2024).
55. Li, X. & Li, Z. xihaoli/STAARpipeline: STAARpipeline_v0.9.7 (v0.9.7). Zenodo. https://doi.org/10.5281/zenodo.10098313 (2023).
56. Li, X. xihaoli/STAAR: STAAR_v0.9.7 (v0.9.7). Zenodo. https://doi.org/10.5281/zenodo.10060210 (2023).
57. Li, X. & Li, Z. xihaoli/STAARpipelineSummary: STAARpipelineSummary_v0.9.7 (v0.9.7). Zenodo. https://doi.org/10.5281/zenodo.10113310 (2023).
58. Li, X. et al. Source data of the MultiSTAAR manuscript "A statistical framework for multi-trait rare variant analysis in large-scale whole-genome sequencing studies". [Data set]. Zenodo. https://doi.org/10.5281/zenodo.14213842 (2024).
59. Gazal, S. et al. Linkage disequilibrium–dependent architecture of human complex traits shows action of negative selection. *Nat. Genet.* **49**, 1421–1427 (2017).
60. Zhou, H. et al. FAVOR: functional annotation of variants online resource and annotator for variation across the human genome. *Nucleic Acids Res.* **51**, D1300–D1311 (2023).
61. Zhou, H., Arapoglou, T., Li, X., Li, Z. & Lin, X. FAVOR Essential Database. V1 edn (Harvard Dataverse, 2022).
62. Moors, J. et al. A Polynesian-specific missense CETP variant alters the lipid profile. *Human Genet. Genomics Adv.* **4**, 100204 (2023).

Acknowledgements

This work was supported by grants R35-CA197449, U19-CA203654, U01-HG012064 and U01-HG009088 (X. Lin), NHLBI TOPMed Fellowship 75N92021F00229 (X. Li and M.S.S.), 1R01AG086379-01 (Z. Liu), R01-HL142711 and R01-HL127564 (P.N. and G.M.P.), R00HG012956-02 (Z.Y.), 75N92020D00001, HHSN268201500003I, N01-HC-95159, 75N92020D00005, N01-HC-95160, 75N92020D00002, N01-HC-95161, 75N92020D00003, N01-HC-95162, 75N92020D00006, N01-HC-95163, 75N92020D00004, N01-HC-95164, 75N92020D00007, N01-HC-95165, N01-HC-95166, N01-HC-95167, N01-HC-95168, N01-HC-95169, UL1-TR-000040, UL1-TR-001079, UL1-TR-001420, UL1-TR001881, DK063491, R01-HL071051, R01-HL071205, R01-HL071250, R01-HL071251, R01-HL071258, R01-HL071259 and UL1-RR033176 (J.I.R.), HHSN268201800001I and U01-HL137162 (K.M.R.), DK078616 and HL151855 (J.B.M.), 1R35-HL135818, R01-HL113338 and HL046389 (S.R.), HL105756 (B.M.P.), HHSN268201600018C, HHSN268201600001C, HHSN268201600002C, HHSN268201600003C and HHSN268201600004C (C.K.), R01-MD012765 and R01-DK117445 (N.F.), R01-HL153805 and R03-HL154284 (B.E.C.), HHSN268201700001I, HHSN268201700002I, HHSN268201700003I, HHSN268201700005I and HHSN268201700004I (E.B.), U01-HL072524, R01-HL104135-04S1, U01-HL054472, U01-HL054473, U01-HL054495, U01-HL054509 and R01-HL055673-18S1 (D.K.A.) and U01-HL72518, HL087698, HL49762, HL59684, HL58625, HL071025, HL112064, NR0224103 and M01-RR000052 (to the Johns Hopkins General Clinical Research Center). The Diabetes Heart Study (DHS) was supported by R01 HL92301, R01 HL67348, R01 NS058700, R01 AR48797, R01 DK071891 and R01 AG058921, the General Clinical Research Center of the Wake Forest University School of Medicine (M01 RR07122, F32 HL085989), the American Diabetes Association and a pilot grant from the Claude Pepper Older Americans Independence Center of Wake Forest University Health Sciences (P60 AG10484). The Framingham Heart Study (FHS) acknowledges the support of contracts NO1-HC-25195, HHSN268201500001I, 75N92019D00031, 1R01HL064753, R01HL076784 and 1R01AG028321 from the National Heart, Lung and Blood Institute and grant supplement R01 HL092577-06S1 for

this research. We also acknowledge the dedication of the FHS study participants, without whom this research would not be possible. R.S.V. is supported in part by the Evans Medical Foundation and the Jay and Louis Coffman Endowment from the Department of Medicine, Boston University Chobanian & Avedisian School of Medicine. The Jackson Heart Study (JHS) is supported and conducted in collaboration with Jackson State University (HHSN268201800013I), Tougaloo College (HHSN268201800014I), the Mississippi State Department of Health (HHSN268201800015I) and the University of Mississippi Medical Center (HHSN268201800010I, HHSN268201800011I and HHSN268201800012I) contracts from the National Heart, Lung and Blood Institute (NHLBI) and the National Institute on Minority Health and Health Disparities (NIMHD). We also thank the staff and participants of the JHS. Support for GENOA was provided by the NHLBI (U01HL054457, U01HL054464, U01HL054481, R01HL119443 and R01HL087660) of the National Institutes of Health (NIH). Collection of the San Antonio Family Study data was supported in part by NIH grants P01 HL045522, MH078143, MH078111 and MH083824, and whole-genome sequencing of SAFS subjects was supported by U01 DK085524 and R01 HL113323. Molecular data for the Trans-Omics in Precision Medicine (TOPMed) program was supported by the NHLBI. Core support, including centralized genomic read mapping and genotype calling, along with variant quality metrics and filtering were provided by the TOPMed Informatics Research Center (3R01HL-117626-02S1; contract no. HHSN268201800002I). Core support, including phenotype harmonization, data management, sample-identity quality control and general program coordination were provided by the TOPMed Data Coordinating Center (R01HL-120393; U01HL-120393; 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 Information.

Author contributions

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

Competing interests

Z.R.M. and R.D. are employees of Insitro. S.M.G. is an employee of Regeneron Genetics Center. M.E.M. receives research funding from Regeneron Pharmaceutical Inc., unrelated to this project. B.M.P. serves on the Steering Committee of the Yale Open Data Access Project funded by Johnson & Johnson. L.M.R. and S.S.R. are consultants for the TOPMed Administrative Coordinating Center (via Westat). P.N. reports research grants from Allelica, Amgen, Apple, Boston Scientific, Genentech/Roche and Novartis, personal fees from Allelica, Apple, AstraZeneca, Blackstone Life Sciences, Creative Education Concepts, CRISPR Therapeutics, Eli Lilly & Co, Esperion Therapeutics, Foresite Capital, Foresite Labs, Genentech/Roche, GV, HeartFlow, Magnet Biomedicine, Merck, Novartis, TenSixteen Bio and Tourmaline Bio, equity in Bolt, Candela, Mercury, MyOme, Parameter Health, Preciseli and TenSixteen Bio, and spousal employment at Vertex Pharmaceuticals, all unrelated to the present work. X. Lin is a consultant of AbbVie Pharmaceuticals and Verily Life Sciences. The other authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s43588-024-00764-8.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43588-024-00764-8.

Correspondence and requests for materials should be addressed to Zilin Li, Zhonghua Liu or Xihong Lin.

Peer review information *Nature Computational Science* thanks Yuehua Cui, Yukinori Okada and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editor: Ananya Rastogi, in collaboration with the *Nature Computational Science* team. Peer reviewer reports are available.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Xihao Li[1,2], Han Chen[3], Margaret Sunitha Selvaraj[4,5,6], Eric Van Buren[7], Hufeng Zhou[7], Yuxuan Wang[8], Ryan Sun[9], Zachary R. McCaw[1], Zhi Yu[4,5,10], Min-Zhi Jiang[2,11], Daniel DiCorpo[8], Sheila M. Gaynor[7], Rounak Dey[7], Donna K. Arnett[12], Emelia J. Benjamin[13,14,15], Joshua C. Bis[16], John Blangero[17], Eric Boerwinkle[3,18], Donald W. Bowden[19], Jennifer A. Brody[16], Brian E. Cade[5,20,21], April P. Carson[22], Jenna C. Carlson[23], Nathalie Chami[24,25], Yii-Der Ida Chen[26], Joanne E. Curran[17], Paul S. de Vries[3], Myriam Fornage[3,27], Nora Franceschini[28], Barry I. Freedman[29], Charles Gu[30], Nancy L. Heard-Costa[15,31], Jiang He[32,33], Lifang Hou[34], Yi-Jen Hung[35], Marguerite R. Irvin[36], Robert C. Kaplan[37,38], Sharon L. R. Kardia[39], Tanika N. Kelly[40], Iain Konigsberg[41], Charles Kooperberg[40], Brian G. Kral[42], Changwei Li[32,33], Yun Li[1,2], Honghuang Lin[43], Ching-Ti Liu[8], Ruth J. F. Loos[24,44], Michael C. Mahaney[17], Lisa W. Martin[45], Rasika A. Mathias[42], Braxton D. Mitchell[46], May E. Montasser[46], Alanna C. Morrison[3], Take Naseri[47,48], Kari E. North[28], Nicholette D. Palmer[19], Patricia A. Peyser[39], Bruce M. Psaty[16,49,50], Susan Redline[20,21], Alexander P. Reiner[38,49], Stephen S. Rich[51], Colleen M. Sitlani[16], Jennifer A. Smith[39], Kent D. Taylor[26], Hemant K. Tiwari[52], Ramachandran S. Vasan[15,53],

Satupa'itea Viali[54,55,56], Zhe Wang[24], Jennifer Wessel[57,58], Lisa R. Yanek[42], Bing Yu[3], NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium*, Josée Dupuis[8,59], James B. Meigs[5,6,60], Paul L. Auer[61], Laura M. Raffield[2], Alisa K. Manning[6,62,63], Kenneth M. Rice[64], Jerome I. Rotter[26], Gina M. Peloso[8], Pradeep Natarajan[4,5,6], Zilin Li[7] ✉, Zhonghua Liu[65] ✉ & Xihong Lin[5,7,66] ✉

[1]Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. [2]Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. [3]Human Genetics Center, Department of Epidemiology, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX, USA. [4]Center for Genomic Medicine and Cardiovascular Research Center, Massachusetts General Hospital, Boston, MA, USA. [5]Program in Medical and Population Genetics, Broad Institute of Harvard and MIT, Cambridge, MA, USA. [6]Department of Medicine, Harvard Medical School, Boston, MA, USA. [7]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA. [8]Department of Biostatistics, Boston University School of Public Health, Boston, MA, USA. [9]Department of Biostatistics, University of Texas MD Anderson Cancer Center, Houston, TX, USA. [10]Clinical and Translational Epidemiology Unit, Department of Medicine, Massachusetts General Hospital, Boston, MA, USA. [11]Department of Biostatistics, The Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA. [12]Provost Office, University of South Carolina, Columbia, SC, USA. [13]Section of Cardiovascular Medicine, Boston Medical Center, Boston University Chobanian & Avedisian School of Medicine, Boston, MA, USA. [14]Department of Epidemiology, Boston University School of Public Health, Boston, MA, USA. [15]Framingham Heart Study, Framingham, MA, USA. [16]Cardiovascular Health Research Unit, Department of Medicine, University of Washington, Seattle, WA, USA. [17]Department of Human Genetics and South Texas Diabetes and Obesity Institute, School of Medicine, The University of Texas Rio Grande Valley, Brownsville, TX, USA. [18]Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX, USA. [19]Department of Biochemistry, Wake Forest University School of Medicine, Winston-Salem, NC, USA. [20]Division of Sleep and Circadian Disorders, Brigham and Women's Hospital, Boston, MA, USA. [21]Division of Sleep Medicine, Harvard Medical School, Boston, MA, USA. [22]Department of Medicine, University of Mississippi Medical Center, Jackson, MS, USA. [23]Department of Human Genetics and Department of Biostatistics and Health Data Science, University of Pittsburgh, Pittsburgh, PA, USA. [24]The Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [25]The Mindich Child Health and Development Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [26]The Institute for Translational Genomics and Population Sciences, Department of Pediatrics, The Lundquist Institute for Biomedical Innovation at Harbor-UCLA Medical Center, Torrance, CA, USA. [27]Brown Foundation Institute of Molecular Medicine, McGovern Medical School, The University of Texas Health Science Center at Houston, Houston, TX, USA. [28]Department of Epidemiology, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. [29]Department of Internal Medicine, Nephrology, Wake Forest University School of Medicine, Winston-Salem, NC, USA. [30]Division of Biology & Biomedical Sciences, Washington University School of Medicine, St. Louis, MO, USA. [31]Department of Neurology, Boston University Chobanian & Avedisian School of Medicine, Boston, MA, USA. [32]Department of Epidemiology, Tulane University School of Public Health and Tropical Medicine, New Orleans, LA, USA. [33]Translational Science Institute, Tulane University, New Orleans, LA, USA. [34]Department of Preventive Medicine, Northwestern University, Chicago, IL, USA. [35]Department of Internal Medicine, Tri-Service General Hospital, National Defense Medical Center, Taipei, Taiwan. [36]Department of Epidemiology, School of Public Health, University of Alabama at Birmingham, Birmingham, AL, USA. [37]Department of Epidemiology and Population Health, Albert Einstein College of Medicine, Bronx, NY, USA. [38]Division of Public Health Sciences, Fred Hutchinson Cancer Center, Seattle, WA, USA. [39]Department of Epidemiology, School of Public Health, University of Michigan, Ann Arbor, MI, USA. [40]Department of Medicine, Division of Nephrology, University of Illinois Chicago, Chicago, IL, USA. [41]Department of Biomedical Informatics, University of Colorado, Aurora, CO, USA. [42]Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, MD, USA. [43]Department of Medicine, University of Massachusetts Chan Medical School, Worcester, MA, USA. [44]Novo Nordisk Foundation Center for Basic Metabolic Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark. [45]School of Medicine and Health Sciences, George Washington University, Washington, DC, USA. [46]Department of Medicine, University of Maryland School of Medicine, Baltimore, MD, USA. [47]Naseri & Associates Public Health Consultancy Firm and Family Health Clinic, Apia, Samoa. [48]Department of Epidemiology, Brown University, Providence, RI, USA. [49]Departments of Epidemiology, University of Washington, Seattle, WA, USA. [50]Department of Health Systems and Population Health, University of Washington, Seattle, WA, USA. [51]Department of Genome Sciences, University of Virginia, Charlottesville, VA, USA. [52]Department of Biostatistics, School of Public Health, University of Alabama at Birmingham, Birmingham, AL, USA. [53]Department of Quantitative and Qualitative Health Sciences, UT Health San Antonio School of Public Health, San Antonia, TX, USA. [54]School of Medicine, National University of Samoa, Apia, Samoa. [55]Department of Chronic Disease Epidemiology, Yale University School of Public Health, New Haven, CT, USA. [56]Oceania University of Medicine, Apia, Samoa. [57]Department of Epidemiology, Fairbanks School of Public Health, Indiana University, Indianapolis, IN, USA. [58]Diabetes Translational Research Center, Indiana University, Indianapolis, IN, USA. [59]Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, QC, Canada. [60]Division of General Internal Medicine, Massachusetts General Hospital, Boston, MA, USA. [61]Division of Biostatistics, Data Science Institute, and Cancer Center, Medical College of Wisconsin, Milwaukee, WI, USA. [62]Metabolism Program, The Broad Institute of MIT and Harvard, Cambridge, MA, USA. [63]Clinical and Translational Epidemiology Unit, Mongan Institute, Massachusetts General Hospital, Boston, MA, USA. [64]Department of Biostatistics, University of Washington, Seattle, WA, USA. [65]Department of Biostatistics, Mailman School of Public Health, Columbia University, New York, NY, USA. [66]Department of Statistics, Harvard University, Cambridge, MA, USA. *A list of authors and their affiliations appears at the end of the paper. ✉e-mail: li@hsph.harvard.edu; zl2509@cumc.columbia.edu; xlin@hsph.harvard.edu

NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium

Namiko Abe[67], Gonçalo Abecasis[68], Francois Aguet[69], Christine Albert[70], Laura Almasy[71], Alvaro Alonso[72], Seth Ament[73], Peter Anderson[74], Pramod Anugu[75], Deborah Applebaum-Bowden[76], Kristin Ardlie[69], Dan Arking[77], Allison Ashley-Koch[78], Stella Aslibekyan[79], Tim Assimes[80], Dimitrios Avramopoulos[77], Najib Ayas[81], Adithya Balasubramanian[18], John Barnard[82], Kathleen Barnes[83], R. Graham Barr[84], Emily Barron-Casella[77], Lucas Barwick[85], Terri Beaty[77], Gerald Beck[86], Diane Becker[87], Lewis Becker[77], Rebecca Beer[88], Amber Beitelshees[73], Takis Benos[89], Marcos Bezerra[90], Larry Bielak[68], Thomas Blackwell[68], Nathan Blue[91], Russell Bowler[92], Ulrich Broeckel[93], Jai Broome[74], Deborah Brown[94], Karen Bunting[67], Esteban Burchard[95], Carlos Bustamante[96], Erin Buth[64],

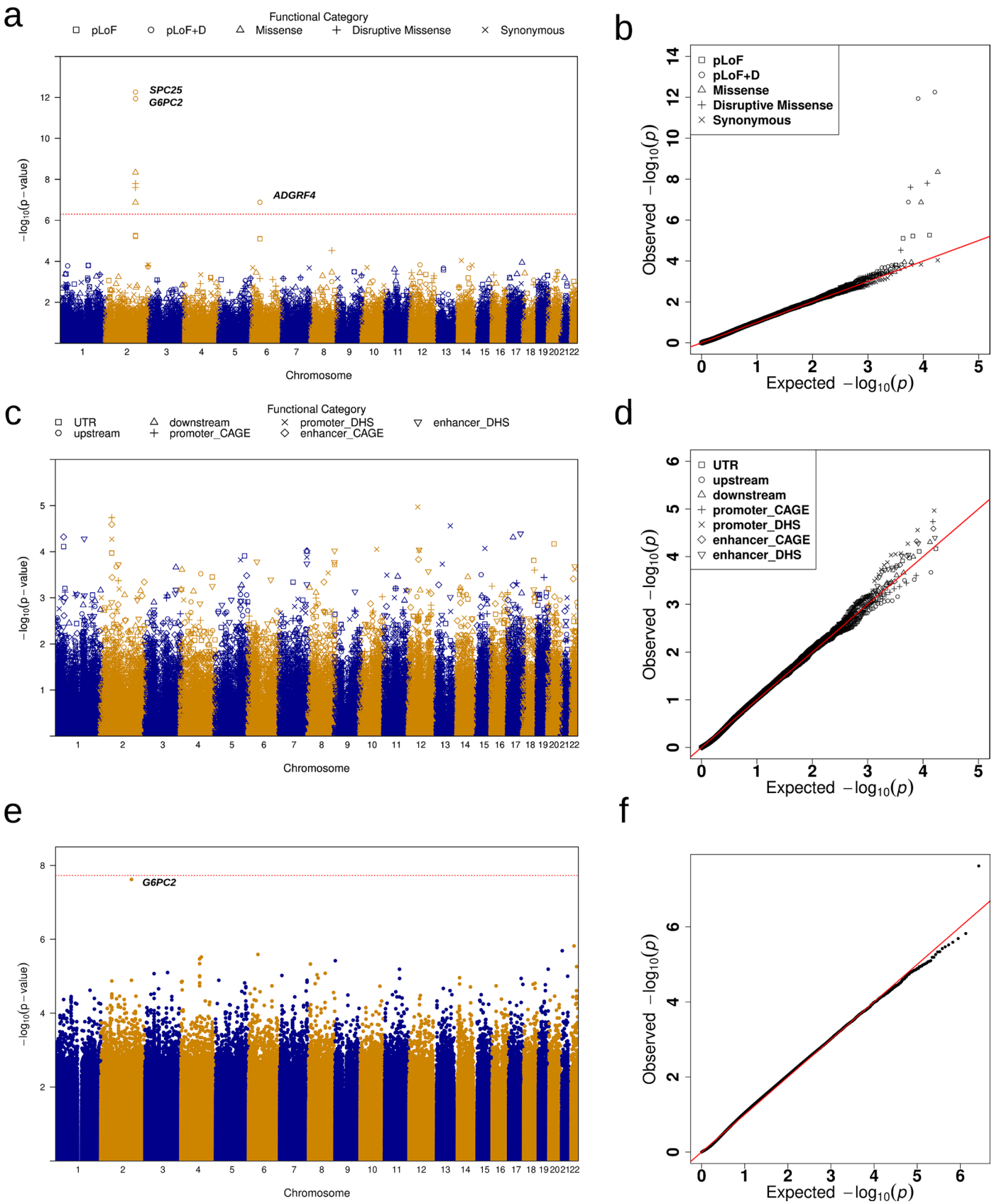
Jonathan Cardwell[97], Vincent Carey[98], Julie Carrier[99], Cara Carty[100], Richard Casaburi[101], Juan P. Casas Romero[98], James Casella[77], Peter Castaldi[102], Mark Chaffin[69], Christy Chang[73], Yi-Cheng Chang[103], Daniel Chasman[104], Sameer Chavan[97], Bo-Juen Chen[67], Wei-Min Chen[105], Michael Cho[98], Seung Hoan Choi[69], Lee-Ming Chuang[106], Mina Chung[82], Ren-Hua Chung[107], Clary Clish[108], Suzy Comhair[109], Matthew Conomos[64], Elaine Cornell[110], Adolfo Correa[111], Carolyn Crandall[101], James Crapo[112], L. Adrienne Cupples[113], Jeffrey Curtis[114], Brian Custer[115], Coleen Damcott[73], Dawood Darbar[116], Sean David[117], Colleen Davis[74], Michelle Daya[97], Michael DeBaun[118], Dawn DeMeo[98], Ranjan Deka[119], Scott Devine[73], Huyen Dinh[18], Harsha Doddapaneni[18], Qing Duan[120], Shannon Dugan-Perez[18], Ravi Duggirala[121], Jon Peter Durda[122], Susan K. Dutcher[123], Charles Eaton[124], Lynette Ekunwe[75], Adel El Boueiz[125], Patrick T. Ellinor[126], Leslie Emery[74], Serpil Erzurum[127], Charles Farber[105], Jesse Farek[18], Tasha Fingerlin[128], Matthew Flickinger[68], Chris Frazar[74], Mao Fu[73], Stephanie M. Fullerton[74], Lucinda Fulton[129], Stacey Gabriel[69], Weiniu Gan[88], Shanshan Gao[97], Yan Gao[75], Margery Gass[130], Heather Geiger[131], Bruce Gelb[132], Mark Geraci[133], Soren Germer[67], Robert Gerszten[134], Auyon Ghosh[98], Richard Gibbs[18], Chris Gignoux[80], Mark Gladwin[133], David Glahn[135], Stephanie Gogarten[74], Da-Wei Gong[73], Harald Goring[136], Sharon Graw[137], Kathryn J. Gray[138], Daniel Grine[97], Colin Gross[68], Yue Guan[73], Xiuqing Guo[139], Namrata Gupta[69], Jeff Haessler[130], Michael Hall[140], Yi Han[18], Patrick Hanly[141], Daniel Harris[142], Nicola L. Hawley[143], Ben Heavner[64], Susan Heckbert[144], Ryan Hernandez[95], David Herrington[145], Craig Hersh[146], Bertha Hidalgo[79], James Hixson[147], Brian Hobbs[148], John Hokanson[97], Elliott Hong[73], Karin Hoth[149], Chao Agnes Hsiung[150], Jianhong Hu[18], Haley Huston[151], Chii Min Hwu[152], Rebecca Jackson[153], Deepti Jain[74], Cashell Jaquish[88], Jill Johnsen[154], Andrew Johnson[88], Craig Johnson[74], Rich Johnston[72], Kimberly Jones[77], Hyun Min Kang[155], Shannon Kelly[95], Eimear Kenny[132], Michael Kessler[73], Alyna Khan[74], Ziad Khan[18], Wonji Kim[156], John Kimoff[157], Greg Kinney[158], Barbara Konkle[159], Holly Kramer[160], Christoph Lange[161], Ethan Lange[97], Leslie Lange[162], Cathy Laurie[74], Cecelia Laurie[74], Meryl LeBoff[98], Jonathon LeFaive[68], Jiwon Lee[98], Sandra Lee[18], Wen-Jane Lee[152], David Levine[74], Daniel Levy[88], Joshua Lewis[73], Xiaohui Li[139], Henry Lin[139], Simin Liu[163], Yongmei Liu[164], Yu Liu[165], Steven A. Lubitz[126], Kathryn L. Lunetta[166], James Luo[88], Ulysses Magalang[167], Barry Make[77], Ani Manichaikul[105], JoAnn Manson[98], Melissa Marton[131], Susan Mathai[97], Susanne May[64], Patrick McArdle[73], Merry-Lynn McDonald[168], Sean McFarland[156], Stephen McGarvey[48], Daniel McGoldrick[169], Caitlin McHugh[64], Becky McNeil[170], Hao Mei[75], Vipin Menon[18], Luisa Mestroni[137], Ginger Metcalf[18], Deborah A. Meyers[171], Emmanuel Mignot[172], Julie Mikulla[88], Nancy Min[75], Mollie Minear[173], Ryan L. Minster[133], Matt Moll[102], Zeineen Momin[18], Courtney Montgomery[174], Donna Muzny[18], Josyf C. Mychaleckyj[105], Girish Nadkarni[132], Rakhi Naik[77], Sergei Nekhai[175], Sarah C. Nelson[64], Bonnie Neltner[97], Caitlin Nessner[18], Deborah Nickerson[169], Osuji Nkechinyere[18], Jeff O'Connell[176], Tim O'Connor[73], Heather Ochs-Balcom[177], Geoffrey Okwuonu[18], Allan Pack[178], David T. Paik[179], James Pankow[180], George Papanicolaou[88], Cora Parker[181], Juan Manuel Peralta[121], Marco Perez[80], James Perry[73], Ulrike Peters[182], Lawrence S. Phillips[72], Jacob Pleiness[68], Toni Pollin[73], Wendy Post[183], Julia Powers Becker[184], Meher Preethi Boorgula[97], Michael Preuss[132], Pankaj Qasba[88], Dandi Qiao[98], Zhaohui Qin[72], Nicholas Rafaels[185], Mahitha Rajendran[18], D. C. Rao[129], Laura Rasmussen-Torvik[186], Aakrosh Ratan[105], Robert Reed[73], Catherine Reeves[131], Elizabeth Regan[187], Muagututi'a Sefuiva Reupena[188], Rebecca Robillard[189], Nicolas Robine[131], Dan Roden[190], Carolina Roselli[69], Ingo Ruczinski[77], Alexi Runnels[131], Pamela Russell[97], Sarah Ruuska[74], Kathleen Ryan[73], Ester Cerdeira Sabino[191], Danish Saleheen[84], Shabnam Salimi[192], Sejal Salvi[18], Steven Salzberg[77], Kevin Sandow[193], Vijay G. Sankaran[194], Jireh Santibanez[18], Karen Schwander[129], David Schwartz[97], Frank Sciurba[133], Christine Seidman[195], Jonathan Seidman[196], Vivien Sheehan[197], Stephanie L. Sherman[198], Amol Shetty[73], Aniket Shetty[97], Wayne Hui-Heng Sheu[152], M. Benjamin Shoemaker[199], Brian Silver[200], Edwin Silverman[98], Robert Skomro[201], Albert Vernon Smith[68], Josh Smith[74], Nicholas Smith[144], Tanja Smith[67], Sylvia Smoller[202], Beverly Snively[203], Michael Snyder[204], Tamar Sofer[134], Nona Sotoodehnia[74], Adrienne M. Stilp[74], Garrett Storm[205], Elizabeth Streeten[73], Jessica Lasky Su[206], Yun Ju Sung[129], Jody Sylvia[98], Adam Szpiro[74], Frédéric Sériès[207], Daniel Taliun[68], Hua Tang[204], Margaret Taub[77], Matthew Taylor[137], Simeon Taylor[73], Marilyn Telen[78], Timothy A. Thornton[74], Machiko Threlkeld[169], Lesley Tinker[208], David Tirschwell[74], Sarah Tishkoff[209], Catherine Tong[64], Russell Tracy[122], Michael Tsai[180], Dhananjay Vaidya[77], David Van Den Berg[210], Peter VandeHaar[68], Scott Vrieze[180], Tarik Walker[97], Robert Wallace[149], Avram Walts[97], Fei Fei Wang[74], Heming Wang[211], Jiongming Wang[68], Karol Watson[101], Jennifer Watt[18], Daniel E. Weeks[212], Joshua Weinstock[155], Bruce Weir[74], Scott T. Weiss[213], Lu-Chen Weng[126], Cristen Willer[114], Kayleen Williams[64], L. Keoki Williams[214], Scott Williams[215], Carla Wilson[98], James Wilson[216], Lara Winterkorn[131], Quenna Wong[74], Baojun Wu[217], Joseph Wu[179], Huichun Xu[73], Ivana Yang[97], Ketian Yu[68], Seyedeh Maryam Zekavat[69], Yingze Zhang[218], Snow Xueyan Zhao[112], Wei Zhao[219], Xiaofeng Zhu[220], Elad Ziv[221], Michael Zody[67], Sebastian Zoellner[68], Mariza de Andrade[222] & Lisa de las Fuentes[223]

[67]New York Genome Center, New York, NY, USA. [68]University of Michigan, Ann Arbor, MI, USA. [69]Broad Institute, Cambridge, MA, USA. [70]Cedars-Sinai Medical Center, Boston, MA, USA. [71]Children's Hospital of Philadelphia, University of Pennsylvania, Philadelphia, PA, USA. [72]Emory University, Atlanta, GA, USA. [73]University of Maryland, Baltimore, MD, USA. [74]University of Washington, Seattle, WA, USA. [75]University of Mississippi, Jackson, MS, USA. [76]National Institutes of Health, Bethesda, MD, USA. [77]Johns Hopkins University, Baltimore, MD, USA. [78]Duke University, Durham, NC, USA. [79]University of Alabama,

Birmingham, AL, USA. [80]Stanford University, Stanford, CA, USA. [81]Department of Medicine, Providence Health Care, Vancouver, British Columbia, Canada. [82]Cleveland Clinic, Cleveland, OH, USA. [83]Tempus, University of Colorado Anschutz Medical Campus, Aurora, CO, USA. [84]Columbia University, New York, NY, USA. [85]The Emmes Corporation, LTRC, Rockville, MD, USA. [86]Cleveland Clinic, Quantitative Health Sciences, Cleveland, OH, USA. [87]Department of Medicine, Johns Hopkins University, Baltimore, MD, USA. [88]National Heart, Lung, and Blood Institute, National Institutes of Health, Bethesda, MD, USA. [89]Department of Epidemiology, University of Florida, Gainesville, FL, USA. [90]Fundação de Hematologia e Hemoterapia de PernambucoHemope, Recife, Brazil. [91]Department of Obstetrics and Gynecology, University of Utah, Salt Lake City, UT, USA. [92]National Jewish Health, National Jewish Health, Denver, CO, USA. [93]Department of Pediatrics, Medical College of Wisconsin, Milwaukee, WI, USA. [94]Department of Pediatrics, University of Texas Health at Houston, Houston, TX, USA. [95]University of California at San Francisco, San Francisco, CA, USA. [96]Department of Biomedical Data Science, Stanford University, Stanford, CA, USA. [97]University of Colorado at Denver, Denver, CO, USA. [98]Brigham & Women's Hospital, Boston, MA, USA. [99]University of Montreal, Montreal, Canada. [100]Washington State University at Pullman, Pullman, WA, USA. [101]University of California at Los Angeles, Los Angeles, CA, USA. [102]Department of Medicine, Brigham & Women's Hospital, Boston, MA, USA. [103]National Taiwan University, Taipei, Taiwan. [104]Division of Preventive Medicine, Brigham & Women's Hospital, Boston, MA, USA. [105]University of Virginia, Charlottesville, VA, USA. [106]National Taiwan University Hospital, National Taiwan University, Taipei, Taiwan. [107]National Health Research Institute Taiwan, Miaoli County, Taiwan. [108]Metabolomics Platform, Broad Institute, Cambridge, MA, USA. [109]Department of Immunity and Immunology, Cleveland Clinic, Cleveland, OH, USA. [110]University of Vermont, Burlington, VT, USA. [111]Department of Population Health Science, University of Mississippi, Jackson, MS, USA. [112]National Jewish Health, Denver, CO, USA. [113]Department of Biostatistics, Boston University, Boston, MA, USA. [114]Department of Internal Medicine, University of Michigan, Ann Arbor, MI, USA. [115]Vitalant Research Institute, San Francisco, CA, USA. [116]University of Illinois at Chicago, Chicago, IL, USA. [117]University of Chicago, Chicago, IL, USA. [118]Vanderbilt University, Nashville, TN, USA. [119]University of Cincinnati, Cincinnati, OH, USA. [120]University of North Carolina, Chapel Hill, NC, USA. [121]University of Texas Rio Grande Valley School of Medicine, Edinburg, TX, USA. [122]Department of Pathology and Laboratory Medicine, University of Vermont, Burlington, VT, USA. [123]Department of Genetics, Washington University in St Louis, St Louis, MO, USA. [124]Brown University, Providence, RI, USA. [125]Channing Division of Network Medicine, Harvard University, Cambridge, MA, USA. [126]Massachusetts General Hospital, Boston, MA, USA. [127]Lerner Research Institute, Cleveland Clinic, Cleveland, OH, USA. [128]Center for Genes, Environment and Health, National Jewish Health, Denver, CO, USA. [129]Washington University in St Louis, St Louis, MO, USA. [130]Fred Hutchinson Cancer Research Center, Seattle, WA, USA. [131]New York Genome Center, New York City, NY, USA. [132]Icahn School of Medicine at Mount Sinai, New York, NY, USA. [133]University of Pittsburgh, Pittsburgh, PA, USA. [134]Beth Israel Deaconess Medical Center, Boston, MA, USA. [135]Department of Psychiatry, Boston Children's Hospital, Harvard Medical School, Boston, MA, USA. [136]University of Texas Rio Grande Valley School of Medicine, San Antonio, TX, USA. [137]University of Colorado Anschutz Medical Campus, Aurora, CO, USA. [138]Department of Obstetrics and Gynecology, Mass General Brigham, Boston, MA, USA. [139]Lundquist Institute, Torrance, CA, USA. [140]Department of Cardiology, University of Mississippi, Jackson, MS, USA. [141]Department of Medicine, University of Calgary, Calgary, Canada. [142]Department of Genetics, University of Maryland, Philadelphia, PA, USA. [143]Department of Chronic Disease Epidemiology, Yale University, New Haven, CT, USA. [144]Department of Epidemiology, University of Washington, Seattle, WA, USA. [145]Wake Forest Baptist Health, Winston-Salem, NC, USA. [146]Channing Division of Network Medicine, Brigham & Women's Hospital, Boston, MA, USA. [147]University of Texas Health at Houston, Houston, TX, USA. [148]Regeneron Genetics Center, Boston, MA, USA. [149]University of Iowa, Iowa City, IO, USA. [150]Institute of Population Health Sciences, National Health Research Institute Taiwan, Miaoli County, Taiwan. [151]Blood Works Northwest, Seattle, WA, USA. [152]Taichung Veterans General Hospital Taiwan, Taichung City, Taiwan. [153]Divisions of Endocrinology, Diabetes and Metabolism and Internal Medicine, Oklahoma State University Medical Center, Columbus, OH, USA. [154]Department of Medicine, University of Washington, Seattle, WA, USA. [155]Department of Biostatistics, University of Michigan, Ann Arbor, MI, USA. [156]Harvard University, Cambridge, MA, USA. [157]McGill University, Montréal, Quebec, Canada. [158]Department of Epidemiology, University of Colorado at Denver, Aurora, CO, USA. [159]Department of Medicine, Blood Works Northwest, Seattle, WA, USA. [160]Department of Public Health Sciences, Loyola University, Maywood, IL, USA. [161]Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA. [162]Department of Medicine, University of Colorado at Denver, Aurora, CO, USA. [163]Department of Epidemiology and Medicine, Brown University, Providence, RI, USA. [164]Department of Cardiology, Duke University, Durham, NC, USA. [165]Cardiovascular Institute, Stanford University, Stanford, CA, USA. [166]Boston University, Boston, MA, USA. [167]Department of Critical Care and Sleep Medicine, Division of Pulmonary, The Ohio State University, Columbus, OH, USA. [168]University of Alabama at Birmingham, Birmingham, AL, USA. [169]Department of Genome Sciences, University of Washington, Seattle, WA, USA. [170]RTI International, Research Triangle Park, NC, USA. [171]University of Arizona, Tucson, AZ, USA. [172]Center For Sleep Sciences and Medicine, Stanford University, Palo Alto, CA, USA. [173]National Institute of Child Health and Human Development, National Institutes of Health, Bethesda, MD, USA. [174]Department of Genes and Human Disease, Oklahoma Medical Research Foundation, Oklahoma City, OK, USA. [175]Howard University, Washington, DC, USA. [176]University of Maryland, Balitmore, MD, USA. [177]University at Buffalo, Buffalo, NY, USA. [178]Division of Sleep Medicine/Department of Medicine, University of Pennsylvania, Philadelphia, PA, USA. [179]Stanford Cardiovascular Institute, Stanford University, Stanford, CA, USA. [180]University of Minnesota, Minneapolis, MN, USA. [181]Biostatistics and Epidemiology Division, RTI International, Research Triangle Park, NC, USA. [182]Fred Hutchinson Cancer Research Center and UW Medicine, Seattle, WA, USA. [183]Departments of Cardiology and Medicine, Johns Hopkins University, Baltimore, MD, USA. [184]Department of Medicine, University of Colorado at Denver, Denver, CO, USA. [185]Colorado Center for Personalized Medicine, University of Colorado at Denver, Denver, CO, USA. [186]Northwestern University, Chicago, IL, USA. [187]Department of Medicine, National Jewish Health, Denver, CO, USA. [188]Lutia I Puava Ae Mapu I Fagalele, Apia, Samoa. [189]Sleep Research Unit and Institute for Mental Health Research, University of Ottawa, Ottawa, Ontario, Canada. [190]Departments of Medicine, Pharmacology and Biomedical Informatics, Vanderbilt University, Nashville, TN, USA. [191]Faculdade de Medicina, Universidade de Sao Paulo, Sao Paulo, Brazil. [192]Department of Pathology, University of Maryland, Seattle, WA, USA. [193]Lundquist Institute, TGPS, Torrance, CA, USA. [194]Division of Hematology and Oncology, Harvard University, Boston, MA, USA. [195]Department of Genetics, Harvard Medical School, Boston, MA, USA. [196]Harvard Medical School, Boston, MA, USA. [197]Department of Pediatrics, Emory University, Atlanta, GA, USA. [198]Department of Human Genetics, Emory University, Atlanta, GA, USA. [199]Departments of Medicine and Cardiology, Vanderbilt University, Nashville, TN, USA. [200]UMass Memorial Medical Center, Worcester, MA, USA. [201]University of Saskatchewan, Saskatoon, Saskatchewan, Canada. [202]Albert Einstein College of Medicine, New York, NY, USA. [203]Department of Biostatistical Sciences, Wake Forest Baptist Health, Winston-Salem, NC, USA. [204]Department of Genetics, Stanford University, Stanford, CA, USA. [205]Department of Genomic Cardiology, University of Colorado at Denver, Aurora, CO, USA. [206]Channing Department of Medicine, Brigham & Women's Hospital, Boston, MA, USA. [207]Université Laval, Quebec City, Quebec, Canada. [208]Cancer Prevention Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA, USA. [209]Department of Genetics, University of Pennsylvania, Philadelphia, PA, USA. [210]USC Methylation Characterization Center, University of Southern California, Los Angeles, CA, USA. [211]Brigham & Women's Hospital, Mass General Brigham, Boston, MA, USA. [212]Department of Human Genetics, University of Pittsburgh, Pittsburgh, PA, USA. [213]Department of Medicine and Channing Division of Network Medicine, Brigham & Women's Hospital, Boston, MA, USA. [214]Henry Ford Health System, Detroit, MI, USA. [215]Case Western Reserve University, Cleveland, OH, USA. [216]Department of Cardiology, Beth Israel Deaconess Medical Center, Cambridge, MA, USA. [217]Department of Medicine, Henry Ford Health

System, Detroit, MA, USA. [218]Department of Medicine, University of Pittsburgh, Pittsburgh, PA, USA. [219]Department of Epidemiology, University of Michigan, Ann Arbor, MI, USA. [220]Department of Population and Quantitative Health Sciences, Case Western Reserve University, Cleveland, OH, USA. [221]Department of Medicine, University of California at San Francisco, San Francisco, CA, USA. [222]Health Quantitative Sciences Research, Mayo Clinic, Rochester, MI, USA. [223]Cardiovascular Division, Department of Medicine, Washington University in St Louis, St Louis, MO, USA.

https://doi.org/10.1038/s43588-024-00764-8

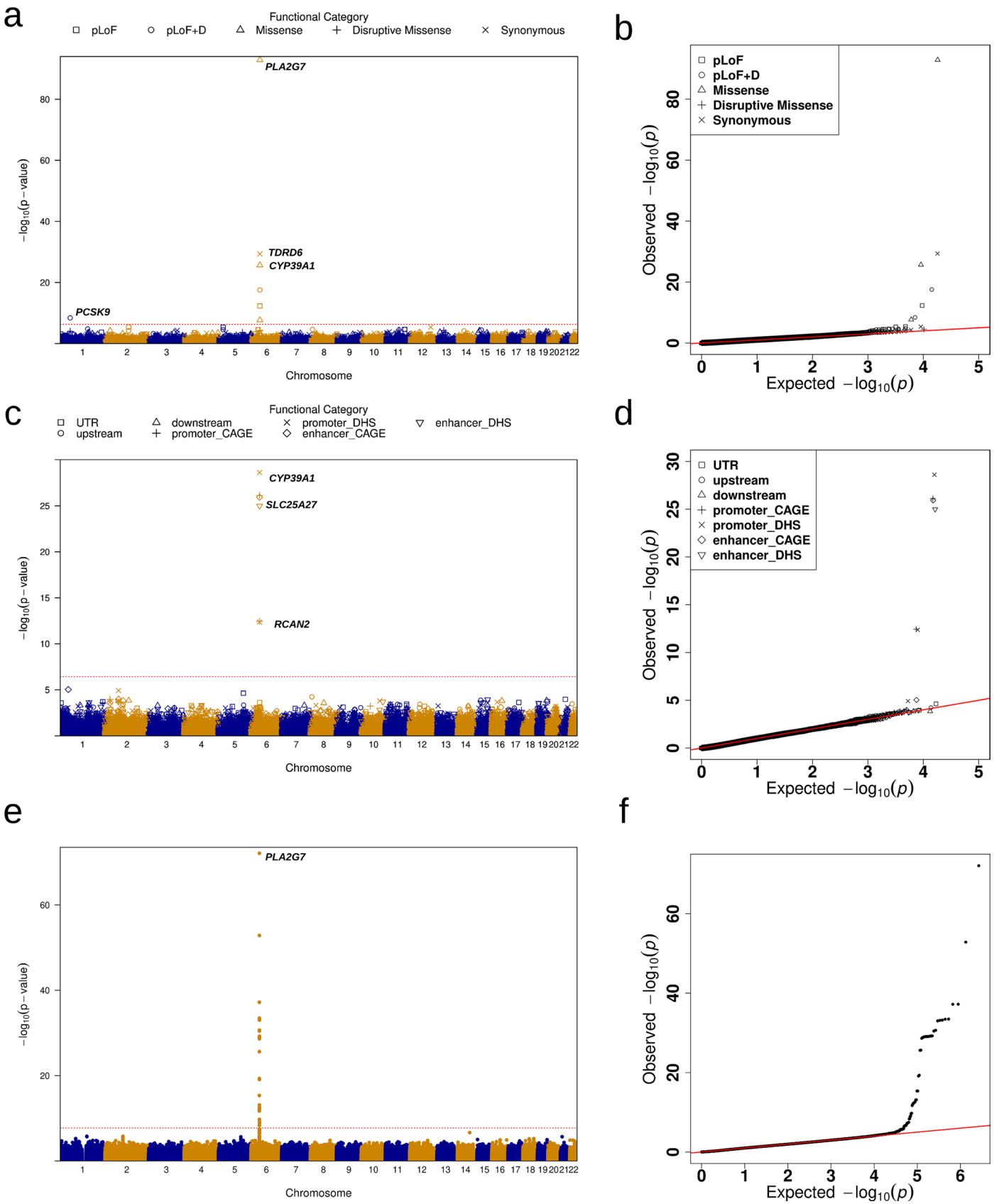


Extended Data Fig. 1 | See next page for caption.

https://doi.org/10.1038/s43588-024-00764-8

Extended Data Fig. 1 | Manhattan plots and Q-Q plots for unconditional gene-centric coding, noncoding and genetic region (2-kb sliding window) multi-trait analysis of fasting glucose (FG) and fasting insulin (FI) using TOPMed data (*n* = 21,731). a, Manhattan plots for unconditional gene-centric coding analysis of protein-coding genes. The horizontal line indicates a genome-wide MultiSTAAR-O *P* value threshold of 5.00×10^{-7}. The significant threshold is defined by multiple comparisons using the Bonferroni correction ($0.05/(20,000 \times 5) = 5.00 \times 10^{-7}$). Different symbols represent the MultiSTAAR-O *P* value of the protein-coding gene using different functional categories (putative loss-of-function, putative loss-of-function and disruptive missense, missense, disruptive missense, synonymous). **b**, Quantile-quantile plots for unconditional gene-centric coding analysis of protein-coding genes. Different symbols represent the MultiSTAAR-O *P*-value of the gene using different functional categories. **c**, Manhattan plots for unconditional gene-centric noncoding analysis of protein-coding genes. The horizontal line indicates a genome-wide MultiSTAAR-O *P* value threshold of 3.57×10^{-7}. The significant threshold is defined by multiple comparisons using the Bonferroni correction ($0.05/(20,000 \times 7) = 3.57 \times 10^{-7}$). Different symbols represent the MultiSTAAR-O *P* value of the protein-coding gene using different functional categories (upstream, downstream, UTR, promoter_CAGE, promoter_DHS, enhancer_CAGE, enhancer_DHS). Promoter_CAGE and promoter_DHS are the promoters with overlap of Cap Analysis of Gene Expression (CAGE) sites and DNase hypersensitivity (DHS) sites for a given gene, respectively. Enhancer_CAGE and enhancer_DHS are the enhancers in GeneHancer-predicted regions with the overlap of CAGE sites and DHS sites for a given gene, respectively. **d**, Quantile-quantile plots for unconditional gene-centric noncoding analysis of protein-coding genes. Different symbols represent the MultiSTAAR-O *P*-value of the gene using different functional categories. **e**, Manhattan plot showing the associations of 2.68 million 2-kb sliding windows versus $-\log_{10}(P)$ of MultiSTAAR-O. The horizontal line indicates a genome-wide *P* value threshold of 1.86×10^{-8}. **f**, Quantile-quantile plot of 2-kb sliding window MultiSTAAR-O *P* values. In panels, **a**, **c** and **e**, the chromosome number are indicated by the colors of dots. In all panels, MultiSTAAR-O is a two-sided test.

https://doi.org/10.1038/s43588-024-00764-8



Extended Data Fig. 2 | See next page for caption.

https://doi.org/10.1038/s43588-024-00764-8

Extended Data Fig. 2 | Manhattan plots and Q-Q plots for unconditional gene-centric coding, noncoding and genetic region (2-kb sliding window) multi-trait analysis of C-reactive protein (CRP), interleukin-6 (IL-6), lipoprotein-associated phospholipase A2 (Lp-PLA2) activity, and lipoprotein-associated phospholipase A2 (Lp-PLA2) mass using TOPMed data (n = 9,380). **a**, Manhattan plots for unconditional gene-centric coding analysis of protein-coding genes. The horizontal line indicates a genome-wide MultiSTAAR-O P value threshold of 5.00×10^{-7}. The significant threshold is defined by multiple comparisons using the Bonferroni correction ($0.05/(20,000 \times 5) = 5.00 \times 10^{-7}$). Different symbols represent the MultiSTAAR-O P value of the protein-coding gene using different functional categories (putative loss-of-function, putative loss-of-function and disruptive missense, missense, disruptive missense, synonymous). **b**, Quantile-quantile plots for unconditional gene-centric coding analysis of protein-coding genes. Different symbols represent the MultiSTAAR-O P-value of the gene using different functional categories. **c**, Manhattan plots for unconditional gene-centric noncoding analysis of protein-coding genes. The horizontal line indicates a genome-wide MultiSTAAR-O P value threshold of 3.57×10^{-7}. The significant threshold is defined by multiple comparisons using the Bonferroni correction ($0.05/(20,000 \times 7) = 3.57 \times 10^{-7}$). Different symbols represent the MultiSTAAR-O P value of the protein-coding gene using different functional categories (upstream, downstream, UTR, promoter_CAGE, promoter_DHS, enhancer_CAGE, enhancer_DHS). Promoter_CAGE and promoter_DHS are the promoters with overlap of Cap Analysis of Gene Expression (CAGE) sites and DNase hypersensitivity (DHS) sites for a given gene, respectively. Enhancer_CAGE and enhancer_DHS are the enhancers in GeneHancer predicted regions with the overlap of CAGE sites and DHS sites for a given gene, respectively. **d**, Quantile-quantile plots for unconditional gene-centric noncoding analysis of protein-coding genes. Different symbols represent the MultiSTAAR-O P-value of the gene using different functional categories. **e**, Manhattan plot showing the associations of 2.67 million 2-kb sliding windows versus $-\log_{10}(P)$ of MultiSTAAR-O. The horizontal line indicates a genome-wide P value threshold of 1.87×10^{-8}. **f**, Quantile-quantile plot of 2-kb sliding window MultiSTAAR-O P values. In panels, **a**, **c** and **e**, the chromosome number are indicated by the colors of dots. In all panels, MultiSTAAR-O is a two-sided test.

nature portfolio

Corresponding author(s): Zilin Li, Zhonghua Liu, Xihong Lin

Last updated by author(s): Nov 3, 2024

Reporting Summary

Nature Portfolio 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 Portfolio policies, see our Editorial Policies 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	The software was used for downloading the data as follows: Wget v1.21.4 (https://www.gnu.org/software/wget/wget.html).
Data analysis	Data analysis was performed in R (4.1.0). STAAR v0.9.7 and MultiSTAAR v0.9.7 were used in simulation and real data analysis and implemented as open-source R packages available at https://github.com/xihaoli/STAAR and https://github.com/xihaoli/MultiSTAAR. These two packages have been archived on Zenodo using xxx. GraphTyper (May 2023) was used to perform genotype calling of the UK Biobank 200K WGS data.

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 and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio 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 description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

This paper used the TOPMed Freeze 8 WGS data and lipids phenotype data. Genotype and phenotype data are both available in database of Genotypes and

Phenotypes. The TOPMed WGS data were from the following twenty study cohorts (accession numbers provided in parentheses): Old Order Amish (phs000956.v1.p1), Atherosclerosis Risk in Communities Study (phs001211), Mt Sinai BioMe Biobank (phs001644), Coronary Artery Risk Development in Young Adults (phs001612), Cleveland Family Study (phs000954), Cardiovascular Health Study (phs001368), Diabetes Heart Study (phs001412), Framingham Heart Study (phs000974), Genetic Study of Atherosclerosis Risk (phs001218), Genetic Epidemiology Network of Arteriopathy (phs001345), Genetic Epidemiology Network of Salt Sensitivity (phs001217), Genetics of Lipid Lowering Drugs and Diet Network (phs001359), Hispanic Community Health Study - Study of Latinos (phs001395), Hypertension Genetic Epidemiology Network and Genetic Epidemiology Network of Arteriopathy (phs001293), Jackson Heart Study (phs000964), Multi-Ethnic Study of Atherosclerosis (phs001416), San Antonio Family Heart Study (phs001215), Genome-wide Association Study of Adiposity in Samoans (phs000972), Taiwan Study of Hypertension using Rare Variants (phs001387), and Women's Health Initiative (phs001237). The sample sizes, ancestry and phenotype summary statistics of these cohorts are given in Supplementary Table 2. The UK Biobank analyses were conducted using the UK Biobank resource under application 52008.

The functional annotation data are publicly available and were downloaded from the following links: GRCh38 CADD v1.4 (https://cadd.gs.washington.edu/download); ANNOVAR dbNSFP v3.3a (https://annovar.openbioinformatics.org/en/latest/user-guide/download); LINSIGHT (https://github.com/CshlSiepelLab/LINSIGHT); FATHMM-XF (http://fathmm.biocompute.org.uk/fathmm-xf); FANTOM5 CAGE (https://fantom.gsc.riken.jp/5/data); GeneCards (https://www.genecards.org; v4.7 for hg38); and Umap/Bismap (https://bismap.hoffmanlab.org; 'before March 2020' version). In addition, recombination rate and nucleotide diversity were obtained from Gazal et al. The whole-genome individual functional annotation data was assembled from a variety of sources and the computed annotation principal components are available at the Functional Annotation of Variant-Online Resource (FAVOR) site (https://favor.genohub.org) and the FAVOR database (https://doi.org/10.7910/DVN/1VGTJI).

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	Sex/gender was defined using a combination of self-reported sex/gender (from participant questionnaires) and study recruitment information.
Population characteristics	The TOPMed data consist of ancestrally diverse and multi-ethnic related samples. The data analyzed in this paper include 38,744 (62.7%) females; 15,636 (25.3%) Black or African-American, 27,439 (44.4%) White, 4,461 (7.2%) Asian American, 13,138 (21.2%) Hispanic/Latino American, and 1,164 (1.9%) Samoans. Race/ethnicity was defined using a combination of self-reported race/ethnicity (from participant questionnaires) and study recruitment information. The average age of the study participants is 52 with a standard deviation of 15.
Recruitment	The TOPMed Freeze 8 lipids data included whole genome sequencing data of 61,838 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 Supplemental Table 2. Detailed information of participant recruitment of each study cohort can be found in Supplementary Note. More details can be found at https://topmed.nhlbi.nih.gov.
Ethics oversight	This study relied on analyses of genetic data from TOPMed cohorts. The study has been approved by the TOPMed Publications Committee, TOPMed Lipids Working Group and all the participating cohorts, including Old Order Amish (phs000956.v1.p1), Atherosclerosis Risk in Communities Study (phs001211), Mt Sinai BioMe Biobank (phs001644), Coronary Artery Risk Development in Young Adults (phs001612), Cleveland Family Study (phs000954), Cardiovascular Health Study (phs001368), Diabetes Heart Study (phs001412), Framingham Heart Study (phs000974), Genetic Study of Atherosclerosis Risk (phs001218), Genetic Epidemiology Network of Arteriopathy (phs001345), Genetic Epidemiology Network of Salt Sensitivity (phs001217), Genetics of Lipid Lowering Drugs and Diet Network (phs001359), Hispanic Community Health Study - Study of Latinos (phs001395), Hypertension Genetic Epidemiology Network and Genetic Epidemiology Network of Arteriopathy (phs001293), Jackson Heart Study (phs000964), Multi-Ethnic Study of Atherosclerosis (phs001416), San Antonio Family Heart Study (phs001215), Genome-wide Association Study of Adiposity in Samoans (phs000972), Taiwan Study of Hypertension using Rare Variants (phs001387), and Women's Health Initiative (phs001237), where the accession numbers are provided in parenthesis. The use of human genetics data from TOPMed cohorts was approved by the Harvard T.H. Chan School of Public Health IRB (IRB13-0353).

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

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 analysis consists of all available 61,838 samples from twenty approved study cohorts of TOPMed Freeze 8. No sample size calculation was performed.

Data exclusions	For TOPMed data, failed variants were excluded in the quality control (QC) procedure.
Replication	The analysis of TOPMed data identified five new lipid trait associations, which were conditionally significant in our multi-trait analyses but undetected in single-trait analysis using TOPMed data. All five assocations were replicated using UK Biobank data using Bonferroni correction, two were detected by gene-centric analysis and three by sliding window analysis. Experimental replication was not attempted.
Randomization	Both TOPMed and UK Biobank are observational studies. For TOPMed data, age, age2, sex, eleven ancestry principal components, ethnicity group indicators, and a variance component for empirically derived sparse kinship matrix were used to account for potential confounding factors such as population structures and sample relatedness. For UK Biobank data, age, age2, sex, ten ancestry principal components, and a variance component for empirically derived sparse kinship matrix were used to account for potential confounding factors such as population structures and sample relatedness. No randomization was used in the study design.
Blinding	We used de-identified coded data for analysis, and hence were blinded.

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 and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

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