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

Zilin Li[1,46]*, Xihao Li[1,46], Hufeng Zhou[1], Sheila M. Gaynor[1], Margaret S. Selvaraj[2,3,4], Theodore Arapoglou[1], Corbin Quick[1], Yaowu Liu[5], Han Chen[6,7], Ryan Sun[8], Rounak Dey[1], Donna K. Arnett[9], Lawrence F. Bielak[10], Joshua C. Bis[11], Thomas W. Blackwell[12], John Blangero[13], Eric Boerwinkle[6,14], Donald W. Bowden[15], Jennifer A. Brody[11], Brian E. Cade[16,17], Matthew P. Conomos[11], Adolfo Correa[18], L. Adrienne Cupples[19,20], Joanne E. Curran[13], Paul S. de Vries[6], Ravindranath Duggirala[13], Barry I. Freedman[21], Harald H. H. Göring[13], Xiuqing Guo[22], Rita R. Kalyani[23], Charles Kooperberg[24], Brian G. Kral[23], Leslie A. Lange[25], Ani Manichaikul[26], Lisa W. Martin[27], Braxton D. Mitchell[28,29], May E. Montasser[30], Alanna C. Morrison[6], Take Naseri[31], Jeffrey R. O'Connell[28], Nicholette D. Palmer[15], Patricia A. Peyser[10], Bruce M. Psaty[11,32], Laura M. Raffield[33], Susan Redline[16,17,34], Alexander P. Reiner[24,35], Muagututi'a Sefuiva Reupena[36], Kenneth M. Rice[37], Stephen S. Rich[26], Jennifer A. Smith[10,38], Kent D. Taylor[22], Ramachandran S. Vasan[20,39], Daniel E. Weeks[40], James G. Wilson[41], Lisa R. Yanek[23], Wei Zhao[10], NHLBI Trans-Omics for Precision Medicine (TOPMed) Consortium, TOPMed Lipids Working Group, Jerome I. Rotter[22], Christen J. Willer[42,43,44], Pradeep Natarajan[2,3,4], Gina M. Peloso[19,20] and Xihong Lin[1,3,45]*

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

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

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

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

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

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

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

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

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

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

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

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

2

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

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

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

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

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

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

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

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

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

[22]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.

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

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

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

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

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

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

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

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

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

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

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

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

[35]Department of Epidemiology, University of Washington, Seattle, WA, USA.

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

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

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

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

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

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

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

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

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

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

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

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

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

Abstract

Large-scale whole-genome sequencing studies have enabled analysis of noncoding rare variants' (RVs) associations with complex human traits. Variant set analysis is a powerful approach to study RV association, and a key component of it is constructing RV sets for analysis. However, existing methods have limited ability to define analysis units in the noncoding genome. Furthermore, there is a lack of robust pipelines for comprehensive and scalable noncoding RV association analysis. Here we propose a computationally-efficient noncoding RV association-detection framework that uses STAAR (variant-set test for association using annotation information) to group noncoding variants in gene-centric analysis based on functional categories. We also propose SCANG (scan the genome)-STAAR, which uses dynamic window sizes and incorporates multiple functional annotations, in a non-gene-centric analysis. We furthermore develop STAARpipeline to perform flexible noncoding RV association analysis, including gene-centric analysis as well as fixed-window-based and dynamic-window-based non-gene-centric analysis. We apply STAARpipeline to identify noncoding RV sets associated with four quantitative lipid traits in 21,015 discovery samples from the Trans-Omics for Precision Medicine (TOPMed) program and replicate several noncoding RV associations in an additional 9,123 TOPMed samples.

Introduction

Genome-wide association studies (GWASs) have successfully identified thousands of common genetic variants for complex diseases and traits; however, these common variants only explain a small fraction of heritability[1]. Recent studies suggest that the missing heritability of complex traits and diseases and causal variants may be accounted for in part by RVs (minor allele frequency (MAF) < 1%)[2-4]. Although whole-exome sequencing (WES) studies have identified exome-wide significant RV associations for complex diseases and traits[5,6], more than 98% of the genetic variants are located in the noncoding genome[6]. Many common variants identified by GWAS as being associated with phenotypes are located in noncoding regions[7,8]. Further, the ENCODE project shows that a significant fraction of noncoding regions are functionally active[9,10], indicating that rare noncoding regions may have an effect on diseases or traits.

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

Single-variant analyses are not appropriate for analysis of rare variants because in realistic settings they lack power[11-13]. To improve power, variant set tests have been proposed that assess the effects of sets of multiple RVs jointly. These tests include burden, SKAT, and most recently STAAR (variant-set test for association using annotation information), which incorporates multiple functional annotations for genetic variants[14-16]. A key challenge of these approaches is the selection of RVs to form variant sets. Several methods have been proposed to create coding and noncoding variant sets for RV association analysis of WGS/WES studies[16-20]. However, these methods have limited utility for defining analysis units in the noncoding genome[21]. For example, for gene-centric analysis, STAAR has been used with two noncoding genetic categories of regulatory regions (masks): using promoters and enhancers in GeneHancer[22] overlaid with Cap Analysis of Gene Expression (CAGE) sites[23,24]; for non-gene-centric analysis, fixed-size sliding windows can be used to scan the genome. As the signal regions (variant-phenotype-association regions) are unknown in practice and their sizes vary across the genome, the fixed-size sliding window approach is likely to lead to power loss when the prespecified window sizes are too big or too small compared with the actual sizes of signal regions. Furthermore, it is often knowledge- and effort-demanding to functionally annotate variants from a WGS/WES study of interest. Limited tools exist for multi-faceted functional annotation and analytic integration of WGS/WES data for rare variant association tests (RVATs). Finally, there is a lack of robust pipelines to perform scalable and comprehensive noncoding RV association analysis in large-scale WGS data with

hundreds of millions of noncoding RVs that have been sequenced across the genome. Much uncertainty remains on the best practices for computationally-efficient RV analysis at the scale of large WGS studies.

To respond to the aforementioned needs, we propose a computationally-efficient noncoding rare variant association-detection framework for WGS data by making three new contributions toward automatically selecting interpretable and powerful variant sets. First, in gene-centric analysis, we propose additional strategies for grouping noncoding variants based on functional annotations, including untranslated regions, upstream regions, downstream regions, promoters, enhancers of protein-coding genes, and long noncoding RNA genes within STAAR. For promoters and enhancers, we offer additional options of overlaying promoters and GeneHancer-based enhancers with not only CAGE sites but also with DNase Hypersensitivity (DHS) sites[9]. Second, in non-gene-centric analysis, instead of using fixed-size sliding windows in STAAR we propose SCANG-STAAR, a flexible data-adaptive window size RVAT method that extends the SCANG (scan the genome) method[18] by incorporating multiple functional annotations through STAAR[16], while accounting for both relatedness and population structure through a generalized linear mixed model framework[25] for quantitative and dichotomous traits[26,27]. Third, we develop *STAARpipeline*, a pipeline that (1) functionally annotates both noncoding and coding variants of a WGS study and builds an annotated genotype dataset using the multi-faceted functional annotation database FAVOR[16] (Functional Annotations of Variants - Online Resource),

9

through FAVORannotator; and (2) performs RVATs using the proposed methods for both gene-centric analysis and non-gene-centric analysis.

We applied the proposed framework to detect noncoding RVs associated with four quantitative lipid traits: low-density lipoprotein cholesterol (LDL-C); high-density lipoprotein cholesterol (HDL-C); triglycerides (TG) and total cholesterol (TC) using 21,015 discovery samples and 9,123 replication samples from the NHLBI TOPMed Freeze 5 WGS data. We performed conditional analysis by conditioning on known lipids-associated variants and identified several novel replicated RVs sets associated with lipids.

Results

Overview of Noncoding RVATs

We propose a computationally-efficient noncoding RVAT framework for phenotype-genotype association analyses of whole-genome sequencing data, focusing on rare variant association analysis in the noncoding genome. This regression-based framework allows adjusting for covariates, population structure, and relatedness by fitting linear and logistic mixed models for quantitative and dichotomous traits[26,27]. A central component of it is the development of strategies to aggregate noncoding rare variants using both flexible gene-centric and non-gene-centric approaches to empower RVATs. For the gene-centric approach, we group noncoding RVs for each gene using eight genetic categories of regulatory regions provided by functional annotations and apply STAAR, which incorporates

multiple *in-silico* variant functional annotation scores that prioritize functional variants using multi-dimensional variant biological functions[16]. For the non-gene-centric analysis, instead of using sliding windows with fixed sizes, we propose SCANG-STAAR, a procedure using dynamic windows with data-adaptive sizes and incorporating multi-dimensional functional annotations. We also perform analytical follow-up to dissect RV association signals independent of a given set of known variants via conditional analysis (**Figure 1**).

Gene-centric analysis of the noncoding genome

In gene-centric analysis of noncoding variants, we provide eight genetic categories of regulatory regions to aggregate noncoding rare variants: (1) promoter RVs overlaid with CAGE sites, (2) promoter RVs overlaid with DHS sites, (3) enhancer RVs overlaid with CAGE sites, (4) enhancer RVs overlaid with DHS sites, (5) untranslated region (UTR) RVs, (6) upstream region RVs, (7) downstream region RVs and (8) noncoding RNA (ncRNA) RVs. The promoter RVs are defined as RVs in the +/- 3-kilobase (kb) window of transcription start sites with the overlap of CAGE sites or DHS sites. The enhancer RVs are defined as RVs in GeneHancer predicted regions with the overlap of CAGE sites or DHS sites[9,22-24]. We define the UTR, upstream, downstream, and ncRNA RVs by GENCODE VEP categories[28,29]. For the UTR mask, we include RVs in both 5' and 3' UTR regions. For the ncRNA mask, we include the exonic and splicing ncRNA RVs. We consider the protein-coding gene for the first seven categories provided by Ensembl[30] and the ncRNA genes provided by GENCODE[28,29].

For each noncoding mask, we calculate its *P* value using the STAAR method that empowers RVATs by incorporating multiple variant functional annotation scores[16]. Functional annotations consist of diverse biological information of genomic elements. Incorporating this external biological information provided by functional annotations can increase the association analysis power[31]. For example, annotation principal components (aPCs) provide multi-dimensional summaries of variant annotations and capture the multi-faceted biological impact, calculated by the first principal component of the set of individual functional annotation scores interpreting similar biological functionality[16]. We incorporate ten aPCs and three integrative scores (CADD[32], LINSIGHT[33], and FATHMM-XF[34]) as weights in constructing STAAR statistics[16]. Details of these 13 functional annotations are given in **Supplementary Table 1**. Specifically, we calculate the *P* value of each variant set using STAAR-O[16], an omnibus test aggregating multiple annotation-weighted burden test[14], SKAT[15], and ACAT-V[35] in the STAAR framework.

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

We improve the STAAR-based fixed-size sliding window RVAT[16,17] by proposing a dynamic window based SCANG-STAAR method, which extends the procedure SCANG[18] by incorporating multi-dimensional functional annotations to flexibly detect the locations and the sizes of signal windows across the genome. Specifically, as location of regions associated with a disease or trait are often unknown in advance and their sizes may vary across the genome, RVAT's

12

default use of a pre-specified fixed-size sliding window method can lead to power loss, if the pre-specified window sizes do not align with the true signal window sizes.

The dynamic window RVAT method, SCANG[18], overcomes the limitation of the fixed-size sliding window method using scan statistics that flexibly detect the sizes and the locations of RV association by scanning the whole genome continuously while allowing for overlapping windows of different sizes by shifting forward a given size window by a small number of variants each time and selecting the windows that maximize the test power, while controlling for the genome-wise (family-wise) error rate by accounting for the correlations of tests from overlapping windows. However, SCANG does not incorporate variant functional annotations and may therefore lose power if annotation information helps identify true signals. We propose SCANG-STAAR by extending SCANG to incorporate multi-dimensional variant functional annotations using STAAR to ameliorate power loss.

In dynamic window analysis, we extend the SCANG-SKAT procedure (SCANG-S) to SCANG-STAAR-S by using the STAAR-SKAT (STAAR-S) P value, which in each overlapping window incorporates multiple variant functional annotations, instead of using just the MAF-based SKAT P value. In SCANG-STAAR-S we first calculate a threshold that controls the genome-wise type I error at a given α level, based on the minimum value of the STAAR-S P value from all moving

13

windows of different sizes in a range of windows (**Online Methods**). The procedure then selects the candidate significant windows whose set-based *P* value beats that threshold. When this results in multiple overlapping windows, we localize the detected significant window as the window whose *P* value is smaller than both the threshold and any window that overlaps with it. We then calculate the genome-wide *P* value of the detected windows by accounting for multiple comparisons of overlapping windows and controlling for the genome-wise (family-wise) error rate (**Online Methods)**.

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

ACAT-V test is designed for sparse alternatives. Hence, it always detects the region with the smallest size that contains the most significant variant in the dynamic window procedure.

Analytical follow-up via conditional analysis

We also perform conditional analysis as an analytical follow-up to identify RV association signals independent of known single variant associations. We first select a list of known variants by including the previously identified trait-associated variants, for example, variants indexed in the GWAS Catalog[36]. We then perform stepwise selection to select the subset of independent variants from the known variants list to be used in the conditional analysis. We perform iterative conditional association analysis until the P values of all variants in the known variant list are larger than a cut-off (1×10^{-4}, **Online Methods**). Instead of adjusting for all known trait-associated variants in the entire chromosome, we adjust for variants in an extended region of the specific variant, for example, a +/- 1-megabase (Mb) window beyond the variant of interest. Finally, we perform conditional analysis of each variant set by fitting the regression model adjusting for the selected known variants near the variant set (for example, in a +/- 1-Mb window).

STAARpipeline and computation cost

Our R package *STAARpipeline* performs scalable phenotype-genotype association analyses of functionally annotated WGS data using the developed

RVAT methods. A further package, *STAARpipelineSummary* summarizes the rare variant findings generated by *STAARpipeline*, including results of both unconditional and conditional analysis and visualization of analysis results.

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

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

STAARpipeline and provides analytical follow-up via conditional analysis. Summarizing the genome-wide TOPMed results took 24 hours using one core with 25 Gb memory.

Association analysis of lipid traits in the TOPMed WGS data

We applied *STAARpipeline* to identify RV-sets associated with four quantitative lipid traits (LDL-C, HDL-C, TG and TC) using TOPMed WGS data[4,16,20]. DNA samples were sequenced at >30X target coverage[4]. The discovery phase consisted of six study cohorts with 21,015 samples sequenced in TOPMed Freeze 5. The replication phase consisted of eight remaining study cohorts with 9,123 samples in TOPMed Freeze 5 (**Supplementary Note**, **Supplementary Table 2**). Sample-level and variant-level quality control (QC) were performed[4,20]. Race/ethnicity was defined using a combination of self-reported race/ethnicity and study recruitment information[37]. The discovery cohorts consisted of 5,849 (27.8%) Black or African American, 12,313 (58.6%) White, 675 (3.2%) Asian American, 1,075 (5.1%) Hispanic/Latino American, and 1,103 (5.3%) Samoan participants. Among all samples in the discovery phase, 3,610 (17.2%) had first degree relatedness, 546 (2.6%) had second degree relatedness, and 472 (2.2%) had third degree relatedness (**Supplementary Figure 1**). There were 215 million single-nucleotide variants (SNVs) observed in the discovery phase, and 205 million (94.9%) were rare variants (MAF < 1%). Among these 205 million rare variants, 202 million (98.8%) were noncoding variants defined by GENCODE VEP. Details of the study-specific demographics, summaries of lipid levels, and

variant number distributions are given in **Supplementary Tables 2-3** and

Supplementary Figure 2.

For each phenotype, we applied rank-based inverse normal transformation of

phenotypes. We adjusted for age, age^2, sex, race/ethnicity, study, and the first 10

ancestral PCs, and controlled for relatedness through heteroscedastic linear

mixed models with sparse genetic relatedness matrices (GRMs) plus study-

race/ethnicity-specific group-specific residual variance components (**Online**

Methods). We accounted for the presence of medications of LDL-C and TC as

before[20]. We tested for an association between lipid traits and RVs (MAF < 1%)

in each variant set. In gene-centric analysis, we defined the eight analysis units

as the previously-described: seven noncoding genetic categories of protein-

coding genes and one category for ncRNA genes. In non-gene-centric analysis,

we performed a 2-kb sliding window analysis with 1-kb skip length and a dynamic

window analysis using SCANG-STAAR-S of all moving windows containing 40 to

300 variants[18]. In unconditional analysis we used Bonferroni-corrected genome-

wide significance thresholds of $\alpha = 0.05/(20,000 \times 7) = 3.57 \times 10^{-7}$ accounting

for 7 different noncoding masks across protein-coding genes; $\alpha = 0.05/20,000 =$

2.50×10^{-6} accounting for ncRNA genes, and $\alpha = 0.05/(2.66 \times 10^6) =$

1.88×10^{-8} accounting for 2.66 million 2-kb sliding windows across the genome.

We controlled the genome-wise (family-wise) error rate for SCANG-STAAR-S

dynamic window analysis at $\alpha = 0.05$ level[18]. We selected individual variants to

be adjusted for in conditional analysis from the list of phenotype-associated

common and low-frequency variants (MAF $\geq$ 1%) indexed in GWAS Catalog[36].

Then we obtained the independent known variants using the algorithm described

before in the analytical follow-up via conditional analysis section (**Online**

Methods, **Supplementary Table 4**).

In gene-centric noncoding unconditional analysis of the discovery samples,

STAARpipeline identified 43 genome-wide significant associations with at least

one of the four lipid levels (**Supplementary Table 5**, **Supplementary Figures**

3a-d, 4a-d, 5a-d, 6a-d). After conditioning on known lipid-associated variants, 14

out of the 43 associations remained significant at the Bonferroni-corrected level

$0.05/43 = 1.16 \times 10^{-3}$ (**Table 1**). In the replication data, and adjusting for known

lipid-associated variants, 4 of these 14 associations achieved significance at

Bonferroni-corrected level $0.05/14 = 3.57 \times 10^{-3}$. These included enhancer DHS

RVs in *APOA1* and HDL-C, promoter CAGE RVs in *APOE* and TG, and

enhancer CAGE or DHS RVs in *APOE* and TG. After further adjustment for

known individual rare variants (minor allele count, MAC $\geq$ 20, **Supplementary**

Table 6), none of the associations remained significant at the same significance

level of 3.57×10^{-3} (**Supplementary Table 7**).

In unconditional analysis of the discovery data, using the 2-kb sliding window

procedure we identified 140 windows as genome-wide significant

(**Supplementary Table 8**, **Supplementary Figures 3e-f, 4e-f, 5e-f, 6e-f**).

Among these 140 significant sliding windows, 14 are located in noncoding

19

regions and, after conditioning on known lipid-associated variants, all remained significant at the Bonferroni-corrected level $0.05/140 = 3.57 \times 10^{-4}$ (**Table 2**). In replication data 9 of the 14 associations were significant at the Bonferroni-corrected level $0.05/14 = 3.57 \times 10^{-3}$ after adjusting for known phenotype-specific variants. When we further adjusted these 9 associations for known individual rare variants (MAC $\geq$ 20), associations for two intronic sliding windows (*PAFAH1B2* and TG) remained significant at the same level of 3.57×10^{-3} (**Supplementary Table 9**).

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

Discussion

We developed a comprehensive association analysis framework for detecting noncoding rare variant set associations in large-scale WGS studies. Crucially, our framework explicitly solves the problem of defining variant sets, which is a significant challenge in practical analysis but not often discussed in other set-based inference methodology work. Our approach allows for continuous and binary traits and accounts for both population structure and relatedness through generalized linear mixed models using gene-centric analysis and non-gene-centric analysis. For gene-centric analysis, we proposed several strategies to define analysis units of rare variants in the noncoding genome, including seven genetic categories of regulatory regions for protein-coding genes, ncRNA genes, and perform RVATs of each noncoding mask using STAAR. For non-gene-centric analysis, to overcome the limitations of fixed-size sliding windows, we proposed SCANG-STAAR, a data-adaptive-size dynamic window scan procedure that incorporates multi-faceted functional annotations. We proposed *STAARpipeline* to perform RVATs using these methods for both noncoding and coding variants using unconditional analysis, as well as conditional analyses, which provides an analytical follow-up to distinguish novel RV association signals independent of known variants.

We developed *STAARpipeline*, a fast and resource-efficient tool for RV association analysis of WGS data that scales linearly on hundreds of thousands of samples, for both quantitative and dichotomous phenotypes. *STAARpipeline* allows researchers to conveniently functionally annotate a WGS/WES study

using the variant functional annotation database FAVOR and the FAVORannotator workflow. *STAARpipeline* optimizes computational feasibility of RV association analysis in two steps. First, *STAARpipeline* reduces the computation burden of fitting the null mixed model using the estimated sparse GRM[16,38]. Second, *STAARpipeline* performs the RV association tests by taking advantage of sparse genotype dosages of RVs[39].

In a WGS RV analysis of lipid traits in TOPMed, we identified and replicated using our *STAARpipeline* several conditional associations with lipid traits in the noncoding genome, including RVs in an intronic region of *PAFAH1B2* and TG, RVs in an intronic region of *SIDT2* and TG, and RVs in an intronic region of *CEP164* and TG, which were not detected by previous analysis of TOPMed Freeze 3 data[16,20]. Several coding rare variants in *PAFAH1B2* have been previously detected associated with TG[40], our findings detected additionally significant RV association in the noncoding region of *PAFAH1B2*. Two intronic common variants in *SIDT2* have been reported associated with TG[41], additional intronic rare variant association in *SIDT2* was detected using *STAARpipeline*.

For non-gene-centric analysis, we proposed improvements to the sliding window analysis using the dynamic window analysis of SCANG-STAAR. Compared with sliding window analysis using a fixed window size and skip length, SCANG-STAAR can increase power by considering all possible sub-windows of different sizes and selecting those windows that maximize power, while incorporating

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

In addition to noncoding rare variants association analysis, *STAARpipeline* also provides single variant analysis for common and low-frequency variants and gene-centric analysis for coding rare variants. The single variant analysis in *STAARpipeline* provides individual *P* values of variants given a MAF or MAC cut-off, for example, MAC $\geq$ 20. The gene-centric coding analysis provides five genetic categories to aggregate coding rare variants of each protein-coding gene: (1) putative loss of function (stop gain, stop loss and splice) RVs, (2) missense RVs, (3) disruptive missense RVs, (4) putative loss of function and disruptive missense RVs, and (5) synonymous RVs. The putative loss of function, missense, and synonymous RVs are defined by GENCODE VEP categories[29,30]. The disruptive variants are further defined by MetaSVM[42], which

measures the deleteriousness of missense mutations. As in the noncoding RV association analysis, single variant and gene-centric coding analyses also scale well in computation time and memory for large-scale WGS data. Using 30,138 related TOPMed samples these two analyses respectively took 3 hours and 5 hours for 100 cores with 6 Gb memory. Thus, *STAARpipeline* provides an efficient and comprehensive analysis tool for both coding and noncoding variant association discovery in large-scale sequencing studies.

With the emergence of large-scale WGS data, there is a pressing need to identify genetic components of complex traits in the noncoding genome. Here we introduce a powerful and scalable framework, *STAARpipeline*, for noncoding RV association detection across the genome. *STAARpipeline* provides several strategies to aggregate noncoding rare variants to empower RV association analysis in the noncoding region. We demonstrate the computational efficiency of *STAARpipeline* in application to the WGS association analysis of lipid traits on 30,138 TOPMed samples. The optimization approaches of *STAARpipeline* make it scalable for even larger data sets. Thus, our framework provides an essential solution for noncoding RV association detection in large-scale WGS data analysis and dissects the genetic contribution of noncoding rare variants to complex diseases.

URLs

STAARpipeline (version 0.9.6), https://github.com/xihaoli/STAARpipeline

and https://content.sph.harvard.edu/xlin/software.html.

STAARpipelineSummary (version 0.9.6),

https://github.com/xihaoli/STAARpipelineSummary

and https://content.sph.harvard.edu/xlin/software.html.

FAVOR, http://favor.genohub.org/.

FAVORannotator, https://github.com/zhouhufeng/FAVORannotator.

References

1. Manolio, T.A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747-753 (2009).
2. Wainschtein, P. *et al.* Recovery of trait heritability from whole genome sequence data. *BioRxiv*, 588020 (2019).
3. Hernandez, R.D. *et al.* Ultrarare variants drive substantial cis heritability of human gene expression. *Nature genetics* **51**, 1349-1355 (2019).
4. Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290-299 (2021).
5. Flannick, J. *et al.* Exome sequencing of 20,791 cases of type 2 diabetes and 24,440 controls. *Nature* **570**, 71-76 (2019).
6. Van Hout, C.V. *et al.* Exome sequencing and characterization of 49,960 individuals in the UK Biobank. *Nature* **586**, 749-756 (2020).
7. Zhang, F. & Lupski, J.R. Non-coding genetic variants in human disease. *Human molecular genetics* **24**, R102-R110 (2015).
8. Khurana, E. *et al.* Role of non-coding sequence variants in cancer. *Nature Reviews Genetics* **17**, 93-108 (2016).
9. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57 (2012).
10. Moore, J.E. *et al.* Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699-710 (2020).
11. Bansal, V., Libiger, O., Torkamani, A. & Schork, N.J. Statistical analysis strategies for association studies involving rare variants. *Nature Reviews Genetics* **11**, 773 (2010).
12. Lee, S., Abecasis, G.R., Boehnke, M. & Lin, X. Rare-variant association analysis: study designs and statistical tests. *The American Journal of Human Genetics* **95**, 5-23 (2014).
13. Kiezun, A. *et al.* Exome sequencing and the genetic basis of complex traits. *Nature genetics* **44**, 623 (2012).

14.   Li, B. & Leal, S.M. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *The American Journal of Human Genetics* **83**, 311-321 (2008).

15.   Wu, M.C. *et al.* Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics* **89**, 82-93 (2011).

16.   Li, X. *et al.* Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale. *Nature genetics* **52**, 969-983 (2020).

17.   Morrison, A.C. *et al.* Practical approaches for whole-genome sequence analysis of heart-and blood-related traits. *The American Journal of Human Genetics* **100**, 205-215 (2017).

18.   Li, Z. *et al.* Dynamic scan procedure for detecting rare-variant association regions in whole-genome sequencing studies. *The American Journal of Human Genetics* **104**, 802-814 (2019).

19.   He, Z., Xu, B., Buxbaum, J. & Ionita-Laza, I. A genome-wide scan statistic framework for whole-genome sequence data analysis. *Nature communications* **10**, 1-11 (2019).

20.   Natarajan, P. *et al.* Deep-coverage whole genome sequences and blood lipids among 16,324 individuals. *Nature communications* **9**, 1-12 (2018).

21.   Bocher, O. & Génin, E. Rare variant association testing in the non-coding genome. *Human Genetics* **139**, 1345-1362 (2020).

22.   Fishilevich, S. *et al.* GeneHancer: genome-wide integration of enhancers and target genes in GeneCards. *Database* **2017**(2017).

23.   Fantom Consortium. A promoter-level mammalian expression atlas. *Nature* **507**, 462 (2014).

24.   Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455-461 (2014).

25.   Breslow, N.E. & Clayton, D.G. Approximate inference in generalized linear mixed models. *Journal of the American statistical Association* **88**, 9-25 (1993).

26.   Chen, H. *et al.* Control for population structure and relatedness for binary traits in genetic association studies via logistic mixed models. *The American Journal of Human Genetics* **98**, 653-666 (2016).

27.   Chen, H. *et al.* Efficient variant set mixed model association tests for continuous and binary traits in large-scale whole-genome sequencing studies. *The American Journal of Human Genetics* (2019).

28.   Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome research* **22**, 1760-1774 (2012).

29.   Frankish, A. *et al.* GENCODE reference annotation for the human and mouse genomes. *Nucleic acids research* **47**, D766-D773 (2019).

30.   Kinsella, R.J. *et al.* Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database* **2011**(2011).

31.   Povysil, G. *et al.* Rare-variant collapsing analyses for complex traits: guidelines and applications. *Nature Reviews Genetics* **20**, 747-759 (2019).

32. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nature genetics* **46**, 310 (2014).

33. Huang, Y.-F., Gulko, B. & Siepel, A. Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. *Nature genetics* **49**, 618-624 (2017).

34. Rogers, M.F. *et al.* FATHMM-XF: accurate prediction of pathogenic point mutations via extended features. *Bioinformatics* **34**, 511-513 (2017).

35. Liu, Y. *et al.* ACAT: A Fast and Powerful P-value Combination Method for Rare-variant Analysis in Sequencing Studies. *bioRxiv*, 482240 (2018).

36. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic acids research* **47**, D1005-D1012 (2019).

37. 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. *American Journal of Epidemiology* (2021).

38. Gogarten, S.M. *et al.* Genetic association testing using the GENESIS R/Bioconductor package. *Bioinformatics* **35**, 5346-5348 (2019).

39. Zheng, X. & Davis, J.W. SAIGEgds—an efficient statistical tool for large-scale PheWAS with mixed models. *Bioinformatics* (2020).

40. Peloso, G.M. *et al.* Association of low-frequency and rare coding-sequence variants with blood lipids and coronary heart disease in 56,000 whites and blacks. *The American Journal of Human Genetics* **94**, 223-232 (2014).

41. Moon, S., Lee, Y., Won, S. & Lee, J. Multiple genotype–phenotype association study reveals intronic variant pair on SIDT2 associated with metabolic syndrome in a Korean population. *Human genomics* **12**, 1-10 (2018).

42. Dong, C. *et al.* Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. *Human molecular genetics* **24**, 2125-2137 (2015).

Acknowledgments

Author contributions

Z.L., X. Li and X. Lin designed the experiments., Z.L., X. Li, H.Z. and X. Lin performed the experiments. Z.L., X. Li, H.Z., S.M.G., M.S.S., T.A., C.Q., Y.L., H.C., R.S., R.D., D.K.A., L.F.B., J.C.B., T.W.B, J.B., E.B., D.W.B., J.A.B., B.E.C.,

Additional information

Supplementary information for this paper includes **Supplementary Figures** (6 figures), **Supplementary Tables** (11 tables) and **Supplementary Note**.

Online Methods

Notations and model

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

We consider the Generalized Linear Model for unrelated samples,

30

$$g(\mu_i) = \alpha_0 + X_i^T \alpha + G_i^T \beta, \tag{1}$$

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

We consider the following Generalized Linear Mixed Model[25,26,43] for related samples,

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

where the random effects b_i account for remaining population structure unaccounted by ancestral principal components and relatedness. Let $b = (b_1, \dots, b_n)^T \sim N(0, \theta\Phi)$ with variance components θ and a genetic relatedness matrix Φ[16,38]. Our goal is testing the null hypothesis of whether the variant-set is associated with the phenotype, adjusting for covariates and relatedness, which corresponds to $H_0: \beta = 0$, that is, $\beta_1 = \beta_2 = \cdots = \beta_p = 0$.

Variant set test using STAAR

The *STAARpipeline* calculates the variant set P value of each analysis unit using the STAAR method that incorporates multiple variant functional annotation scores[16]. Assume there are K annotations and $\hat{\pi}_{jk} = \frac{rank(A_{jk})}{M}$, where A_{jk} is the kth annotation for the jth variant ($k = 1, \cdots, K; j = 1, \cdots, p$). For $k = 0$, we assume $\hat{\pi}_{j0} = 1$. Assume $w_{jl} = Beta(\text{MAF}_j; a_{1l}, a_{2l})$, where $(a_{11}, a_{21}) = (1,25)$, $(a_{12}, a_{22}) = (1,1)$ and MAF_j is the MAF of the jth variant ($j = 1, \cdots, p$). The burden test statistic

using kth variant functional annotation and lth beta density as the weight is given by

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

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

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

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

$$Q_{ACAT-V,l,k} = \overline{\hat{\pi}_{\cdot k} w_{\cdot l}^2 \text{MAF}(1-\text{MAF})} \tan\left((0.5 - p_{0,k})\pi\right)$$

$$+ \sum_{j=1}^{p'} \hat{\pi}_{jk} w_{jl}^2 \text{MAF}_j\left(1 - \text{MAF}_j\right) \tan\left((0.5 - p_j)\pi\right),$$

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

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

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

$$+ tan\{(0.5 - p_{STAAR-ACAT-V})\pi\}],$$

and the corresponding *P*-value is calculated by

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

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

Dynamic window analysis using SCANG-STAAR

The *STAARpipeline* performs dynamic window analysis using the SCANG-STAAR procedure, which extends the dynamic window rare variant test procedure SCANG by incorporating multiple variant functional annotations using the STAAR method. Under the global null hypothesis, there is no variant associated with the phenotype across the genome. Under the alternative hypothesis, there exists at least one region associated with the phenotype. SCANG-STAAR procedure provides a valid test by using the minimum value of the *P* value of all candidate moving windows of different sizes

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

where $p(I)$ is the *P* value of region *I*, $|I|$ is the number of variants in a window *I*, and L_{min} and L_{max} are the smallest and largest number of variants in the searching windows, respectively. For SCANG-STAAR-S and SCANG-STAAR-B procedures, $p(I)$ is the STAAR-S and STAAR-B *P* value of window *I*, respectively. For SCANG-STAAR-O, $p(I)$ is the omnibus *P* value of STAAR-S

33

and STAAR-B calculated by ACAT method[35]. Similar to the SCANG procedure, SCANG-STAAR controls the genome-wise type I error at a given α level by using the $(1-\alpha)$th quantile of the empirical distribution of p_{min} as an empirical threshold $h(\alpha, p_{min}, L_{min}, L_{max})$[18]. We reject the null hypothesis if the *P* value of any window is smaller than $h(\alpha, p_{min}, L_{min}, L_{max})$. If this results in only one window, the detected window is $\hat{I} = \text{argmin}_{L_{min} \leq |I| \leq L_{max}} p(I)$. If this results in multiple overlapping windows, we localize the signals as the window whose *P* value is smaller than both the threshold and the windows that overlap with it.

Conditional analysis

The *STAARpipeline* performs conditional analysis to identify RV association independent of known variants. We first select a list of known variants by including the trait-associated variants identified in literature, for example, variants indexed in GWAS Catalog[36] or significant variants in large-scale GWAS. The significant variants detected in individual analysis using the same data could also be added into the known variants list to ensure the RV signals are not captured by the significant individual variants. We then use the following stepwise selection strategy to select a subset of independent variants representing the known variant list as the variants adjusted in the conditional analysis:

1. Calculate the individual *P* value of all variants in the known variants list and select the most significant variant.

2. For each step, calculate the *P* values of all the remaining variants conditional on the variant(s) that have already been selected. For each

variant, we only condition on the selected variants within a specified region of that variant, such as the +/- 1-Mb window.

3. Select the variant with minimum conditional *P* value that is lower than the cutoff *P* value, for example, 1×10^{-4}.

4. Repeat steps 2-3 until no variants can be selected.

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

Statistical analysis of lipid traits in the TOPMed data

The TOPMed WGS data consist of ancestrally diverse and multi-ethnic related samples[4,44]. Race/ethnicity was defined using a combination of self-reported race/ethnicity and study recruitment information (**Supplementary Note**)[37]. The discovery cohorts consist of 5,849 (27.8%) Black or African American, 12,313 (58.6%) White, 675 (3.2%) Asian American, 1,075 (5.1%) Hispanic/Latino American and 1,103 (5.3%) Samoans. The replication cohorts consist of 2,265 (24.8%) Black or African American, 5,615 (61.5%) White, and 1,243 (13.6%) Hispanic/Latino American.

We applied STAARpipeline to identify RV sets associated with four quantitative lipid traits (LDL-C, HDL-C, TG and TC) using the TOPMed WGS data. LDL-C and TC were adjusted for the presence of medications as before[20]. Linear regression model adjusting for age, age^2, sex was first fit for each study-

race/ethnicity-specific group. In addition, for Old Order Amish, we also adjusted for *APOB* p.R3527Q in LDL-C and TC analyses and adjusted for *APOC3* p.R19Ter in TG and HDL-C analyses[20]. The residuals were rank-based inverse normal transformed and rescaled by the standard deviation of the original phenotype within each group. We then fit a heteroscedastic linear mixed model (HLMM) for the rank normalized residuals, adjusting for 10 ancestral PCs, study-ethnicity group indicators, and a variance component for empirically derived kinship matrix plus separate group-specific residual variance components to account for population structure and relatedness. The output of HLMM was then used to perform following variant set analyses for rare variants (MAF < 1%) by scanning the genome, including gene-centric analysis using seven variant categories (promoter RVs overlaid with CAGE sites, promoter RVs overlaid with DHS sites, enhancer RVs overlaid with CAGE sites, enhancer RVs overlaid with DHS sites, UTR RVs, upstream RVs and downstream RVs) for each protein coded gene, ncRNA RVs, 2-kb sliding windows with 1-kb skip length, and dynamic windows with variants number between 40 and 300. The WGS RVAT analysis was performed using R packages STAAR (version 0.9.6), STAARpipeline (version 0.9.6) and STAARpipelineSummary (version 0.9.6).

Genome build

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

Code availability

STAARpipeline is implemented as an open-source R package available at

https://github.com/xihaoli/STAARpipeline

and https://content.sph.harvard.edu/xlin/software.html. *STAARpipelineSummary*

is implemented as an open-source R package available at

https://github.com/xihaoli/STAARpipelineSummary

and https://content.sph.harvard.edu/xlin/software.html.

Data availability

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

sample sizes, ethnicity and phenotype summary statistics of these cohorts are given in **Supplementary Table 3**.

The functional annotation data are publicly available and were downloaded from the following links: GRCh38 CADD v1.4 (https://cadd.gs.washington.edu/download), ANNOVAR dbNSFP v3.3a (https://annovar.openbioinformatics.org/en/latest/user-guide/download), LINSIGHT (https://github.com/CshlSiepelLab/LINSIGHT), FATHMM-XF (http://fathmm.biocompute.org.uk/fathmm-xf), CAGE (https://fantom.gsc.riken.jp/5/data), GeneHancer (https://www.genecards.org), and Umap/Bismap (https://bismap.hoffmanlab.org). In addition, recombination rate and nucleotide diversity were obtained from Gazal et al[45]. The tissue-specific functional annotations were downloaded from ENCODE (https://www.encodeproject.org/report/?type=Experiment).

Methods-only references

43. Chen, H. *et al.* Efficient variant set mixed model association tests for continuous and binary traits in large-scale whole-genome sequencing studies. *The American Journal of Human Genetics* **104**, 260-274 (2019).
44. Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *BioRxiv*, 563866 (2019).
45. Gazal, S. *et al.* Linkage disequilibrium–dependent architecture of human complex traits shows action of negative selection. *Nature genetics* **49**, 1421 (2017).

FIGURES

Figure 1. Workflow of *STAARpipeline*. (a) Prepare the input data of *STAARpipeline*, including genotypes, phenotypes and covariates. (b) Annotate all variants in the genome using FAVORannotator through FAVOR database and calculate the (sparse) genetic relatedness matrix. (c) Define analysis units in the noncoding genome: eight genetic categories of regulatory regions, sliding windows and dynamic windows using SCANG. (d) Obtain genome-wide significant associations and perform analytical follow-up via conditional analysis.
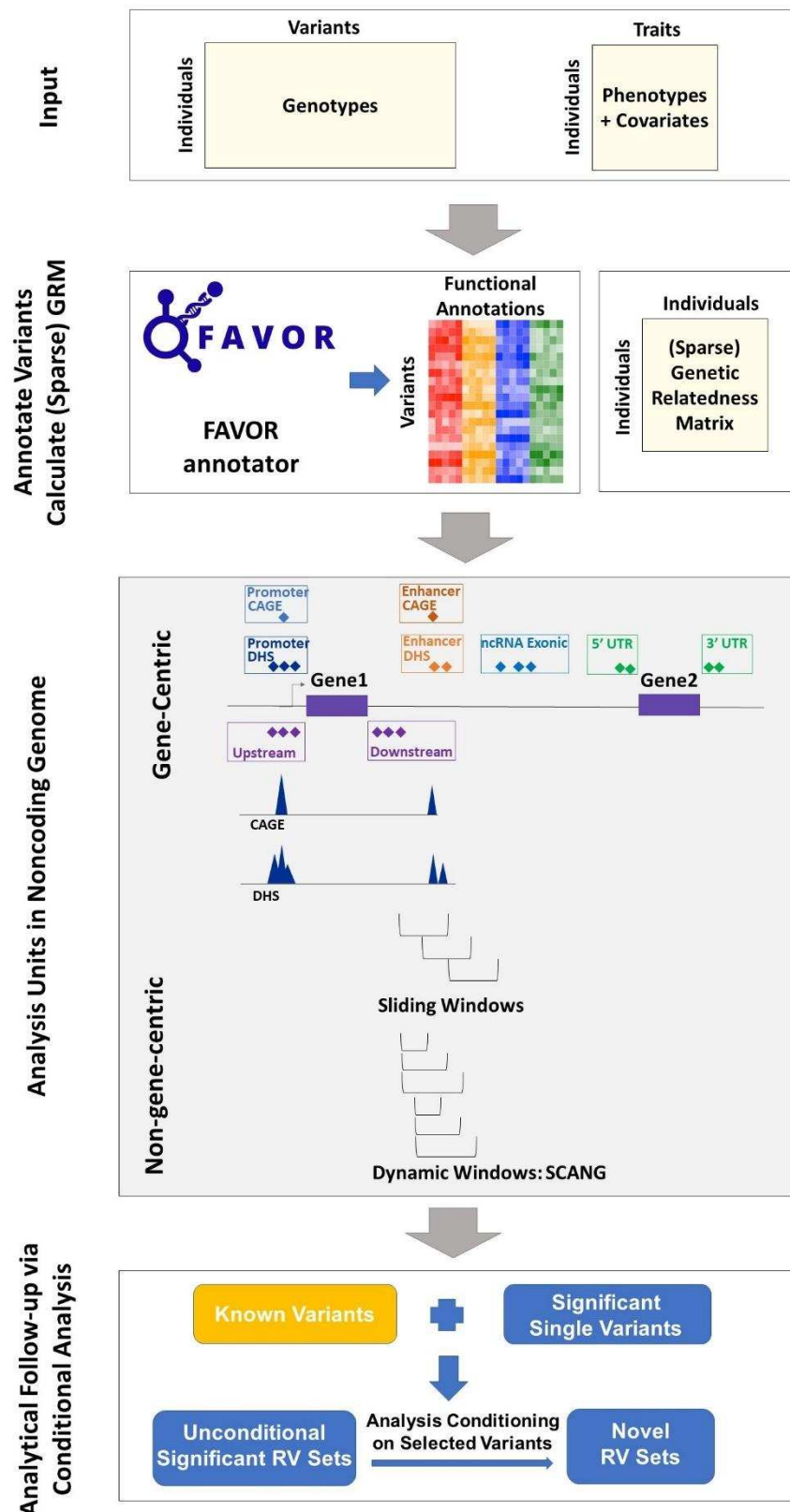
TABLES

Table 1. Gene-centric noncoding analysis results of both unconditional analysis and analysis conditional on known common and low-frequency variants. 21,015 discovery samples and 9,123 replication samples from the NHLBI Trans-Omics for Precision Medicine (TOPMed) program are considered in the analysis. Results for the conditionally significant genes (unconditional STAAR-O $P < 3.57 \times 10^{-7}$ and conditional STAAR-O $P < 1.16 \times 10^{-3}$ for 7 different noncoding masks across protein-coding genes; unconditional STAAR-O $P < 2.50 \times 10^{-6}$ and conditional STAAR-O $P < 1.16 \times 10^{-3}$ for ncRNA genes) using discovery samples are presented in the table. Chr (Chromosome); Category (Functional category); #SNV (Number of rare variants (MAF < 1%) of the particular functional category in the gene); STAAR-O (STAAR-O P value); HDL-C (High-density lipoprotein cholesterol); LDL-C (Low-density lipoprotein cholesterol); TG (Triglycerides); TC (Total cholesterol); Variants Adjusted (Adjusted variants in conditional analysis); n/a, no variant adjusted in the conditional analysis.

Trait	Gene	Chr	Category	Discovery			Replication			Variants Adjusted
				#SNV	STAAR-O (Unconditional)	STAAR-O (Conditional)	#SNV	STAAR-O (Unconditional)	STAAR-O (Conditional)	
HDL-C	APOA1	11	enhancer_DHS	1862	2.19E-07	7.67E-07	1005	1.50E-03	3.17E-03	rs964184, rs12269901
LDL-C	LDLR	19	upstream	68	2.35E-17	4.24E-04	27	5.58E-01	6.31E-01	rs12151108, rs688, rs6511720
	LDLR	19	promoter_CAGE	131	1.88E-17	3.37E-04	56	2.51E-02	9.50E-02	rs12151108, rs688, rs6511720
	APOE	19	promoter_CAGE	91	1.45E-11	4.88E-12	35	1.86E-01	4.36E-02	rs7412, rs429358, rs35136575
	LDLR	19	promoter_DHS	257	4.03E-17	7.21E-04	113	5.74E-02	2.27E-01	rs12151108, rs688, rs6511720
	APOE	19	promoter_DHS	162	9.81E-11	3.41E-12	64	7.45E-02	3.42E-02	rs7412, rs429358, rs35136575
	LDLR	19	enhancer_CAGE	150	2.82E-17	5.01E-04	71	1.20E-02	4.05E-02	rs12151108, rs688, rs6511720
	APOE	19	enhancer_DHS	239	9.84E-11	2.03E-11	112	2.55E-01	1.34E-01	rs7412, rs429358, rs35136575
	CTC-527H23.4	16	ncRNA	32	1.15E-06	1.15E-06	17	9.12E-01	9.12E-01	n/a
TG	APOE	19	promoter_CAGE	92	4.45E-12	7.48E-06	36	9.45E-06	3.53E05	rs12721054, rs5112, rs429358
	APOA5	11	promoter_DHS	175	2.39E-08	3.47E-05	84	1.19E-04	8.78E-03	rs964184, rs9804646, rs3135506, rs2266788
	APOE	19	promoter_DHS	163	1.80E-11	9.56E-06	65	2.96E-06	1.13E-05	rs12721054, rs5112, rs429358
	COL18A1	21	enhancer_CAGE	256	1.92E-07	1.92E-07	147	4.57E-02	4.57E-02	n/a
	APOE	19	enhancer_DHS	241	2.02E-11	8.44E-05	116	1.12E-05	4.15E-05	rs12721054, rs5112, rs429358

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

Trait	Chr	Start Location	End Location	Gene	Discovery			Replication			Variants Adjusted
					#SNV	STAAR-O (Unconditional)	STAAR-O (Conditional)	#SNV	STAAR-O (Unconditional)	STAAR-O (Conditional)	
HDL-C	8	57,071,644	57,073,643	Intergenic (IMPAD1)	111	1.79E-08	1.79E-08	53	8.38E-01	8.38E-01	n/a
	11	116,802,930	116,804,929	Intergenic (ZPR1)	135	1.25E-08	4.31E-08	76	9.49E-05	2.02E-04	rs964184, rs12269901
	11	117,146,930	117,148,929	Intron (PAFAH1B2)	165	5.98E-09	8.28E-08	98	6.02E-04	1.12E-03	rs964184, rs12269901
	11	117,147,930	117,149,929	Intron (PAFAH1B2)	168	8.85E-09	1.22E-07	96	8.72E-04	1.64E-03	rs964184, rs12269901
	16	56,760,029	56,762,028	Intron (NUP93)	132	1.38E-08	9.65E-06	68	2.45E-01	1.15E-01	rs247616, rs5883, rs7499892, rs17231520, rs5880
	16	56,761,029	56,763,028	Intron (NUP93)	141	1.50E-08	1.09E-05	73	5.87E-01	2.26E-01	rs247616, rs5883, rs7499892, rs17231520, rs5880
LDL-C	1	55,333,498	55,335,497	Intergenic (GOT2P1)	171	6.66E-16	5.81E-07	95	1.27E-06	5.81E-07	rs11591147, rs28362263, rs505151, rs12117661, rs472495
	1	55,334,498	55,336,497	Intergenic (GOT2P1)	148	5.55E-16	5.49E-07	81	1.20E-06	5.49E-07	rs11591147, rs28362263, rs505151, rs12117661, rs472495
TG	11	117,146,930	117,148,929	Intron (PAFAH1B2)	164	7.81E-19	4.13E-18	93	2.17E-17	5.66E-17	rs964184, rs9804646, rs3135506, rs2266788
	11	117,147,930	117,149,929	Intron (PAFAH1B2)	165	1.15E-18	6.11E-18	94	3.47E-17	9.13E-17	rs964184, rs9804646, rs3135506, rs2266788
	19	44,882,528	44,884,527	Intron (NECTIN2)	145	1.06E-08	2.18E-07	88	2.71E-02	8.07E-01	rs12721054, rs5112, rs429358
TC	1	55,333,498	55,335,497	Intergenic (GOT2P1)	175	1.98E-13	3.83E-14	101	5.84E-07	1.88E-07	rs11591147, rs28362263, rs505151, rs12117661, rs2495477
	1	55,334,498	55,336,497	Intergenic (GOT2P1)	149	1.80E-13	3.49E-14	90	5.53E-07	1.78E-07	rs11591147, rs28362263, rs505151, rs12117661, rs2495477
	19	44,894,528	44,896,527	Intron (TOMM40)	180	2.73E-10	8.95E-08	97	2.68E-03	4.22E-01	rs7412, rs429358, rs12721054

Table 3. Dynamic window analysis results of unconditional analysis and analysis conditional on known common and low-frequency variants. 21,015 discovery samples and 9,123 replication samples from the NHLBI Trans-Omics for Precision Medicine (TOPMed) program are considered in the analysis. Results for the conditionally significant sliding windows (unconditional genome-wide error rate $GWER < 0.05$; conditional STAAR-S $P < 5.56 \times 10^{-4}$) using discovery samples are presented in the table. Chr (Chromosome); Start Location (Start location of the dynamic window); End Location (End location of the dynamic window); #SNV (Number of rare variants (MAF < 1%) in the dynamic window; GWER (genome-wide error rate); STAAR-S (STAAR-S P value); HDL-C (High-density lipoprotein cholesterol); LDL-C (Low-density lipoprotein cholesterol); TG (Triglycerides); TC (Total cholesterol); Variants Adjusted (Adjusted variants in conditional analysis). Physical positions of each window are on build hg38.

Trait	Chr	Start Location	End Location	Gene	Discovery				Replication			Variants Adjusted
					#SNV	GWER	STAAR-S (Unconditional)	STAAR-S (Conditional)	#SNV	STAAR-S (Unconditional)	STAAR-S (Conditional)	
HDL-C	11	116,866,780	116,867,288	Intron (SIK3)	40	0.0295	2.24E-09	8.45E-09	19	2.22E-05	5.46E-05	rs964184, rs12269901
	11	116,928,564	116,929,045	Intron (SIK3)	40	0.0025	1.50E-10	4.43E-10	18	7.81E-04	1.06E-03	rs964184, rs12269901
LDL-C	1	55,335,150	55,335,701	Intergenic (GOT2P1)	40	<0.0005	8.58E-18	7.49E-19	21	9.29E-07	4.80E-07	rs11591147, rs28362263, rs505151, rs12117661, rs472495
	19	11,319,992	11,320,870	Intron (TSPAN16)	60	0.02	1.44E-09	3.16E-05	41	5.04E-01	5.10E-01	rs12151108, rs688, rs6511720
TG	11	117,147,061	117,148,086	Intron (PAFAH1B2)	80	<0.0005	5.10E-16	8.55E-15	41	9.48E-19	3.44E-18	rs964184, rs9804646, rs3135506, rs2266788
	11	117,182,856	117,183,310	Intron (SIDT2)	40	<0.0005	3.96E-12	1.08E-11	15	3.77E-14	6.53E-14	rs964184, rs9804646, rs3135506, rs2266788
	11	117,349,560	117,350,171	Intron (CEP164)	50	0.013	1.08E-09	1.26E-09	29	4.12E-11	6.39E-11	rs964184, rs9804646, rs3135506, rs2266788
TC	1	55,291,905	55,293,502	Intergenic (GOT2P1)	140	0.0055	3.17E-10	8.77E-05	68	4.76E-01	2.30E-01	rs11591147, rs28362263, rs505151, rs12117661, rs2495477
	1	55,335,119	55,335,584	Intergenic (GOT2P1)	40	<0.0005	1.63E-15	4.44E-16	26	2.23E-07	7.03E-08	rs11591147, rs28362263, rs505151, rs12117661, rs2495477
	19	11,319,627	11,320,925	Intron (TSPAN16)	110	<0.0005	2.95E-12	2.32E-05	75	3.40E-01	5.90E-01	rs73015024, rs688, rs2278426, rs6511720