

Spring 5-2020

STATISTICAL METHODS FOR FUNCTIONAL ANNOTATION-BASED RARE VARIANT ASSOCIATION ANALYSIS

YIDING MA

UTHealth School of Public Health

Follow this and additional works at: https://digitalcommons.library.tmc.edu/uthsph_dissertsopen



Part of the [Community Psychology Commons](#), [Health Psychology Commons](#), and the [Public Health Commons](#)

Recommended Citation

MA, YIDING, "STATISTICAL METHODS FOR FUNCTIONAL ANNOTATION-BASED RARE VARIANT ASSOCIATION ANALYSIS" (2020). *UT School of Public Health Dissertations (Open Access)*. 128.
https://digitalcommons.library.tmc.edu/uthsph_dissertsopen/128

This is brought to you for free and open access by the School of Public Health at DigitalCommons@TMC. It has been accepted for inclusion in UT School of Public Health Dissertations (Open Access) by an authorized administrator of DigitalCommons@TMC. For more information, please contact digitalcommons@library.tmc.edu.

STATISTICAL METHODS FOR FUNCTIONAL ANNOTATION-BASED RARE VARIANT
ASSOCIATION ANALYSIS

by

YIDING MA, BS, MS

APPROVED:



WENYAW CHAN, PHD



PENG WEI, PHD



JOHN M. SWINT, PHD



BING YU, PHD



DEAN, THE UNIVERSITY OF TEXAS
SCHOOL OF PUBLIC HEALTH

Copyright
by
Yiding Ma, BS, MS, PHD
2020

STATISTICAL METHODS FOR FUNCTIONAL ANNOTATION-BASED RARE
VARIANT ASSOCIATION ANALYSIS

by

YIDING MA
BS, Peking University, 2008
MS, Michigan State University, 2011
MS, University of Michigan, 2013

Presented to the Faculty of The University of Texas

School of Public Health

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS
SCHOOL OF PUBLIC HEALTH
Houston, Texas
May 2020

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my dissertation chair, Dr. Peng Wei for his perfect guidance and valuable suggestions, as well as the generous financial support he provided for this dissertation work. Particularly, I appreciate his understanding for the difficulty in my life. It was great to work with him in the past years and I really appreciate all his help and supports.

I am deeply grateful to my committee chair, Dr. Wenyaw Chan; my minor advisor, Dr. John Swint; my breadth advisor, Dr. Bing Yu and my external reviewer, Dr. Han Chen. I felt very grateful that I had all of them as my committee members. Without them, I would never have completed my PhD degree in Biostatistics. I would also like to thank Dr. Taylor Maxwell for our collaborative project and fun conversations.

Finally, I want to acknowledge my parents. And I owe my deepest gratitude to my friends: Tianzhong, Ying, Yue-ming, Ziqiao and so many others, for their encouragement and support in my most difficult times.

STATISTICAL METHODS FOR FUNCTIONAL ANNOTATION-BASED RARE VARIANT ASSOCIATION ANALYSIS

Yiding Ma, BS, MS, PHD
The University of Texas
School of Public Health, 2020

Dissertation Chair: Peng Wei, PHD

Despite ongoing large-scale population-based whole-genome sequencing (WGS) projects such as the TOPMed program, WGS-based association analysis of complex traits remains a tremendous challenge. External biological knowledge, such as functional annotations based on the ENCODE, Epigenomics Roadmap and GTEx projects, may be helpful in distinguishing causal rare variants from neutral ones; however, each functional annotation can only provide certain aspects of the biological functions. Our knowledge for selecting informative annotations *a priori* is limited and incorporating non-informative annotations will introduce noise and lose power. In the first part of this dissertation, we propose FunSPU, a versatile and adaptive test that incorporates multiple biological annotations. In addition to extensive simulations, we illustrate our proposed test using the TWINSUK cohort of UK10K WGS data based on six functional annotations. We identified genome-wide significant genetic loci on chromosome 19 near gene *TOMM40* and *APOC4-APOC2* associated with low-density lipoprotein (LDL), which are replicated in the UK10K ALSPAC cohort (n=1,497). These replicated LDL-associated loci were missed by existing rare variant association tests that either ignore external biological information or rely on a single source of biological knowledge.

Individual-level genetic data is not always accessible due to privacy concerns. Instead, summary association statistics are widely available based on large-scale meta-analysis of genome-wide association studies (GWASs). We further extend adaptive tests incorporating functional annotations to summary statistics (FunSPUs) in the second part of this dissertation. We show that our test can identify more significant genes compared to the corresponding annotation-ignorant tests. Moreover, we obtained several genome-wide significant loci associated with high-density lipoprotein (HDL) levels from a smaller meta-analysis of GWASs ($n=94,595$) which were reported by a follow-up meta-analysis with a larger sample size ($n=188,577$).

In the third part, we propose to evaluate the performance of functional annotations by partitioning the heritability of complex traits. We focused on rare variants from WGS data. Our proposed method is phenotype-specific and no “gold standard” variants are required. We used the Atherosclerosis Risk in Communities Study (ARIC) WGS data to estimate heritability and evaluated the performance of 12 functional annotations including conservation scores and ensemble deleteriousness prediction scores.

TABLE OF CONTENTS

List of Tables	i
List of Figures	ii
Chapter 1: Background	1
Literature Review.....	1
Whole genome sequencing (WGS) data	1
Functional annotations.....	3
Public Health Significance.....	4
Specific Aims.....	5
Chapter 2: A versatile and adaptive multiple functional annotation-based association test of whole-genome sequencing data	6
Introduction.....	7
Materials and Methods.....	9
Notations	9
Review of the data-adaptive aSPU test.....	10
New test: FunSPU - a data-adaptive test incorporating multiple annotations	11
Alternative approaches to incorporating multiple functional annotations: aSPU_minP and aSPU_Fisher	13
wtFunSPU: extension of FunSPU to allow for global weighting of multiple annotations.....	15
Results.....	16
Simulation setups	16
Simulation results.....	18
Application to the UK10K WGS data	20
Discussion	24
Supporting Information.....	34
Reference	55
Chapter 3: Summary statistics-based association analysis incorporating functional annotations	58
Introduction.....	58
Materials and Methods.....	61
Implement: Monte-Carlo Simulation.....	62
Simulation setups	62
Data availability	64
Results.....	64
Simulation results.....	64

Application to the lipid GWAS data	65
Discussion	67
Supporting Information	73
Reference	75
Chapter 4: A novel framework for comparing functional annotations of rare variants without gold standard	77
Introduction	77
Methods and Materials	79
Partition of the heritability and evaluation framework	79
Functional annotations and WGS data	80
Results	82
Discussion	86
Supporting Information	95
Reference	98
Chapter 5: Conclusion and Future Work	100
References	102

LIST OF TABLES

Table 2.1: Empirical type I error rates of various tests	28
Table 2.2: Genome-wide significant sliding windows identified by various tests in the UK10K TWINSUK cohort and replication in the ALSPAC cohort of UK10K.	29
Table 2.3: Computational time needed (mean and standard deviation (SD) with 32 rare variants (RVs) in an RV-set) for selected methods under comparison in the simulation study of power.....	34
Table 2.4: List of heritability of RVs by each category of functional annotation (TWINSUK cohort).	35
Table 2.5: Summary and comparison of the proposed tests.	39
Table 2.6: List of heritability of common variants (CVs; MAF>5%) by each category of functional annotation (TWINSUK WGS cohort) estimated by LD score regression.	40
Table 3.1: Empirical type I error rates of various tests at $\alpha = 0.005$	73
Table 4.1: Regression slopes and Pearson's correlations coefficients calculated by partitioning the heritability into various categories	95

LIST OF FIGURES

Figure 2.1: Empirical power of various tests for eight causal RVs and increasing number of nonassociated RVs (Scenario A).....	30
Figure 2.2: Empirical power of various tests for eight causal RVs and increasing number of nonassociated RVs (Scenario B).	31
Figure 2.3: Rescaled scores of functional annotations.....	32
Figure 2.4: Association test results for LDL at the locus around gene <i>APOC4-APOC2</i>	33
Figure 2.5: Heritability per SNV ($h^2/\#SNV$) of HDL sorted by category of functional annotation score (TWINSUK cohort).	42
Figure 2.6: Heritability per SNV ($h^2/\#SNV$) of LDL sorted by category of functional annotation score (TWINSUK cohort).	43
Figure 2.7: Heritability per SNV ($h^2/\#SNV$) of BMI sorted by category of functional annotation score (TWINSUK cohort).	43
Figure 2.8: Heritability per SNV ($h^2/\#SNV$) of SBP sorted by category of functional annotation score (TWINSUK cohort).	45
Figure 2.9: Distributions of genome-wide functional scores for rare variants (MAF < 5%) in the UK10K TWINSUK cohort.....	46
Figure 2.10. Pairwise correlation of rescaled scores of functional annotations across the whole genome.	47
Figure 2.11: Global quantile-quantile (QQ) plots for association analysis of rare variants with HDL in the UK10K TWINSUK cohort	48
Figure 2.12: Global QQ plots for association analysis of rare variants with LDL in the UK10K TWINSUK cohort	49
Figure 2.13: Global QQ plots for association analysis of rare variants with BMI in the UK10K TWINSUK cohort	50
Figure 2.14: Global QQ plots for association analysis of rare variants with SBP in the UK10K TWINSUK cohort	51
Figure 2.15: LocusZoom plots of association test results for LDL at the locus around <i>TOMM40</i> and <i>APOC4-APOC2</i> in the UK10K TWINSUK cohort	52
Figure 2.16: Association test results for coding rare variants (MAF < 5%) in association with LDL around genes <i>TOMM40</i> , <i>APOE</i> , and <i>APOC4-APOC2</i>	54
Figure 3.1: Empirical type I error rates of various summary statistics-based tests at $\alpha = 0.05$	69

Figure 3.2: Empirical power of various tests for eight causal SNPs and increasing number of neutral RVs.....	70
Figure 3.3: Manhattan plots for the association test results for trait HDL based on the 2013 lipid data.....	71
Figure 3.4: Manhattan plots for the association test results for trait HDL based on the 2010 lipid data.....	72
Figure 3.5: The Q-Q plots for the association analysis of simulated data	73
Figure 3.6: The Q-Q plots for association tests of the summary statistics with HDL in the 2013 lipid data.....	74
Figure 4.1: Illustration of the proposed framework for comparing functional annotations of rare variants	89
Figure 4.2: Phenotype-specific rank of regression slopes estimated by partitioning ARIC RVs for conservation scores	90
Figure 4.3: Phenotype-specific rank of regression slopes estimated by partitioning UK10K TwinsUK RVs for conservation scores	91
Figure 4.4: Phenotype-specific rank of regression slopes estimated by partitioning ARIC RVs for ensemble functional annotations	92
Figure 4.5: Regression slopes (vertical axis) and Pearson's correlations coefficients (horizontal axis) calculated by partitioning the heritability for apo A1 levels.	93
Figure 4.6: Regression slopes (vertical axis) and Pearson's correlations coefficients (horizontal axis) calculated by partitioning the heritability for BMIs	94
Figure 4.7: Phenotype-specific rank of regression slopes estimated by partitioning ARIC RVs for all functional scores	95
Figure 4.8: Phenotype-specific rank of regression slopes estimated by partitioning UK10K TwinsUK RVs for all functional scores	96
Figure 4.9: Phenotype-specific rank of regression slopes estimated by partitioning UK10K ALSPAC RVs for all functional scores.....	97

CHAPTER 1: BACKGROUND

Literature Review

Whole genome sequencing (WGS) data

DNA sequencing is the act of determining the nucleotide sequence of given DNA molecules. The earliest DNA sequences were obtained in the early 1970s, following by the first revolution in the DNA sequencing techniques by Frederick Sanger and colleagues [1,2] . The Sanger's Method dominated DNA sequencing for the following decades and lead to the determination of a complete human reference genome sequence [3,4].

In the 2000's, the next-generation sequencing (NGS) technique was developed. Their efficient design with respect to the labor and reagents has triggered a steady drop in sequencing cost. Numerous NGS platforms [5,6] have been launched or are announced. By 2007, it was possible to sequence over 500Mb a day on a single machine [7], and that was when the 1000 Genomes Project was founded to perform low-coverage (2-4X) sequencing on up to 2,500 human genomes. In 2010, HiSeq 2000 sequencing system by Illumina can generate two billion paired-end reads and 200Gb of quality filtered data in a single run, which allows researchers to obtain 30-fold coverage of two human genomes in a single run. With rapidly falling price of sequencing, genotyping is now shifting from the traditional, targeted approach to whole genome sequencing.

The desire to study low frequency and rare variants (LFVs and RVs respectively) in a genome-wide fashion was met by fast development in sequencing technologies. In recent years, large-scale whole-exome sequencing and whole-genome sequencing (WGS) data have been generated, such as those in the Exome Sequencing Project [8], the UK10K project [9]

and the ongoing NIH NHLBI Trans-Omics for Precision Medicine (TOPMed) WGS program [10],

Complex diseases, such as obesity and type 2 diabetes, have a considerable burden on population health. [11]. Genome-wide association studies (GWAS) have identified thousands of genetic loci associated with a wide range of complex diseases and trait, and common variants identified by GWAS have proven highly informative to identify novel biological processes underlying common disease [12]. However, common variants discovered from early GWAS explained only a small proportion of phenotypic variance for most common traits. Therefore, including GWAS signals on top of risk factors was not able to improve the predictive value in clinical usage.

The missing heritability theory [13] hypothesized that GWAS might have missed variants that have large effects but too low frequency to be detected by SNP array. This is also supported by the evolution theory that alleles susceptible to diseases and their risks are likely to be deleterious and could not reach high frequency due to purifying selection [14] [15]. Some clinical studies have shown that rare variants contribute to several complex neurodevelopmental disorders [16] [17]. The variants with low to rare frequency could be where a large proportion of missing heritability resides. With the improvement of technologies and the increased capacity to identify rare variants, WGS-based association studies are expected to provide further opportunities for the discovery of variants that have larger and even causal effects.

Functional annotations

The definition of genome annotation is supplementing the DNA sequences with additional layers of information[11,12]. It can be classified as structural annotation, which identifies important genomic elements such as genes, the precise localization of genes within the genome and the elucidation of exon/intron structures, and functional annotation, which deals with the biological function, regulation and expression analysis of these elements. We have already discovered that nearly 99% of the ~3.3 billion nucleotides that constitute the human genome do not code for proteins [3]. Recent genome-wide association studies (GWAS) also pointed out that many trait-associated loci, including ones that contribute to human diseases, also lie outside protein-coding regions [13-18]. These studies indicated the noncoding regions of the human genome is an essential resource of functionally significant elements with diverse gene regulatory and other functions.

It is noteworthy that there is no universal definition of function. Specifically, the geneticists aim to establish the biological relevance of a DNA segment. This approach is often considered as a gold standard for defining function although limited by high cost and low prevalence of some phenotypes. Alternatively, the evolutionary biologists evaluate selective constraint to show preferential conservation across evolutionary time. Finally, the molecular biologists focus on the evidence of molecular activity. Gene regulation studies have already elucidated many functional noncoding elements, including promoters, enhancers, silencers and insulators. RNA metabolism studies also defined noncoding RNA genes such as microRNAs, piRNAs, structural RNAs, and regulatory RNAs [19-22]. These noncoding functional elements are associated with distinctive chromatin structures that

display signature patterns of histone modifications, DNA methylation, DNase accessibility, and transcription factor occupancy [23]. Overall, each approach above provides complementary information in some respects, and functional elements identified by each approach are often quantitatively enriched for each other.

Public Health Significance

The majority of diseases are complex diseases and many of them have high incidence rate in the U.S. and worldwide, including CHD, hypertension, type 2 diabetes, Alzheimer's diseases, and more [24]. Complex diseases have been a big burden in public health. The development of complex diseases involves many genetic factors, environmental and behavioral factors, and their interactions. Although genetic studies of complex diseases have promising progress in last two decades, identification of missing heritability from rare variants remains a big challenge.

I will develop statistical methods that will provide researchers with more powerful and robust data-adaptive association tests for either rare variants (RVs) or summary statistics incorporating with functional annotations. Furthermore, an R package implementing the methods will be released as part of the methodology development, which will greatly facilitate the research community to use the new methods in real data analysis.

In conclusion, my dissertation work will provide useful methods and tools which lead to better understanding of the underlying genetic factors and explain the heritability of human complex diseases. In the long run, it may contribute to the prevention, diagnosis and cure of complex diseases.

Specific Aims

Both functional annotations and whole genome sequencing (WGS) datasets are growing rapidly in recent years. However, limited work has been done to incorporate functional annotations into association analysis of WGS data. To address the current needs and challenges, I propose the following three specific aims for my dissertation.

Aim 1: Develop a multiple functional annotation-based association test for rare variants. The proposed test is adaptive at both the annotation and variant levels simultaneously. It has potential to maintain high power even in the presence of many noise annotations and neutral RVs.

Aim 2: Incorporating functional annotations into summary statistics-based association analysis. In this aim, I will extend the data-adaptive test to the case with GWAS summary statistics incorporating functional annotations. No individual-level genotype or phenotype data is required in this approach.

Aim 3: Develop a novel framework for comparing functional annotations of rare variants without golden standard. This proposed framework will be used to select the potential global-informative functional annotations for association analysis or determine deleterious genetic variants in clinical practice. To my knowledge, no trait-specific comparison of functional annotations of rare variants has been developed.

**CHAPTER 2: A VERSATILE AND ADAPTIVE MULTIPLE FUNCTIONAL
ANNOTATION-BASED ASSOCIATION TEST OF WHOLE-GENOME
SEQUENCING DATA
JOURNAL ARTICLE**

Title of Journal Article

FunSPU: a versatile and adaptive multiple functional annotation-based association test of whole-genome sequencing data

Name of Journal Accepted

PLoS Genetics

Introduction

In recent years, large-scale whole-exome sequencing and whole-genome sequencing (WGS) data have been generated, such as those in the Exome Sequencing Project [1], the UK10K project [2] and the ongoing NIH NHLBI Trans-Omics for Precision Medicine (TOPMed) WGS program [3], providing unprecedented opportunities to investigate low-frequency variants (minor allele frequency [MAF] between 1% and 5%) and rare variants (RVs; $MAF < 1\%$) in association with complex diseases and traits. However, WGS-based association analysis of complex traits remains a tremendous challenge due to the large number of RVs, many of which are non-trait-associated neutral variants. External biological knowledge, such as functional annotations, might be informative to distinguish causal RVs from neutral ones. Some recent large-scale functional genomic studies, such as ENCODE [4], NIH Roadmap Epigenomics [5] and GTEx [6] projects, provide rich resources to use in characterizing the functional consequences of single nucleotide variants (SNVs), especially those in non-coding regions. Many approaches have been developed for functional annotations by integrating these data, e.g., CADD [7], GenoSkyline [8] and Eigen [9]; see Liu et al for a recent comparative review [10]. In WGS analysis, investigators may filter a subset of SNVs by annotations [2,11], or use a single source of functional scores as weights in association tests to boost the statistical power [12-14]; however, each functional annotation can only provide a certain aspect of the biological functions, e.g., sequence conservation across species or biochemical activity of non-coding regions in a tissue. Our *a priori* knowledge to select the informative annotation(s) regarding a phenotype and genomic regions of interest is limited, and incorporating noninformative annotations will introduce noise and lose power.

To address this analytical challenge, we propose a family of versatile and powerful tests called “FunSPU” that allow for incorporating multiple functional annotations simultaneously in the adaptive sum of powered score (aSPU) test framework [15]. The fundamental idea of aSPU is to construct a general class of association tests, each of which is the most powerful under varying, yet unknown, local genetic architecture, then data-adaptively select the most significant test. Since each functional annotation system contains limited biological knowledge, multiple sources of functional annotations may provide complementary information. Therefore, a test that integrates multiple functional annotations simultaneously is potentially powerful. The proposed test is adaptive at both the annotation and variant levels and thus maintains high power even in the presence of noninformative annotations and a large number of neutral RVs. We also propose minimum p -value (minP) and Fisher’s meta-analysis-like approaches to combine the p -values with respect to multiple annotations. Moreover, to further increase the statistical power, we propose to incorporate a trait-specific global weight for each annotation based on partitioning the heritability.

Using extensive simulations and application to the UK10K WGS data [2], we compared our proposed FunSPU tests with the corresponding annotation-ignorant aSPU test as well as some existing RV association tests, such as the T5 burden test and SKAT [16]. We also compared our method with a recently published multiple functional annotation-based association test called functional score test (FST) [17]. Using the UK10K TWINSUK WGS cohort as the discovery sample ($n=1,752$), we considered six functional annotations, CADD [7], RegulomeDB [18], FunSeq [19], Funseq2 [20], GERP++ [21] and GenoSkyline [8], and four quantitative traits, low-density lipoprotein (LDL), high-density lipoprotein (HDL), body

mass index (BMI) and systolic blood pressure (SBP). We identified genome-wide significant genetic loci on chromosome 19 near gene *TOMM40* and *APOC4-APOC2* that are associated with LDL, which are replicated in the UK10K ALSPAC WGS cohort (n=1,497). These replicated LDL-associated loci were missed by existing RV association tests that either ignore external biological information or rely on a single source of biological knowledge. We have implemented the proposed test in an R package “FunSPU”.

Materials and Methods

Notations

Suppose that for subject $i = 1, \dots, n$, $\mathbf{Y} = (Y_1, \dots, Y_n)$ is the vector of a trait, and $\mathbf{X}_i = (X_{i1}, \dots, X_{ik})'$ is the vector of the genotype scores of k RVs, for example, from a gene or some genomic region. Here, we use additive coding for each RV; that is, X_{ij} is the count of the minor allele at RV j for subject i . For simplicity, we ignore other covariates in our model. We consider a generalized linear model (GLM):

$$g(E(Y_i)) = \beta_0 + \sum_{j=1}^k X_{ij} \beta_j,$$

where g is a link function; for continuous Y_i , g is the identity link $g(\mu) = \mu$ and the GLM is reduced to a linear model, whereas g is the logit link $g(\mu) = \log(\frac{\mu}{1-\mu})$ for binary Y_i . For the purpose of exposition, we introduce our proposed tests in the linear model framework with a quantitative trait and no covariates, though the methods can be similarly extended to binary traits, and adjusted for covariates in the GLM and score function framework [15,22,23].

We test the null hypothesis $H_0: \beta = (\beta_1, \dots, \beta_k)' = 0$, that is, there is no association between any of the RVs and the trait under H_0 . Our proposed tests are based on the score vector $\mathbf{U} = (U_1, \dots, U_k)'$ for β and its covariance matrix \mathbf{V} ,

$$U = \sum_{i=1}^n (Y_i - \bar{Y})X_i, \quad V = \text{Cov}(U|H_0) = \bar{Y}(1 - \bar{Y}) \sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})',$$

where \bar{Y} and \bar{X} are the sample means of the Y_i 's and X_i 's, respectively.

Review of the data-adaptive aSPU test

Pan et al. [15] proposed a new adaptive test that retains high power across a wide range of varying, yet unknown, genetic architecture for the analysis of RVs. This test is based on a class of the SPU test:

$$T_{SPU(\gamma)}(U) = \sum_{j=1}^k U_j^\gamma,$$

where $\gamma \geq 1$ is a positive integer. Suppose that we have a set of candidate values of γ in Γ , e.g., $\Gamma = \{1, 2, \dots, 8, \infty\}$, as used in our later experiments. It is known that SPU(1) is equivalent to the burden test, while SPU(2) is a variance-component score test equivalent to SKAT with a linear kernel. Importantly, as γ increases (as an even integer), the $SPU(\gamma)$ test puts more weights on the larger component of U while gradually ignoring the remaining component. In particular, we have $T_{SPU(\gamma)} \propto \|U\|_\gamma = (\sum_{j=1}^k |U_j|^\gamma)^{\frac{1}{\gamma}} \rightarrow \|U\|_\infty = \max_{1 \leq j \leq k} |U_j|$, as $\gamma \rightarrow \infty$. The SPU(∞) is closely related to the minP test (but ignores possibly varying variances of U_j 's); the two tests often perform similarly [24]. Since the power of an $SPU(\gamma)$ test depends on the choice of γ while the optimal choice of γ depends on the unknown true association pattern of the RVs to be tested, it would be desirable to data-adaptively choose the value of γ . To this end, the aSPU test takes the minimum p -value of the SPU(γ) tests as its test statistic: $T_{aSPU} =$

$\min_{\gamma \in \Gamma} p_{SPU(\gamma)}$. In this case, T_{aSPU} is no longer a genuine p -value; we use resampling

approaches such as residual permutation or parametric bootstrap to obtain its p -value.

New test: FunSPU - a data-adaptive test incorporating multiple annotations

Our proposed test is in the data-adaptive aSPU test framework. Importantly, the proposed test is adaptive at both the annotation and SNV levels. Suppose that we have the score vector $U = (U_1, \dots, U_k)'$ for k RVs from a gene region or sliding window based on a linear regression model. Let $0 \leq w_{lj} \leq 1$ denote the functional score from the l th of m properly scaled annotations for the j th of k RVs. The proposed functional annotation-based SPU test is

$$T_{SPU-F}(\gamma_a, \gamma) = \sum_{l=1}^m \left[\left(\sum_{j=1}^k (w_{lj} U_j)^\gamma \right)^{\frac{1}{\gamma}} \right]^{\gamma_a},$$

where two positive integers $\gamma \geq 1$ and $\gamma_a \geq 1$ respectively control the individual variants' and annotations' relative contributions to the overall test statistic; e.g., $\gamma_a = 1$ treats all annotations equally, while $\gamma_a = \infty$ only considers the most significant annotation. The inner sum of weighted U_j with power γ is the weighted SPU, and they are normalized to the power of $1/\gamma$ before being subjected to the outer sum with power γ_a . Since the number of the RVs in this test statistic is identical across all m annotations, it is not necessary to further normalize the weighted SPU test by the number of RVs.

The intuition to use γ_a as the powers of the weighted SPU is similar to that for γ . In general, a smaller γ_a , e.g., $\gamma_a = 1$, is more effective when there are more informative annotations, each of which is roughly equally discriminative regarding the deleteriousness of the RVs for the trait of interest. In contrast, a larger γ_a is preferred if there is only one or

fewer informative annotations that can well distinguish causal variants from neutral ones for the trait. As $\gamma_a \rightarrow \infty$, only the most significant weighted SPU is considered.

We aim to perform powerful tests when there are unknown association patterns of RVs and unknown informativeness of functional annotations. In practice, since we have no *a priori* knowledge about choosing γ and γ_a , we need to conduct a grid search over a set of possible values of both γ and γ_a . However, searching too many values will introduce extra variability and lead to reduced power. This effect was later confirmed when we used $\gamma_a \in \{1, 2, 3, \dots, 8, \infty\}$ and $\gamma \in \{1, 2, 3, \dots, 8, \infty\}$ in some preliminary simulations. Based on the results of aSPU tests [15] and the feature of annotations, we decided to use $\gamma_a \in \Gamma_a = \{1, 2, 4, 8, \infty\}$ and $\gamma \in \Gamma = \{1, 2, 3, \dots, 6\}$ for the rest of the study. We retained $\gamma_a = \infty$ as an approximation to the minP test and ignored some higher values of γ since the results tend to be similar to $\gamma = 6$.

Given a set of γ and γ_a , e.g., $\gamma \in \Gamma = \{1, 2, 3, \dots, 6\}$ and $\gamma_a \in \Gamma_a = \{1, 2, 4, 8, \infty\}$, the proposed data-adaptive FunSPU test statistic is defined as

$$T_{FunSPU} = \min_{\gamma \in \Gamma, \gamma_a \in \Gamma_a} p_{SPU-Fun(\gamma_a, \gamma)},$$

where $p_{SPU-Fun(\gamma_a, \gamma)}$ is calculated by the resampling methods detailed below. Although the score vector U has an asymptotic normal distribution $N(0, V)$, it is not easy to derive the asymptotic distribution of T_{FunSPU} . Therefore, we propose using a single layer of permutations (without covariates) or residual permutations (with covariates) to obtain p -values as done in aSPU [15, 22]. Specifically, we first permute the original set of trait Y to obtain a new set of $Y^{(b)}$ for $b = 1, \dots, B$. Then, we calculate the null score vector $U^{(b)}$ and the corresponding test statistic $T_{SPU-Fun(\gamma_a, \gamma)}^{(b)} = T_{SPU-Fun(\gamma_a, \gamma)}(U^{(b)})$ as well as their p -values

$p_{SPU-Fun(\gamma_a, \gamma)}^{(b)} = \left[\sum_{b_1 \neq b} I \left(\left| T_{SPU-Fun(\gamma_a, \gamma)}^{(b_1)} \right| \geq \left| T_{SPU-Fun(\gamma_a, \gamma)}^{(b)} \right| \right) + 1 \right] / B$. Therefore, we have

$T_{FunSPU}^{(b)} = \min_{\gamma \in \Gamma, \gamma_a \in \Gamma_a} p_{SPU-F}^{(b)}(\gamma_a, \gamma)$, and the final p -value of the FunSPU test $p_{FunSPU} =$

$$\left[\sum_{b=1}^B I \left(T_{FunSPU}^{(b)} \leq T_{FunSPU} \right) + 1 \right] / (B + 1).$$

In the FunSPU test above, we ignored the possibly different variances of the score function component U_j , for example, due to varying MAF of the RVs. On the other hand, previous research has shown that it may be beneficial to account for the heterogeneity of variances in the SPU framework [24]. Therefore, we further propose an inverse-variance weighted version of FunSPU:

$$T_{SPUW-Fun(\gamma_a, \gamma)} = \sum_{l=1}^m \left[\left(\sum_{j=1}^k \left(w_{lj} U_j / \sqrt{V_{jj}} \right)^\gamma \right)^{\frac{1}{\gamma}} \right]^{\gamma_a},$$

$$T_{FunSPUW} = \min_{\gamma \in \Gamma, \gamma_a \in \Gamma_a} p_{SPUW-Fun(\gamma_a, \gamma)},$$

where V_{jj} is the j th diagonal element of $V = \text{Cov}(U|H_0)$ as given before.

Alternative approaches to incorporating multiple functional annotations: aSPU_minP and aSPU_Fisher

We considered alternative approaches to incorporate multiple functional annotations into the aSPU test. In contrast to the two-level FunSPU approach, we can obtain modified aSPU tests via the score vector U weighted by each functional annotation, i.e., $T_{SPU(\gamma)}^{(l)}(U) = \sum_{j=1}^k (w_{lj} U_j)^\gamma$ and $T_{aSPU}^{(l)} = \min_{\gamma \in \Gamma} p_{SPU(\gamma)}^{(l)}$, for $l = 1, \dots, m$. We can obtain the genuine p -value $p_{aSPU}^{(l)}$ by resampling methods. To combine multiple functional annotations, we can further

employ some general approaches to combine multiple p -values, $p_{aSPU}^{(l)}$. For example, we can simply use $T_{aSPU_minP} = \min_{1 \leq l \leq m} p_{aSPU}^{(l)}$ as the test statistic of m modified aSPU tests. This aSPU_minP test is similar, but not exactly equivalent to the case of FunSPU with $\gamma_a = \infty$: the latter chooses the maximum $|T_{SPU(\gamma)}^{(l)}|$ and then uses resampling methods to obtain a genuine p -value directly, while the aSPU_minP test calculates the empirical p -value $p_{aSPU}^{(l)}$ first, and then uses the minimum p -value T_{aSPU_minP} as the new test statistic and resampling to calculate the final p -value.

Another common method for combining p -values is Fisher's meta-analysis approach, i.e., $T_{aSPU_Fisher} = -2 \sum_{l=1}^m \ln(p_{aSPU}^{(l)})$. If the m p -values were independent, T_{aSPU_Fisher} would follow a chi-squared distribution with $2m$ degrees of freedom. However, our $T_{aSPU}^{(l)}$ tests are correlated via the score vector U . Hence, we also use resampling approaches to calculate the final p -value. We can similarly apply the inverse-variance weighted method to aSPU_minP and aSPU_Fisher tests, respectively denoted as aSPUw_minP and aSPUs_Fisher.

Of note, aSPU_minP is closely related to the FST test [17]. Specifically when we restrict $\gamma = 1, 2$, aSPU_minP is equivalent to FST_minP except for the up-weight of rarer variants and weighted sum approach to combine burden and dispersion test statistics in the latter, as compared to the minP approach in the former. Similarly, aSPU_Fisher is closely related to FST_Fisher.

wtFunSPU: extension of FunSPU to allow for global weighting of multiple annotations

In our proposed FunSPU test, we treated all m functional annotations equally *a priori* and completely relied on the data to adaptively combine multiple annotations in each test unit, for example, a sliding window. This may be less efficient in the presence of overall inferior or superior annotations for a trait of interest, in which case it would be desirable to globally down-weight inferior annotations (and up-weight superior annotations). To this end, we propose to modify the FunSPU test by introducing an annotation-level weight $\rho = (\rho_1, \dots, \rho_m)'$ and denote the modified test as wtFunSPU:

$$T_{wtSPU-Fun(\gamma_a, \gamma)}^* = \sum_{l=1}^m \left[\rho_l \left(\sum_{j=1}^k (w_{lj} U_j)^\gamma \right)^{\frac{1}{\gamma}} \right]^{\gamma_a},$$

$$T_{wtFunSPU} = \min_{\gamma \in \Gamma, \gamma_a \in \Gamma_a} p_{wtSPU-Fun(\gamma_a, \gamma)}.$$

Since we assume no *a priori* knowledge regarding the informativeness of a functional annotation for a given trait, we propose to estimate ρ_l based on some global correlation measure between the annotation weights, genotypes and phenotype. A promising approach is based on partitioning the heritability h^2 by functional annotations [25]: a functional annotation is more informative for the trait of interest if SNVs with higher functional scores contribute to more heritability on average. Specifically, given an annotation, we first partition the genome-wide RVs based on Q discrete functional categories or percentiles of continuous functional scores; we then estimate the heritability h_q^2 for all SNVs in functional category $q = 1, \dots, Q$, using the GCTA tool [26]. We next compute the average per-SNV heritability $h_q^2 / \#SNV^{(q)}$ for each annotation category q and regress $h_q^2 / \#SNV_q$ on q to estimate the

slope: $E(h_q^2/\text{\#SNV}^{(q)}) = \beta_0 + \beta q$, where β is used as the global weight ρ for the corresponding functional annotation in the wtFunSPU test. Prior to this calculation, we transform the functional annotation to positive integers $q = 1, \dots, Q$ such that larger q corresponds to a more likely functional category. If a functional category has a very small number of SNVs or h_q^2 close to zero, this category is combined with a nearby category; see Figures 2.5 to 2.8 and Table 2.4 for details.

Results

Simulation setups

We conducted extensive simulations to evaluate and compare the performance of our proposed functional annotation-based tests with existing association tests for RVs. To make the simulation study representative of real RV data, we randomly selected 200 RVs from chr16:56.8M~57.1M of the UK10K TWINSUK genotype data of 1,718 unrelated individuals. MAFs of the selected RVs were no larger than 1%.

To evaluate power, we generated the simulated phenotypes as follows. First, we simulated 3 sets of informative annotations (w_{1j}, w_{2j}, w_{3j}) and 3 sets of random annotations (w_{4j}, w_{5j}, w_{6j}) independently ($j = 1, 2, \dots, 200$ ordered by genomic positions). We designated the first 100 RVs as causal variants ($j = 1, 2, \dots, 100$) and the remaining 100 RVs as neutral variants ($j = 101, 102, \dots, 200$). The informative annotations were generated from a uniform distribution $U(0.4, 1)$ corresponding to causal variants and from $U(0, 0.6)$ corresponding to neutral variants. All of the random annotations were generated from $U(0, 1)$. Second, we randomly selected $k = k_1 + k_2$ RVs: k_1 causal RVs from $j = 1, 2, \dots, 100$ and k_2 neutral RVs from $j = 101, 102, \dots, 200$. Third, we used only informative annotations to

calculate the effect size $\beta_j = c_\beta(w_{1j} + w_{2j} + w_{3j})$ for each causal RV. Fourth, the simulated phenotype was obtained from $Y_i = \sum_{j=1}^{k_1} X_{ij}\beta_j + \varepsilon_i$, where ε_i followed $N(0,3)$ and $i=1,2,\dots,1718$. Furthermore, to evaluate the globally weighted wtFunSPU test, we calculated the correlations between the sum of the genotypes weighted by each annotation and simulated phenotypes for each of the 1,000 simulation replications, and used the mean of the 1,000 correlations as the global weight of each annotation, i.e., $\rho_l = \overline{\text{cor}}(\tilde{Y}, \sum_{j=1}^k w_{lj} \tilde{X}_j)$ ($l = 1, 2, \dots, 6$), where \tilde{Y} and \tilde{X}_j are the vectors of Y_i and X_{ij} in each replication correspondingly.

We considered two simulation scenarios. In scenario A, we used all three informative annotations, three random annotations and one dummy annotation (1's for all RVs) in functional annotation-based tests (FunSPU, aSPU_minP, and others). To test the effect of more “noisy” annotations, we implemented scenario B, which used only one informative annotation, all three random annotations and one dummy annotation in the tests. In both scenarios A and B, we used identical procedure as above to generate simulated phenotypes Y_i , and fixed $k_1 = 8$ and $k_2 = \{8, 16, 32, 64, 128\}$, respectively. We set $c_\beta = 0.5$ for tests that incorporated global weights and $c_\beta = 1$ for other tests.

To evaluate the type I error rate, we simulated $Y_i \sim N(0,3)$ ($i=1,2,\dots,1718$), independent of k neutral RVs and 6 random annotations all from $U(0,0.6)$ in each replication. We set increasing numbers of neutral RVs with $k = \{8, 16, 32, 64, 128\}$.

The empirical type I error rate was calculated based on 50,000 replications with the significance level $\alpha = 0.005$, while the empirical power was calculated based on 1,000 replications for each scenario with $\alpha = 0.05$. For permutation-based tests, 10,000 and 1,000

resamplings were conducted for each replication to evaluate type I error and power, respectively.

Simulation results

As shown in Table 2.1, all the tests under comparison could control the type I error rate satisfactorily around 0.005, except for aSPU(w)_minP and aSPU(w)_Fisher tests, which were slightly inflated (between 0.006 and 0.007) with fewer number (e.g., 8) of neural variants. Besides Monte Carlo error, one possible reason for the slight inflation was that combining multiple annotations at the level of p-values might be sometimes numerically unstable in the presence of extreme p-values.

Regarding power, we first considered scenario A (Figure 2.1), which was an advantageous scenario for our proposed tests since all three informative annotations together with three random annotations and one dummy annotation were used in the tests. The dummy annotation (constant 1) was supposed to retain the unweighted SPU in the adaptive tests, as in aSPU. Although the simulated annotations for causal and neutral RVs had modest differences, i.e., from $U(0.4, 1)$ and $U(0, 0.6)$, respectively, the tests incorporating functional annotations, such as FunSPU, wtFunSPU, aSPU_minP and FST, always had higher power than tests that ignored functional annotations, such as aSPU, SKAT and T1. The FunSPU test appeared to be less powerful than aSPU_minP, suggesting a lack of efficiency in the former's complete data-adaptive strategy to combine multiple annotations. On the other hand, wtFunSPU and wtFunSPUw outperformed aSPU_minP and FST, supporting the effectiveness of the global weighting scheme. Between the latter two, aSPU_minP had an increasing edge over FST in the presence of larger number of neural variants, due to its going beyond burden (SPU($\gamma = 1$)) and variance-component (SPU($\gamma = 2$)) tests with additional γ

parameters. We also observed that the inverse-variance weighted tests always outperformed the original tests, e.g., wtFunSPUw versus wtFunSPU, and this advantage became more obvious with a higher proportion of neutral RVs. Lastly, the power of the aSPU_Fisher test was similar to that of the aSPU_minP test until the number of neutral variants increased to 64 and 128, when the former became less powerful than the latter.

Next, we considered a weaker scenario for our proposed tests. In scenario B (Figure 2.2), we used only one informative annotation, but all three random annotations and one dummy annotation in the tests. In this case, we had a higher proportion of “noisy” annotations in our tests. We observed that the FunSPU test was marginally more powerful than aSPU, SKAT and T1, but was less powerful than the aSPU_minP test by a large margin. In fact, scenario B was an advantageous scenario for the latter test, which only considered the most informative annotation. Similarly, aSPU_minP was more powerful than aSPU_Fisher when the number of neutral variants exceeded 16, due to the latter treating the one informative and three non-informative annotations equally. Finally, the globally weighted wtFunSPU and wtFunSPUw, especially the latter, were more powerful than the aSPU_minP and FST tests, again suggesting the benefit of globally down-weighting noninformative annotations.

We also compared the computational time needed for different methods. As shown in Table 2.3, FunSPU and aSPU_minP were on par with aSPU, but were more computationally intensive than the asymptotic-based burden and SKAT tests. As shown in the real data analysis later on, by employing a step-up permutation strategy, we were able to perform genome-wide scan of WGS data with FunSPU and related tests.

Application to the UK10K WGS data

To further evaluate the performance of our proposed tests on real data, we applied FunSPU and other state-of-the-art tests, including SKAT, T5 burden test and FST (combined test)[17], to association analysis of the UK10K WGS data with four complex quantitative traits: LDL, HDL, BMI and SBP. We used the TWINSUK samples as the discovery cohort and the ALSPAC samples as the replication cohort with $n=1706/1497$ (TWINSUK/ALSPAC), $1718/1497$, $1752/1792$ and $1740/1796$, respectively, for LDL, HDL, BMI and SBP, after merging WGS genotype and phenotype data. After removing SNVs that did not pass quality control (QC) as done in the original UK10K analysis[2], as well as singletons and INDELs, we had a total of 10,979,027 RVs and low-frequency variants with $MAF < 5\%$ in the discovery cohort. Briefly, the UK10K WGS data QC included various low-level variant calling and filtering QC measures, variant-level QC to exclude variants with Hardy-Weinberg equilibrium (HWE) test $p\text{-value} < 10^{-6}$, and sample-level QC to exclude samples in poor concordance with their corresponding GWAS data[2]. Since the discovery cohort TWINSUK only included women, we adjusted for age at baseline, but not gender, as a covariate in association testing in both discovery and replication cohorts.

We considered six types of functional annotations for RVs. CADD[7], FunSeq[19], FunSeq2[20], RegulomeDB[18] and GERP++[21] were extracted from the precomputed WGS [27] library, and GenoSkyline (blood) annotation was generated from the region-based GenoSkyline library [8]. We re-scaled all annotations to numerical weights within the interval (0, 1), with larger weights corresponding to a greater likelihood of being functional (Figure 2.9). Among the above annotations, rank scores for CADD, Funseq2, GenoSkyline and GERP++ were provided in the WGS library [27], and the re-scaled score was defined

as $w = (\text{raw rank score} - \text{min})/(\text{max} - \text{min})$, where min and max were, respectively, the minimum and maximum raw rank scores for a given functional annotation. The RegulomeDB categories $s = (1, 2, \dots, 6)$ were transformed into $(0, 1)$ by $f(s) = (7-s)/6$, whereas the Funseq categories $s = (0, 1, 2, \dots, 6)$ were transformed by $f(s) = (1+s)/7$. We substituted the missing values or zero values with 0.01 (FunSeq, FunSeq2, RegulomeDB) or 0.0001 (GERP++). There was no missing value in CADD and GenoSkyline for the RVs considered here. Figure 2.10 shows the pairwise correlation coefficients among the 6 annotations: while some annotations were moderately correlated ($r > 0.3$), for example, GERP++ with CADD, and Funseq2 with RegulomeDB/Genoskyline, others were much less correlated. This suggests that multiple annotations may provide complementary information regarding the functional consequence of genetic variants, and it may be beneficial to incorporate them simultaneously in association analysis as proposed in the FunSPU framework here. Following the procedure proposed in Section 2.5, we calculated the phenotype-specific weight for each of the six annotations and used them as global weights in the wtFunSPU test. As shown in Figures 2.5 to 2.8 and Table 2.4, RegulomeDB, Funseq and GenoSkyline tended to have consistently higher weights than GERP++, Funseq2 and CADD, while the numerical values and the relative magnitudes of the weights could vary across phenotypes.

We employed a sliding window approach to group RVs with a window length of 10k base pairs (bp) and a step size of 8.75k bp, resulting in 319,306 windows in total. Using the conservative Bonferroni procedure, we set the family-wise error rate at 0.05 with a significance level $= 0.05/319306 = 1.56e-07$, which equals 6.81 on the $-\log_{10}$ scale. To achieve this genome-wide significance level, we used a step-up permutation strategy [22,28]. We first performed $B=10,000$ permutations for all sliding windows and gradually increased

B ; if those sliding windows with estimated p -values $< 10/B$, we increased B to 10 times the current value and re-estimated the p -values for these sliding windows. The number of permutations in the final stage was $B=10^8$. Of note, the variant-specific score functions in aSPU, aSPU_minP, FunSPU and wtFunSPU were not weighted by MAF, while those in aSPUw, aSPUw_minP, FunSPUw, and wtFunSPUw were inverse-variance weighted, where variants with lower MAF were up-weighted. By default, SKAT and FST used Beta(1,25) weights to up-weight variants with lower MAF [16,17].

As shown in Figures 2.11 to 2.14, the quantile-quantile plots for the proposed FunSPU tests were well behaved, with no discernible indication of global p -value inflation, suggesting that the FunSPU tests could control the type I error rate well in genome-wide scans. Table 2.2 shows all sliding windows with at least one genome-wide significant p -value in the TWINSUK discovery cohort by any of the association tests under consideration. To confirm our findings in the TWINSUK cohort, we performed replication analysis of the genome-wide significant sliding windows in the ALSPAC cohort. As shown in Table 2.2, four sliding windows were replicated for the corresponding phenotypes and association tests with a replication p -value $< 0.05/24 = 2.1\text{e-}3$ based on the Bonferroni correction for 24 sliding windows: 3 by at least one of the functional annotation-based tests (1 by wtFunSPU, 1 by FunSPU and aSPU_minP and 1 by aSPUw_minP) and one by the aSPU test. In contrast, none of the 6 sliding windows identified by the FST test in the discovery cohort was replicated; neither did SKAT nor T5 replicate any sliding window.

Three of the four replicated sliding windows were close to each other on chromosome 19 around *TOMM40*, *APOE* and *APOC4-APOC2* genes. These loci have been previously identified and replicated to be associated with LDL by large-scale meta-analysis of GWAS

common variants [29-31]. Numerous functional and genetic association studies have shown that *APOE* plays a central role in lipoprotein metabolism and neurodegeneration [32-34]. Specifically, *APOE* has three isoforms, 2, 3, and 4: *APOE2* is associated with elevated plasma LDL level and increased cardiovascular disease risk, whereas *APOE4* is associated with increased risk of Alzheimer's disease[34]. While previous large-scale whole-exome sequencing and ExomeChip-based association studies did not identify exonic RVs in *APOE* associated with LDL[35,36], a recent association analysis of 16,324 deep-coverage WGS samples from the TOPMed project identified LDL-associated rare non-coding variants upstream of *APOE*[37]. Here we were able to identify the *TOMM40/APOE* locus and additionally *APOC4-APOC2* locus that harbor LDL-associated RVs with fewer than a couple of thousand samples, suggesting that the power of the FunSPU test was boosted by incorporating external biological knowledge.

We also investigated the effects of multiple annotations on the FunSPU tests. Although some high scores were observed around the *TOMM40* and *APOC4-APOC2* gene regions for Funseq2, Funseq, RegulomeDB and GenoSkyline (Figure 2.3), they did not appear to be obviously different from those scores outside these two loci. Figure 2.4 shows the association signals of selected tests in this genomic region, whereas Figure 2.15 shows all individual annotation-based aSPU tests. As for *APOC4-APOC2*, three of the six annotations, namely, Funseq2, RegulomeDB and GenoSkyline (Figure 2.15-E, H, I), positively contributed to the highly significant *p*-values of wtFunSPU and FunSPU (Figure 2.15-A, B), although none of these individual annotation-based aSPU tests would reach the genome-wide significance threshold, demonstrating the benefit of integrating multiple functional annotations in the FunSPU framework. RegulomeDB and GenoSkyline also had higher

global weights for LDL (Table 2.4 (B)), which further boosted the p -value of the wtFunSPU test to the genome-wide significance level. As for *TOMM40*, Funseq2, CADD and GERP++ (Figure 2.15-E,F,D) positively contributed to the genome-wide significance of FunSPU and aSPU_minP (Figure 2.15-A and Table 2.2); whereas wtFunSPU missed this locus due to its low global weighting of these three annotations (Table 2.6 (B)). This suggests that wtFunSPU, FunSPU and aSPU_minP may complement each other and may be used together in association analysis of WGS data.

To further investigate whether the *TOMM40* and *APOC4-APOC2* loci identified for LDL cholesterol were driven by coding RVs, we only retained nonsynonymous RVs in the original sliding windows in this region (13 nonsynonymous RVs out of total 784 RVs) and applied the aSPU test to each sliding window which had at least two RVs. For a sliding window with a single RV, we merged it with its neighboring sliding window. As shown in Figure 2.16, none of the sliding windows had a p -value < 0.01 , far less significant than the original association testing results (Figure 2.4). We therefore conclude that it is very unlikely the identified associations were driven by coding RVs.

Discussion

We have proposed a versatile and adaptive association test, FunSPU, to exploit multiple sources of biological knowledge in the analysis of WGS data. It is adaptive at both the annotation and variant levels, and thus maintains high statistical power, even in the presence of noninformative annotations and a larger number of neutral variants. We have further proposed a globally weighted wtFunSPU test to more effectively down-weight less informative functional annotations in a trait-specific manner. Using the UK10K WGS data, we demonstrated that our proposed FunSPU test and its extensions, including the wtFunSPU

and aSPU_minP tests, are more powerful tools to identify genome-wide significant loci than existing RV association tests that either ignore external biological information or rely on a single source of biological knowledge. The FunSPU family of tests would thus serve as a powerful and complementary tool for ongoing and future large-scale WGS studies, such as the NHLBI TOPMed project [3] of over 100,000 individuals and the UK Biobank [38] WGS project of 50,000 individuals. We have also summarized and compared the FunSPU family of tests in Table 2.5.

The six functional annotations we considered here are diverse in terms of resources and features. For example, GERP++[21] is a sequence conservation score, whereas other annotations are ensemble scores based on integrating multiple sources of features, such as various functional genomic assays in the ENCODE project[4] and eQTL evidence. As demonstrated in Figure 2.10, a majority of the annotations were only moderately correlated with each other, supporting our proposal to incorporate multiple annotations' approximately orthogonal yet complementary information regarding the functional consequence of RVs in the framework of the FunSPU association test. The FunSPU test can easily incorporate additional functional annotations, including some newly developed ones [10], such as fathmm-MKL[39], Eigen/Eigen-PC[9] and DeepSEA[40].

To further de-noise noninformative annotations, we proposed a novel trait-specific measure based on partitioning the heritability and used it as a global weight for each annotation in the wtFunSPU test. Interestingly, our proposal is along the line of estimating group-specific weights in the context of weighted hypothesis testing [41,42], though the latter is based on the mixture model, in contrast to the mixed model-based heritability partition here. Although it may look counterintuitive at first glance, our proposed data-dependent

global weights actually did not inflate the type I error rates in both simulations (Table 2.1) and the real data analysis, as evidenced by the QQ plots of wtFunSPU (Figures 2.11 to 2.14). The reason is that we used a much larger number of observations, i.e., RVs across the whole genome, to estimate a few annotation category-specific heritability parameters h^2 , based on which we derived a single global weight. This is in line with the “sieve principle”, which justifies using aggregated data to estimate a much smaller number of weights and then using them in subsequent hypothesis testing of small units of data (e.g., genes or sliding windows) with controlled family-wise error rate [42,43]. Our proposed measure also has the potential to be used to compare the discriminative performance of whole-genome annotations for a complex trait of interest, for which known deleterious and neutral variants are rarely available (see Table 2.6). This warrants further investigation. We have also applied the LD score regression method [44] to calculate the weights for common variants (MAF>5%) using the UK10K TWINSUK WGS data. As shown in Table 2.8, the weights were largely qualitatively similar to those derived from RVs, suggesting that our proposed strategy to infer the global weighting of annotations is quite robust.

We have some practical considerations for our proposed tests. First, some functional annotations are not well-defined across the whole genome, resulting in relatively high missing data rates, for example, 68% for Funseq; the missing scores may reduce the reliability of annotation-based association tests. On the other hand, considering multiple complementary functional annotations simultaneously may at least partially remedy the problem of missing information. Second, by employing parallel computing and a step-up residual permutation strategy for the FunSPU family of tests, we are able to perform computationally feasible genome-wide scans for WGS data. For example, in the UK10K

TWINSUK WGS data application, it took 24 hours for 500 computing cores to complete the sliding window-based FunSPU scan in R, including 10^8 residual permutations for the top sliding windows to reach the genome-wide significance threshold.

[Table 2.1: Empirical type I error rates of various tests](#)

All the tests were performed at significance level $\alpha = 0.005$ for increasing number of neutral RVs with 50,000 simulation replications ($B=10,000$ for resampling-based tests). Annotation-based tests were based on six random annotations. aSPU: adaptive sum of powered score test; aSPU_minP: combining multiple p -values of aSPU tests by minimum p approach; aSPU_Fisher: to combining multiple p -values of aSPU tests by Fisher's meta-analysis approach; FunSPU: multiple functional annotation-based SPU test; wtFunSPU: global weighted FunSPU; T1: burden test of variants with MAF smaller than 1%; SKAT: the sequence kernel association test; (w): inverse-variance weighted score function in the SPU framework.

Test	No. of neutral RVs				
	8	16	32	64	128
aSPU	0.0059	0.0057	0.0050	0.0046	0.0043
aSPU_minP	0.0067	0.0058	0.0054	0.0062	0.0043
aSPUw_minP	0.0060	0.0053	0.0044	0.0056	0.0056
aSPU_Fisher	0.0061	0.0054	0.0059	0.0054	0.0047
aSPUw_Fisher	0.0062	0.0062	0.0051	0.0050	0.0047
FunSPU	0.0053	0.0047	0.0043	0.0045	0.0037
FunSPUw	0.0057	0.0062	0.0041	0.0050	0.0037
wtFunSPU	0.0045	0.0046	0.0046	0.0037	0.0042
wtFunSPUw	0.0047	0.0056	0.0039	0.0034	0.0050
T1	0.0052	0.0053	0.0049	0.0053	0.0055
SKAT	0.0051	0.0050	0.0053	0.0045	0.0038

Table 2.2: Genome-wide significant sliding windows identified by various tests in the UK10K TWINSUK cohort and replication in the ALSPAC cohort of UK10K.

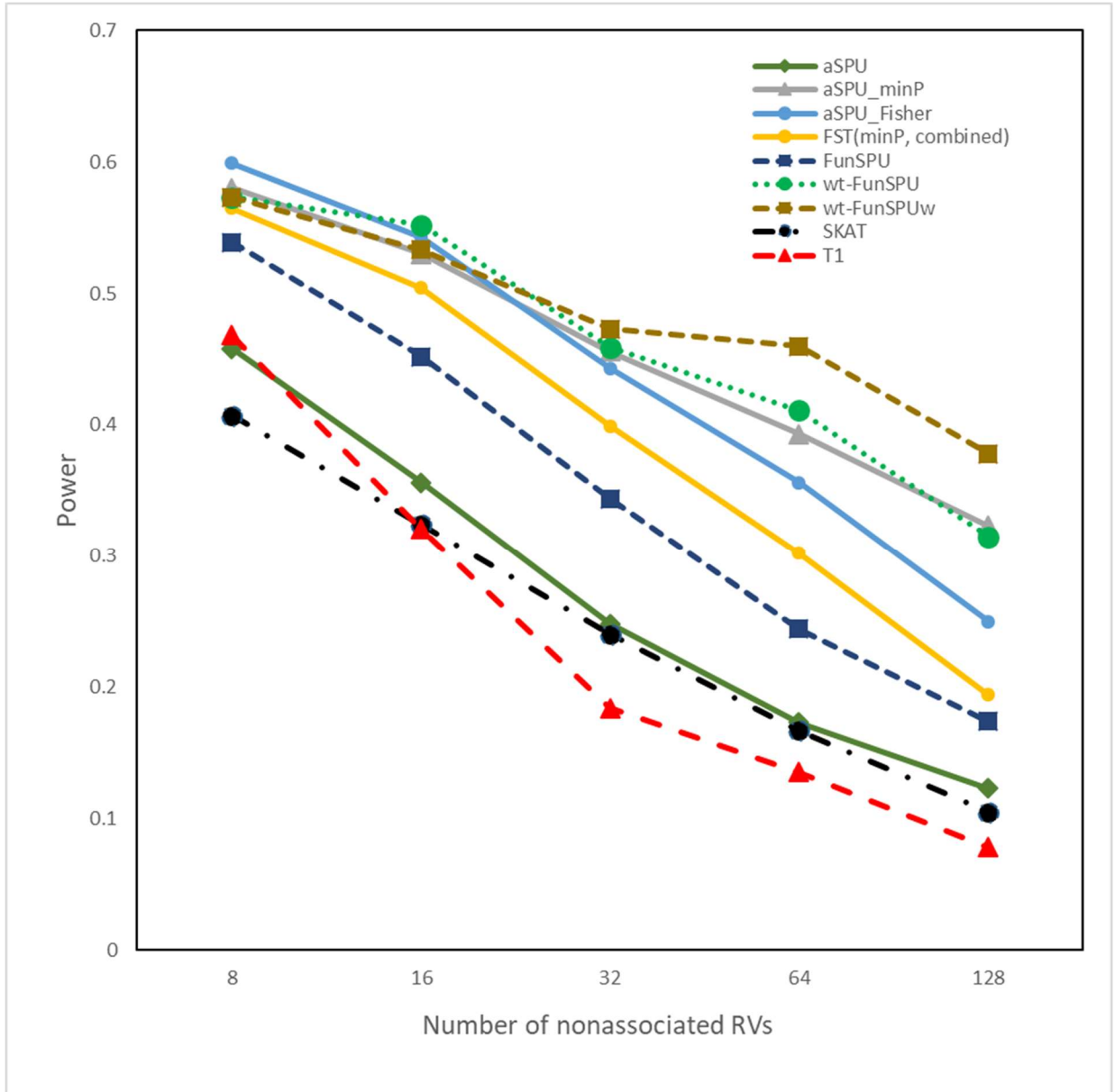
Significant p -values are in boldface; only significant p -values in the ALSPAC cohort were reported (TWINSUK p -value/ALSPAC p -value shaded when both are significant). cMAF: cumulative minor allele frequency. Base pair (bp) position based on reference genome hg19.

Trait	Chr	Start position- Stop position (bp)	Gene(s)	cMAF	# SNVs	p -Values									
						Globally weighted		Not globally weighted		aSPU _minP	aSPUw _minP	aSPU	FST (combined)	SKAT	T5
						wtFun -SPU	wtFun -SPUw	Fun- SPU	Fun- SPUw						
HDL	1	39,070,016-39,080,016	Intergenic	0.14/0.33	39/38	4.3e-2	2.3e-3	3.3e-2	<1e-8	2.0e-2	3.2e-6	3.6e-2	1.2e-3	5.2e-2	2.0e-4
HDL	2	173,903,159-173,913,159	<i>RAPGEF4</i>	0.084/0.21	33/31	<1e-8	<1e-8	2.4e-6	<1e-8	4.9e-6	8.9e-6	2.4e-5	2.5e-6	1.1e-6	1.2e-4
HDL	3	63,804,000-63,814,000	<i>C3orf49</i>	0.02/0.17	37/36	2.9e-3	<1e-8	4.0e-2	1.8e-5	3.1e-2	4.5e-5	0.12	3.2e-3	1.6e-2	5.2e-2
HDL	5	6,548,648-6,558,648	Intergenic	0.001/0.32	59/59	0.13	<1e-8	0.18	<1e-8	6.2e-2/ <1e-4	<1e-8/ <1e-4	0.15	2.0e-5	2.1e-3	0.19
HDL	5	6,557,398-6,567,398	Intergenic	0.34/0.26	48/43	4.3e-2	1.5e-2	0.13	<1e-8	9.5e-3	<1e-8	0.24	1.2e-8	1.8e-2	0.81
LDL	3	102,287,400-102,297,400	Intergenic	0.25/0.19	56/39	4.4e-5	9.9e-3	<1e-8	3.6e-2	9.0e-6	1.3e-2	8.1e-6	1.3e-4	2.5e-5	4.5e-3
LDL	3	102,427,400-102,437,400	Intergenic	0.16/0.27	32/40	<1e-8	5e-8	<1e-8	1.3e-3	1.2e-6/ <1e-4	5.1e-5	2.5e-4	3.3e-7	1.3e-4	0.68
LDL	5	43,259,958-43,269,958	<i>NIM1K</i>	0.43/0.29	41/46	9e-8	<1e-8	6.5e-7	3.0e-7	1.7e-5/ <1e-4	8.2e-6	2.5e-5	1.5e-5	2.4e-3	4.4e-6
LDL	12	13,771,517-13,781,517	<i>GRIN2B</i>	0.18/0.23	43/48	3.5e-3	5.8e-4	0.80	3.5e-5	5.9e-6	1.7e-6	0.60	2.4e-11	0.23	0.96
LDL	12	13,780,267-13,790,267	<i>GRIN2B</i>	0.26/0.27	45/43	6.0e-2	8.0e-4	0.47	1.3e-4	1.0e-6	1.7e-6	0.33	2.0e-11	4.9e-2	0.53
LDL	19	45,387,096-45,397,096	<i>PVRL2/ TOMM40</i>	0.21/ 0.22	33/ 37	5.4e-2/ 1.8e-3	0.10	3.5e-7/ <1e-4	5.0e-3/ 1.6e-3	2.1e-7/ <1e-4	4.5e-3	5.0e-8/ <1e-4	2.4e-4	2.4e-4/ 7.9e-6	0.25
LDL	19	45,395,846-45,405,846	<i>TOMM40</i>	0.42/0.59	65/62	8.6e-3/ <1e-4	8.6e-2	3.0e-8/ <1e-4	1.4e-4/ 1.9e-3	<1e-8/ <1e-4	5.8e-5/ <1e-4	5.0e-7	1.2e-4/ 1.0e-10	4.7e-4/ 1.1e-6	0.28
LDL	19	45,439,596-45,449,596	<i>APOC4- APOC2</i>	0.37/0.18	25/25	<1e-8/ <1e-4	1.1e-4	1.2e-6/ <1e-4	6.9e-4	2.1e-5/ <1e-4	1.4e-4	2.3e-4	4.7e-5/ 2.4e-7	1.1e-3/ 9.0e-4	0.15
BMI	3	35,619,294-35,629,294	Intergenic	0.24/0.24	39/40	<1e-8	4.2e-5	5.0e-5	1.1e-4	3.9e-5	3.6e-5	8.0e-5	2.9e-6	4.4e-5	1.5e-5
BMI	4	22,825,237-22,835,237	Intergenic	0.18/0.24	47/48	1.3e-3	5.2e-5	1.8e-3	6.0e-5	1.6e-4	3.4e-4/ <1e-4	2.8e-3	2.6e-8	1.4e-6	3.0e-5
BMI	10	13,937,041-13,947,041	<i>FRMD4A</i>	0.28/0.30	49/53	8.5e-2	2.4e-3	0.16	1.4e-4	5.9e-4	2.4e-4	0.18	1.9e-8	0.12	7.2e-2
BMI	12	26,179,017-26,189,017	<i>RASSF8</i>	0.38/0.40	37/37	7.5e-2	1.2e-4	0.12	4.4e-4	2.1e-4	3.5e-4/ <1e-4	0.12	7.5e-8	6.6e-2	2.1e-4
BMI	15	42,935,528-42,945,528	<i>STARD9</i>	0.28/0.23	43/56	<1e-8	2.7e-6	<1e-8	1.2e-5	2.1e-6	4.5e-5	1.2e-6	4.8e-7	8.4e-8	1.1e-7
BMI	16	60,159,304-60,169,304	Intergenic	0.27/0.12	29/28	4.3e-4	<1e-8	2.5e-2	2.1e-4	6.1e-4/ <1e-4	6.4e-6/ <1e-4	5.1e-3	1.4e-5	0.15	3.0e-4
BMI	21	40,851,354-40,861,354	<i>SH3BGR</i>	0.26/0.17	32/26	4.4e-4	<1e-8	2.5e-3	1.1e-4	1.3e-3	6.7e-5/ 2.6e-3	1.2e-3	3.9e-4	1.5e-2	3.4e-4
SBP	4	118,855,917-118,865,917	Intergenic	0.36/0.16	32/27	5.5e-2	<1e-8	0.68	8.1e-5	5.2e-2	4.2e-5/ <1e-4	0.75	2.5e-2	0.39	0.34
SBP	6	42,558,580-42,568,580	<i>UBR2</i>	0.96/0.095	28/22	<1e-8	1.3e-4	3.4e-4	5.6e-4	1.1e-6	1.6e-4	3.9e-4	1.7e-6	2.7e-3	7.8e-5
SBP	6	121,241,282-121,251,282	Intergenic	0.50/0.23	53/46	<1e-8	1.8e-4	1.2e-4	1.4e-3	1.6e-4	2.1e-3/ <1e-4	5.5e-4	3.0e-4	1.1e-4	7.0e-4
SBP	11	38,236,659-38,246,659	Intergenic	0.60/0.48	61/61	<1e-8	1.4e-4	8.6e-5	6.4e-6	2.1e-5	1.9e-4	3.8e-3	4.4e-6	8.5e-3	0.36

Significance threshold: $p < 1.56 \times 10^{-7}$ for TWINSUK and $p < 2 \times 10^{-3}$ for ALSPAC

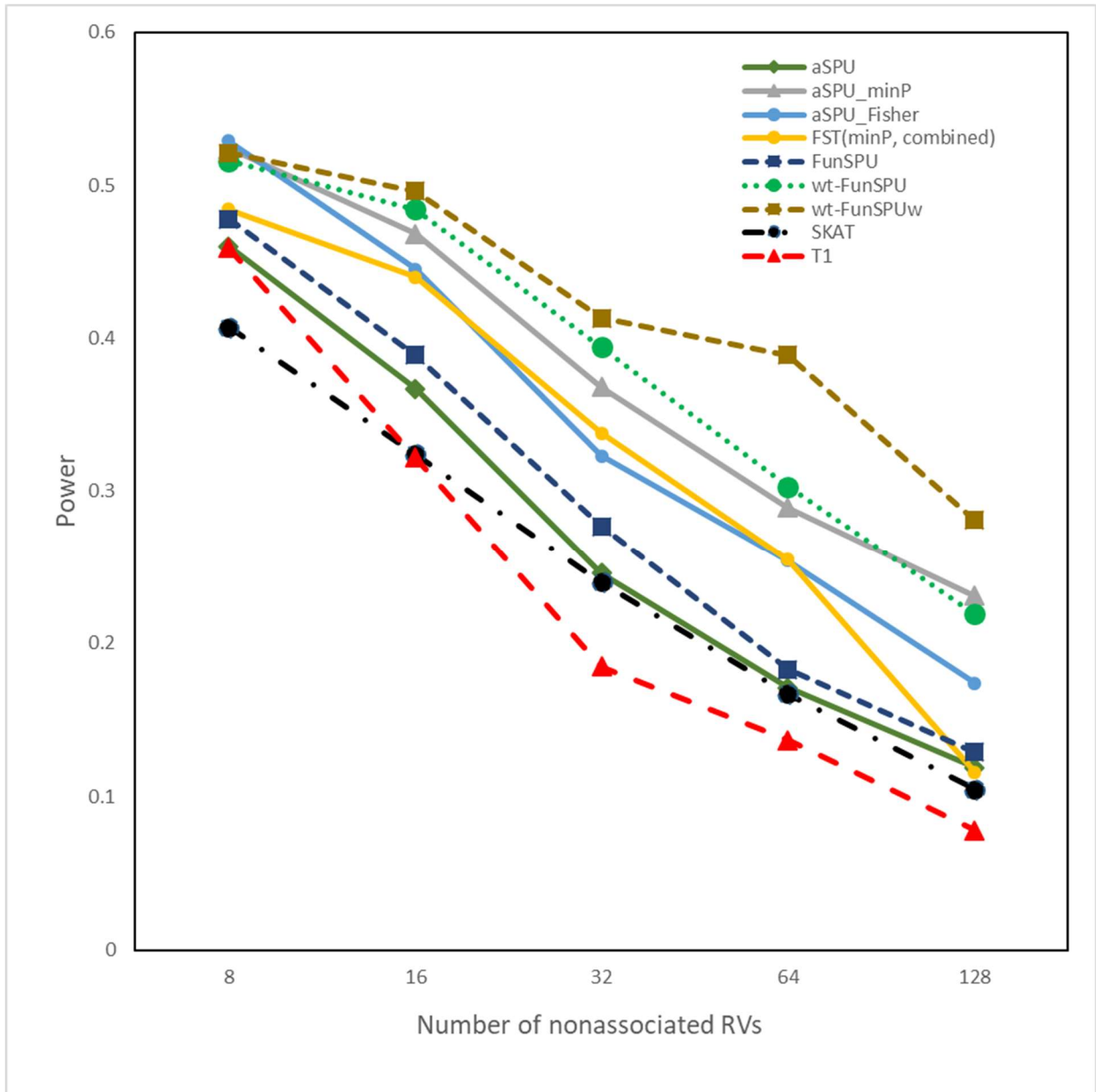
Figure 2.1: Empirical power of various tests for eight causal RVs and increasing number of nonassociated RVs (Scenario A).

Significance level $\alpha = 0.05$. The incorporated annotations for association tests include all three informative annotations and three noninformative annotations (Scenario A). All the results were based on 1000 simulation replications.



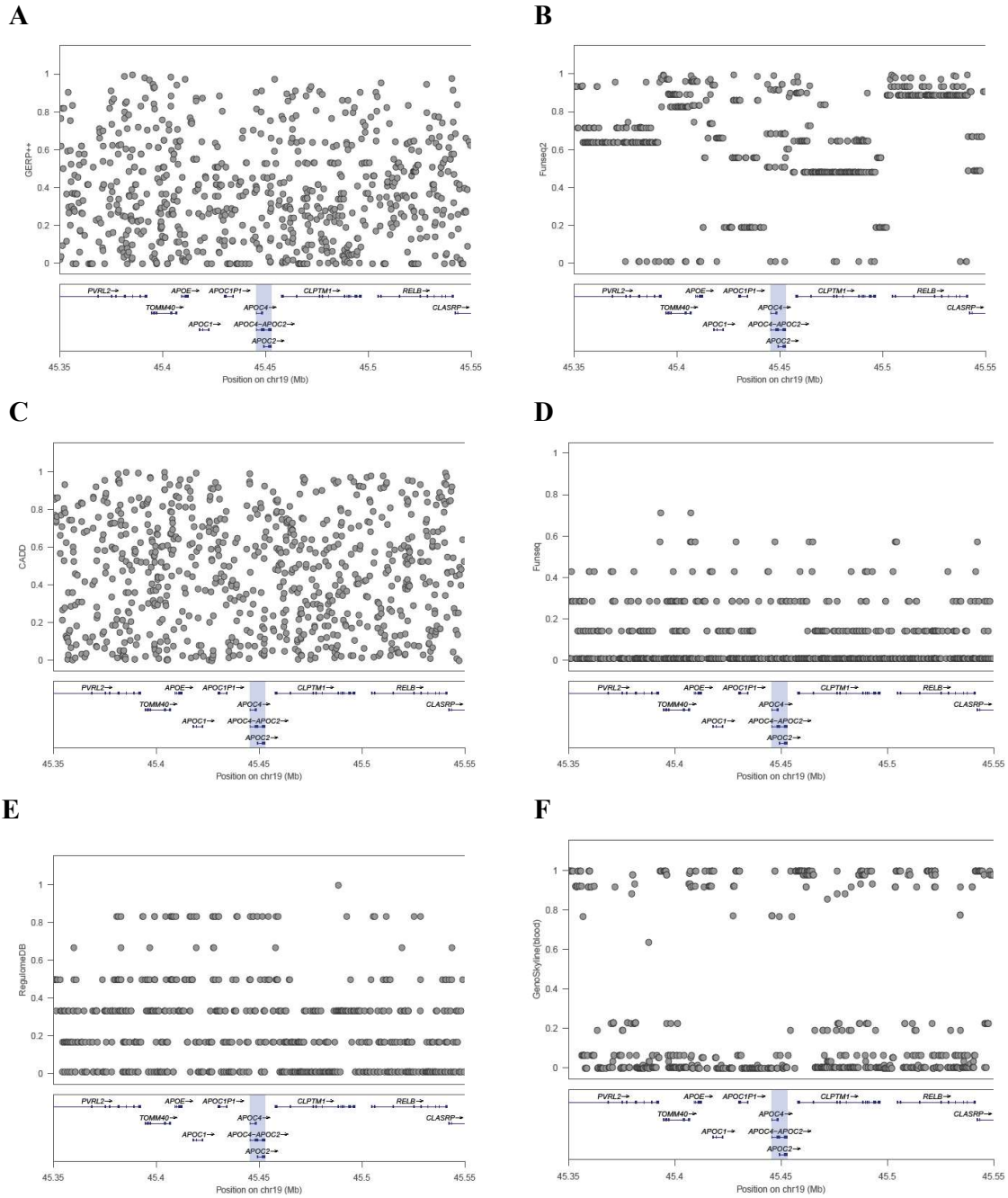
[Figure 2.2: Empirical power of various tests for eight causal RVs and increasing number of nonassociated RVs \(Scenario B\).](#)

Significance level $\alpha = 0.05$. The incorporated annotations for association tests include one out of three informative annotations and three noninformative annotations (Scenario B). All the results were based on 1000 simulation replications.



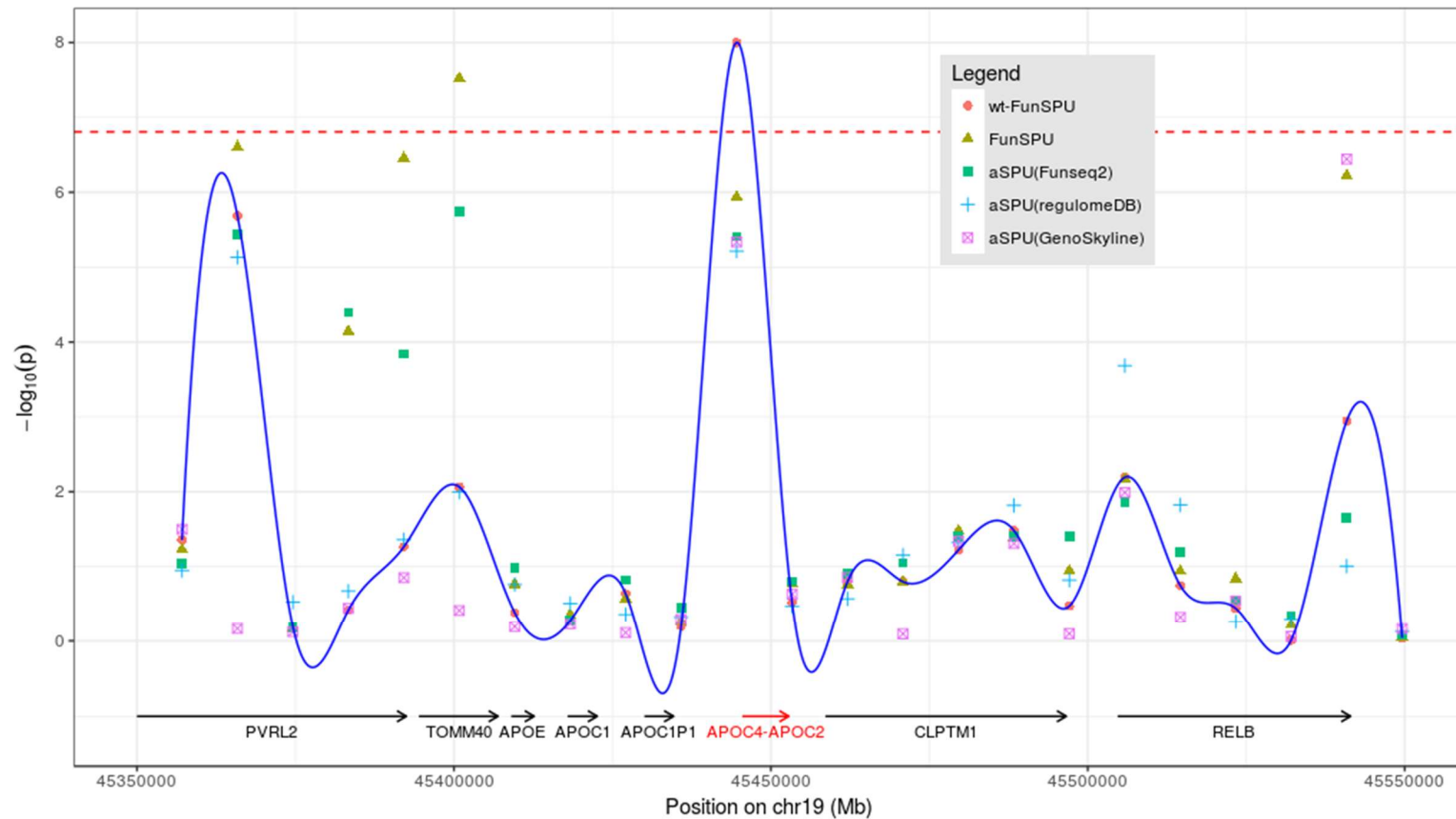
[Figure 2.3: Rescaled scores of functional annotations](#)

(A) GERP++, (B) Funseq2, (C) CADD, (D) Funseq, (E) RegulomeDB, and (F) GenoSkyline (blood) at the locus around gene *APOC4-APOC2*. The scores were rescaled to the interval [0, 1].



[Figure 2.4: Association test results for LDL at the locus around gene *APOC4-APOC2*.](#)

The round points and the trace show the results from the globally weighted wtFunSPU test. Other points correspond to the results of FunSPU and single annotation-based aSPU (Funseq2, RegulomeDB, GenoSkyline), respectively. Dashed line indicates the threshold of genome-wide significance level ($p < 1.56\text{e-}7$).



Supporting Information

Table 2.3: Computational time needed (mean and standard deviation (SD) with 32 rare variants (RVs) in an RV-set) for selected methods under comparison in the simulation study of power (Figure 1; Scenario A for power evaluation).

Computation time (in seconds)	Mean	SD
aSPU	0.268	0.054
aSPU minP	1.923	0.118
FunSPU	0.804	0.070
SKAT	0.033	0.022
Burden (T1)	0.024	0.015

Table 2.4: List of heritability of RVs by each category of functional annotation (TWINSUK cohort).

The *italicized* categories of each annotation were merged as one category. The corresponding phenotypes:

(A) HDL, (B) LDL, (C) BMI, and (D) SBP

(A) High-density lipoprotein (HDL)

Annotation	value	h^2 (RV)	#SNV	h^2 /#SNV	Normalized (by GERP++=1)	Regression Slopes
GERP++	1	0.391	2610225	1.497E-07	1.000	0.204
	2	0.454	2432456	1.868E-07	1.247	
	3	0.399	1878023	2.125E-07	1.419	
	4	0.401	1781008	2.253E-07	1.504	
	5	0.455	1603753	2.834E-07	1.893	
Funseq2	1	0.346	4208678	8.224E-08	0.549	0.177
	2	0.484	2483215	1.949E-07	1.302	
	3	0.396	2311991	1.714E-07	1.145	
	4	0.332	1861326	1.784E-07	1.191	
CADD	1	0.410	2563253	1.600E-07	1.068	0.238
	2	0.427	2446753	1.744E-07	1.165	
	3	0.416	2310764	1.801E-07	1.203	
	4	0.420	1905380	2.204E-07	1.472	
	5	0.552	1752877	3.149E-07	2.103	
RegulomeDB	1	0.438	3083174	1.420E-07	0.948	2.502
	2	0.495	2111757	2.345E-07	1.566	
	3	0.417	556547	7.497E-07	5.006	
	4	0.423	136034	3.110E-06	20.770	
	5	0.224	187380	1.196E-06	7.988	
Funseq	1	0.487	4990794	9.759E-08	0.652	15.170
	2	0.470	1080633	4.347E-07	2.903	
	3	0.412	228382	1.805E-06	12.054	
	4	0.313	29950	1.044E-05	69.685	
	5	0.000	5027	1.989E-10	0.001	
	6	0.000	280	3.571E-09	0.024	
GenoSkyline (blood)	1	0.459	9702179	4.726E-08	0.316	5.359
	2	0.212	245856	8.643E-07	5.772	
	3	0.147	120694	1.222E-06	8.159	
	4	0.348	910298	3.820E-07	2.551	

(B) Low-density lipoprotein (LDL)

Annotation	value	h^2 (RV)	#SNV	h^2 /#SNV	Normalized (by GERP++=1)	Regression Slopes
GERP++	1	0.349	2610225	1.339E-07	1.000	0.046
	2	0.406	2432456	1.669E-07	1.246	
	3	0.371	1878023	1.974E-07	1.474	
	4	0.275	1781008	1.545E-07	1.154	
	5	0.274	1603753	1.709E-07	1.276	
Funseq2	1	0.276	4208678	6.556E-08	0.490	0.154
	2	0.362	2483215	1.456E-07	1.087	
	3	0.299	2311991	1.292E-07	0.965	
	4	0.260	1861326	1.397E-07	1.043	
CADD	1	0.356	2563253	1.390E-07	1.038	0.066
	2	0.373	2446753	1.526E-07	1.139	
	3	0.284	2310764	1.229E-07	0.918	
	4	0.349	1905380	1.833E-07	1.369	
	5	0.294	1752877	1.677E-07	1.252	
RegulomeDB	1	0.281	3083174	9.120E-08	0.681	3.419
	2	0.368	2111757	1.744E-07	1.302	
	3	0.457	556547	8.208E-07	6.130	
	4	0.423	136034	3.106E-06	23.196	
	5	0.343	187380	1.829E-06	13.662	
Funseq	1	0.357	4990794	7.161E-08	0.535	3.162
	2	0.382	1080633	3.534E-07	2.639	
	3	0.301	228382	1.319E-06	9.854	
	4	0.051	29950	1.686E-06	12.595	
	5	0.000	5027	1.989E-10	0.001	
	6	0.008	280	2.986E-05	223.008	
GenoSkyline (blood)	1	0.354	9702179	3.649E-08	0.273	1.276
	2	0.173	245856	7.044E-07	5.261	
	3	0.089	120694	7.349E-07	5.488	
	4	0.221	910298	2.432E-07	1.817	

(C) Body mass index (BMI)

Annotation	value	h^2 (RV)	#SNV	h^2 /#SNV	Normalized (by GERP++=1)	Regression Slopes
GERP++	1	0.126	2610225	4.838E-08	1.000	0.231
	2	0.102	2432456	4.210E-08	0.870	
	3	0.070	1878023	3.745E-08	0.774	
	4	0.182	1781008	1.023E-07	2.115	
	5	0.119	1603753	7.418E-08	1.533	
Funseq2	1	0.163	4208678	3.875E-08	0.801	-0.033
	2	0.158	2483215	6.346E-08	1.312	
	3	0.000	2311991	4.325E-13	0.000	
	4	0.102	1861326	5.458E-08	1.128	
CADD	1	0.103	2563253	4.012E-08	0.829	0.154
	2	0.141	2446753	5.779E-08	1.194	
	3	0.095	2310764	4.110E-08	0.849	
	4	0.138	1905380	7.239E-08	1.496	
	5	0.123	1752877	7.004E-08	1.448	
RegulomeDB	1	0.102	3083174	3.310E-08	0.684	1.633
	2	0.156	2111757	7.404E-08	1.530	
	3	0.060	556547	1.075E-07	2.223	
	4	0.171	136034	1.255E-06	25.944	
	5	0.000	187380	5.337E-12	0.000	
Funseq	1	0.126	4990794	2.516E-08	0.520	5.423
	2	0.123	1080633	1.139E-07	2.354	
	3	0.000	228382	4.379E-12	0.000	
	4	0.000	29950	3.339E-11	0.001	
	5	0.000	5027	1.989E-10	0.004	
	6	0.038	280	1.363E-04	2816.763	
GenoSkyline (blood)	1	0.129	9702179	1.328E-08	0.274	0.940
	2	0.021	245856	8.593E-08	1.776	
	3	0.058	120694	4.840E-07	10.004	
	4	0.054	910298	5.914E-08	1.222	

(D) Systolic blood pressure (SBP)

Annotation	value	h^2 (RV)	#SNV	h^2 /#SNV	Normalized (by GERP++=1)	Regression Slopes
GERP++	1	0.256	2610225	9.804E-08	1.000	0.263
	2	0.327	2432456	1.343E-07	1.370	
	3	0.273	1878023	1.456E-07	1.485	
	4	0.377	1781008	2.117E-07	2.159	
	5	0.302	1603753	1.884E-07	1.922	
Funseq2	1	0.337	4208678	8.015E-08	0.818	0.198
	2	0.224	2483215	9.010E-08	0.919	
	3	0.252	2311991	1.089E-07	1.111	
	4	0.258	1861326	1.386E-07	1.413	
CADD	1	0.294	2563253	1.146E-07	1.169	0.165
	2	0.360	2446753	1.470E-07	1.499	
	3	0.256	2310764	1.108E-07	1.130	
	4	0.372	1905380	1.954E-07	1.993	
	5	0.300	1752877	1.713E-07	1.748	
RegulomeDB	1	0.358	3083174	1.161E-07	1.184	2.840
	2	0.279	2111757	1.320E-07	1.346	
	3	0.151	556547	2.716E-07	2.770	
	4	0.285	136034	2.098E-06	21.400	
	5	0.355	187380	1.895E-06	19.333	
Funseq	1	0.328	4990794	6.575E-08	0.671	14.050
	2	0.328	1080633	3.031E-07	3.092	
	3	0.420	228382	1.837E-06	18.736	
	4	0.159	29950	5.318E-06	54.243	
	5	0.028	5027	5.485E-06	55.948	
	6	0.000	280	3.571E-09	0.036	
GenoSkyline (blood)	1	0.329	9702179	3.387E-08	0.345	1.349
	2	0.066	245856	2.668E-07	2.721	
	3	0.055	120694	4.587E-07	4.679	
	4	0.260	910298	2.861E-07	2.918	

Table 2.5: Summary and comparison of the proposed tests.

Test	Allow multiple functional annotations as weights?	Way to combine multiple annotations/weights	Comment	recommended for genome-wide scan of WGS data
aSPU(w)*	No	N/A	an omnibus RV association test that covers the burden, SKAT, minP tests and beyond	Yes
aSPU(w)_minP	Yes	Minimum p-value of individual annotation-weighted aSPU	Closely related to the FST test (He et al, AJHG 2017)	Yes
aSPU(w)_Fisher	Yes	Fisher's meta-analysis approach to combine p-values of individual annotation-weighted aSPU	Assuming all annotations are roughly equally informative; often less powerful than aSPU_minP	No
FunSPU(w)	Yes	two-layers of gamma parameters to control each variant's and annotation's contribution to the overall test statistic	more general than aSPU_minP and aSPU_Fisher; the main proposal here	Yes
wtFunSPU(w)	Yes	FunSPU with a data-derived global weight for each annotation	When the global weighting is consistent with the relative informativeness of multiple annotations for a given locus, power gain is expected; otherwise, power loss can occur compared to FunSPU	Yes

*: (w) indicates inverse-variance weighted score function in the SPU framework; variants with lower MAF are up-weighted.

[Table 2.6: List of heritability of common variants \(CVs; MAF>5%\) by each category of functional annotation \(TWINSUK WGS cohort\) estimated by LD score regression.](#)

The *italicized* categories of each annotation were merged as one category. The corresponding phenotypes: (A) HDL, (B) LDL

(A) HDL

Annotation	value	h ² (CV)	Proportion of SNPs	h ² /proportion of SNPs	Normalized (by GERP++=1)	Regression Slopes (common variants by LD score regression)	Regression Slopes (TWINSUK, rare variants by GCTA)
GERP++	1	0.247	0.206	1.202	1.000	0.597	0.204
	2	-0.025	0.197	-0.128	-0.106		
	3	-0.055	0.153	-0.356	-0.296		
	4	0.442	0.138	3.194	2.657		
	5	0.355	0.113	3.131	2.605		
Funseq2	1	0.076	0.352	0.217	0.181	0.890	0.177
	2	0.139	0.200	0.695	0.578		
	3	0.210	0.177	1.186	0.987		
	4	0.478	0.132	3.618	3.010		
CADD	1	0.249	0.221	1.126	0.937	0.506	0.238
	2	0.034	0.213	0.159	0.132		
	3	0.116	0.187	0.622	0.518		
	4	0.151	0.138	1.096	0.912		
	5	0.392	0.106	3.698	3.076		
RegulomeDB	1	-0.166	0.254	-0.655	-0.545	2.833	2.502
	2	-0.017	0.158	-0.111	-0.092		
	3,4,5	0.617	0.069	8.987	7.477		
Funseq	1	0.093	0.170	0.547	0.455	5.329	15.170
	2	0.302	0.073	4.155	3.457		
	3	0.041	0.013	3.255	2.708		
	4,5,6	0.054	0.002	26.359	21.929		
Geno-Skyline (blood)	1	0.564	0.775	0.728	0.606	1.422	5.359
	2,3,4	0.377	0.091	4.146	3.449		

(B) LDL

Annotation	value	h^2 (CV)	Proportion of SNPs	h^2 /proportion of SNPs	Normalized (by GERP++=1)	Regression Slopes (common variants by LD score regression)	Regression Slopes (TWINSUK, rare variants by GCTA)
GERP++	1	0.239	0.206	1.161	1	0.0589	0.046
	2	0.344	0.197	1.745	1.503		
	3	0.045	0.153	0.293	0.253		
	4	0.232	0.138	1.679	1.446		
	5	0.174	0.113	1.537	1.323		
Funseq2	1	0.041	0.352	0.118	0.101	0.9571	0.154
	2	0.159	0.2	0.796	0.686		
	3	0.249	0.177	1.411	1.215		
	4	0.478	0.132	3.616	3.115		
CADD	1	0.195	0.221	0.882	0.76	0.5924	0.066
	2	0.104	0.213	0.487	0.42		
	3	0.127	0.187	0.677	0.583		
	4	0.126	0.138	0.915	0.788		
	5	0.435	0.106	4.108	3.538		
Regulome DB	1	-0.015	0.254	-0.06	-0.052	2.7101	3.419
	2	0.181	0.158	1.146	0.987		
	3,4,5	0.617	0.069	8.99	7.744		
Funseq	1	-0.176	0.17	-1.037	-0.893	6.516	3.162
	2	0.562	0.073	7.743	6.67		
	3	0.008	0.013	0.643	0.554		
	4,5,6	0.063	0.002	31.121	26.806		
Geno-Skyline (blood)	1	0.615	0.775	0.086	0.684	1.416	1.276
	2,3,4	0.371	0.091	0.773	3.516		

Figure 2.5: Heritability per SNV ($h^2/\text{\#SNV}$) of HDL sorted by category of functional annotation score (TWINSUK cohort).

Dashed lines represent linear regression results (categories merged).

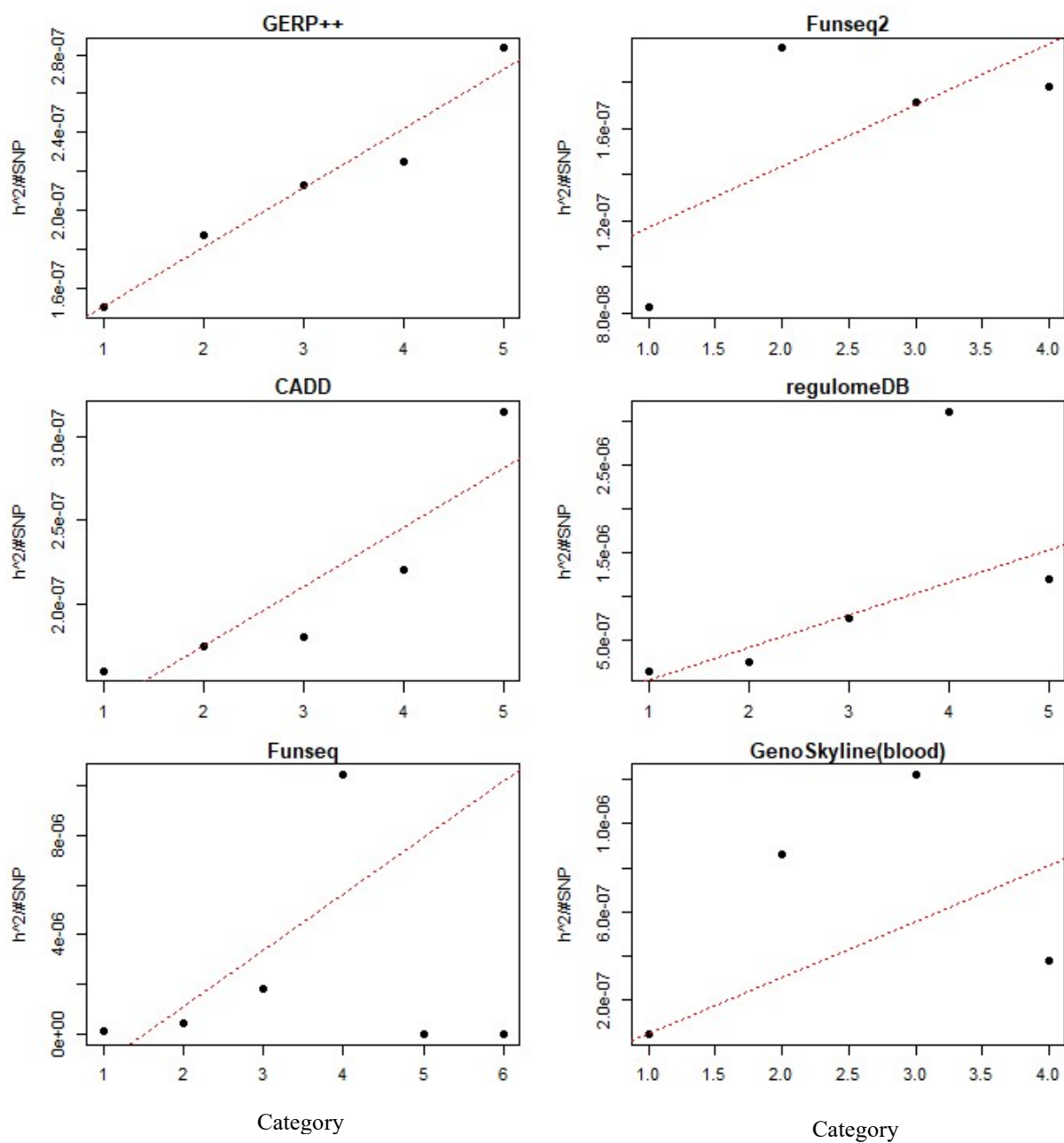


Figure 2.6: Heritability per SNV ($h^2/\text{\#SNV}$) of LDL sorted by category of functional annotation score (TWINSUK cohort).

Dashed lines represent linear regression results (categories merged).

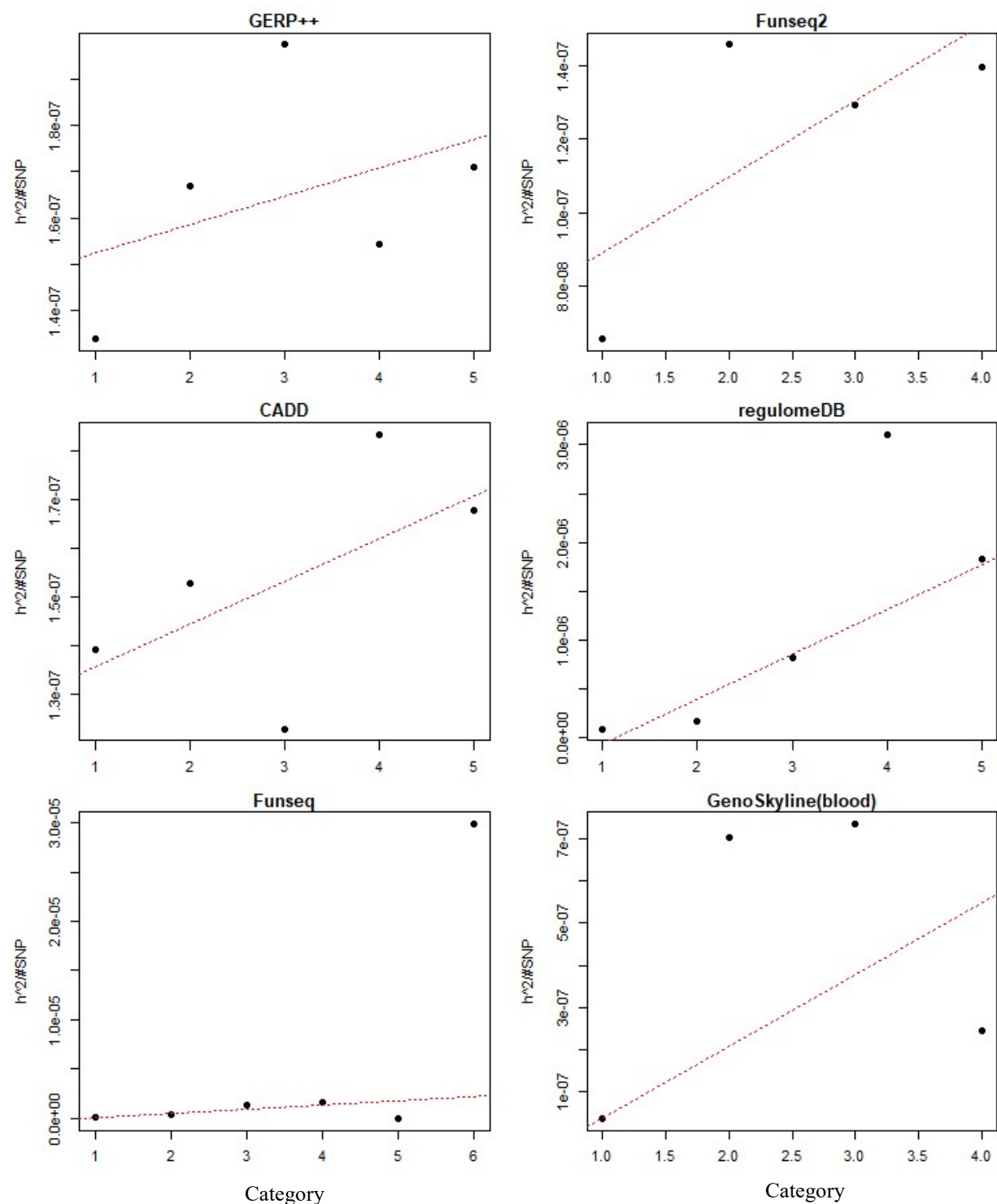


Figure 2.7: Heritability per SNV ($h^2/\text{\#SNV}$) of BMI sorted by category of functional annotation score (TWINSUK cohort).

Dashed lines represent linear regression results (categories merged).

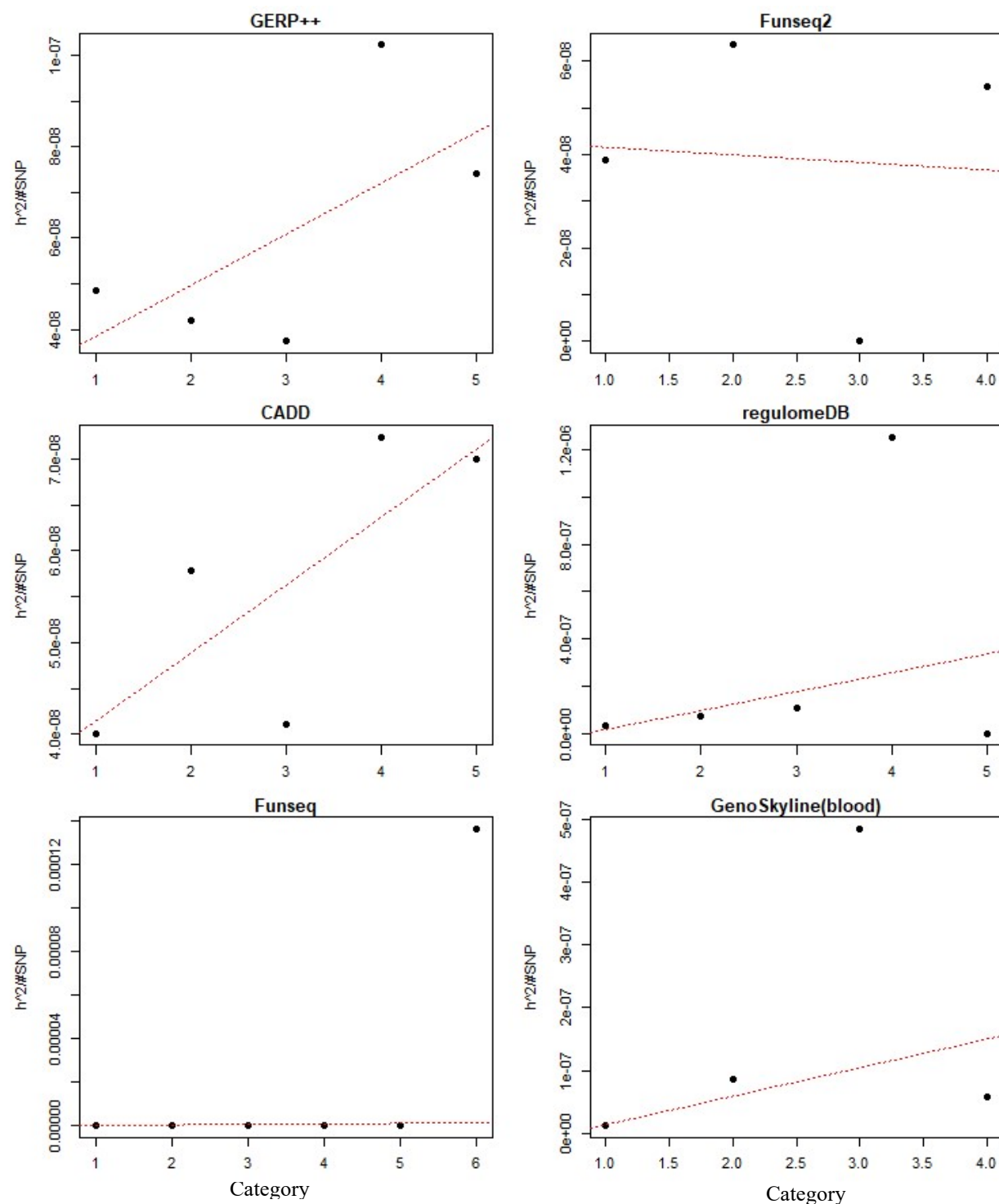


Figure 2.8: Heritability per SNV ($h^2/\text{\#SNV}$) of SBP sorted by category of functional annotation score (TWINSUK cohort).

Dashed lines represent linear regression results (categories merged).

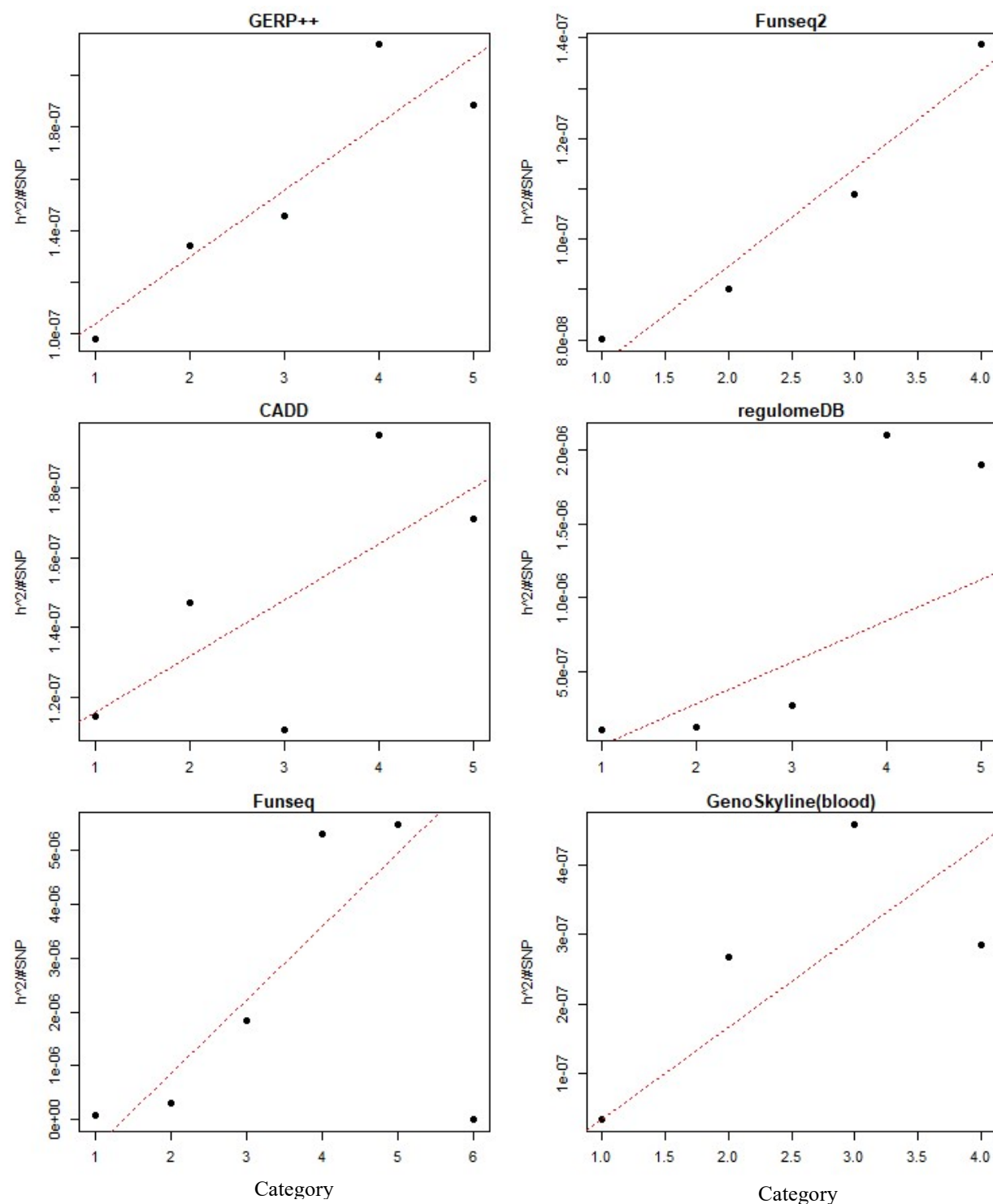


Figure 2.9: Distributions of genome-wide functional scores for rare variants (MAF < 5%) in the UK10K TWINSUK cohort.

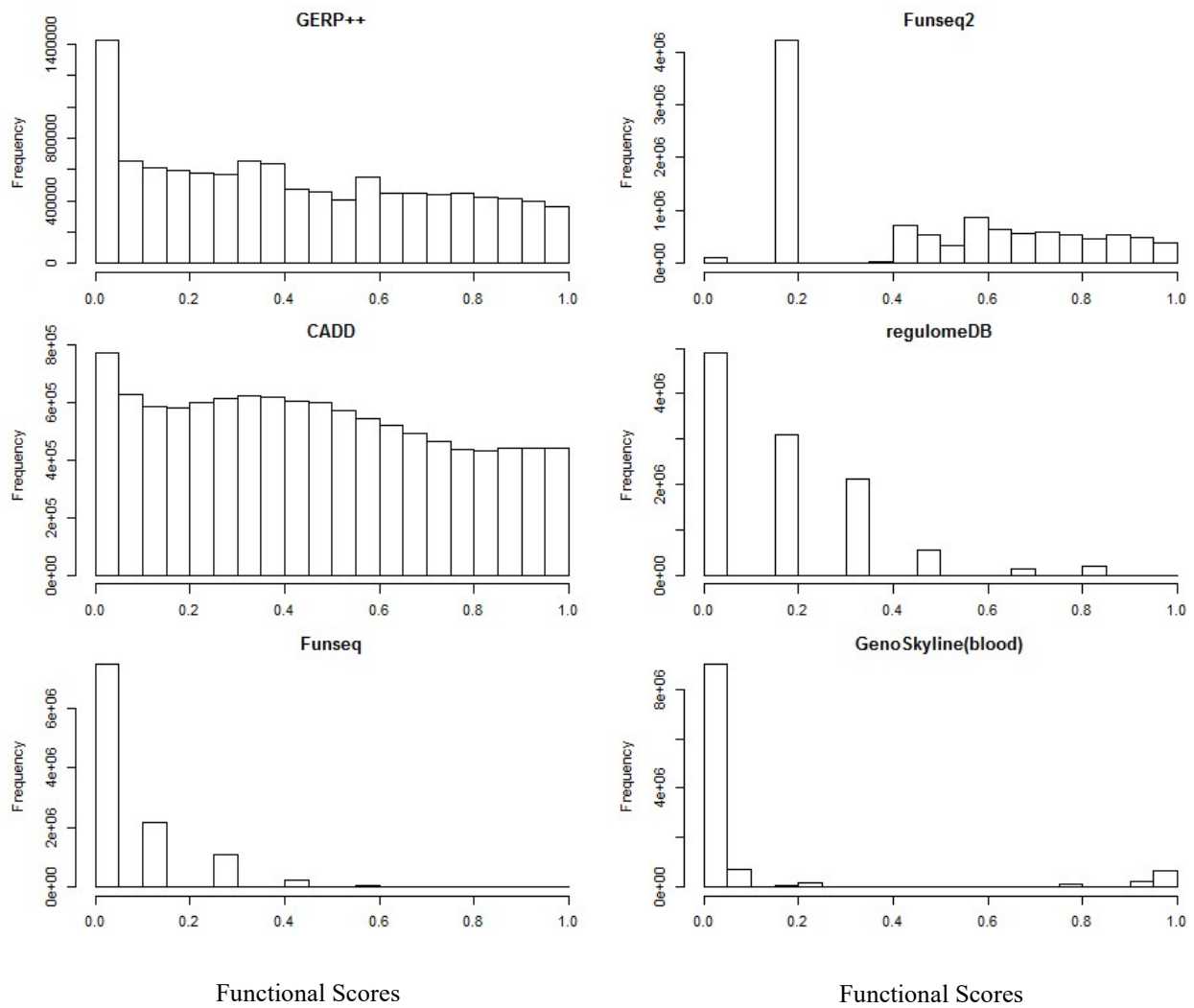


Figure 2.10. Pairwise correlation of rescaled scores of functional annotations across the whole genome.

The correlation plot was generated by R package corrplot.

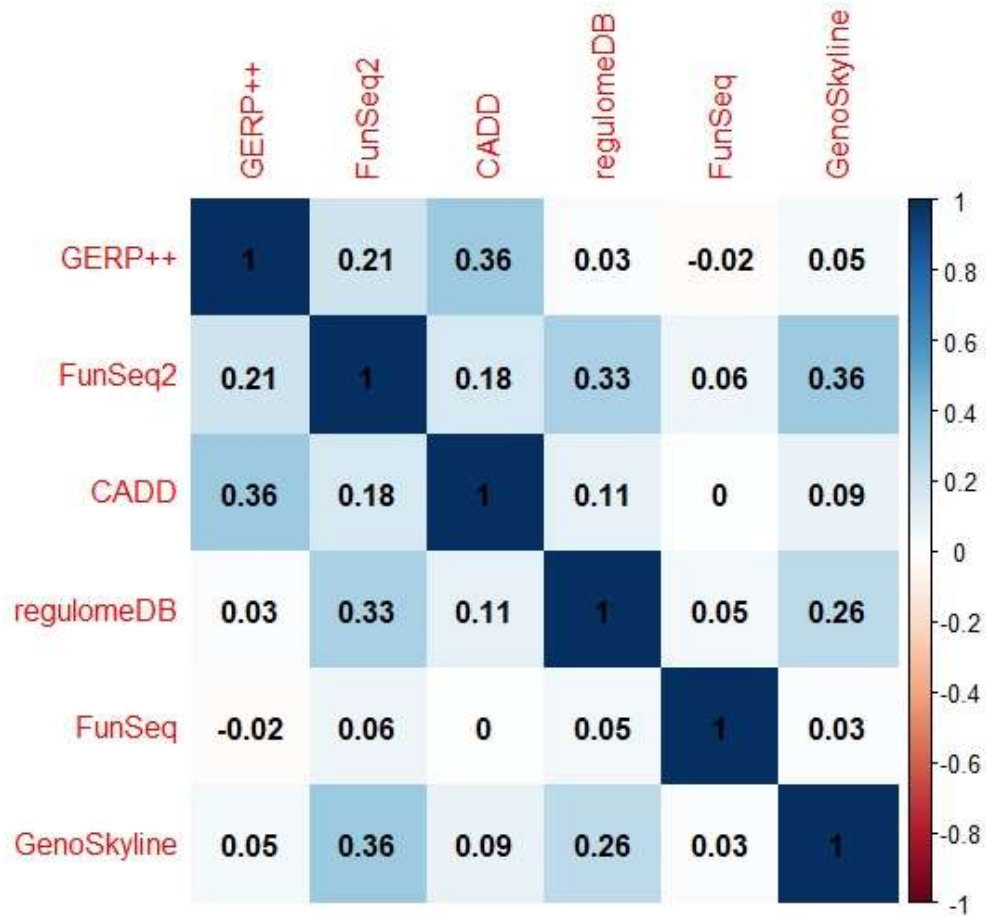
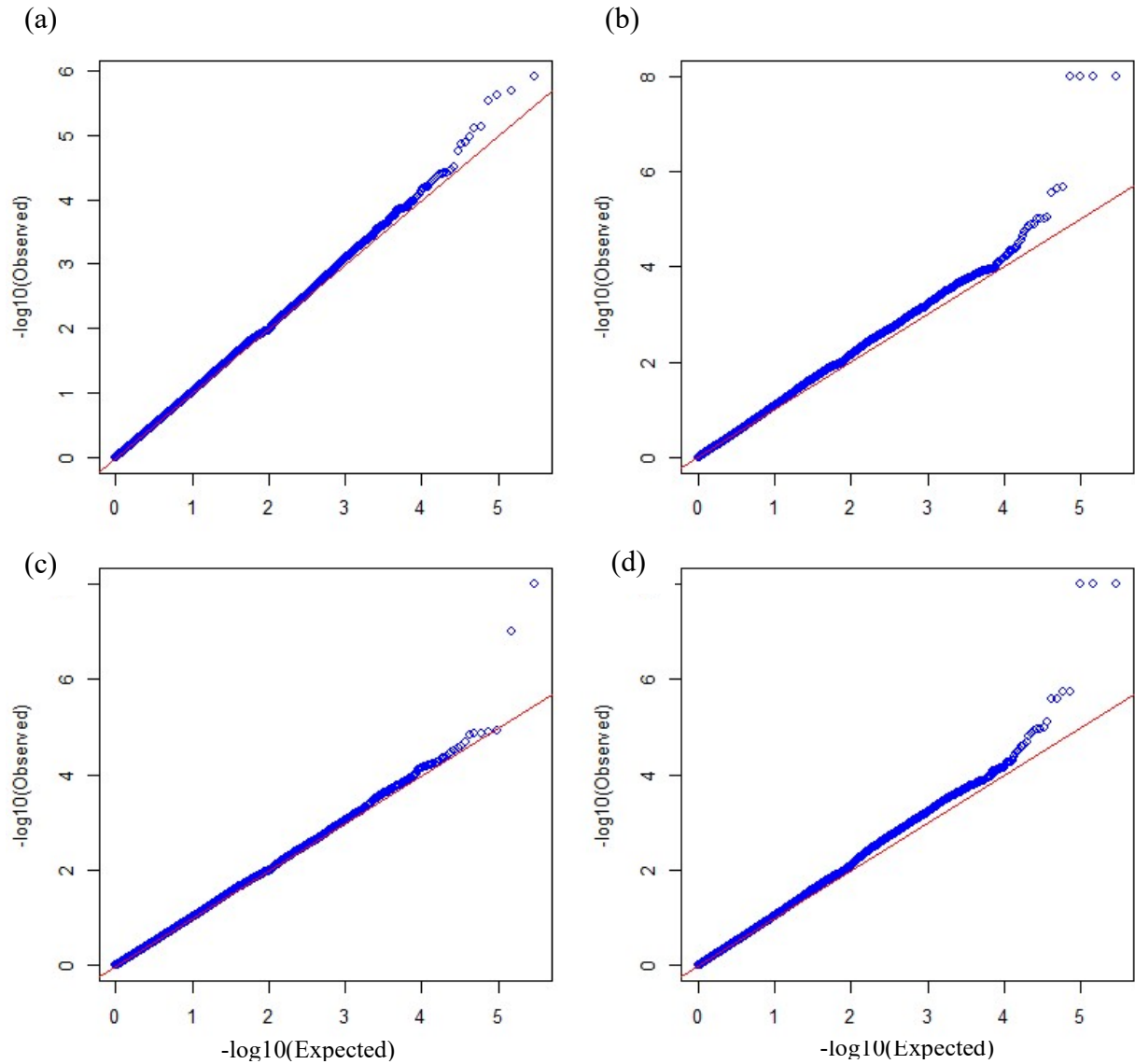


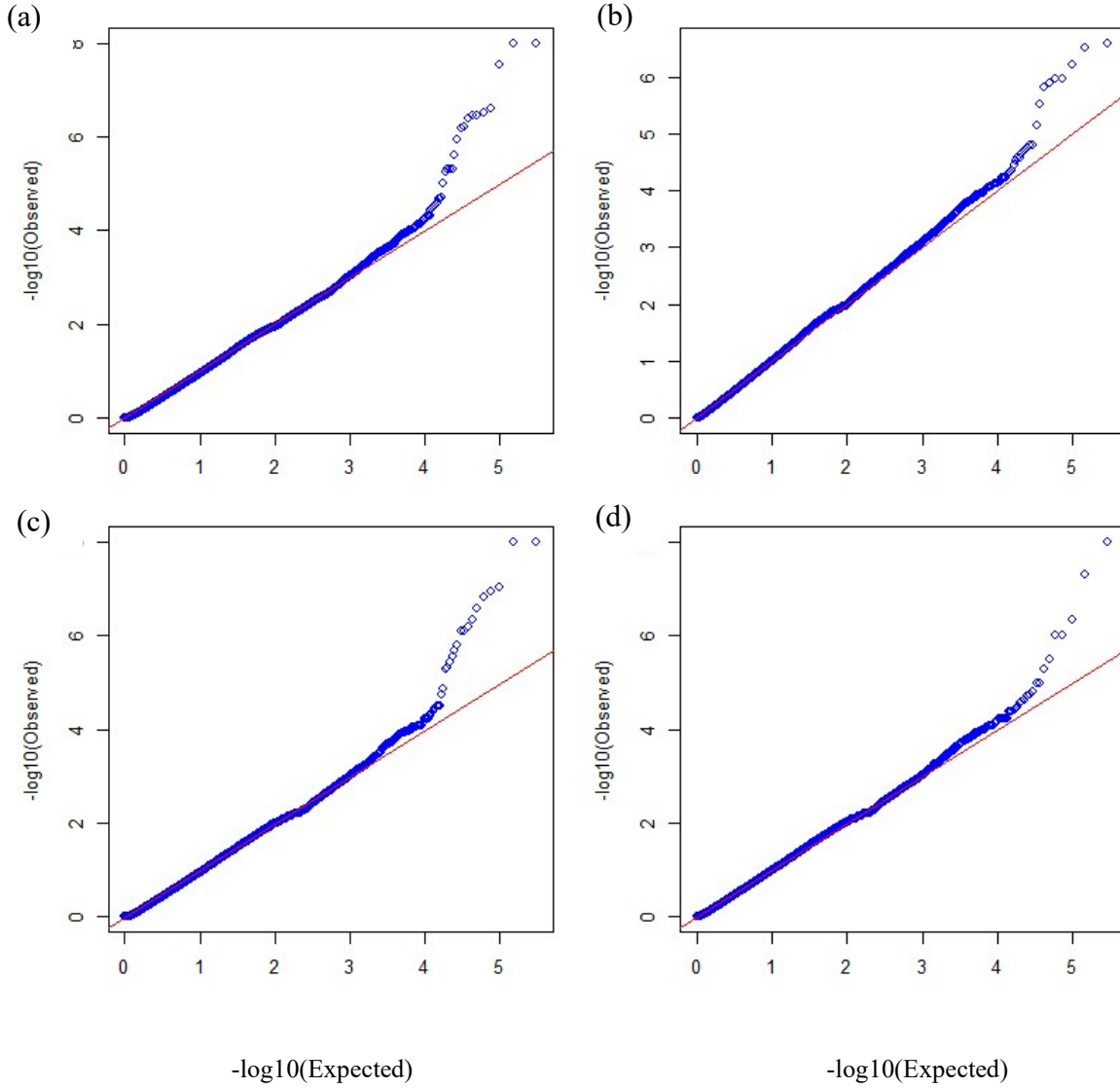
Figure 2.11: Global quantile-quantile (QQ) plots for association analysis of rare variants with HDL in the UK10K TWINSUK cohort

(a) FunSPU (genomic control $\lambda = 1.004$), (b) FunSPUw ($\lambda = 1.076$), (c) wtFunSPU with global weights ($\lambda = 1.004$), and (d) wtFunSPUw with global weights ($\lambda = 1.047$).



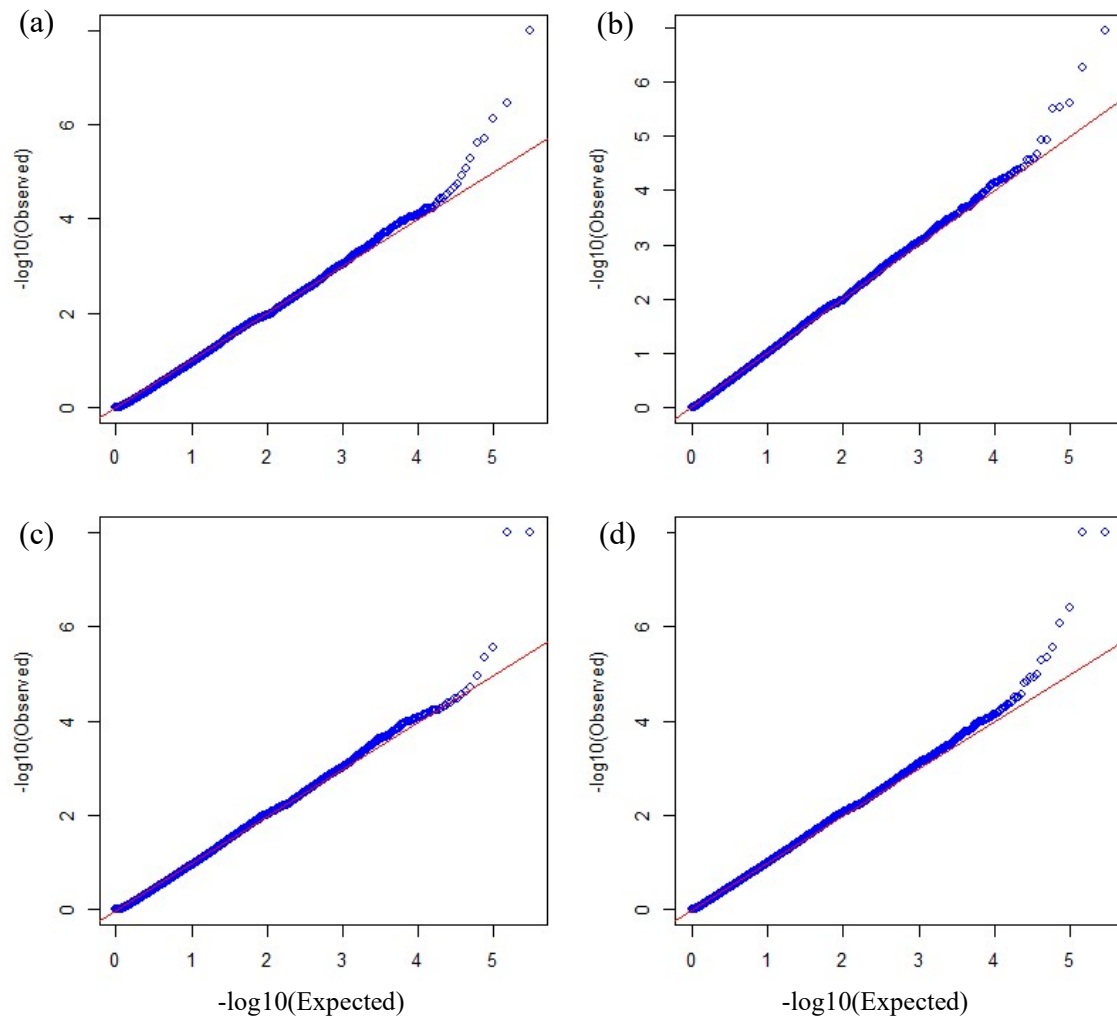
[Figure 2.12: Global QQ plots for association analysis of rare variants with LDL in the UK10K TWINSUK cohort](#)

(a) FunSPU (genomic control $\lambda = 0.611$), (b) FunSPUw ($\lambda = 0.805$),
(c) wtFunSPU with global weights ($\lambda = 0.642$), and (d) wtFunSPUw with global weights ($\lambda = 0.789$).



[Figure 2.13: Global QQ plots for association analysis of rare variants with BMI in the UK10K TWINSUK cohort](#)

(a) FunSPU (genomic control $\lambda = 0.604$), (b) FunSPUw ($\lambda = 0.825$),
(c) wtFunSPU with global weights ($\lambda = 0.635$), and (d) wtFunSPUw with global weights ($\lambda = 0.825$).



[Figure 2.14: Global QQ plots for association analysis of rare variants with SBP in the UK10K TWINSUK cohort](#)

(a) FunSPU (genomic control $\lambda = 0.601$), (b) FunSPUw ($\lambda = 0.821$),
(c) wtFunSPU with global weights ($\lambda = 0.631$), and (d) wt-FunSPUw with global weights ($\lambda = 0.793$).

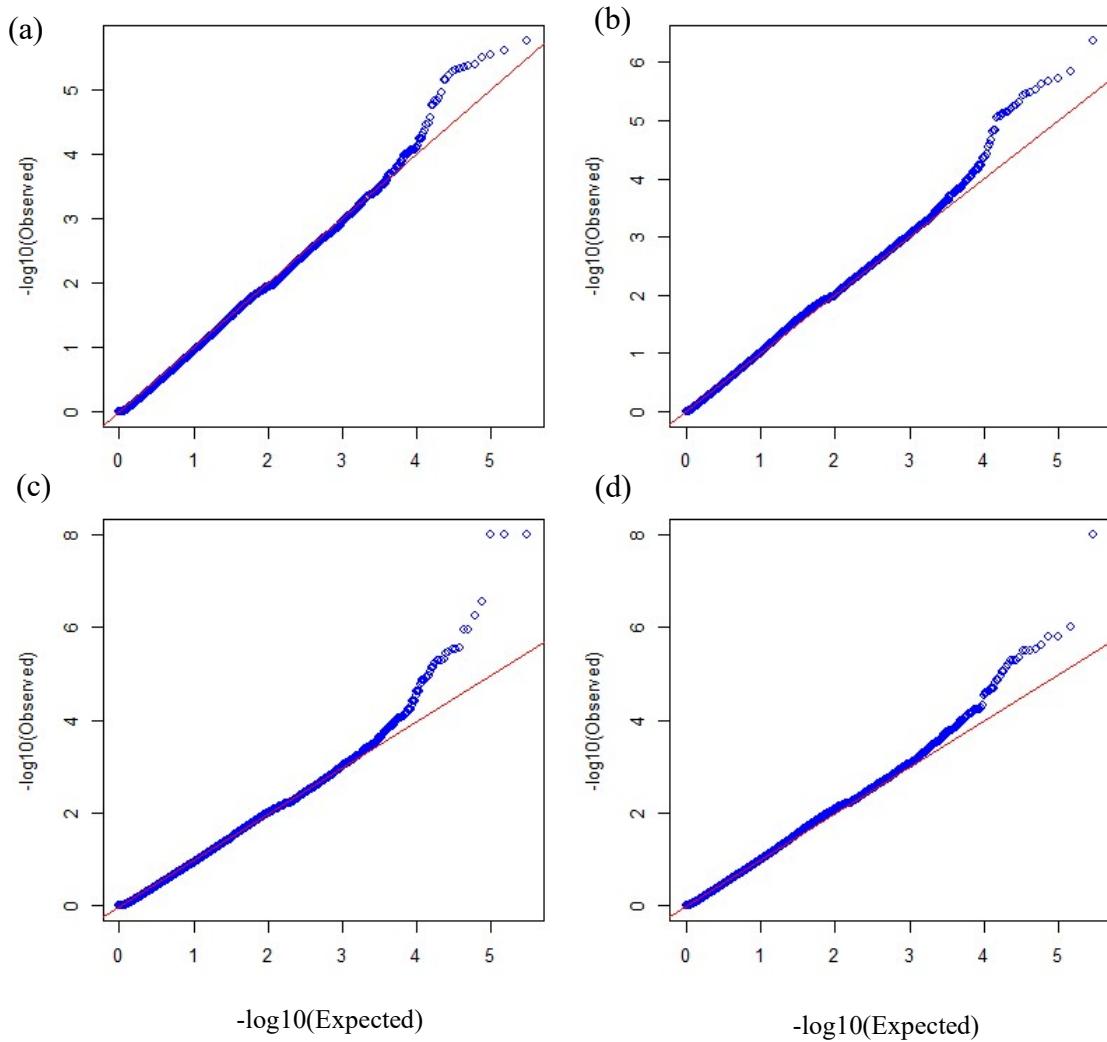
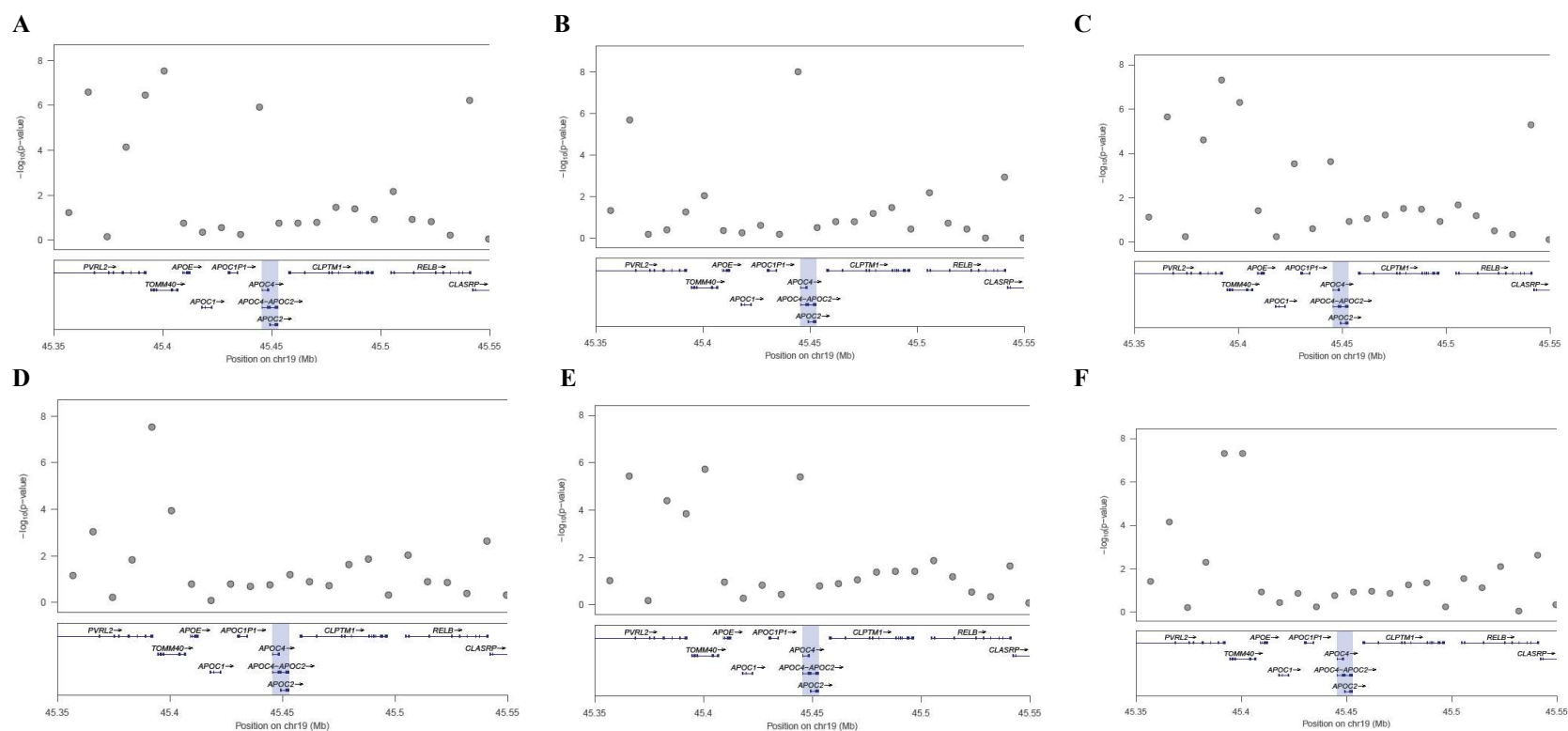


Figure 2.15: LocusZoom plots of association test results for LDL at the locus around *TOMM40* and *APOC4-APOC2* in the UK10K TWINSUK cohort

(A) FunSPU, (B) wtFunSPU incorporating global weights, (C) aSPU, (D)-(I) aSPU incorporating a single functional annotation: (D) GERP++, (E) Funseq2, (F) CADD, (G) Funseq, (H) RegulomeDB, and (I) GenoSkyline (blood).



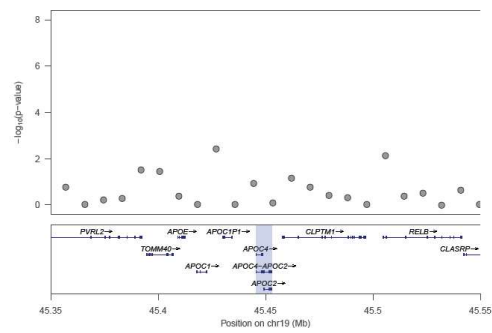
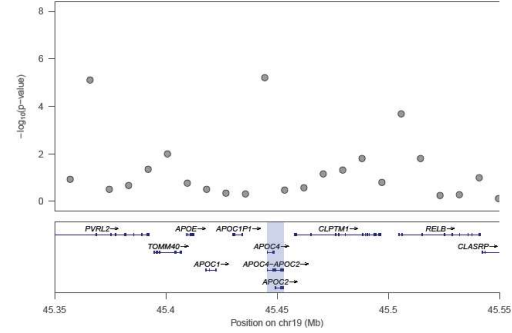
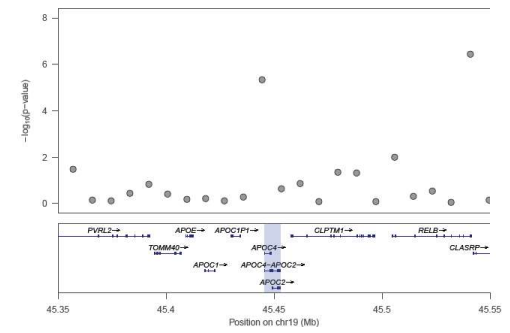
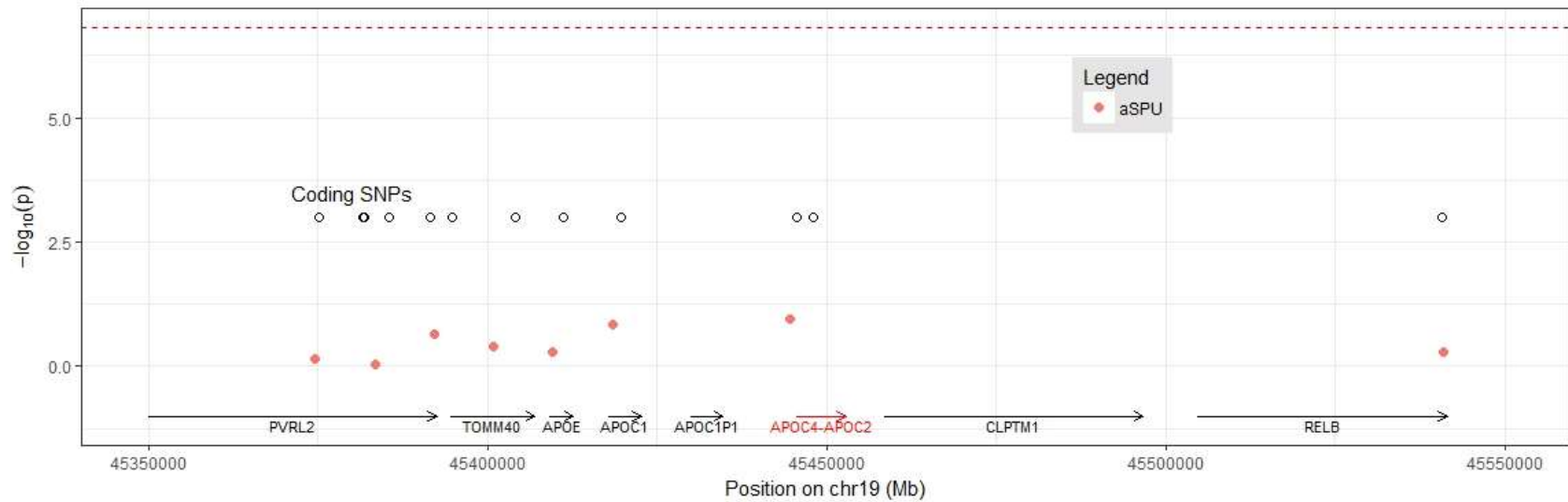
G**H****I**

Figure 2.16: Association test results for coding rare variants (MAF < 5%) in association with LDL around genes *TOMM40*, *APOE*, and *APOC4-APOC2*

Red dots correspond to $-\log_{10}(\text{aSPU p-value})$, and blank dots mark the locations of nonsynonymous rare variants. Dashed line indicates the threshold of genome-wide significance level ($p < 1.56\text{e-}7$).



Reference

1. Crosby J, Peloso GM, Auer PL, Crosslin DR, Stitzel NO, et al. (2014) Loss-of-function mutations in APOC3, triglycerides, and coronary disease. *N Engl J Med* 371: 22-31.
2. The UK10K Project Consortium, Walter K, Min JL, Huang J, Crooks L, et al. (2015) The UK10K project identifies rare variants in health and disease. *Nature* 526: 82-90.
3. Brody JA, Morrison AC, Bis JC, O'Connell JR, Brown MR, et al. (2017) Analysis commons, a team approach to discovery in a big-data environment for genetic epidemiology. *Nat Genet* 49: 1560-1563.
4. Consortium EP (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 489: 57-74.
5. Roadmap Epigenomics C, Kundaje A, Meuleman W, Ernst J, Bilenky M, et al. (2015) Integrative analysis of 111 reference human epigenomes. *Nature* 518: 317-330.
6. GTEx Consortium, Laboratory DA, Coordinating Center -Analysis Working G, Statistical Methods groups-Analysis Working G, Enhancing Gg, et al. (2017) Genetic effects on gene expression across human tissues. *Nature* 550: 204-213.
7. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, et al. (2014) A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* 46: 310-315.
8. Lu Q, Powles RL, Wang Q, He BJ, Zhao H (2016) Integrative Tissue-Specific Functional Annotations in the Human Genome Provide Novel Insights on Many Complex Traits and Improve Signal Prioritization in Genome Wide Association Studies. *PLoS Genet* 12: e1005947.
9. Ionita-Laza I, McCallum K, Xu B, Buxbaum JD (2016) A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nat Genet* 48: 214-220.
10. Liu X, Li C, Boerwinkle E (2017) The performance of deleteriousness prediction scores for rare non-protein-changing single nucleotide variants in human genes. *J Med Genet* 54: 134-144.
11. Werling DM, Brand H, An JY, Stone MR, Zhu L, et al. (2018) An analytical framework for whole-genome sequence association studies and its implications for autism spectrum disorder. *Nat Genet* 50: 727-736.
12. Kim T, Wei P (2016) Incorporating ENCODE information into association analysis of whole genome sequencing data. *BMC Proc* 10: 257-261.
13. Morrison AC, Huang Z, Yu B, Metcalf G, Liu X, et al. (2017) Practical Approaches for Whole-Genome Sequence Analysis of Heart- and Blood-Related Traits. *Am J Hum Genet* 100: 205-215.
14. Su YR, Di CZ, Hsu L, C GEC (2017) A unified powerful set-based test for sequencing data analysis of GxE interactions. *Biostatistics* 18: 119-131.
15. Pan W, Kim J, Zhang Y, Shen X, Wei P (2014) A powerful and adaptive association test for rare variants. *Genetics* 197: 1081-1095.

16. Wu MC, Lee S, Cai T, Li Y, Boehnke M, et al. (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet* 89: 82-93.
17. He Z, Xu B, Lee S, Ionita-Laza I (2017) Unified Sequence-Based Association Tests Allowing for Multiple Functional Annotations and Meta-analysis of Noncoding Variation in MetaboChip Data. *Am J Hum Genet* 101: 340-352.
18. Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, et al. (2012) Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res* 22: 1790-1797.
19. Khurana E, Fu Y, Colonna V, Mu XJ, Kang HM, et al. (2013) Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science* 342: 1235587.
20. Fu Y, Liu Z, Lou S, Bedford J, Mu XJ, et al. (2014) FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. *Genome Biol* 15: 480.
21. Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, et al. (2010) Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput Biol* 6: e1001025.
22. Pan W, Kwak IY, Wei P (2015) A Powerful Pathway-Based Adaptive Test for Genetic Association with Common or Rare Variants. *Am J Hum Genet* 97: 86-98.
23. Yang T, Chen H, Tang H, Li D, Wei P (2018) A powerful and data-adaptive test for rare-variant-based gene-environment interaction analysis. *Stat Med*.
24. Pan W (2009) Asymptotic tests of association with multiple SNPs in linkage disequilibrium. *Genet Epidemiol* 33: 497-507.
25. Gusev A, Lee SH, Trynka G, Finucane H, Vilhjalmsdottir BJ, et al. (2014) Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *Am J Hum Genet* 95: 535-552.
26. Yang J, Lee SH, Goddard ME, Visscher PM (2011) GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* 88: 76-82.
27. Liu X, White S, Peng B, Johnson AD, Brody JA, et al. (2016) WGS: an annotation pipeline for human genome sequencing studies. *J Med Genet* 53: 111-112.
28. Wei P, Cao Y, Zhang Y, Xu Z, Kwak IY, et al. (2016) On Robust Association Testing for Quantitative Traits and Rare Variants. *G3 (Bethesda)* 6: 3941-3950.
29. Kathiresan S, Willer CJ, Peloso GM, Demissie S, Musunuru K, et al. (2009) Common variants at 30 loci contribute to polygenic dyslipidemia. *Nat Genet* 41: 56-65.
30. Waterworth DM, Ricketts SL, Song K, Chen L, Zhao JH, et al. (2010) Genetic variants influencing circulating lipid levels and risk of coronary artery disease. *Arterioscler Thromb Vasc Biol* 30: 2264-2276.
31. Willer CJ, Sanna S, Jackson AU, Scuteri A, Bonnycastle LL, et al. (2008) Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nat Genet* 40: 161-169.
32. Lin YT, Seo J, Gao F, Feldman HM, Wen HL, et al. (2018) APOE4 Causes Widespread Molecular and Cellular Alterations Associated with Alzheimer's Disease Phenotypes in Human iPSC-Derived Brain Cell Types (vol 98, pg 1141, 2018). *Neuron* 98: 1294-1294.

33. Kawashiri MA, Tsukamoto K, Secreto A, Usher DC, Pure E, et al. (2000) Apoe2 and apoE4 are less effective than ApoE3 in inhibiting atherosclerosis in LDL receptor deficient mice. *Circulation* 102: 145-145.
34. Huang YWA, Zhou B, Wernig M, Sudhof TC (2017) ApoE2, ApoE3, and ApoE4 Differentially Stimulate APP Transcription and A beta Secretion. *Cell* 168: 427-+.
35. Liu DJ, Peloso GM, Yu H, Butterworth AS, Wang X, et al. (2017) Exome-wide association study of plasma lipids in > 300,000 individuals. *Nature Genetics* 49: 1758-+.
36. Lange LA, Hu YN, Zhang H, Xue CY, Schmidt EM, et al. (2014) Whole-Exome Sequencing Identifies Rare and Low-Frequency Coding Variants Associated with LDL Cholesterol. *American Journal of Human Genetics* 94: 233-245.
37. Natarajan P, Peloso GM, Zekavat SM, Montasser M, Ganna A, et al. (2018) Deep-coverage whole genome sequences and blood lipids among 16,324 individuals. *Nature Communications* 9.
38. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, et al. (2015) UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med* 12: e1001779.
39. Shihab HA, Rogers MF, Gough J, Mort M, Cooper DN, et al. (2015) An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics* 31: 1536-1543.
40. Zhou J, Troyanskaya OG (2015) Predicting effects of noncoding variants with deep learning-based sequence model. *Nat Methods* 12: 931-934.
41. Roeder K, Devlin B, Wasserman L (2007) Improving power in genome-wide association studies: Weights tip the scale. *Genetic Epidemiology* 31: 741-747.
42. Roeder K, Wasserman L (2009) Genome-Wide Significance Levels and Weighted Hypothesis Testing. *Statistical Science* 24: 398-413.
43. Bickel PJ (1993) Efficient and adaptive estimation for semiparametric models. Baltimore: Johns Hopkins University Press. xix, 560 p. p.
44. Finucane HK, Bulik-Sullivan B, Gusev A, Trynka G, Reshef Y, et al. (2015) Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature Genetics* 47: 1228-+.

CHAPTER 3: SUMMARY STATISTICS-BASED ASSOCIATION ANALYSIS

INCORPORATING FUNCTIONAL ANNOTATIONS

Introduction

Genome-wide association studies (GWAS) have been used to successfully identify tens of thousands of risk loci associated with complex traits and diseases in the past decade. However, due to privacy concerns and other logistical considerations, it is often difficult for the researchers to obtain access to individual-level data. At the same time, many summary association statistics from large GWAS are publicly available [1]. The sample sizes of these studies are in the range of 20,000 to >300, 000, which will compensate the loss of accuracy due to the use of summary statistics when the sample of individual-level data is smaller. In addition, analysis of summary statistics also has potential to reduce computational burden, which does not scale with the sample size as individual-level data. Typically, public summary association statistics consist of per-allele single nucleotide polymorphism (SNP) effect size (or log odds ratios for case-control traits) and standard errors, which can be used to compute z-score. Some repositories of summary statistics also include allele frequencies and linkage disequilibrium (LD) information. SNP effect sizes $\hat{\beta}_i$ and their standard error $se(\hat{\beta}_i)$ are typically estimated by regressing the phenotype on the genotype values at the SNP of interest [2]. At large sample sizes, the vector of z-scores $Z = \frac{\hat{\beta}_i}{se(\hat{\beta}_i)}$ at a locus is approximated by a multivariate normal distribution with mean 0 and variance equal to the linkage disequilibrium (LD) matrix V , e.g.

$$\frac{\hat{\beta}_i}{\text{se}(\hat{\beta}_i)} \sim \text{MVN}(0, V)$$

Many approaches have been developed for summary statistic-based analysis of complex traits. Single-variant association tests, including conditional association and imputation using summary statistics, is applicable to identify new secondary associations at a previously identified GWAS locus, thus complementing the missing heritability for these traits. Researchers reported that conditional and joint association analyses of multiple SNPs can be directly approximated by summary association statistics together with LD information estimated from a population reference panel such as 1000 Genomes [3] and preclude the individual-level data. Another approach to boost association power in GWAS is to leverage LD information from a population reference panel to impute genotypes at untyped variants [4]. The rationale of this method is that correlations between z -scores can be modelled using a multivariate normal (MVN) distribution with the variance equal to the LD correlation matrix [5]. However, conditional association and imputation using summary statistics crucially rely on accurate LD information. Low variant frequency and mismatch between the LD reference panel and original GWAS samples can lead to reducing accuracy of these two approaches.

Fine-mapping is the process to identify the particular genetic variants that are likely to causally influence the examined trait in a trait-associated region determined by a genome-wide association study (GWAS) [6]. A straightforward approach to fine-mapping is to prioritize variants based on the marginal P values [7]. Furthermore, recent study has shown

that integrating association strength (e.g. ranking P values) with functional genomic annotation data can improve fine-mapping accuracy using a multiple causal variant model [8]. In addition, polygenic architectures can also be inferred by identifying tissue-specific functional annotations that are enriched for causal disease signals, or conducting fine-mapping without integrating functional annotation data (typically under a single causal variant assumption) and then overlapping the resulting credible sets with functional annotation data to assess enrichment [9] [10]. Identifying enriched functional annotations using summary statistic has a promising future due to fast growth of both functional annotations and summary statistic datasets. However, increasing number of functional annotations also make computational and statistical challenges in extracting the informative functional annotations among many “noisy” ones.

To address these analytical challenges, we will extend functional annotation-based data-adaptive test (FunSPU) to the case with GWAS summary statistics without individual-level genotype and phenotype data. The new summary statistics-based test “FunSPUs” is also adaptive at both the annotation and variants levels. However, the higher accessibility of summary statistics dataset and less computing burden indicate potential usefulness of our proposed method. We compared FunSPUs test with the corresponding aSPUs test which is also adaptive but ignore the annotations. We also compared our proposed method with existing gene-based summary statistics tests, such as gene-based association test that uses extended Simes procedure (GATES) [11] in simulations studies and application to the Global Lipids Genetics Consortium (GLGC) datasets.

Materials and Methods

Suppose that we do not have individual-level data (Y_i, X_i) 's, but only single SNP-based summary statistics Z-scores $Z = (Z_1, \dots, Z_k)'$ with $Z_j = \hat{\beta}_j / se(\hat{\beta}_j) \approx U_j / se(U_j)$ for $j = 1, 2, \dots, k$. Previous research (Kwak, 2016) has shown the approximation

$$Cov(Z_j, Z_l) = Corr(Z_j, Z_l) \approx Corr(U_j, U_l) \approx Corr(X_{ij}, X_{il})$$

which can be estimated from some reference panel from a similar population e.g. 1000 Genome data.

Based on the FunSPU test with for individual genotype and phenotype data, I propose a new functional annotation-based SPU test with only summary statistics Z :

$$T_{SPUS-Fun(\gamma_a, \gamma)} = \sum_{l=1}^m \left[\left(\sum_{j=1}^k (w_{lj} Z_j)^\gamma \right)^{\frac{1}{\gamma}} \right]^{\gamma_a}$$

where two positive integers γ and γ_a respectively control the individual summary statistics and annotations' relative contributions to the overall test statistic. The optimal combination of γ and γ_a depends on the pattern of trait associations and annotations respectively, as previous discussion on FunSPU. Given a set of γ and γ_a , e.g. $\gamma_a \in \Gamma_a = \{1, 2, 4, 8, \infty\}$ and $\gamma \in \Gamma = \{1, 2, 3, \dots, 8\}$, the corresponding data adaptive test statistic (FunSPUs) is defined as

$$T_{FunSPUS} = \min_{\gamma \in \Gamma, \gamma_a \in \Gamma_a} p_{SPUS-Fun(\gamma_a, \gamma)},$$

where $p_{SPUS-Fun(\gamma_a, \gamma)}$ is calculated by the Monte-Carlo simulation detailed below.

Implement: Monte-Carlo Simulation

Let the correlation of Z-scores to be $R = \text{Corr}(Z_j, Z_l) \approx \text{Corr}(X_{ij}, X_{il})$. By the asymptotic distribution of U , under null hypothesis H_0 , Z follows a multivariate normal distribution $N(0, R)$, R can be estimated from some reference panel e.g. 1000 Genome. Therefore, we propose using Monte Carlo simulations to obtain the P-values of the $T_{SPUS-Fun}$ and $T_{FunSPUS}$ tests. Specifically, we will first generate independent $Z^{(b)} \sim N(0, R)$ for $b = 1, \dots, B$. Then, we will calculate the null test statistic $T_{SPUS-Fun(\gamma_a, \gamma)}^{(b)} = T_{SPUS-Fun(\gamma_a, \gamma)}(Z^{(b)})$ as well as their p -values $p_{SPUS-Fun(\gamma_a, \gamma)}^{(b)} = \left[\sum_{b_1 \neq b} I \left(\left| T_{SPUS-Fun(\gamma_a, \gamma)}^{(b_1)} \right| \geq \left| T_{SPUS-Fun(\gamma_a, \gamma)}^{(b)} \right| + 1 \right) \right] / B$. Therefore, we will have $T_{FunSPUS}^{(b)} = \min_{\gamma \in \Gamma, \gamma_a \in \Gamma_a} p_{SPUS-Fun(\gamma_a, \gamma)}^{(b)}$, and the final p -value of the FunSPUs test $p_{FunSPUS} = \left[\sum_{b=1}^B I \left(T_{FunSPUS}^{(b)} \leq T_{FunSPUS}(Z) \right) + 1 \right] / (B + 1)$.

Simulation setups

In this simulation study, we use real genotype data from UK10K, and generate summary statistics (Z-scores) from linear regressions. We select common variants (CVs) from chr2:173.03M~174.98M of the UK10K TWINSUK genotype data of 1,854 unrelated individuals. MAFs of the selected CVs were larger than 5%. Since some SNPs are highly correlated in this region, we prune out any SNPs when absolute correlations are larger than 0.5 (e.g. $|\text{Corr}(X_{ij}, X_{il})| > 0.5$) and just keep the first SNP by locations, finally resulting in 332 SNPs.

The procedures to generate simulated Z-scores are as follows:

Step 1: we simulate 3 sets of informative annotations (w_{1j}, w_{2j}, w_{3j}) from a uniform distribution $U(0.4, 1)$ corresponding to causal variants and from $U(0, 0.6)$ corresponding to neutral variants, as well as 3 sets of random annotations (w_{4j}, w_{5j}, w_{6j}) from $U(0, 1)$.

Step 2: we designate the first 160 CVs as causal variants ($j = 1, 2, \dots, 160$) and the remaining 172 CVs as neutral variants ($j = 161, 162, \dots, 332$). Then we randomly select $k = k_1 + k_2$ RVs: k_1 causal CVs from $j = 1, 2, \dots, 160$ and k_2 neutral CVs from $j = 161, 162, \dots, 332$.

Step 3: we use only informative annotations to calculate the effect size $\beta_j = 0.02 \times (w_{1j} + w_{2j} + w_{3j})$ for each causal CV.

Step 4: the simulated phenotype is obtained from $Y_i = \sum_{j=1}^{k_1} X_{ij}\beta_j + \varepsilon_i$, where ε_i follows $N(0,1)$ and $i=1,2,\dots,1854$.

Step 5: finally, we obtain Z-scores from coefficients of linear regression of (Y_i, X_i) .

Since we simulate Z-scores from individual-level data, we can use the correlation matrix of real genotype data $R = \text{Corr}(X_{ij}, X_{il})$ as our reference panel in FunSPUs Monte Carlo simulations. We include all three informative annotations, three random annotations and one dummy annotation in all tests (FunSPUs, GATES2 and others). The values of power are fixed as $k_1 = 8$ and $k_2 = \{8, 16, 32, 64, 128\}$, respectively.

The empirical power is calculated based on 1,000 replications for each scenario, while the type I error is calculated based on 5,000 replications for each scenario with the $\alpha = 0.005$. The significance levels are set as $\alpha = 0.05$ for both simulation studies. For permutation-based tests, 5,000 and 1,000 resamplings were conducted for each replication to evaluate type I error and power, respectively. Since the Monte Carlo simulation is much

faster than permutations, we also increase the number of resamplings to 50,000 for the type I error estimation at significance levels $\alpha = 0.005$.

Data availability

The 2013 lipid association scan meta-analysis results can be downloaded at <http://csg.sph.umich.edu/willer/public/lipids2013/>, and the 2010 lipid association scan meta-analysis results can be downloaded at <http://csg.sph.umich.edu/willer/public/lipids2010/>.

Results

Simulation results

We conducted an extensive simulation on FunSPUs and compared the empirical type I error with aSPUs and GATES2. FunSPUs and aSPUs tests could control the type I error rate at significant $\alpha = 0.05$ (SI Figure 3.1). However, type I error of GATE2 were slightly inflated with a greater number of neutral SNPs. When we attempted to control the rate of type I error at a lower level $\alpha = 0.005$, we observed considerable inflation in some cases of FunSPUs (Table 3.1). Due to incorporation of multiple annotations, FunSPUs might be less stable at the low level of p-values.

Next, we compared the empirical powers of FunSPUs, aSPUs and GATE2 (Figure 3.2). As shown in the figure, FunSPUs always had a higher power than aSPUs and GATE2 which disregarded functional annotations. We also observed that this advantage dimmed with a higher proportion of neutral SNPs. The power of FunSPUs was like the power of aSPUs and GATES2 when the number of neutral SNPs increased to 128. Since GATES2 performs as obviously high type I error at large number of neutral SNPs, the power of GATES2 is not reliable in this case.

Application to the lipid GWAS data

First, we applied FunSPUs to a 2013 lipid GWAS summary dataset [12] (~189,000 samples) and select high-density lipoprotein (HDL) as study trait (see Figure 3.3(A)). In the genome-wide study, we used six functional annotations (CADD [13], RegulomeDB [14], FunSeq [15], Funseq2 [16], GERP++ [17] and GenoSkyline [18]) and the 1000 Genomes Project data as the reference sample [19]. There were 24,766 genes generated from UCSC Table Browser [20] RefSeq track data in May 2014. Thus the genome-wide significant level after Bonferroni correction is $0.05/24766 = 2.02 \times 10^{-6} = 10^{-5.69}$. To reach this level, we used a step-up strategy. We started with simulation number $B=10,000$ and gradually increased B ; if those genes with estimated p -values $< 10/B$, we increased B to 10 times the current value and reran the simulation for these genes. The number of simulations in the final stage was $B=10^6$. The Manhattan plots for the results of FunSPUs is shown in Figure 3.3(A).

In total, we identified 268 genome-wide significant genes, eleven of which (MARCH8, ARL15, CMIP, COBLL1, FAM13A, FTO, MLXIPL, RBM5, SLC39A8, STAB1, UBASH3B) have been reported by 2013 lipid GWAS study [12]. It is noteworthy that 5 out of these 11 genes (RBM5, STAB1, FAM13A, MARCH8, FTO) were identified by 2013 study but not by 2010 lipid GWAS study [21] (~100,000 samples). As a comparison, we applied aSPUs test to this 2013 lipid GWAS summary dataset ignoring all functional annotations (see Figure 3.3(B)) and identified 208 significant genes. However, 66 genes identified by FunSPUs would have been missed by aSPUs, including all of 5 newly reported genes in 2013 GWAS study above. In contrast, there were only 6 genes identified by aSPUs missed by FunSPUs, including one newly reported gene (DGAT2) in 2013 GWAS study.

Therefore, FunSPUs incorporating functional annotations demonstrates power to identify more associations than aSPUs. For reference, we also used the GATES2 to perform genome-wide scan on this dataset (see Figure 3.3(c)) and identified 235 genome-wide significant genes of which 188 were overlapped by FunSPUs and 150 by aSPUs.

Next, we applied the FunSPUs and aSPUs to a 2010 lipid dataset [21] (~100,000 samples) with trait HDL and 6 functional annotations same as above. We identified 236 genome-wide significant genes by FunSPUs and the Manhattan plots is shown in Figure 3.4(A). Although the number of identified genes here is less than that obtained from 2013 GWAS summary dataset, 7 out of these 236 genes (ATG7, FAM13A, RSPO3, DAGLB, SNX13, MARCH8, DGAT2) were identified by 2013 GWAS study but not by 2010 GWAS study. Considering that there are only 24 significant genes for HDL were newly reported by 2013 study, FunSPUs has potential to identify more significant genes from a smaller dataset. Even so, we notice that 50 significant genes were identified by FunSPUs in 2013 summary dataset while missed by FunSPUs in 2010 summary dataset, 3 of which (RBM5, STAB1, FTO) were identified by 2013 GWAS study but not by 2010 GWAS study. These results indicated that the dataset with large sample size still has advantage for FunSPUs test. There were 18 significant genes were identified in 2010 summary dataset and missed in 2013 dataset. None of them was included in newly reported genes list of 2013 GWAS study. The aSPUs test also identified 208 significant genes in 2010 summary dataset (see Figure 3.4(B)) which were totally identical with the results obtained from 2013 summary dataset.

Discussion

In summary, we have proposed a new adaptive test for summary statistics incorporating multiple functional annotations. In addition to avoiding the accessibility problem as individual-level genomic data, the sample size of the summary statistics can achieve >100,000 and even larger if summary statistics obtained from meta-analysis. Since there is no need to process individual-level data, and we implement Monte-Carlo simulations rather than permutations to obtain the p value for FunSPUs test, the computing speed of FunSPUs is faster than that of FunSPU and aSPU. However, it still takes several hours per gene to reach the level of $p < 10^{-6}$. If computing resources are limited, we recommend using FunSPUs to validate reported significant loci identified by GWAS studies previously instead of genome-wide scan.

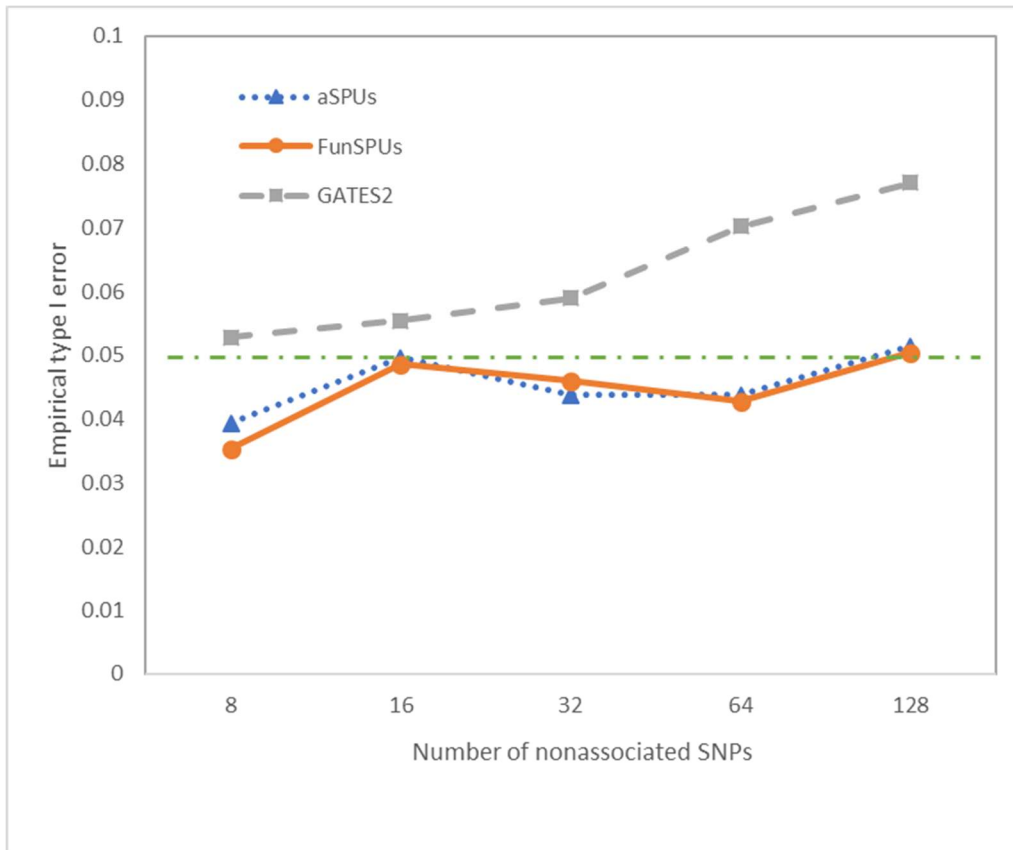
Applying aggregation strategy to association test studies can greatly reduce multiple tests burden and increase the power of tests. Since the summary statistics we used are based on common variants, it is more explicit to interpret functional and biological mechanism underlying complex traits by using gene-based tests rather than sliding window strategy. Further research can include SNVs adjacent to genes to identify *cis*-regulators. In addition, we noticed that aSPUs test omitting any functional annotations also has potential to identify new significant loci which were missed by FunSPUs. Therefore, we can use aSPUs as complementary of FunSPUs in association studies.

Our work still has some limitations. The most critical concern is related to the type I error inflation at low significant level. Simulation results show the inflation of FunSPUs test occurs at about $\alpha < 0.05$. In contrast, the type I error inflation of aSPUs test is lower at

corresponding significant level (SI Figure 3.5). In the real data analysis, we also observed obvious deviation from expected distribution of both FunSPUs and aSPUs tests at even lower significant level (SI Figure 3.6), although SNVs in gene regions are more likely to be trait-associated SNVs and consist of more significant results. Since the genome-wide scan usually includes >20,000 genes, the family-wise significant threshold of p -values may reach 10^{-6} or smaller based on Bonferroni correction. Therefore, the type I error inflation at low significant level may yield a large amount of false positive results in genome-wide association tests. To overcome this challenge, we can implement other summary statistics tests in the study and find the intersection of multiple approaches. Otherwise, we can also incorporate different functional annotations into FunSPUs tests to identify the consistently significant loci.

[Figure 3.1: Empirical type I error rates of various summary statistics-based tests at \$\alpha = 0.05\$](#)

All the tests were performed at significance level $\alpha = 0.05$ for increasing number of neutral SNPs with 5,000 simulation replications. Annotation-based tests (FunSPUs) were based on six random annotations. The horizontal dashed line indicated $\alpha = 0.05$.



[Figure 3.2: Empirical power of various tests for eight causal SNPs and increasing number of neutral RVs.](#)

The incorporated annotations for association tests include three informative annotations and three noninformative annotations. All the results were based on 1,000 simulation replications.

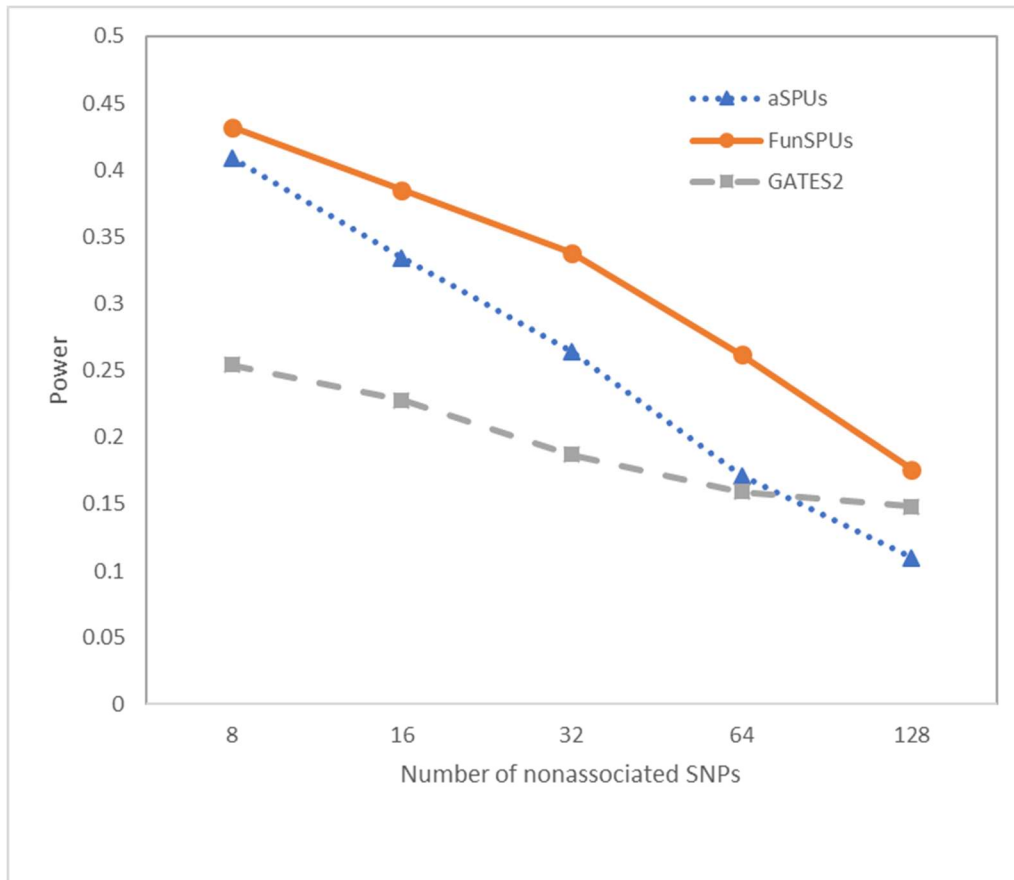


Figure 3.3: Manhattan plots for the association test results for trait HDL based on the 2013 lipid data.

(A) FunSPUs; (B) aSPUs; (C) GATES2. Significance threshold: $p < 4.03 \times 10^{-5}$

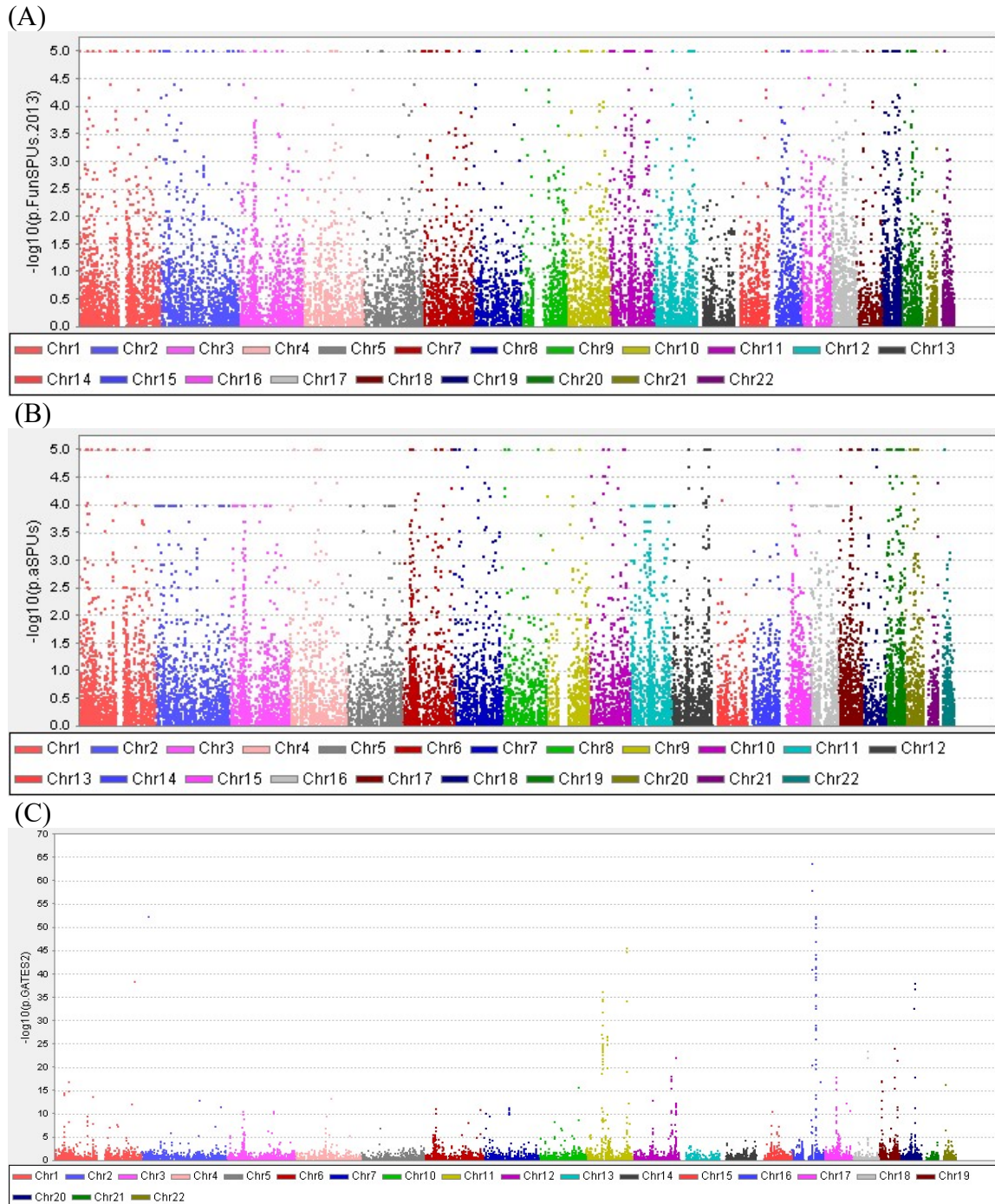
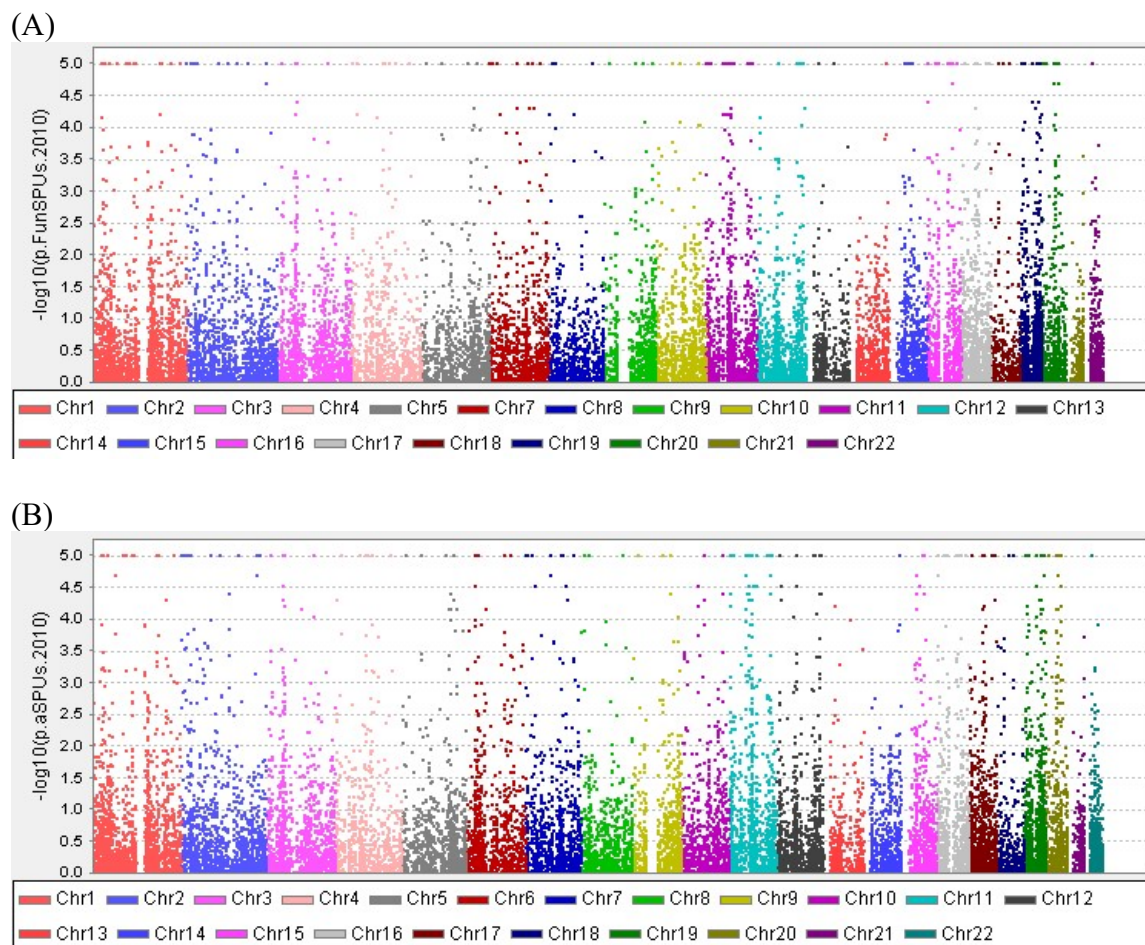


Figure 3.4: Manhattan plots for the association test results for trait HDL based on the 2010 lipid data

(A) FunSPUs; (B) aSPUs. Significance threshold: $p < 4.03 \times 10^{-5}$



Supporting Information

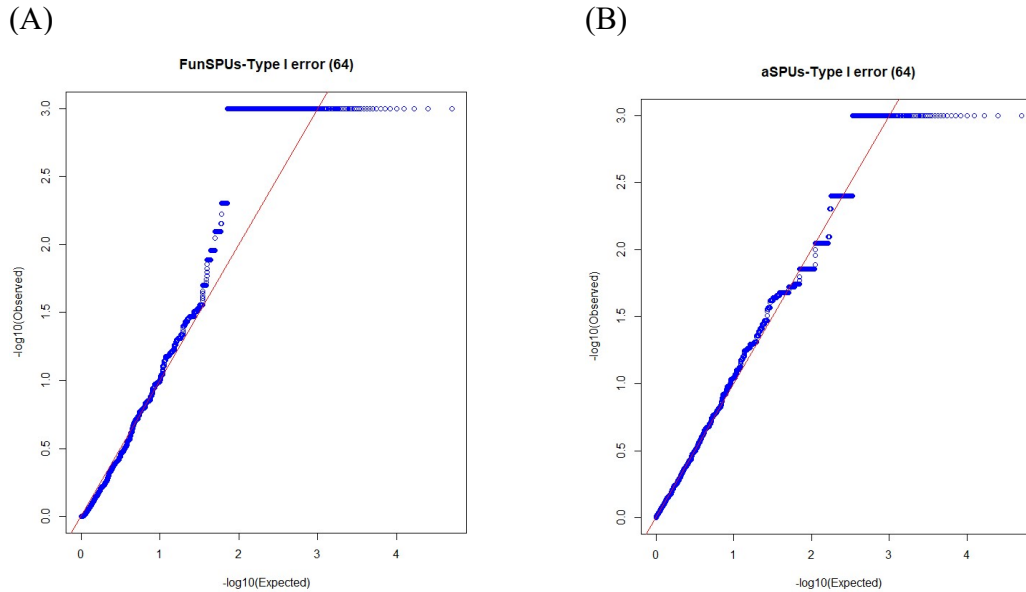
[Table 3.1: Empirical type I error rates of various tests at \$\alpha = 0.005\$](#)

Association tests for increasing number of neutral SNPs with 50,000 simulation replications. Annotation-based tests were based on six random annotations.

Test $\alpha = 0.005$	Number of neutral SNPs				
	8	16	32	64	128
aSPUs	0.00024	0.00942	0.00288	0.00580	0.00310
FunSPUs	0.00056	0.01356	0.00428	0.01664	0.00746

[Figure 3.5: The Q-Q plots for the association analysis of simulated data](#)

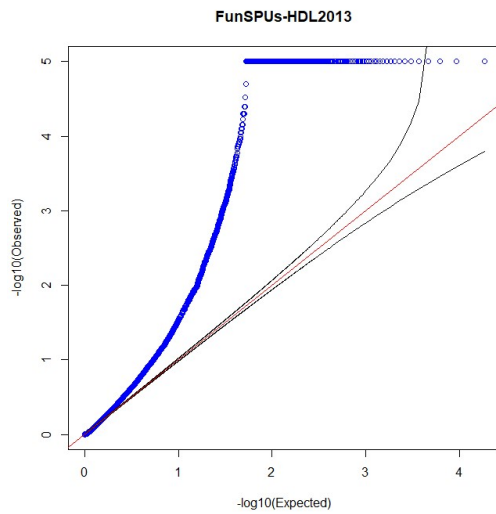
(A) FunSPUs (B) aSPUs tests for 64 neutral SNPs with 1,000 simulation replications



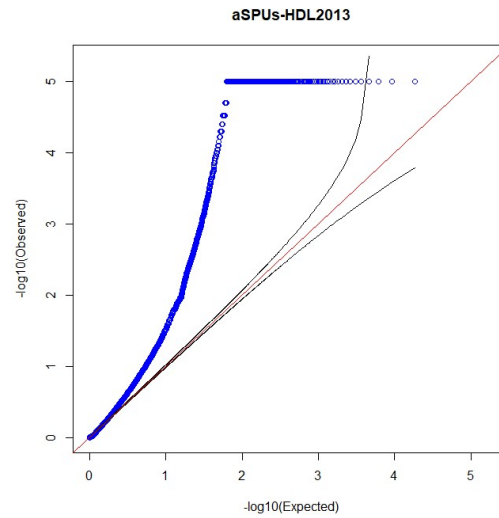
[Figure 3.6: The Q-Q plots for association tests of the summary statistics with HDL in the 2013 lipid data](#)

(A) FunSPUs; (B) aSPUs.

(A)



(B)



Reference

1. Zheng J, Erzurumluoglu AM, Elsworth BL, Kemp JP, Howe L, et al. (2017) LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. *Bioinformatics* 33: 272-279.
2. Pasaniuc B, Price AL (2017) Dissecting the genetics of complex traits using summary association statistics. *Nat Rev Genet* 18: 117-127.
3. Yang J, Ferreira T, Morris AP, Medland SE, Genetic Investigation of ATC, et al. (2012) Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat Genet* 44: 369-375, S361-363.
4. Marchini J, Howie B (2010) Genotype imputation for genome-wide association studies. *Nat Rev Genet* 11: 499-511.
5. Conneely KN, Boehnke M (2007) So many correlated tests, so little time! Rapid adjustment of P values for multiple correlated tests. *Am J Hum Genet* 81: 1158-1168.
6. Schaid DJ, Chen W, Larson NB (2018) From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat Rev Genet* 19: 491-504.
7. Faye LL, Machiela MJ, Kraft P, Bull SB, Sun L (2013) Re-ranking sequencing variants in the post-GWAS era for accurate causal variant identification. *PLoS Genet* 9: e1003609.
8. Kichaev G, Yang WY, Lindstrom S, Hormozdiari F, Eskin E, et al. (2014) Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS Genet* 10: e1004722.
9. Farh KK, Marson A, Zhu J, Kleinewietfeld M, Housley WJ, et al. (2015) Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* 518: 337-343.
10. Liu X, Li C, Boerwinkle E (2017) The performance of deleteriousness prediction scores for rare non-protein-changing single nucleotide variants in human genes. *J Med Genet* 54: 134-144.
11. Li MX, Gui HS, Kwan JS, Sham PC (2011) GATES: a rapid and powerful gene-based association test using extended Simes procedure. *Am J Hum Genet* 88: 283-293.
12. Willer CJ, Schmidt EM, Sengupta S, Peloso GM, Gustafsson S, et al. (2013) Discovery and refinement of loci associated with lipid levels. *Nat Genet* 45: 1274-1283.
13. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, et al. (2014) A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* 46: 310-315.
14. Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, et al. (2012) Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res* 22: 1790-1797.
15. Khurana E, Fu Y, Colonna V, Mu XJ, Kang HM, et al. (2013) Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science* 342: 1235587.
16. Fu Y, Liu Z, Lou S, Bedford J, Mu XJ, et al. (2014) FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. *Genome Biol* 15: 480.

17. Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, et al. (2010) Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput Biol 6: e1001025.
18. Lu Q, Powles RL, Wang Q, He BJ, Zhao H (2016) Integrative Tissue-Specific Functional Annotations in the Human Genome Provide Novel Insights on Many Complex Traits and Improve Signal Prioritization in Genome Wide Association Studies. PLoS Genet 12: e1005947.
19. Genomes Project C, Abecasis GR, Altshuler D, Auton A, Brooks LD, et al. (2010) A map of human genome variation from population-scale sequencing. Nature 467: 1061-1073.
20. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, et al. (2002) The human genome browser at UCSC. Genome Res 12: 996-1006.
21. Teslovich TM, Musunuru K, Smith AV, Edmondson AC, Stylianou IM, et al. (2010) Biological, clinical and population relevance of 95 loci for blood lipids. Nature 466: 707-713.

CHAPTER 4: A NOVEL FRAMEWORK FOR COMPARING FUNCTIONAL ANNOTATIONS OF RARE VARIANTS WITHOUT GOLD STANDARD

Introduction

Despite more potentially deleterious single nucleotide variants (SNVs) were identified and reported in whole genome sequencing studies, accurate interpretation of these candidate SNVs remains a challenge for clinical use. The American College of Medical Genetics and Genomics (ACMG) has developed an instructive guideline for classification of sequence variants using criteria informed by expert opinion and empirical data [1]. In this guideline, a variety of computational (in silico) predictive tools can be used to aid in the interpretation of sequence variants. Although computational tools have been widely used in screening and prioritizing candidate variants, expert opinion and empirical data are still critical factors to determine the deleteriousness of SNVs in clinical practice.

In recent years, next generation sequencing technology has crucially reduced the cost of whole genome sequencing (WGS), making the WGS data feasible for clinical applications. Some genome-wide studies provide adequate resources for new functional genomic deleteriousness predictive scores designed specifically for missense or splicing variants. The primary method used to identify candidate functional elements in genomic sequences has been based on conservation and acceleration, such as phastCons [2], GERP++ [3] and phyloP [4]. Furthermore, a new group of algorithms have been proposed for predicting the deleteriousness of a variant in genes or intergenic regions, such as CADD [5], Funseq2 [6], Eigen/Eigen-PC [7], fitCons [8] and fathmm-MKL[9]. However, evaluating the predictive performance of these “genome-level” deleteriousness prediction scores, especially compared

to conservation scores, is still a challenging question, which is an obstacle to applying whole genome sequencing to clinical diagnosis.

A recent study proposed an approach to investigate the prediction accuracy of genome-level scores for rare non-protein-changing single nucleotide variants (npcSNVs) in and near human genes [10]. In this study, npcSNVs which cause the canonical transcript of a pathogenic gene from the HGMD database [11] was determined as the “gold standard” of deleterious variants. Totally 23 deleteriousness prediction scores and conservation scores were compared using receiver operating characteristic (ROC) and area under the ROC curve (AUC), and fathmm-MKL coding score was found to have the best prediction accuracy. However, comparison of genome-level prediction scores for intergenic npcSNVs is yet available due to the lack of gold standard, i.e., known deleterious and neutral npcSNVs, especially those not near genes. To address this challenge, we propose a trait-specific framework for comparing functional annotations of rare variants without gold standard. The fundamental idea is based on the partition of heritability, which is a data-driven process. The rank of the resulting correlation coefficient ρ between functional annotation deleteriousness category and per-SNV heritability within each category weights provides a practical estimation of predictive performance of functional annotations. Furthermore, our proposed measure ρ is trait-specific, in contrast to existing methods based on known deleterious SNVs by pooling together a large number of heterogeneous Mendelian diseases, such as those based on the HGMD database.

Methods and Materials

Partition of the heritability and evaluation framework

Since we assume that there is no gold standard regarding the informativeness of a functional annotation for a given trait, we suppose to create a framework to estimate global correlation measure between the annotation scores, genotypes and phenotype. Our proposed approach is based on partitioning the heritability h^2 by functional annotations [12]. Suppose that the values of rank scores or deleteriousness categories have ordinal association with biological function strength, i.e., a functional annotation should be more informative for the trait of interest if SNVs with higher functional scores contribute to more heritability on average. Therefore, we propose the following procedures to estimate the global correlation measure (Figure 1):

Step 1: Prior to this calculation, we transform the functional annotation to positive integers $q = 1, \dots, Q$ such that larger q corresponds to a more likely functional category. If a functional category has a very small number of SNVs or the category-specific heritability h_q^2 close to zero, this category is combined with a nearby category.

Step 2: Given an annotation, we first partition the genome-wide RVs based on Q discrete functional categories or percentiles of continuous functional scores.

Step 3: We then estimate the heritability h_q^2 for all SNVs in functional category $q = 1, \dots, Q$, using the GCTA tool [13]. Specifically, the GCTA tool estimate genetic relationship matrix (GRM) between individuals from the set of SNVs in a category via a linear mixed model approach, then perform REML (restricted maximum likelihood) analysis to estimate the

variance explained by the SNVs that were used to estimate the GRM, a.k.a. the heritability of this set of SNVs. The estimation is adjusted by age and top 10 PCs to capture phenotypic variance due to population substructure.

Step 4: We compute the average per-SNV heritability $h_{(q)}^* = h_q^2 / \#SNV^{(q)}$ for each annotation category q .

Step 5: If there are more than two categories ($Q > 2$), we regress $h_q^2 / \#SNV_q$ on q to estimate the slope: $E(h_q^2 / \#SNV^{(q)}) = \beta_0 + \beta q$, where β is used as the global correlation measure ρ for the corresponding functional annotation and trait. We also compare the regression slopes β with the Pearson correlation coefficients between $h_q^2 / \#SNV_q$ and numerical categories (1, 2, ..., Q) to determine the reliability of the global correlation results. If there are only two categories (i.e., $Q = 2$), we directly calculate the prediction accuracy of per-SNV heritability as the global correlation measure ρ for the corresponding functional annotation and trait:

$$\rho = \frac{h_2^2 / \#SNV^{(2)}}{h_1^2 / \#SNV^{(1)} + h_2^2 / \#SNV^{(2)}}$$

We can also calculate Matthew correlation coefficient between two categories as ρ value.

Functional annotations and WGS data

We used the NHLBI Trans-Omics for Precision Medicine (TOPMed) Atherosclerosis Risk in Communities Study (ARIC) [14] whole genome sequencing (WGS) samples as the discovery dataset and UK10K samples (TwinsUK and ALSPAC cohorts) as the replication cohorts. To achieve consistency among study cohorts, we only selected the European American subjects from ARIC dataset with a sample size $n=3,369$, and the sample sizes of

TwinsUK and ALSPAC cohorts are $n=1,754$ and $n=1,868$. The genomic coordinates are based on GRCh38. After removing singletons, INDELs and duplicated SNVs from the ARIC WGS data, we extracted a total of 22,935,967 rare and low frequency variants with $MAF < 5\%$. The RVs from TwinsUK and ALSPAC were extracted following the same procedure and the total numbers of RVs were 10,977,173 and 35,054,869 respectively.

The SNVs of ARIC and UK10K were annotated using the WGS pipeline [15]. We extracted totally 36 functional annotations from the precomputed WGS library including 8 conservation scores, 14 deleteriousness scores and 14 tissue-specific GenoSkyline Plus scores. Since category-specific-heritability calculations by the GCTA were not numerically reliable for some annotations (i.e. heritability of more than one categories is too small), we selected totally 12 functional annotations for the comparison herein, including CADD, DANN, fathmmMKL_non.coding, fathmmMKL_coding, GERP++, GenoCanyon, phastCons100way_vertebrate, phastCons46way_placental, phastCons46way_primate, phyloP100way_vertebrate, phyloP46way_placental and phyloP46way_primate. Rank (percentile) scores are available for all the above annotations in the WGS pre-computed dataset. We considered 10 traits in the ARIC WGS data: low-density lipoprotein (LDL), total cholesterol (TC), triglycerides (Trig), Apolipoprotein A1 (ApoA1), Apolipoprotein B (ApoB), body mass index (BMI), standing height (Height), diastolic blood pressure (diastolic_BP), systolic blood pressure (systolic_BP) and fasting insulin level (Insulin). We omitted ApoA1 and ApoB in the analysis of TwinsUK and ALSPAC WGS data due to the small values of the overall estimated heritability.

Results

The regression slope β generated by above method is used for the trait-specific comparison of functional annotations. Since the overall heritability of each trait is fixed, partition of heritability by annotation scores indicates the prediction performance of corresponding functional annotation. If the heritability per SNV ($h_q^2/\text{\#SNV}^{(q)}$) is significantly higher in the more deleterious functional score groups, we can conclude that this functional annotation is more predictive of the specific phenotype. Otherwise, if $h_q^2/\text{\#SNV}^{(q)}$ is practically invariable across functional score groups, we can infer low informativeness of the functional score on this phenotype. When we compare the performance of functional scores across phenotypes, we need to tackle the diverse scale of the overall heritability for each phenotype. Therefore, it is not meaningful to directly compare regression slope β 's among phenotypes. Instead, we ranked the functional annotations by the order of slope β 's from low to high (i.e., rank #1 has the lowest slope) for each phenotype, then we could compare the ranks of each annotations across phenotypes to evaluate the phenotype-specific prediction performance. We grouped the phenotypes based on biological functions to discern the potential pattern of ranks (SI Figure 4.7-4.9).

First, we compared conservation scores by the regression slope β 's derived from heritability of partitioning the ARIC RV data (Figure 4.2). PhyloP scores almost outperformed phastCons scores across phenotypes regardless of alignment sets, i.e., primate, placental mammals or vertebrate species. Of note, phyloP measures both conservation and acceleration, while phastCons measures conservation only. When we inspect the

conservation tracks in UCSC Genome Browser, phastCons and phyloP scores are roughly similar under low resolution. However, phyloP scores capture more variation at finer resolution [4]. Thus, phyloP scores have the potential to locate functional non-coding SNV sites more accurately. On the other hand, the performance of phastCons scores depends on alignment sets, i.e., primate>placental>vertebrate (Figure 4.2), which could be intuitively interpreted as distance between alignment animal species and Homo sapiens. This pattern was not significant among the phyloP scores.

In general, the performance of GERP++ was between phyloP and phastCons scores for all the phenotypes. We know that phyloP scores implemented four statistical, phylogenetic tests including the precursor of GERP++, the genomic evolutionary rate profiling (GERP) test. Therefore, phyloP scores may include information complementary to GERP++ [4] and outperform it.

The rank trend on the ARIC RV data was partially replicable in the TwinsUK and ALSPAC cohorts of UK10K (SI Figure 4.8 and 4.9) including BMI, diastolic blood pressure and fasting insulin level, despite observing noticeable difference for some phenotypes (Figure 4.3). The sample size of each UK10K cohort is notably smaller than that of ARIC, and the ranks might be less reliable, due to the larger standard errors of estimated h^2 especially for the phenotype with small underlying h^2 .

Next, we evaluated some ensemble deleteriousness prediction scores of non-coding variants (Figure 4.4). We found that fathmm-MKL coding score (trained with coding disease-causing and 10 feature groups) always outperformed fathmm-MKL non-coding score (trained with non-coding disease-causing and top 4 feature groups) except for ApoA1 level and

diastolic blood pressure, and both were comparable to CADD. Previous study indicates that the prediction accuracy of fathmm-MKL coding and non-coding scores exceeds CADD on rare non-protein-changing SNVs (npcSNVs) [10]. However, our proposed method demonstrates phenotype-specific performance of fathmm-MKL non-coding and CADD.

Our proposed measure ρ marginally ranked fathmm-MKL non-coding higher than CADD on BMI and diastolic blood pressure (DBP). As for lipid traits, such as LDL and triglycerides, CADD significantly outperformed fathmm-MKL non-coding and even fathmm-MKL coding score. These results help us better understand functional annotations in the context of specific phenotypes.

On the other hand, DANN had the lowest performance among ensemble prediction scores. In fact, the original study did not directly compare to CADD due to a much smaller training set of SNVs for DANN [16]. We speculate that the underlying feature of deep neural network (DNN) may have impeded training a viable prediction score. In addition, GenoCanyon[17], as the pioneer of functional annotations of the non-coding regions, also provide poor prediction performance across all phenotypes. In fact, some new tissue-specific functional annotations developed by the similar strategy, e.g. GenoSkyline[18] and GenoSkyline-plus [19], demonstrate better functional predictivity.

The regression slope obtained as above can be used to measure the association between a functional annotation and a specific phenotype, i.e., per-SNV heritability increase with the deleteriousness category increasing by 1. While this measure is informative in the presence of linear association between $h_q^2/\#\text{SNV}^{(q)}$ and q , it is less reliable in the presence of

non-linear trend. For example, if only the highest annotation group (e.g., rank scores between 0.8 and 1) of SNVs contribute to the majority of heritability, the regression slope will be inaccurately estimated compared to the annotations linearly associated with $h_q^2/\text{\#SNV}^{(q)}$, although the former annotation also has the potential to screen out deleterious SNVs.

Therefore, we plot the regression slopes versus correlations between annotation categories and $h_q^2/\text{\#SNV}^{(q)}$, the latter of which measures the linear dependence. As shown in Figure 4.5, for the partitioning of the heritability for apo A1 levels, we observed that both regression slope and correlation of fathmm-MKL non-coding score were, suggesting that β captured the linear trend. In contrast, the regression slope of fathmm-MKL coding score was lower than that of fathmm-MKL non-coding score, but the correlation of coding score was also low, indicating that β was not a reliable measure for its predictive performance. Thus, we cannot arbitrarily determine that non-coding scores had a better chance to filter out SNVs not related to apoA1 level than coding score, simply based on the β 's. In contrast, the correlations of both fathmm-MKL coding and non-coding scores were quite high for BMI (Figure 4.6), so we can conclude that coding score should have better predictive performance due to the higher regression slope. In general, we can trust the evaluation of a functional annotation based on the regression slope when the correlation between deleteriousness category q and $h_q^2/\text{\#SNV}^{(q)}$ is a high (e.g., > 0.75) value. If this correlation is relatively low, we need to scrutinize the results of partitioning the heritability by deleteriousness categories.

Discussion

The method we have proposed here provides a general framework to compare multiple functional annotations for complex traits. Due to the lack of reported disease-causing SNVs, in particular non-coding ones, as gold standard, the existing comparisons in the literature only focus on SNVs in or near a gene ($\pm 5\text{kb}$) [10]. In fact, a number of recently developed genome-wide functional scores have already covered a large portion of non-coding variants. Our method is based on partitioning the heritability by annotation scores in principle, so it does not depend on databases of disease-causing SNVs such as ClinVar or HGMD. Furthermore, our proposed method can provide insights into phenotype-specific comparative performance of various functional annotations, which is applicable to a wide range of research and clinical settings covering diverse complex diseases and traits. By applying our proposed method to the ARIC WGS data and 10 complex traits, we were able to rank the performance of 12 functional annotations including conservation scores and ensemble deleteriousness prediction scores for each trait.

In addition, we validated the robustness of our method with regard to the number of deleteriousness score categories. We reclassified CADD rank scores into 2, 3, 4 or 5 groups at equal intervals and partitioned the UK10K TwinsUK RVs accordingly. Then we estimated h^2 of HDL and calculated the corresponding regression slopes and Pearson's correlation coefficients (SI Table 4.1). Except for $Q = 2$, the regression slopes of $Q = 3, 4, 5$ were quite consistent, and the correlation values were very high (≥ 0.90). These results show that the number of categories has a limited effect on the estimation of regression slopes.

To make scores of different functional annotations comparable, we would use rank scores (percentiles) whenever available in the WGS database or convert raw scores to the interval [0, 1] in our calculation. If a score is bimodal, we can directly convert it to binary categories (Q=2) and calculate the global correlation measure $\rho = \frac{h_2^2/\#SNV^{(2)}}{h_1^2/\#SNV^{(1)}+h_2^2/\#SNV^{(2)}}$.

We employed the GCTA to estimate the overall heritability of all autosomal (?) chromosomes. Due to high LD, the total heritability is smaller than the sum of the h^2 of each deleteriousness score category. For EA subjects from the ARIC cohort after removing singleton SNVs, the range of overall h^2 was from 0.15 (DBP) to 0.63 (BMI). The accuracy of our proposed method depends on the standard error (SE) of the estimated h^2 . For a given set of SNVs, $SE(h^2)$ is inversely proportional to the sample size of the GWAS/WGS data but independent of the estimated h^2 value[20]. The theoretical derivation demonstrates that estimated $SE(h^2)$ for the genome-wide common variants is approximately $316/N$ [21]. Based on this result, we can obtain that $SE(h^2)=0.094$ for ARIC ($n=3,369$). As for UK10K TwinsUK with a smaller sample size ($n=1,754$), $SE(h^2)=0.180$. If we want to reduce $SE(h^2)$ to 0.01, the corresponding sample size must reach 30,000. Meanwhile, the variance in genetic relatedness will decrease when more RVs are included, thereby making $SE(h^2)$ higher.

When using GCTA to estimate heritability from WGS data, bias may be encountered due to MAF or LD stratification of underlying casual variants. Therefore, the GREML-LDMS method is proposed to estimate heritability which adjusts for difference in both MAF and LD between variants [22]. Since we only choose RVs (MAF <5%), both MAF and LD are at lower values and it is not necessary to use this method to estimate h^2 .

Our proposed method has some limitations. First, our method requires that the heritability of the phenotype of interest should be significantly higher than 0. Otherwise, the evaluation of any functional annotation for that phenotype is not feasible. If the estimated h^2 is too small, based on the discussion above, it is not meaningful to evaluate the phenotype by our method. In fact, our method is more appropriate for phenotypes for which a larger proportion of h^2 can be explained by RVs, such as height or BMI [22]. Second, since the accuracy of our proposed method depends on the sample size, a bottleneck is the availability of WGS datasets of large sample sizes, which has become of less concern thanks to the completed and ongoing large-scale WGS projects, such as the recently completed NHLBI TOPMed project of more than 135,000 individuals and the ongoing UK Biobank WGS project of 500,000 individuals. Despite these caveats, our proposed procedure based on partitioning the heritability provides a novel and general framework to compare multiple functional annotations for complex traits without gold standard.

Figure 4.1: Illustration of the proposed framework for comparing functional annotations of rare variants

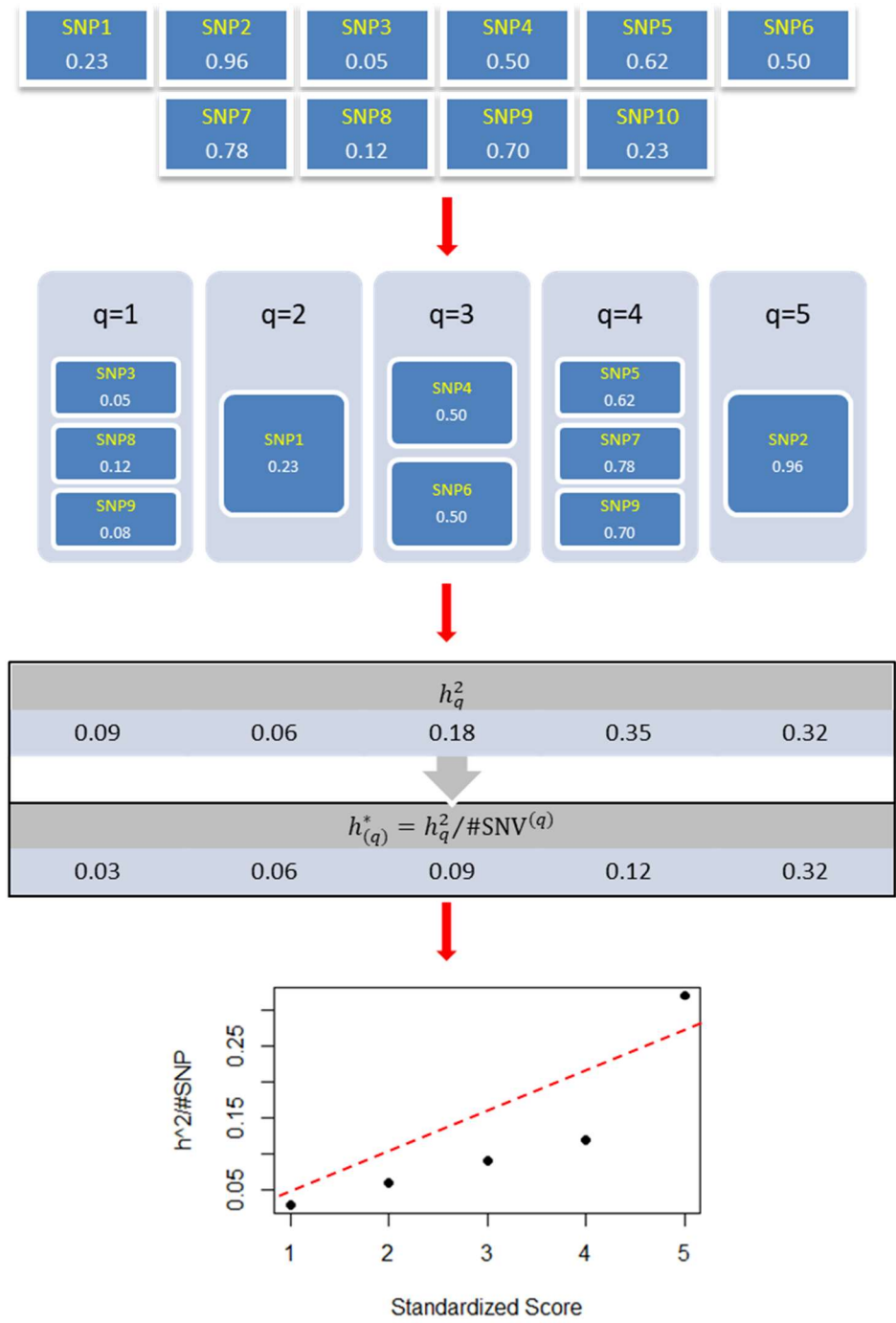
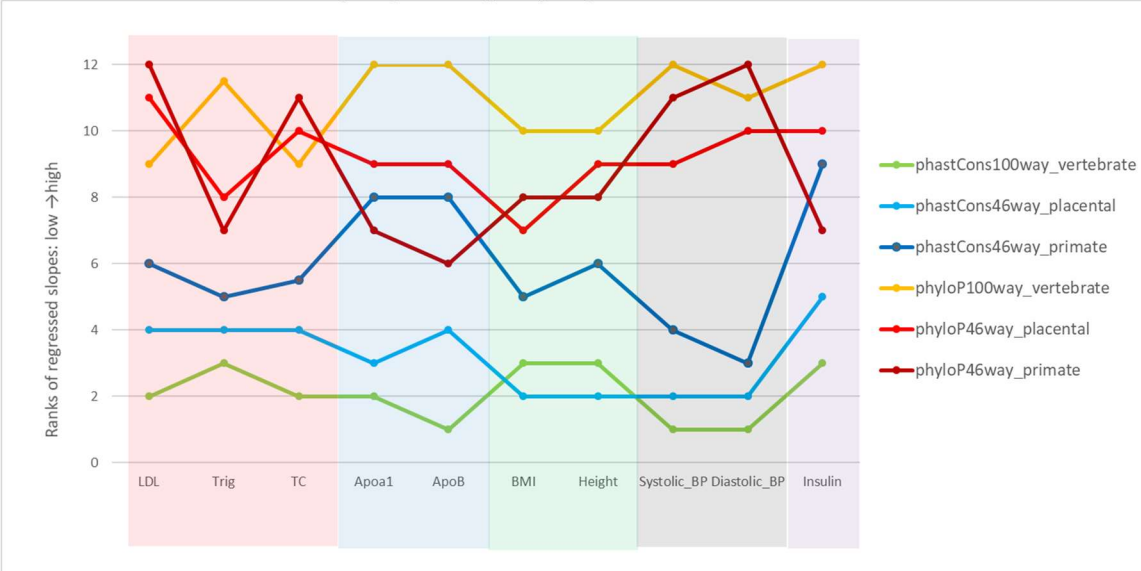


Figure 4.2: Phenotype-specific rank of regression slopes estimated by partitioning ARIC RVs for conservation scores

Higher rank indicates stronger global correlations for the corresponding functional annotation and trait. Shaded colors highlight the groupings of traits.



[Figure 4.3: Phenotype-specific rank of regression slopes estimated by partitioning UK10K](#)

[TwinsUK RVs for conservation scores](#)

Higher rank indicates stronger global correlations for the corresponding functional annotation and trait. Shaded colors highlight the groupings of traits.

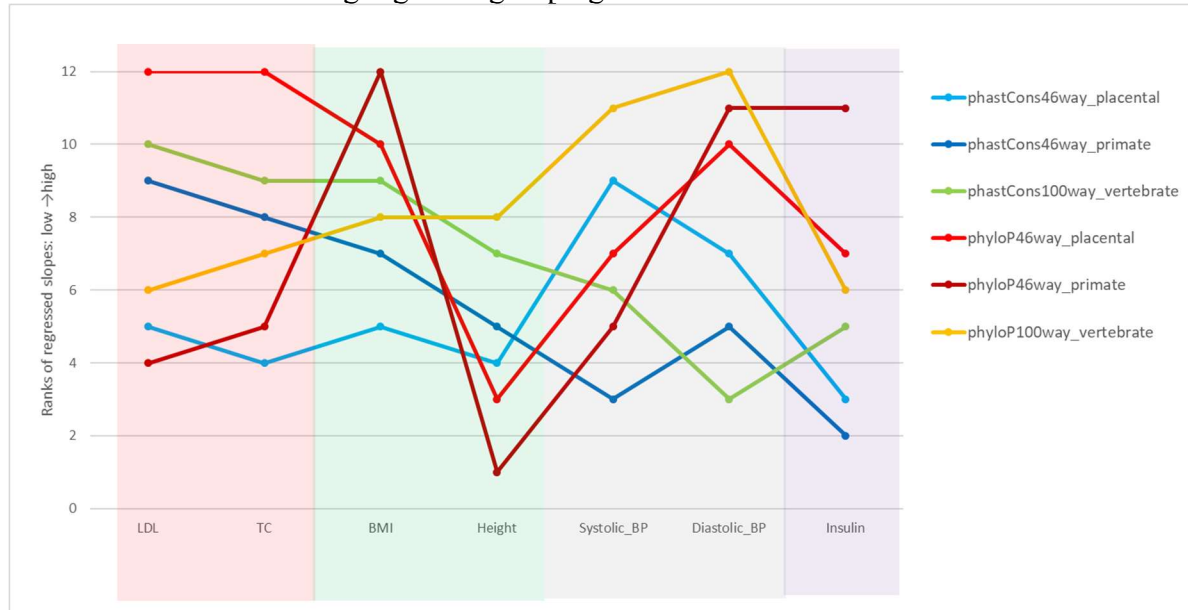


Figure 4.4: Phenotype-specific rank of regression slopes estimated by partitioning ARIC RVs for ensemble functional annotations

Higher rank indicates stronger global correlations for the corresponding functional annotation and trait. Shaded colors highlight the groupings of traits.

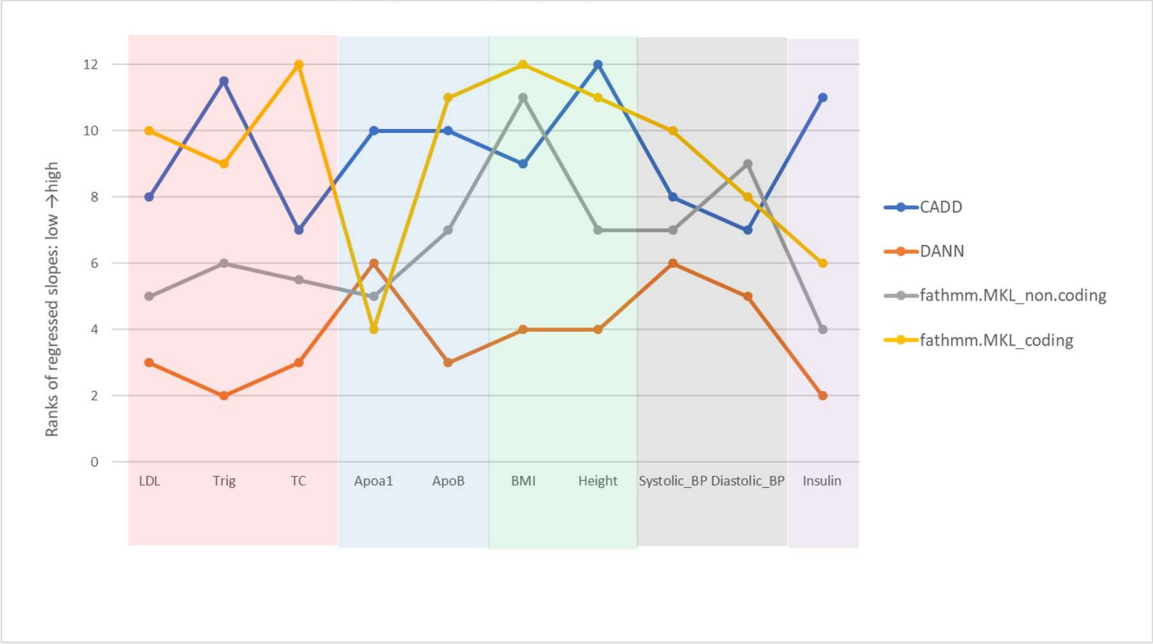


Figure 4.5: Regression slopes (vertical axis) and Pearson’s correlations coefficients (horizontal axis) calculated by partitioning the heritability for apo A1 levels.

Heritability estimation was based on ARIC RVs.

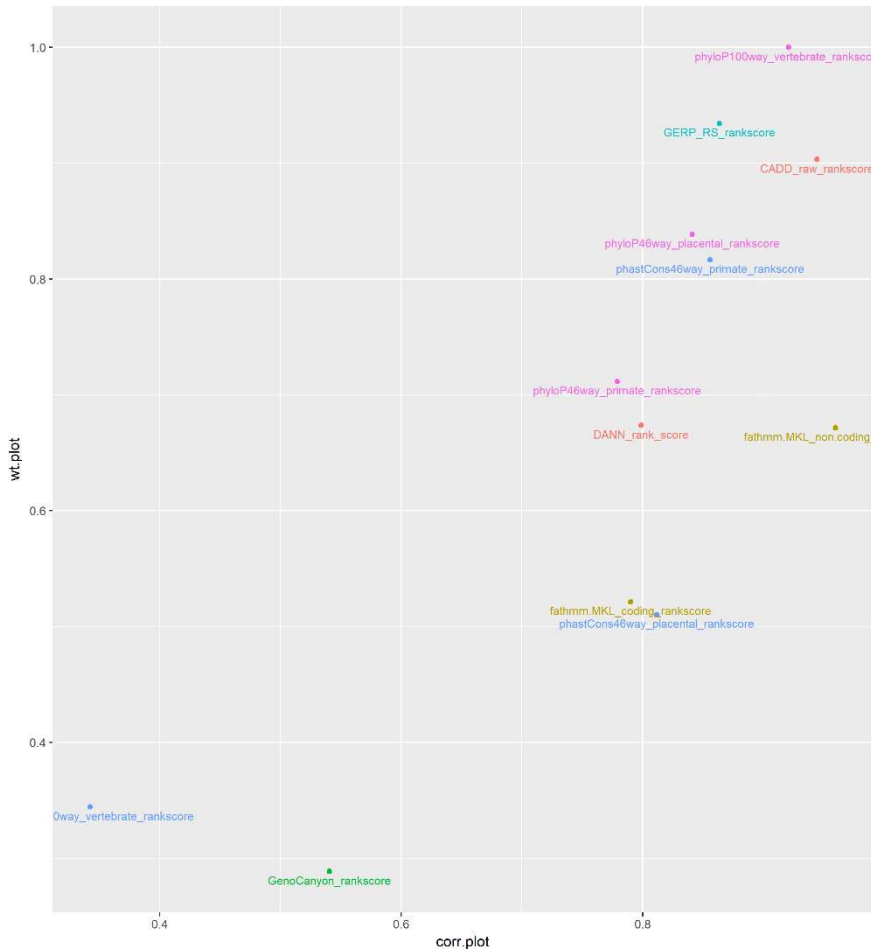
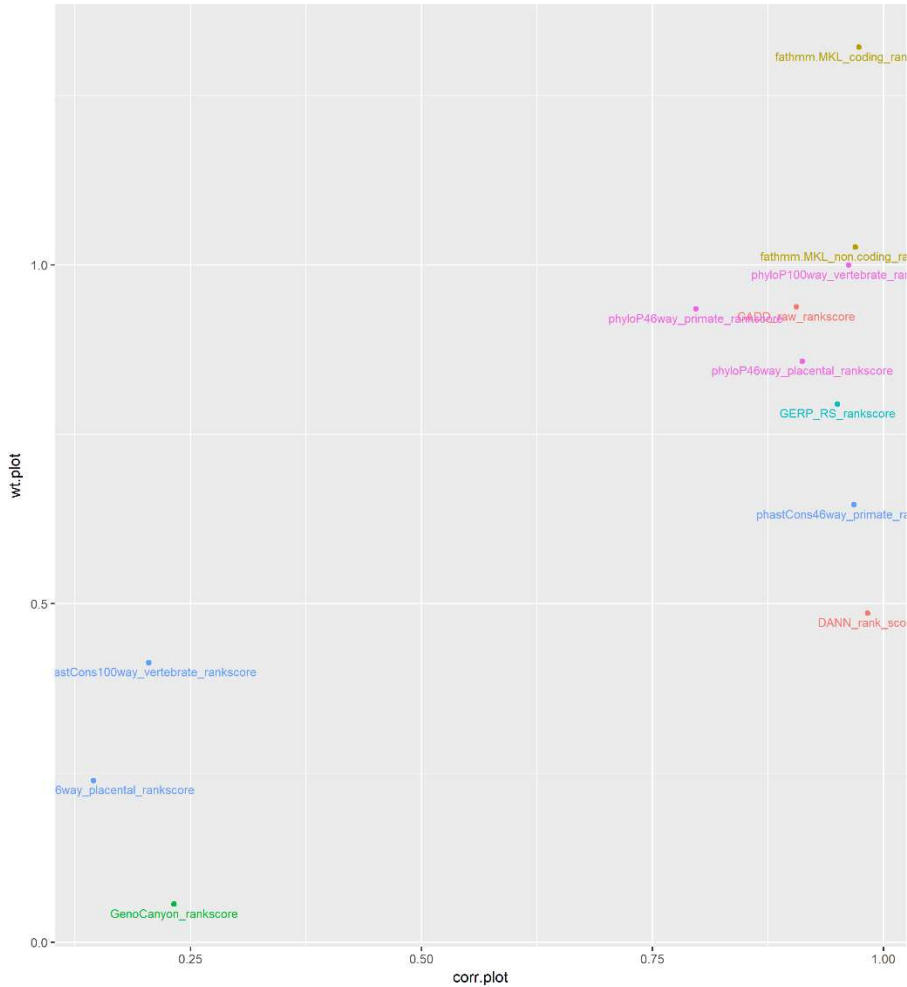


Figure 4.6: Regression slopes (vertical axis) and Pearson’s correlations coefficients (horizontal axis) calculated by partitioning the heritability for BMIs

Heritability estimation was based on ARIC RVs.



Supporting Information

[Table 4.1: Regression slopes and Pearson's correlations coefficients calculated by partitioning the heritability into various categories](#)

Heritability was estimated for HDL in UK10K TwinsUK RVs partitioned by CADD scores. Normalized slopes set the slope of 5 categories as unit 1.

# of categories	Slope	Normalized slope	Pearson correlation
2	4.47E-08	1.25	1
3	3.27E-08	0.92	0.98
4	3.67E-08	1.03	0.96
5	3.56E-08	1	0.90

[Figure 4.7: Phenotype-specific rank of regression slopes estimated by partitioning ARIC RVs for all functional scores](#)

Higher rank indicates stronger global correlations for the corresponding functional annotation and trait. Shaded colors highlight the groupings of traits.

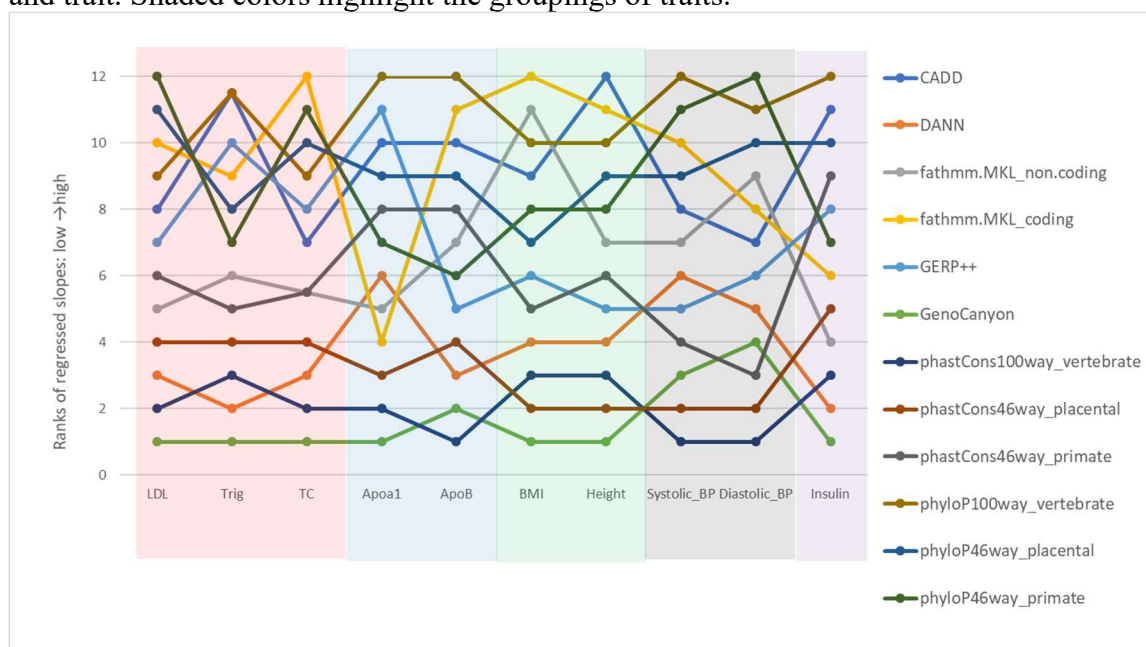
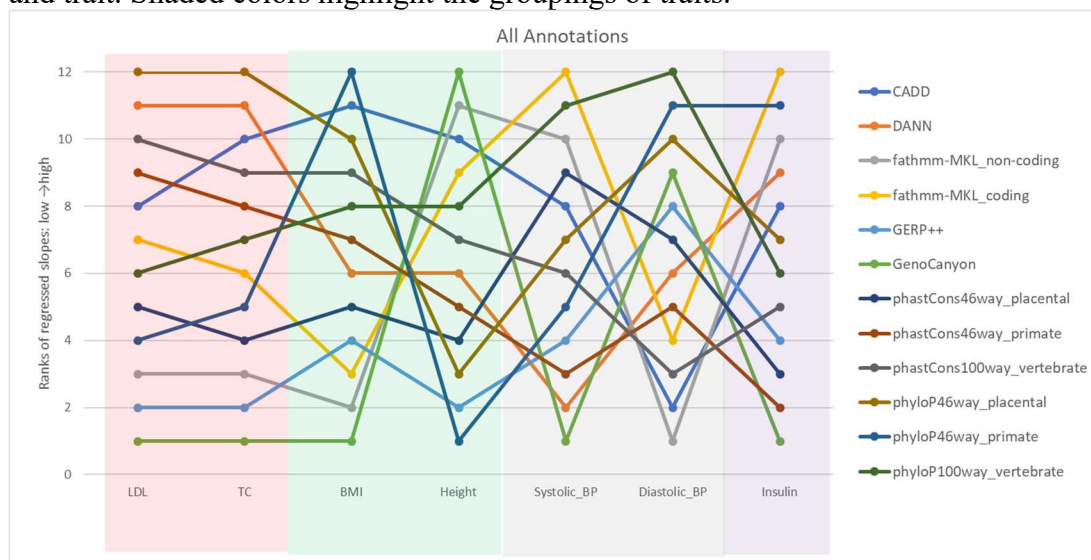


Figure 4.8: Phenotype-specific rank of regression slopes estimated by partitioning UK10K

TwinsUK RVs for all functional scores

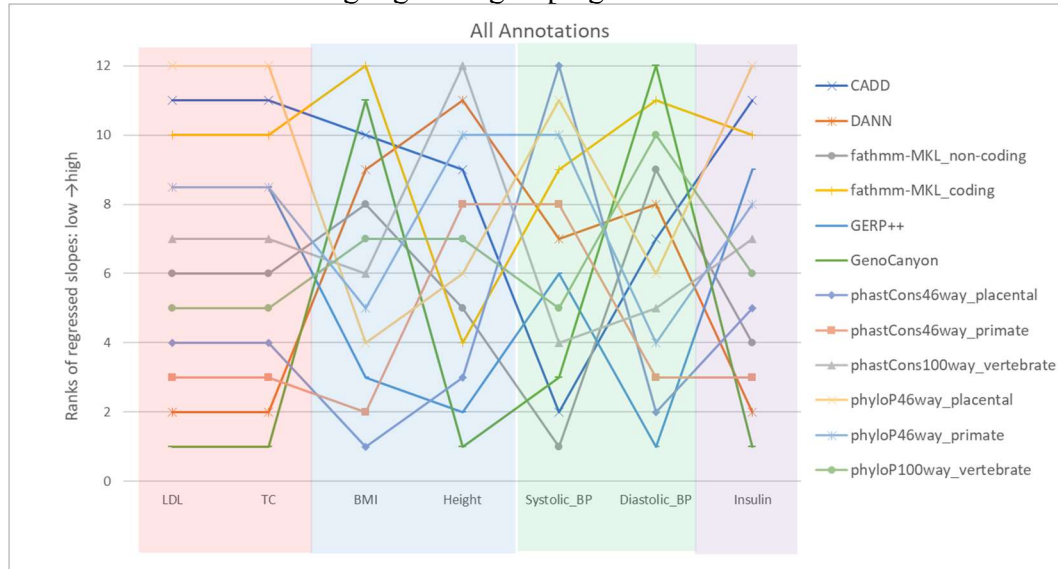
Higher rank indicates stronger global correlations for the corresponding functional annotation and trait. Shaded colors highlight the groupings of traits.



[Figure 4.9: Phenotype-specific rank of regression slopes estimated by partitioning UK10K](#)

[ALSPAC RVs for all functional scores](#)

Higher rank indicates stronger global correlations for the corresponding functional annotation and trait. Shaded colors highlight the groupings of traits.



Reference

1. Richards S, Aziz N, Bale S, Bick D, Das S, et al. (2015) Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med* 17: 405-424.
2. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, et al. (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* 15: 1034-1050.
3. Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, et al. (2010) Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput Biol* 6: e1001025.
4. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A (2010) Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res* 20: 110-121.
5. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, et al. (2014) A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* 46: 310-315.
6. Fu Y, Liu Z, Lou S, Bedford J, Mu XJ, et al. (2014) FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. *Genome Biol* 15: 480.
7. Ionita-Laza I, McCallum K, Xu B, Buxbaum JD (2016) A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nat Genet* 48: 214-220.
8. Gulko B, Hubisz MJ, Gronau I, Siepel A (2015) A method for calculating probabilities of fitness consequences for point mutations across the human genome. *Nat Genet* 47: 276-283.
9. Shihab HA, Rogers MF, Gough J, Mort M, Cooper DN, et al. (2015) An integrative approach to predicting the functional effects of non-coding and coding sequence variation. *Bioinformatics* 31: 1536-1543.
10. Liu X, Li C, Boerwinkle E (2017) The performance of deleteriousness prediction scores for rare non-protein-changing single nucleotide variants in human genes. *J Med Genet* 54: 134-144.
11. Stenson PD, Mort M, Ball EV, Shaw K, Phillips A, et al. (2014) The Human Gene Mutation Database: building a comprehensive mutation repository for clinical and molecular genetics, diagnostic testing and personalized genomic medicine. *Hum Genet* 133: 1-9.
12. Gusev A, Lee SH, Trynka G, Finucane H, Vilhjalmsen BJ, et al. (2014) Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *Am J Hum Genet* 95: 535-552.
13. Yang J, Lee SH, Goddard ME, Visscher PM (2011) GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* 88: 76-82.
14. (1989) The Atherosclerosis Risk in Communities (ARIC) Study: design and objectives. The ARIC investigators. *Am J Epidemiol* 129: 687-702.

15. Liu X, White S, Peng B, Johnson AD, Brody JA, et al. (2016) WGSa: an annotation pipeline for human genome sequencing studies. *J Med Genet* 53: 111-112.
16. Quang D, Chen Y, Xie X (2015) DANN: a deep learning approach for annotating the pathogenicity of genetic variants. *Bioinformatics* 31: 761-763.
17. Lu Q, Hu Y, Sun J, Cheng Y, Cheung KH, et al. (2015) A statistical framework to predict functional non-coding regions in the human genome through integrated analysis of annotation data. *Sci Rep* 5: 10576.
18. Lu Q, Powles RL, Wang Q, He BJ, Zhao H (2016) Integrative Tissue-Specific Functional Annotations in the Human Genome Provide Novel Insights on Many Complex Traits and Improve Signal Prioritization in Genome Wide Association Studies. *PLoS Genet* 12: e1005947.
19. Lu Q, Powles RL, Abdallah S, Ou D, Wang Q, et al. (2017) Systematic tissue-specific functional annotation of the human genome highlights immune-related DNA elements for late-onset Alzheimer's disease. *PLoS Genet* 13: e1006933.
20. Visscher PM, Hemani G, Vinkhuyzen AA, Chen GB, Lee SH, et al. (2014) Statistical power to detect genetic (co)variance of complex traits using SNP data in unrelated samples. *PLoS Genet* 10: e1004269.
21. Vinkhuyzen AA, Wray NR, Yang J, Goddard ME, Visscher PM (2013) Estimation and partition of heritability in human populations using whole-genome analysis methods. *Annu Rev Genet* 47: 75-95.
22. Yang J, Bakshi A, Zhu Z, Hemani G, Vinkhuyzen AA, et al. (2015) Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat Genet* 47: 1114-1120.

CHAPTER 5: CONCLUSION AND FUTURE WORK

Genetics association analysis of rare variants is still a challenge nowadays. In the first part of this dissertation, we proposed a new method to boost the statistical power by integrating multiple functional annotations. In the real data analysis, we identified some new genome-wide significant loci which were missed by existing rare variant association tests that either ignore external biological information or rely on a single source of biological knowledge, and these significant results can be replicated in a replication cohort. Due to the rapid growth of published functional annotations and WGS datasets, we expected widely applications of our proposed methods, and our proposed method can be simply expanded to case-control data. We expect that further implementation of the core functions in the C language should reduce the computational burden to a more affordable level. In addition, it would be desirable to develop some asymptotic theory and test to save the computational time. Of note, some asymptotic theory has been developed for the aSPU test in the context of testing two high-dimensional means for common variants [1]; however, extension to rare variants and the FunSPU test proposed here is not trivial and warrants future research.

In fact, it is not easy for researchers to access individual-level genetic data. In the second part, we proposed FunSPUs, an association test incorporating multiple functional annotations for GWAS summary statistics. FunSPUs is more powerful than the corresponding aSPUs test so that we can identify some family-wised significant loci at a relative smaller sample size. However, the inflation of type I error is still a concern when we applied FunSPUs to genome-wide scan. In particular, we found notable inflation at low

significant level (e.g. $\alpha = 0.005$). Our current research is still in preliminary stage and need to be compared with existing association tests for summary statistics. Since FunSPUs tests may yield a large amount of false positive results, our main challenge in subsequent work is how to identify true positive associations.

In the third part, we developed a phenotype-specific approach to evaluate the performances of functional annotation scores. The general idea is based on the partition of genome-wide SNVs by annotation scores, and estimate the heritability of each category to obtain a correlation measure between the annotation and the corresponding phenotype. Our proposed method focuses on RVs due to most non-coding SNVs are RVs. Since the number of reported deleterious non-coding variants is limited, it is not adequate to use these reported SNVs as golden standard to evaluate functional annotations. Therefore, data driven methods as our proposed one are the only feasible methods. For the rank scores that adhere to the uniform distribution, we can try to partition the scores into more categories. This work will be easier to implement if we compose an effective pipeline for our proposed method. It is also potential to develop a new functional score based on our estimation results. For example, we can compute a weighted sum of multiple functional scores which weights are corresponding regressed slopes. We also look forward to validating our results if there are more reported deleterious non-coding SNVs as “gold standard”.

We have implemented the proposed FunSPU test and its extensions in an R package “FunSPU”, available at <https://github.com/sputnik1985/FunSPU>, and to be posted to R/CRAN.

REFERENCES

1. Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A* 74: 5463-5467.
2. Smith M, Brown NL, Air GM, Barrell BG, Coulson AR, et al. (1977) DNA sequence at the C termini of the overlapping genes A and B in bacteriophage phi X174. *Nature* 265: 702-705.
3. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, et al. (2001) Initial sequencing and analysis of the human genome. *Nature* 409: 860-921.
4. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, et al. (2001) The sequence of the human genome. *Science* 291: 1304-1351.
5. Metzker ML (2010) Sequencing technologies - the next generation. *Nat Rev Genet* 11: 31-46.
6. Pareek CS, Smoczynski R, Tretyn A (2011) Sequencing technologies and genome sequencing. *J Appl Genet* 52: 413-435.
7. Mardis ER (2008) Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet* 9: 387-402.
8. Tg, Hdl Working Group of the Exome Sequencing Project NHL, Blood I, Crosby J, Peloso GM, et al. (2014) Loss-of-function mutations in APOC3, triglycerides, and coronary disease. *N Engl J Med* 371: 22-31.
9. The UK10K Project Consortium, Walter K, Min JL, Huang J, Crooks L, et al. (2015) The UK10K project identifies rare variants in health and disease. *Nature* 526: 82-90.
10. Brody JA, Morrison AC, Bis JC, O'Connell JR, Brown MR, et al. (2017) Analysis commons, a team approach to discovery in a big-data environment for genetic epidemiology. *Nat Genet* 49: 1560-1563.
11. Stein L (2001) Genome annotation: from sequence to biology. *Nat Rev Genet* 2: 493-503.
12. Brent MR (2005) Genome annotation past, present, and future: how to define an ORF at each locus. *Genome Res* 15: 1777-1786.
13. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, et al. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences* 106: 9362-9367.
14. Kleinjan DA, Lettice LA (2008) Long - range gene control and genetic disease. *Advances in genetics* 61: 339-388.
15. Kleinjan DA, van Heyningen V (2005) Long-range control of gene expression: emerging mechanisms and disruption in disease. *The American Journal of Human Genetics* 76: 8-32.
16. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, et al. (2012) Systematic localization of common disease-associated variation in regulatory DNA. *Science* 337: 1190-1195.
17. Schaub MA, Boyle AP, Kundaje A, Batzoglou S, Snyder M (2012) Linking disease associations with regulatory information in the human genome. *Genome research* 22: 1748-1759.

18. Ward LD, Kellis M (2011) HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic acids research* 40: D930-D934.
19. Aravin AA, Hannon GJ, Brennecke J (2007) The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race. *science* 318: 761-764.
20. Bartel DP (2009) MicroRNAs: target recognition and regulatory functions. *cell* 136: 215-233.
21. Olovnikov I, Aravin AA, Toth KF (2012) Small RNA in the nucleus: the RNA-chromatin ping-pong. *Current opinion in genetics & development* 22: 164-171.
22. Rinn JL, Chang HY (2012) Genome regulation by long noncoding RNAs. *Annual review of biochemistry* 81: 145-166.
23. Kellis M, Wold B, Snyder MP, Bernstein BE, Kundaje A, et al. (2014) Defining functional DNA elements in the human genome. *Proc Natl Acad Sci U S A* 111: 6131-6138.
24. Craig J (2008) Complex diseases: Research and applications. *Nature Education* 1: 184.
25. Xu GJ, Lin LF, Wei P, Pan W (2016) An adaptive two-sample test for high-dimensional means. *Biometrika* 103: 609-624.