Genetics and population analysis

Powerful and efficient SNP-set association tests across multiple phenotypes using GWAS summary data

Bin Guo and Baolin Wu*

Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, MN 55455, USA

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on January 3, 2018; revised on August 29, 2018; editorial decision on September 16, 2018; accepted on September 18, 2018

Abstract

Motivation: Many GWAS conducted in the past decade have identified tens of thousands of disease related variants, which in total explained only part of the heritability for most traits. There remain many more genetics variants with small effect sizes to be discovered. This has motivated the development of sequencing studies with larger sample sizes and increased resolution of genotyped variants, e.g., the ongoing NHLBI Trans-Omics for Precision Medicine (TOPMed) whole genome sequencing project. An alternative approach is the development of novel and more powerful statistical methods. The current dominating approach in the field of GWAS analysis is the “single trait single variant” association test, despite the fact that most GWAS are conducted in deeply-phenotyped cohorts with many correlated traits measured. In this paper, we aim to develop rigorous methods that integrate multiple correlated traits and multiple variants to improve the power to detect novel variants. In recognition of the difficulty of accessing raw genotype and phenotype data due to privacy and logistic concerns, we develop methods that are applicable to publicly available GWAS summary data.

Results: We build rigorous statistical models for GWAS summary statistics to motivate novel multi-trait SNP-set association tests, including variance component test, burden test and their adaptive test, and develop efficient numerical algorithms to quickly compute their analytical *P*-values. We implement the proposed methods in an open source R package. We conduct thorough simulation studies to verify the proposed methods rigorously control type I errors at the genome-wide significance level, and further demonstrate their utility via comprehensive analysis of GWAS summary data for multiple lipids traits and glycemic traits. We identified many novel loci that were not detected by the individual trait based GWAS analysis.

Availability and implementation: We have implemented the proposed methods in an R package freely available at <http://www.github.com/baolinwu/MSKAT>.

Contact: baolin@umn.edu

Supplementary information: [Supplementary data](#) are available at *Bioinformatics* online.

1 Introduction

The genome-wide association studies (GWAS) have been extremely successful identifying tens of thousands of variants associated with various human diseases and traits in the past decade (Visscher *et al.*, 2017), and providing valuable insights into their underlying genetic

architecture. However, significant gaps remain: most identified variants have small effects and in total they explained only a small part of the overall heritability for most traits (Manolio *et al.*, 2009). This has motivated the formation of various mega-scale consortia pooling many GWAS and the ongoing large-scale whole-genome sequencing

project to increase the resolution of genotyped variants. All these tremendous efforts are expected to identify more variants and further improve our understanding of the genetic architecture. Alternatively a cost-effective approach is to develop novel and powerful statistical methods that can leverage the existing GWAS data to identify novel variants. Many methods have exploited the polygenic nature of most traits and leveraged the existing GWAS with richly measured phenotypes to develop joint test of multiple traits and multiple variants. They have convincingly demonstrated their utility to identify more genetic variants (see, e.g. [Broadaway et al., 2016](#); [Ferreira and Purcell, 2009](#); [He et al., 2013](#); [Liu et al., 2010](#); [Maity et al., 2012](#); [O'Reilly et al., 2012](#); [Seoane et al., 2014](#); [Tang and Ferreira 2012](#); [van der Sluis et al., 2013](#); [Wu and Pankow 2015, 2016](#); [Wu et al., 2010](#); [Yang et al., 2010](#)). A common feature of these methods is that they need raw genotype and phenotype data. However, due to various policy and privacy concerns, it is generally hard to access raw GWAS data. It has now become a common practice to share GWAS summary association statistics, which include, e.g., the SNP id, MAF, effect allele, regression coefficient and associated standard error, P -value, and sample size. Hence, it is desirable to develop novel association test methods that need only GWAS summary data ([Pasaniuc and Price, 2017](#)).

We have seen active development of such statistical methods recently. For example, [Bakshi et al. \(2016\)](#) proposed a method to test SNP-set association with a single trait using GWAS summary data. [Stephens \(2013\)](#) and [Zhu et al. \(2015\)](#) proposed novel methods for multi-trait association test with a single variant. Their insightful observation is that the across trait test Z -statistics have the same correlation as the traits. This has motivated using the empirical correlation matrix of Z -statistics to estimate trait correlation matrix widely adopted in most existing methods ([Cichonska et al., 2016](#); [Kwak and Pan, 2016](#)). However, this approach implicitly assumes that majority of variants are null and hence has limitations for most traits of polygenic nature, since many variants will have effects, although small, but in aggregation they have large overall effects (as reflected in their large genetic heritability), which will convolute the empirical correlation matrix (with genetic correlation) and lead to biased estimates (see next section for more discussions). Furthermore, due to the complicated dependence across traits and variants, it is very challenging to analytically account for their dependence and efficiently and accurately calculate significance P -values for most existing joint test methods ([Cichonska et al., 2016](#); [Kwak and Pan, 2016](#); [Van der Sluis et al., 2014](#)). As a result, they have relied on ad hoc approach or crude Monte Carlo approximation that is computationally intensive and not scalable to genome-wide association test.

In this paper we build rigorous statistical models and develop sound and powerful multi-trait SNP-set association test methods using GWAS summary data. We made several contributions compared to the existing methods: (i) we adopt the recently developed LD score regression approach to more accurately estimate the trait correlation matrix using the summary statistics; (ii) we develop a novel variance components test approach to motivate our proposed methods based on a rigorous multivariate normal regression model derived for the summary statistics; (iii) for the proposed tests, we derive analytical formula and develop efficient numerical algorithms to calculate their P -values accurately and efficiently without the need of resampling or permutation; (iv) we rigorously verify that our proposed methods properly control the type I errors at the genome-wide significance level through very large-scale simulation studies; and further demonstrate their utility through comprehensive analysis of GWAS summary data for multiple lipids traits and glycemic traits.

The structure of this paper is as follows. We discuss the proposed multi-trait SNP-set association test methods using GWAS summary data in Section 2. In Section 3, we investigate the performance of proposed methods through simulation studies. We conduct comprehensive analysis of GWAS summary data for multiple lipids traits and glycemic traits in Section 4. We identified many novel loci that were not detected by the current individual trait based tests. We end the paper with a discussion in Section 5. All technical details are provided in the [Supplementary Materials](#).

2 Methods

Assume we have the GWAS summary data for K quantitative traits measured on a cohort of unrelated individuals. Denote the marginal trait correlation matrix as $\Sigma = (\sigma_{kj})$. For simplicity of notation, consider the association test Z -statistics (standardized regression coefficient estimates) across all traits for a SNP, denoted as $(z_1, \dots, z_K)$. [Bulik-Sullivan et al. \(2015\)](#) showed that $\mathbb{E}(z_k z_j) = \sigma_{kj} + g_{kj} \ell$ under a polygenic model, where ℓ is the SNP's average LD score (computed as the average of its LD r^2 with all other SNPs), and g_{kj} quantifies the trait genetic correlation. We can thus obtain unbiased estimate of σ_{kj} by regressing $z_k z_j$ on the LD scores, which are pre-computed using the 1K Genomes samples (1000 Genomes Project Consortium, 2012). In contrast, the empirical correlation matrix of summary Z -statistics will produce biased estimates of Σ for polygenic traits due to the variant LD and trait genetic correlation. For simplicity of notation, we still use Σ to denote the estimated trait correlation matrix throughout the discussion.

In the following, we develop methods to test the joint association of a given set of SNPs (in a gene region or pathway) with all K traits simultaneously. Denote the association test Z -statistic as z_{kj} for the k -th trait and j -th SNP, $k = 1, \dots, K$; $j = 1, \dots, m$. Let $Z = (z_{kj})$. Our null hypothesis is that Z have zero means.

2.1 Multi-trait SNP-set association test using GWAS summary data

Assume that the Z -statistics are derived based on genotypes coding the copies of minor alleles (i.e., the minor allele is the effect allele; if not, we switch the sign of Z -statistics). We can derive a multivariate normal regression model for the summary statistics Z and show that asymptotically $\text{Var}(\text{vec}(Z)) = R \otimes \Sigma$, where R is the $m \times m$ LD correlation matrix, $\text{vec}()$ is the vector operator stacking the columns of a matrix into a vector, and $\otimes$ denotes the Kronecker product.

Let $Z_j = (z_{1j}, \dots, z_{Kj})^T$ denote the K summary Z -statistics for the j -th SNP. For a single variant, the chi-square statistic $Z_j^T \Sigma^{-1} Z_j$ can be used to test for the effect of variant j across all traits. To summarize the overall SNP-set effects, we consider $Q = \sum_{j=1}^m Z_j^T \Sigma^{-1} Z_j$. Q generally has robust performance with mixed variant effects. When all variants have similar effects, we can gain more power by testing the collapsed genotype scores. We can readily check that a burden type test using GWAS summary data is $B = (\sum_{j=1}^m Z_j)^T \Sigma^{-1} (\sum_{j=1}^m Z_j)$. We rigorously derive Q and B via a variance components test approach under the derived multivariate normal regression model for the summary statistics. We further show that under the null of no SNP-set effects, Q is distributed as the weighted sums of independent χ_K^2 random variables, and B is a scaled χ_K^2 random variable.

2.2 Adaptive multi-trait SNP-set association test

To combine the strength of Q and B , we consider their weighted average, $Q_\rho = (1 - \rho)Q + \rho B$, and use the minimum p -value,

$T = \min_{\rho \in [0,1]} Pval(Q_\rho)$, over a finite grid of ρ 's as an adaptive test statistic. We note that Q_ρ can also be derived as a variance components test. It is very challenging to derive the null distribution of T to compute its P -value. A working solution is the crude Monte Carlo approximation, which is however too computationally intensive. We develop efficient algorithms to calculate the p -value of T without the need of resampling or permutation (See [Supplementary Materials](#) for technical details).

Next we investigate the performance of our proposed methods through numerical studies and applications. We will refer to our proposed methods as MSATS (for Multi-trait SNP-set Association Tests using Summary data) and denote B as MBT, Q as MQT, and T as MAT in the following discussion. When comparing the proposed MSATS to alternative methods, we mainly focus on computationally efficient statistical methods and do not include those computationally intensive approaches. We also view our proposed methods as a complementary approach to single trait single variant association tests. The MSATS tests are direct generalizations of the corresponding single trait GWAS summary data based SNP-set test methods ($K=1$) as studied in [Guo and Wu \(2018\)](#). We denote the single trait burden test, sum of squares test and their adaptive test, as SBT, SQT and SAT respectively in the following discussion. We use their Bonferroni corrected p -values based on the number of traits.

3 Results

3.1 Simulations

We conducted simulation studies to evaluate the performance of the proposed methods. We consider three traits and set the trait correlation matrix: $\Sigma_{12} = -0.1081$, $\Sigma_{13} = -0.3579$ and $\Sigma_{23} = 0.2313$, which are the outcome correlation matrix based on three lipids traits GWAS with around 100,000 European individuals ([Teslovich et al., 2010](#); see next section). We set R as the LD correlation matrix of 22 SNPs in the gene KLHL8.

We first evaluate the type I errors of the proposed MSATS tests. We simulate 10^8 null summary statistics from $\mathcal{N}(0, R \otimes \Sigma)$ to estimate the type I errors at significance levels $\alpha = 10^{-4}$, 10^{-5} , and 2.5×10^{-6} . [Table 1](#) summarizes the results. Overall we can see that all proposed tests have well-controlled type I errors.

To evaluate the power of proposed methods, we simulate 10^5 summary statistics from $\mathcal{N}(A \otimes \Delta, R \otimes \Sigma)$, where A is a vector of length 22 and Δ is a vector of length 3. We assume 4 SNPs with association signals by randomly setting 4 A_i as 1 or -1, and setting the rest of A as zero. [Table 2](#) shows the estimated power at 2.5×10^{-6} significance level under four combinations of A for four settings of Δ : two fixed values of Δ and randomly simulated Δ uniformly from $U(-3, 3)$ and normally from $\mathcal{N}(0, 1)$. Overall the proposed MSATS tests perform better than the single trait based tests. Among the three multi-trait tests, when we have the same effects across all 4 signal

Table 1. Ratio of empirical type I errors divided by the significance level α estimated over 10^8 simulations

		10^{-4}	10^{-5}	2.5×10^{-6}
MSATS	MQT	0.98 (0.01)	0.99 (0.03)	1.00 (0.06)
	MBT	0.99 (0.01)	1.04 (0.03)	1.05 (0.06)
	MAT	0.99 (0.01)	0.97 (0.03)	0.93 (0.06)

Note: Listed within parentheses are the standard errors. MAT, MQT and MBT are the proposed multi-trait adaptive test, quadratic test and burden test, respectively.

SNPs (the first and last combinations), the MBT test performs the best while MQT has the worst performance. When we have balanced mixing of positive and negative SNP effects (the third combination), not surprisingly, MBT has the worst performance since directly summing the summary statistics cancels out the overall signals, while MQT performs the best. Very interestingly, with unbalanced mixing of positive and negative SNP effects (the second combination), the adaptive test MAT truly combines the strength of both MQT and MBT and has the best performance with much improved power compared to MQT and MBT. The adaptive test MAT performs robustly across all scenarios and has the overall best performance. We have done more power simulations by varying number of signal SNPs and simulating various values of Δ , and obtained the same conclusions. The complete results are available at the [Supplementary Materials](#).

3.2 Application to GWAS summary data for lipids traits

We conduct a comprehensive analysis of the GWAS summary data for high-density lipoprotein cholesterol (HDL), low-density lipoprotein cholesterol (LDL), and Triglycerides (TG) conducted by the Global Lipids Consortium with around 100,000 European individuals ([Teslovich et al., 2010](#)). The GWAS summary data are downloaded from <http://csg.sph.umich.edu/abecasis/public/lipids2010>. With the LD score regression approach, the estimated correlation matrix are: $\Sigma_{12} = -0.1081$, $\Sigma_{13} = -0.3579$, $\Sigma_{23} = 0.2313$.

We first remove those SNPs with $MAF < 0.05$ and then group all SNPs within 20 kb of a gene region into a set. For each SNP-set, we further perform LD pruning so that all pairwise LD $r^2 \leq 0.8$. As we view the proposed MSATS tests as a complementary approach to the traditional single variant single trait association test, and for illustrating the relative performance of different methods, we further remove those genome-wide significant SNPs ($P \leq 5 \times 10^{-8}$ for any trait). In total we obtain 18,754 SNP-sets with at least two SNPs for downstream analysis. [Willer et al. \(2013\)](#) conducted a followup study of those promising SNPs identified by [Teslovich et al. \(2010\)](#) based on around 190,000 European participants. We use their summary results as partial validation in our analysis. Here we mainly

Table 2. Estimated test power (%) under 2.5×10^{-6} significance level

Δ	nonzero A_i 's	MAT	MQT	MBT	SAT	SQT	SBT
(3, 2.5, -1.5)	(1, 1, 1, 1)	36.5	0.3	37.8	5.7	0.0	8.5
	(1, 1, 1, -1)	59.6	41.0	7.7	5.0	1.7	1.3
	(1, 1, -1, -1)	27.0	40.8	0.0	0.8	1.8	0.0
	(-1, -1, -1, -1)	36.1	0.3	37.3	5.4	0.0	8.2
(4, 2, 2)	(1, 1, 1, 1)	49.1	0.5	48.6	6.9	0.0	9.6
	(1, 1, 1, -1)	51.8	33.3	6.4	3.3	1.1	0.8
	(1, 1, -1, -1)	21.2	33.2	0.0	0.4	1.1	0.0
	(-1, -1, -1, -1)	49.0	0.5	48.6	6.7	0.0	9.3
$\mathcal{N}(0, 1)$	(1, 1, 1, 1)	35.6	8.6	35.4	22.2	3.6	22.7
	(1, 1, 1, -1)	34.7	30.3	11.8	22.2	18.7	5.5
	(1, 1, -1, -1)	27.1	30.3	0.0	15.4	18.7	0.0
	(-1, -1, -1, -1)	35.8	8.5	35.5	22.2	3.6	22.9
$U(-3, 3)$	(1, 1, 1, 1)	44.8	6.5	44.3	17.3	0.1	19.7
	(1, 1, 1, -1)	49.1	41.7	13.2	18.2	10.6	2.6
	(1, 1, -1, -1)	36.1	41.9	0.0	5.6	10.7	0.0
	(-1, -1, -1, -1)	44.8	6.4	44.3	17.1	0.1	19.5

Note: Data are simulated from $\mathcal{N}(A \otimes \Delta, R \otimes \Sigma)$. A has 4 nonzero A_i 's with different signs. MAT, MQT and MBT are the proposed multi-trait adaptive test, quadratic test, and burden test; and SAT, SQT and SBT are the corresponding single-trait based test methods.

Fig. 1. Venn diagram of number of significant genes identified by proposed MSATS tests for the three lipids traits

Table 3. Significant genes identified only by MAT, missed by MQT and MBT

Gene	MQT	MBT	MAT	minP-2010	minP-2013
ANUBL1	1.57e-05	3.71e-06	1.72e-06	5.36e-07	2.28e-06
CCDC18	4.70e-06	4.36e-06	1.57e-06	4.10e-07	3.56e-10
CTSA	3.10e-06	6.68e-05	9.62e-07	3.37e-21	3.98e-36
LCAT	3.82e-06	6.45e-06	2.40e-06	1.22e-31	3.30e-52
SPI1	1.68e-05	1.36e-05	1.07e-06	3.30e-15	1.40e-31

Note: We listed the MSATS test *P*-values and the minimum *P*-value across all SNPs in the gene and all traits from [Teslovich et al. \(2010\)](#) study (denoted as minP-2010) and [Willer et al. \(2013\)](#) study (denoted as minP-2013).

focus on the results for the proposed MSATS tests. The single trait based test methods have identified much less significant genes. We list their detailed results in the [Supplementary Materials](#).

At the Bonferroni corrected SNP-set significance level 2.67×10^{-6} , the proposed MSATS tests identified a total of 237 significant genes. [Figure 1](#) shows the Venn diagram comparing the number of significant genes identified by the proposed tests. The MBT identified 101 significant genes, MQT identified 191 significant genes, and MAT identified 218 significant genes. The adaptive test MAT captured the majority of significant SNP-sets identified by MQT and MBT, and it further identified five additional significant genes which were not identified by MQT and MBT. [Table 3](#) shows the MSATS test *p*-values for these five genes and the minimum *p*-value across all SNPs in the gene and all traits from [Teslovich et al. \(2010\)](#) and [Willer et al. \(2013\)](#). Among them four genes harbor genome-wide significant SNPs in the meta analysis of [Willer et al. \(2013\)](#). The novel gene ANUBL1 has been implicated in lipids metabolism ([Demetz et al., 2014](#); [Burkhardt et al., 2015](#)).

Among the 237 significant genes, 209 genes harbored genome-wide significant SNPs from the meta-analysis results of [Willer et al. \(2013\)](#); and 45 of them were “novel” genes in the sense that the meta-analysis of [Teslovich et al. \(2010\)](#) did not identify any significant SNPs in these 45 genes. [Table 4](#) lists 10 representative genes identified by the proposed tests. The [Supplementary Materials](#) listed the detailed information for all identified genes.

Table 4. Some “novel” significant genes

Gene	MQT	MBT	MAT	minP-2010	minP-2013
NIPSNAP3B	1.89e-07	5.66e-03	4.84e-07	7.71e-07	1.63e-18
RAB11B	9.18e-08	1.10e-06	7.75e-08	6.63e-08	6.66e-18
ZDHHHC18	3.77e-08	2.77e-09	3.51e-09	6.11e-07	9.74e-16
NUDC	2.34e-07	1.15e-06	1.85e-07	1.40e-07	2.00e-15
OBP2B	4.76e-09	6.08e-06	6.75e-09	1.15e-07	4.32e-15
MIR148A	1.59e-10	2.01e-08	6.03e-11	7.93e-07	3.95e-14
GPR61	2.92e-07	2.12e-04	4.97e-07	6.35e-08	5.60e-14
GP2N	1.62e-06	2.22e-04	2.20e-06	1.40e-07	1.44e-13
NR0B2	5.38e-08	5.26e-07	3.31e-08	1.40e-07	1.44e-13
LRRC36	1.17e-07	5.36e-02	1.66e-07	1.67e-07	1.05e-12

We listed MSATS test *P*-values and the minimum *P*-value across all SNPs in the gene and all traits from [Teslovich et al. \(2010\)](#) study (denoted as minP-2010) and [Willer et al. \(2013\)](#) study (denoted as minP-2013).

Table 5. MSATS test *P*-values for 27 novel genes

Gene	MQT	MBT	MAT	minP-2010	minP-2013
ACADS	3.44e-02	1.48e-06	5.12e-06	3.87e-04	1.29e-03
ACHE	4.13e-02	1.37e-08	2.86e-08	9.53e-05	4.30e-05
ANUBL1	1.57e-05	3.71e-06	1.72e-06	5.36e-07	2.28e-06
AURKB	1.12e-06	1.77e-06	1.74e-06	3.01e-07	2.46e-07
BACE1	3.81e-07	8.44e-02	4.65e-07	1.59e-05	5.69e-08
CBFB	1.52e-06	8.28e-01	4.59e-06	2.70e-07	5.66e-08
CD320	8.43e-02	1.44e-07	4.89e-07	1.79e-03	1.19e-04
CENPA	7.82e-03	3.51e-07	5.40e-07	1.01e-03	1.02e-05
CSH2	1.92e-01	2.36e-06	4.35e-06	3.43e-03	1.45e-02
ETNK1	1.20e-06	1.15e-01	1.88e-06	4.93e-07	1.42e-06
GH1	8.52e-01	8.13e-14	1.55e-13	5.35e-03	4.17e-03
KIAA1949	2.05e-07	2.53e-02	5.90e-07	1.65e-06	3.41e-06
MAFB	2.28e-06	1.61e-01	5.35e-06	1.78e-05	4.65e-07
NICN1	1.18e-01	2.85e-07	3.90e-07	7.46e-03	9.85e-03
NPEPPS	6.87e-09	4.39e-01	1.28e-08	1.15e-07	1.36e-06
NSUN5	9.29e-06	7.82e-09	2.83e-08	9.58e-05	5.95e-07
OR8J1	1.36e-03	2.52e-06	3.36e-06	8.21e-07	1.73e-06
RABEP2	2.14e-02	1.04e-37	1.75e-37	4.39e-03	7.40e-05
RNF39	1.85e-06	8.31e-02	3.79e-06	5.62e-05	1.05e-06
SFN	2.48e-08	1.31e-07	1.25e-08	1.66e-06	2.30e-06
SFRS16	2.02e-11	2.01e-02	5.85e-11	2.74e-07	8.91e-08
SNORD19	2.78e-03	2.58e-09	6.19e-09	6.74e-04	9.75e-05
SPCS1	2.86e-03	1.70e-09	4.13e-09	7.26e-04	9.75e-05
TPPP3	1.70e-07	4.79e-02	2.40e-07	5.84e-07	3.90e-07
TRIM50	9.39e-06	1.58e-08	2.02e-08	9.58e-05	8.12e-08
TRIP6	3.86e-02	1.42e-06	2.88e-06	2.97e-05	4.30e-05
ZDHHHC1	7.54e-07	5.42e-03	1.22e-06	5.84e-07	3.90e-07

We listed MSATS test *P*-values and the minimum *P*-value across all SNPs in the gene and all traits from [Teslovich et al. \(2010\)](#) study (denoted as minP-2010) and [Willer et al. \(2013\)](#) study (denoted as minP-2013).

In total we identified 27 novel genes: they did not harbor any significant SNPs in [Teslovich et al. \(2010\)](#) and [Willer et al. \(2013\)](#). [Table 5](#) summarizes their information. For each novel gene, we listed the MSATS test *P*-values together with the minimum *P*-value of all SNPs in the gene across all traits for both studies.

Some identified novel genes have been reported to be significantly associated with lipids traits or harbor some genome-wide significant SNPs in some recent GWAS. For example, MAFB is close to a significant SNP identified in a LDL GWAS ([Kathiresan et al., 2009](#)), and it has been shown to increase LDL cholesterol levels by negatively regulating the LDL receptor transcription activity ([Petersen](#)

et al., 2004). Through cis-eQTL analysis, Yao *et al.* (2015) showed that the SNP rs7206971 in NPEPPS is genome-wide significantly associated with LDL cholesterol. In a GWAS meta-analysis, Postmus *et al.* (2014) showed that the SNP rs445925 (0.1Mb upstream of SFRS16 on chromosome 19) is significantly associated with LDL response after statin treatment. The SNP rs2240466, which is close to NSUN5, was found genome-wide significantly associated with TG (Aulchenko *et al.*, 2009; Folkersen *et al.*, 2010). TPPP3 is a protein coding gene which was found to be a signal for HDL in the analysis of Charlesworth *et al.* (2009). Andreassen *et al.* (2015) showed that three of the identified novel genes (CBFB, ETNK1, KIAA1949) contained risk SNPs shared between the immune-mediated diseases and TG or HDL. In addition, several identified novel genes have been reported to be involved in some diseases or traits which are strongly related to the lipids traits. For example, coronary artery disease (CAD) is a common disease showing strong relationship with lipids traits (Nair *et al.*, 2009; Wilson, 1990). Weng *et al.* (2016) found that RNF39 might cause those genes involved in cardio-dysfunction to contribute to risk of nonobstructive CAD in Caucasian women, while the SNP rs2301753 in RNF39 can reduce the likelihood of nonobstructive CAD. For NICN1, LeBlanc *et al.* (2015) showed that it might have some eQTL effect on CAD in the blood tissue. Waist-to-hip ratio (WHR) was shown to be significantly correlated with serum lipids concentrations (especially HDL concentration) in obese women (Komiya and Masuda, 1989). ZDHHC1 has been found to be significantly associated with WHR in women of both African and European ancestry (Ng *et al.*, 2017). Similarly, the SNP rs6784615 near SNORD19 was also associated with WHR (Heid *et al.*, 2010). ACADS is the plausible candidate gene underlying total-cholesterol quantitative trait loci on Chromosome 5 in male mice, and this finding provided insight into the genetic mechanisms of plasma lipid metabolism (Suto and Kojima, 2017). Hietaniemi *et al.* (2005) found that the GH1 was associated with LDL cholesterol. SPCS1 is a protein coding gene, and Suzuki *et al.* (2013) found that it participated in the assembly of Hepatitis C Virus (HCV) which was associated with cholesterol and lipoproteins (Felmlee *et al.*, 2013). For the RABEP2 gene, copy number variations of a genomic region including this gene have been associated with severe obesity which is a common risk factor of HDL (Bochukova *et al.*, 2010). Micale *et al.* (2008) showed that TRIM50 encoded an E3 ubiquitin ligase, and research has shown that E3 ubiquitin ligase IDOL induced the degradation of the low density lipoprotein receptor (Hong *et al.*, 2010) and helped control cholesterol (Brown and Hsieh, 2016). Another protein coding gene BACE1 was found to play an important role in linking lipids to Alzheimer's Disease (Di Paolo and Kim, 2011) and interact with scaffolding proteins of lipid rafts (Hattori *et al.*, 2006). All these evidences suggest that these identified novel genes are likely to be associated with lipids. Further research is needed on the possible roles of these identified SNPs/genes.

3.3 Application to GWAS summary data for glycemic traits

We also jointly analyze the GWAS meta-analysis summary data for the fasting glucose, fasting insulin, indices of β -cell function and insulin resistance conducted by the MAGIC consortium with 46, 186 non-diabetic European participants (Dupuis *et al.*, 2010). We follow the previous procedure to process the summary data. Figure 2 shows the Venn diagram comparing the number of significant genes identified by the proposed tests. The MSATS tests identified a total of 19 significant genes, and 11 of them are novel genes. They perform

Fig. 2. Venn diagram of number of significant genes identified by the proposed MSATS tests for the four glycemic traits

better than the single trait based tests. The adaptive test MAT is the most powerful. We provide detailed analysis results at the [Supplementary Materials](#).

4 Discussion

Various multi-trait testing methods have been developed along the line of Stephens (2013) and Zhu *et al.* (2015), though generally they require computing-intensive Monte Carlo or MCMC sampling (see, e.g., Shim *et al.*, 2015; Zhu and Stephens, 2017). Our proposed methods provide an alternative approach with extreme scalability and robust performance. In the same spirit as these existing approaches, the proposed methods need only the publicly available GWAS summary statistics without the need to access raw genotype and phenotype data. We expect that these newly developed GWAS summary data based joint testing methods can potentially identify more novel variants and shed more lights into the underlying disease mechanism.

We note that the Bayesian modeling is an alternative and very useful approach (see, e.g., Shim *et al.*, 2015; Zhu and Stephens, 2017). Compared to frequentist approach, they are often more flexible to accommodate various useful outside information (e.g., as a form of informative prior) and can easily account for the complicated dependence across traits and variants. For example, Zhu and Stephens (2017) recently developed a novel approach to simultaneously model the summary statistics across all variants. However, a potential drawback is that they often need intensive MCMC sampling to make inferences, which partially hinders their applicability, especially when we prototype different methods and want to make quick inferences on millions of variants or tens of thousands of SNP-sets. Therefore there is a pressing need to develop powerful, flexible, yet efficient statistical methods that are built on rigorous statistical models and can integrate various information. Our developed methods showcase that novel integration of multiple related traits and variants can identify more novel variants.

Throughout the discussions, we have mainly focused on multiple GWAS summary data from the same cohort. The proposed methods can be readily extended to GWAS with partially overlapped samples

(see [Supplementary Materials](#)). We have implemented the proposed methods in an R package publicly available online.

Acknowledgements

The authors are grateful to the University of Minnesota Supercomputing Institute for assistance with the computations. They would like to thank the associate editor and reviewers for their constructive comments, which have greatly improved the presentation of the paper. Data on glycemic traits have been contributed by MAGIC investigators and have been downloaded from www.magicinvestigators.org.

Funding

This work was supported in part by NIH grants GM083345 and CA134848.

References

- 1000 Genomes Project Consortium. (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**, 56–65.
- Andreassen, O.A. *et al.* (2015) Abundant genetic overlap between blood lipids and immune-mediated diseases indicates shared molecular genetic mechanisms. *PLoS One*, **10**, e0123057.
- Aulchenko, Y.S. *et al.* (2009) Loci influencing lipid levels and coronary heart disease risk in 16 European population cohorts. *Nat. Genet.*, **41**, 47–55.
- Bakshi, A. *et al.* (2016) Fast set-based association analysis using summary data from GWAS identifies novel gene loci for human complex traits. *Sci. Rep.*, **6**, 32894.
- Bochukova, E.G. *et al.* (2010) Large, rare chromosomal deletions associated with severe early-onset obesity. *Nature*, **463**, 666–670.
- Broadaway, K.A. *et al.* (2016) A statistical approach for testing cross-phenotype effects of rare variants. *Am. J. Hum. Genet.*, **98**, 525–540.
- Brown, A.J. *et al.* (2016) Foiling IDOL to help control cholesterol. *Circ. Res.*, **118**, 371–373.
- Bulik-Sullivan, B. *et al.* (2015) An atlas of genetic correlations across human diseases and traits. *Nat. Genet.*, **47**, 1236–1241.
- Burkhardt, R. *et al.* (2015) Integration of genome-wide SNP data and gene-expression profiles reveals six novel loci and regulatory mechanisms for amino acids and acylcarnitines in whole blood. *PLoS Genet.*, **11**.
- Charlesworth, J.C. *et al.* (2009) Toward the identification of causal genes in complex diseases: a gene-centric joint test of significance combining genomic and transcriptomic data. *BMC Proc.*, **3**, S92.
- Cichonska, A. *et al.* (2016) metacCA: summary statistics-based multivariate meta-analysis of genome-wide association studies using canonical correlation analysis. *Bioinformatics*, **32**, 1981–1989.
- Demetz, E. *et al.* (2014) The arachidonic acid metabolome serves as a conserved regulator of cholesterol metabolism. *Cell Metab.*, **20**, 787–798.
- Di Paolo, G., and Kim, T.-W. (2011) Linking lipids to Alzheimer's disease: cholesterol and beyond. *Nat. Rev. Neurosci.*, **12**, 284.
- Dupuis, J. *et al.* (2010) New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. *Nat. Genet.*, **42**, 105–116.
- Felmler, D.J. *et al.* (2013) Hepatitis C virus, cholesterol and lipoproteins' impact for the viral life cycle and pathogenesis of liver disease. *Viruses*, **5**, 1292–1324.
- Ferreira, M.A.R., and Purcell, S.M. (2009) A multivariate test of association. *Bioinformatics*, **25**, 132–133.
- Folkersen, L. *et al.* (2010) Association of genetic risk variants with expression of proximal genes identifies novel susceptibility genes for cardiovascular disease. *Circulation*, **3**, 365–373.
- Guo, B., and Wu, B. (2018) Statistical methods to detect novel genetic variants using publicly available GWAS summary data. *Comput. Biol. Chem.*, **74**, 76–79.
- Hattori, C. *et al.* (2006) Bace1 interacts with lipid raft proteins. *J. Neurosci. Res.*, **84**, 912–917.
- He, Q. *et al.* (2013) A general framework for association tests with multivariate traits in large-scale genomics studies. *Genet. Epidemiol.*, **37**, 759–767.
- Heid, I.M. *et al.* (2010) Meta-analysis identifies 13 new loci associated with waist-hip ratio and reveals sexual dimorphism in the genetic basis of fat distribution. *Nat. Genet.*, **42**, 949–960.
- Hietaniemi, M. *et al.* (2005) Igf-I concentrations are positively associated with carotid artery atherosclerosis in women. *Ann. Med.*, **37**, 373–382.
- Hong, C. *et al.* (2010) The E3 ubiquitin ligase IDOL induces the degradation of the low density lipoprotein receptor family members VLDLR and ApoE2. *J. Biol. Chem.*, **285**, 19720–19726.
- Kathiresan, S. *et al.* (2009) Common variants at 30 loci contribute to polygenic dyslipidemia. *Nat. Genet.*, **41**, 56–65.
- Komiya, S., and Masuda, T. (1989) Relationship of the waist to hip ratio with serum lipids in women. *Ann. Physiol. Anthropol.*, **8**, 239.
- Kwak, I.-Y., and Pan, W. (2017) Gene- and pathway-based association tests for multiple traits with GWAS summary statistics. *Bioinformatics*, **33**, 64–71.
- LeBlanc, M. *et al.* (2016) Identifying novel gene variants in coronary artery disease and shared genes with several cardiovascular risk factors. *Circ. Res.*, **118**, 83–94.
- Liu, J.Z. *et al.* (2010) A versatile gene-based test for genome-wide association studies. *Am. J. Hum. Genet.*, **87**, 139–145.
- Maity, A. *et al.* (2012) Multivariate phenotype association analysis by marker-set kernel machine regression. *Genet. Epidemiol.*, **36**, 686–695.
- Manolio, T.A. *et al.* (2009) Finding the missing heritability of complex diseases. *Nature*, **461**, 747–753.
- Micale, L. *et al.* (2008) Williams–Beuren syndrome trim50 encodes an E3 ubiquitin ligase. *Eur. J. Hum. Genet.*, **16**, 1038–1049.
- Nair, D. *et al.* (2009) Association of total cholesterol/high-density lipoprotein cholesterol ratio with proximal coronary atherosclerosis detected by multislice computed tomography. *Prevent. Cardiol.*, **12**, 19–26.
- Ng, M.C. *et al.* (2017) Discovery and fine-mapping of adiposity loci using high density imputation of genome-wide association studies in individuals of African ancestry: African ancestry anthropometry genetics consortium. *PLoS Genet.*, **13**, e1006719.
- O'Reilly, P.F. *et al.* (2012) MultiPhen: joint model of multiple phenotypes can increase discovery in GWAS. *PLoS ONE*, **7**, e34861.
- Pasaniuc, B., and Price, A.L. (2017) Dissecting the genetics of complex traits using summary association statistics. *Nat. Rev. Genet.*, **18**, 117–127.
- Petersen, H.H. *et al.* (2004) Low-density lipoprotein receptor-related protein interacts with Mafk, a regulator of hindbrain development. *FEBS Lett.*, **565**, 23–27.
- Postmus, I. *et al.* (2014) Pharmacogenetic meta-analysis of genome-wide association studies of LDL cholesterol response to statins. *Nat. Commun.*, **5**, 5068.
- Seoane, J.A. *et al.* (2014) Canonical correlation analysis for gene-based pleiotropy discovery. *PLoS Comput. Biol.*, **10**, e1003876.
- Shim, H. *et al.* (2015) A multivariate genome-wide association analysis of 10 LDL subfractions, and their response to statin treatment, in 1868 Caucasians. *PLoS One*, **10**, e0120758.
- Stephens, M. (2013) A unified framework for association analysis with multiple related phenotypes. *PLoS One*, **8**, e65245.
- Suto, J.-i., and Kojima, M. (2017) Identification of quantitative trait loci that determine plasma total-cholesterol and triglyceride concentrations in ddd/sgn and c57bl/6j inbred mice. *Cholesterol*, 3178204.
- Suzuki, R. *et al.* (2013) Signal peptidase complex subunit 1 participates in the assembly of hepatitis C virus through an interaction with e2 and ns2. *PLoS Pathogens*, **9**, e1003589.
- Tang, C.S., and Ferreira, M.A.R. (2012) A gene-based test of association using canonical correlation analysis. *Bioinformatics*, **28**, 845–850.
- Teslovich, T.M. *et al.* (2010) Biological, clinical, and population relevance of 95 loci for blood lipids. *Nature*, **466**, 707–713.
- Van der Sluis, S. *et al.* (2013) TATES: efficient multivariate genotype-phenotype analysis for genome-wide association studies. *PLoS Genet.*, **9**, e1003235.
- Van der Sluis, S. *et al.* (2014) Mgas: a powerful tool for multivariate gene-based genome-wide association analysis. *Bioinformatics*, **31**, 1007–1015.
- Visscher, P.M. *et al.* (2017) 10 Years of GWAS discovery: biology, function, and translation. *Am. J. Hum. Genet.*, **101**, 5–22.
- Weng, L. *et al.* (2016) Genetic loci associated with nonobstructive coronary artery disease in Caucasian women. *Physiol. Genomics*, **48**, 12–20.

- Willer,C.J. *et al.* (2013) Discovery and refinement of loci associated with lipid levels. *Nat. Genet.*, **45**, 1274–1283.
- Wilson,P.W. (1990) High-density lipoprotein, low-density lipoprotein and coronary artery disease. *Am. J. Cardiol.*, **66**, A7–A10.
- Wu,B., and Pankow,J.S. (2015) Statistical methods for association tests of multiple continuous traits in genome-wide association studies. *Ann. Hum. Genet.*, **79**, 282–293.
- Wu,B., and Pankow,J.S. (2016) Sequence kernel association test of multiple continuous phenotypes. *Genet. Epidemiol.*, **40**, 91–100.
- Wu,M.C. *et al.* (2010) Powerful SNP-set analysis for case-control genome-wide association studies. *Am. J. Hum. Genet.*, **86**, 929–942.
- Yang,Q. *et al.* (2010) Analyze multivariate phenotypes in genetic association studies by combining univariate association tests. *Genet. Epidemiol.*, **34**, 444–454.
- Yao,C. *et al.* (2015) Integromic analysis of genetic variation and gene expression identifies networks for cardiovascular disease phenotypes. *Circulation*, **131**, 536–549.
- Zhu,X. *et al.* (2015) Meta-analysis of correlated traits via summary statistics from GWASs with an application in hypertension. *Am. J. Hum. Genet.*, **96**, 21–36.
- Zhu,X., and Stephens,M. (2017) Bayesian large-scale multiple regression with summary statistics from genome-wide association studies. *Ann. Appl. Stat.*, **11**, 1561–1592.