HHS Public Access

Author manuscript

Comput Biol Chem. Author manuscript; available in PMC 2019 June 01.

Published in final edited form as:

Comput Biol Chem. 2018 June ; 74: 76–79. doi:10.1016/j.compbiolchem.2018.02.016.

Statistical methods to detect novel genetic variants using publicly available GWAS summary data

Bin Guo and **Baolin Wu***

Division of Biostatistics, School of Public Health, University of Minnesota

Keywords

GWAS; SNP-set association test; Summary statistics

We propose statistical methods to detect novel genetic variants just using genome-wide association studies (GWAS) summary data without access to raw genotype and phenotype data. With more and more summary data being posted for public access in the post GWAS era, the proposed methods are practically very useful to identify additional interesting genetic variants and shed lights on the underlying disease mechanism. We illustrate the utility of our proposed methods with application to GWAS meta-analysis results of fasting glucose from the international MAGIC consortium. We found several novel genome-wide significant loci that are worth further study.

1 Introduction

In the past decade, the genome-wide association studies (GWAS) have been very successful in identifying thousands of common genetic variants that are associated with various traits and diseases (Visscher *et al.*, 2017). These GWAS are primarily based on the paradigm of single variant single trait association tests, and have typically made publicly available the association test summary statistics, which include, e.g., the minor allele frequency (MAF), the estimated effect sizes with their standard errors, and significance p-values for each single nucleotide polymorphism (SNP) analyzed in a GWAS. Since it is generally much harder to access the individual-level GWAS phenotype and genotype data due to privacy concerns and various logistical considerations, it has motivated tremendous interest in developing new methods for further analyzing GWAS association test summary data (Pasaniuc and Price, 2017). For example, for the single variant based association test, the GWAS meta-analysis (Evangelou and Ioannidis, 2013) is typically conducted based on the association test summary statistics, which can be as efficient as analyzing individual-level data across all studies (Lin and Zeng, 2010). Similar methods have been developed for meta-analysis of the rare variant set association across studies (Hu *et al.*, 2013; Lee *et al.*, 2014). Specifically Hu *et al.* (2013) showed that the score statistics across variants approximately followed the same

*baolin@umn.edu.

correlation as the rare variant genotype scores, and developed a novel Wald chi-square test for the rare variant set association just using single rare variant association score statistics. For joint association test of a single variant with multiple correlated traits, Stephens (2013) and Zhu *et al.* (2015) proposed methods using only individual GWAS summary statistics and GWAS meta-analysis summary results. The key insight of these approaches is that for a single variant, the association test Z-statistics across different traits share the same correlation as the traits (Stephens, 2013). Most identified common variants have small effect sizes and contribute a small proportion to the overall heritability (Manolio *et al.*, 2009), it often helps to aggregate signals across multiple variants to boost the detection power (Wu et al., 2010). In this work, we study methods for genome-wide variant set association test at the gene level using only GWAS summary data. We provide a transparent derivation showing that the correlation of GWAS (meta-analysis) association test Z-statistics across variants can be computed based on the variant linkage disequilibrium (LD) matrix. Hence we can leverage LD information from a population reference panel to estimate the correlation of Z-statistics and conduct variant set association test. We further develop and post publicly available R programs that can very efficiently and accurately compute p-values for the summary data based association tests. The proposed methods are practically very useful to further mine the vast amount of public GWAS summary data to identify additional interesting genetic variants. We illustrate the utility of our proposed methods with application to GWAS meta-analysis results of fasting glucose from the international MAGIC consortium.

2 Materials and Methods

Consider the single variant association test for a continuous trait Y based on the following linear regression model $Y = X\alpha_j + G_j\beta_j + \in_j$ for the variant G_j, $j = 1,\ldots,M$. Here X is a vector of p covariates to be adjusted (including the intercept, age, and gender, e.g.), and α_j and β_j are the regression coefficients. ϵ_j is assumed to follow a normal distribution $\boldsymbol{N}(0,\sigma_j^2)$. The random error ϵ_j has been indexed depending on each variant, since in principle ϵ_j consists of two parts: variation due to random measurement error and polygenic contributions. The random measurement error can be assumed identical across all variants. But the polygenic contribution could be different since part of it will be captured by $G_j\beta_j$ for those risk variants. Therefore null variants will have identical errors with the same variance, while the variances are potentially different for risk variants and those nearby variants in linkage disequilibrium (LD) with the risk variants.

Given the observations of n unrelated individuals, denote $\boldsymbol{X}$ as the $n\times p$ design matrix, $\mathbf{Y}$ the n-vector of outcomes, G_j the n-vector of genotypes for the j-th variant. Denote projection matrix $\boldsymbol{P} = \boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$, where $\boldsymbol{I}_n$ is a n-th order identity matrix. We can check that

$$\widehat{\beta_j} = \frac{Y^T PG_j}{G_j^T PG_j}, \quad \upsilon\mathrm{ar}\left(\hat{\beta}_j\right) = \frac{\sigma_j^2}{G_j^T PG_j}.$$

The GWAS summary Z-statistics are computed as the standardized genetic regression coefficients by their estimated standard errors where the unknown variance σ_j^2 is replaced by their estimate $\hat{\sigma}_j^2$

$$z_j = \frac{\widehat{\beta}_j}{\sqrt{\widehat{var}\left(\widehat{\beta}_j\right)}} = \frac{Y^T \widetilde{G}_j}{\hat{\sigma}_j}, \quad \widetilde{G}_j = \frac{PG_j}{\sqrt{G_j^T PG_j}}.$$

Here $\widetilde{G}_j$ is essentially the vector of standardized genotypes adjusting for other covariates. We can check that asymptotically $Var\left(z_j\right) = 1$, $Cov\left(z_j, z_l\right) = \rho_{jl}$, where $\rho_{jl} = \frac{G_j^T PG_l}{\sqrt{(G_j^T PG_j)(G_l^T PG_l)}}$ (see appendix for details). When the adjusted covariates are all independent of tested variants (e.g., age and gender), we can unbiasedly estimate the covariance between the summary Z-statistics using $cor(G_j, G_l)$, i.e., the LD correlation matrix (see appendix for details). Therefore we can compute the null covariance matrix of summary Z-statistics by leveraging the LD information from a population reference panel, e.g., the 1000 Genomes Project (Abecasis *et al.*, 2012), or some existing GWAS data of similar ancestry. As argued by Hu *et al.* (2013), for weakly informative covariates, the LD can still provide a very good approximation to the correlation of score statistics, which is proportional to the Z-statistics. Therefore in general we expect the LD correlation matrix provides a good estimate of the Z-statistic correlations. The previous results also hold for GWAS meta-analysis summary results (see appendix for details).

For simplicity of notation, consider a set of m variants in a gene region, denote their summary Z-statistics as $(z_1,\ldots,z_m)$, and $R = (r_{ij})$ the estimated null correlation matrix computed based on the variant LD. We consider the following three SNP-set association tests: (1) sum test (ST), $B = \sum_{j=1}^{m} z_j$, which is a type of burden test statistic (Madsen and Browning, 2009); (2) squared sum test (S2T), $Q = \sum_{j=1}^{m} z_j^2$, which is a type of SKAT statistic (Wu et al., 2010); and (3) adaptive test (AT), based on the minimum p-value $T = \min_{\rho\in[0,1]} P(Q_\rho)$, where $Q_\rho = (1-\rho)Q + \rho B^2$ and $P(Q_\rho)$ denotes its p-value. The AT is in the same vein as the SKAT-O statistic (Lee *et al.*, 2012). We can readily check that $B^2/\left(\sum_{i,j} r_{ij}\right)$ asymptotically follows the x_1^2 distribution; and Q is asymptotically distributed as the weighted sum of independent x_1^2 random variables with weights being the eigenvalues of R. We follow the approach of Wu *et al.* (2016) to efficiently and accurately compute the p-value for the AT based on an one-dimensional numerical integration. In practice we search over $\rho \in (0, 0.1^2, 0.2^2, 0.3^2, 0.4^2, 0.5^2, 0.5, 1)$for the minimum p-value. The ST has good performance when all variants have effects of same direction and approximately equal size, but is sensitive to the direction of variant effects. The S2T has better performance than the ST under a mix of protective and deleterious variants. The AT could adapt to the data and generally has more consistent and robust performance with good detection power across a wide range of scenarios.

3 Results

3.1 Simulation study

We first conduct a simulation study to assess the type I errors of the three tests. We simulate 10^8 random vectors from $\boldsymbol{N}(0,R)$ to estimate their type I errors at significance levels $\alpha = 10^{-4}, 10^{-5}$, and 2.5×10^{-6}. We consider a set of 20 SNPs in the NPHS2 gene and take their LD matrix as R. Table 1 summarizes the results. Overall we can see that all three tests have well-controlled type I errors.

3.2 Application to fasting glucose GWAS meta-analysis summary results

We analyze the summary data from the GWAS meta-analysis of fasting glucose conducted by the international MAGIC consortium (Dupuis *et al.*, 2010). The association results are based on 21 GWAS with around 46,186 non-diabetic participants of European descent who are informative for fasting glucose. The summary data is publicly available at ftp://ftp.sanger.ac.uk/pub/magic/MAGIC_FastingGlucose.txt. The summary data consists of the MAF, effect size estimate and its associated standard error, and p-value for 2,470,476 SNPs. For illustration, we first remove 290 genome-wide significant SNPs with p-value less than 5×10^{-8} and then filter out those SNPs with MAF < 0.05. We download the list of genes and their coordinates (transcription start and end positions based on the hg19/GRChB37 reference genome) from the UCSC genome browser (Kent *et al.*, 2002). We take all SNPs that are located in or near a gene as a set to be analyzed for joint association. Specifically following Wu et al. (2010), we group all SNPs from 20 kb upstream of a gene to 20 kb downstream of a gene. For each SNP set corresponding to a gene, we also perform LD pruning: we remove those SNPs that have pairwise LD $r^2 > 0.8$ with other SNPs. Using these criteria, we obtain 18,725 SNP sets that have at least two SNPs. We set our genome-wide SNP set significance level as 2.67×10^{-6}, which is the Bonferroni corrected significance level for the total number of tested SNP sets.

We note that the MAGIC consortium has performed a followup replication study using a Metabochip consisting of a small panel of promising SNPs and a much larger sample size from 66 fasting glucose GWAS with around 133,010 non-diabetic participants (Scott *et al.*, 2012). The summary data contains the results for 64,493 pre-selected SNPs, and is available at ftp://ftp.sanger.ac.uk/pub/magic/MAGIC_Metabochip_Public_data_release_25Jan.zip. We use them as partial validation for our analysis of Dupuis *et al.* (2010) summary data.

When applying the three SNP-set tests to the summary data of Dupuis *et al.* (2010), the ST identified 12 significant genes, S2T identified 20 genes, and AT identified 22 significant genes. The adaptive test AT captures the majority of significant genes identified by ST and S2T. An interesting case is the significant gene FADS1 identified only by the adaptive test AT with a significance p-value of 2.39×10^{-6}, and ST and S2T reported p-values of 1.40×10^{-5} and 4.41×10^{-6}, respectively. This gene harbored genome-wide significant SNPs in the study of Dupuis *et al.* (2010) and the followup study of Scott *et al.* (2012). Figure 1 shows the Venn diagram comparing the number of significant genes identified by the proposed tests. Among all 25 significant genes identified by the three SNP-set tests, 10 are novel genes which didn't harbor any significant SNPs in the study of Dupuis *et al.* (2010). These

novel genes contain promising SNPs (often with small or modest effect sizes) worth further study. And among these 10 novel genes, 6 genes have been found to contain genome-wide significant SNPs in the followup study of Scott *et al.* (2012). Table 2 lists the test p-values for some novel genes together with the minimum p-values across all SNPs in the gene in the two studies. The complete results are available at the supplementary materials.

For illustration, we also performed SNP-set tests using all SNPs and obtained similar conclusions. The complete results are available at the supplementary materials.

4 Discussion

In summary, we have proposed SNP-set tests using the GWAS summary data. The proposed methods are efficient and scalable to analyze summary data for millions of genome-wide SNPs. As more and more summary data are now being posted for public access in the post GWAS era, these methods will be practically very useful to identify more genetic variants associated with various diseases.

Our previous discussions have implicitly assumed equal weights for variants. We note that we can readily incorporate variant weights into the proposed tests as follows. Denote the associated variant weights as $\boldsymbol{W} = (w_1,\ldots,w_m)$, which can be determined by the variant MAFs $(p_1,\ldots,p_m)$. We then consider weighted summary statistics $(w_1z_1,\ldots,w_mz_m)$, and their associated covariance matrix is then $\mathrm{diag}(\boldsymbol{W})R\mathrm{diag}(\boldsymbol{W})$. Note that the Z-statistics are inherently standardized: the Z-statistic is roughly proportional to the score statistic scaled by the genotype standard error. So setting the constant weight corresponds to using weight $w_j = 1/\sqrt{p_j(1-p_j)}$ following the approach of Wu et al. (2010) and Madsen and Browning (2009). There have been some recent research that leverages additional functional annotation information and public GWAS summary data to further identify novel genetic variants (Gusev *et al.*, 2016; Mancuso *et al.*, 2017), which can be approached as incorporating variant weights $\boldsymbol{W}$ under the proposed framework. We are currently exploring this approach and will report the results in the future.

We want to remark that it is more productive to treat the proposed SNP-set tests as a complementary instead of competing approach to the traditional single variant based association test. Therefore we recommend applying the proposed tests to genes that do not harbor genome-wide significant SNPs to further identify more novel variants. We have implemented the proposed methods in an R package available at http://www.github.com/baolinwu/mkatr. We provide sample R codes to install and use the package at the supplementary materials.

Acknowledgement

This research was supported in part by NIH grant GM083345 and CA134848. We want to thank the editor and reviewers for their constructive comments that have greatly improved the presentation of the paper. We are grateful to the University of Minnesota Supercomputing Institute for assistance with the computations. Data on glycemic traits have been contributed by MAGIC investigators and have been downloaded from www.magicinvestigators.org. We want to thank Dr. James Pankow for pointing out the MAGIC data source to us.

Appendix

GWAS summary Z-statistics

Denote $\boldsymbol{Y}$ as the continuous outcome observed for a cohort of n unrelated individuals. Denote the $n \times p$ covariate matrix as $\boldsymbol{X}$, which includes the intercept and other commonly adjusted covariates (e.g., age and gender). Here p denotes the total number of covariates.Under the marginal working regression model for variant j, $\boldsymbol{Y} = \boldsymbol{X}\alpha_j + \boldsymbol{G}_j\beta_j + \epsilon_j$, where $\in_j \sim \mathcal{N}(0.\sigma_j^2 I_n)$, we have $\hat{\beta}_j = G_j^T PY / \left\{G_j^T PG_j\right\}$, where $P = In - X(X^TX)^{-1}X^T$ is the projection matrix. Hence $\left(\widehat{\beta_j}, \widehat{\beta_l}\right) = cor(G_j^T PY, G_l^T PY) = r_{jl}$, where $r_{jl} = \frac{G_j^T PG_j}{\sqrt{(G_j^T PG_j)}\sqrt{(G_l^T PG_l)}}$.

When the covariates $\boldsymbol{X}$ do not contain ancestry covariates (e.g., for the homogeneous European population), $\boldsymbol{G}_j$ is independent of $\boldsymbol{X}$: hence r_{jl} equals to the LD correlation between the two variants, $cor(\boldsymbol{G}_j, \boldsymbol{G}_l)$. Let z_j denote the standardized $\widehat{\beta_j}$ (i.e., the summary Z-statistic)

$z_j = \frac{\widehat{\beta_j}}{\sqrt{\widehat{Var}\left(\widehat{\beta_j}\right)}} = \frac{Y^T \widetilde{G}_j}{\hat{\sigma}_j}, \quad \widetilde{G}_j = \frac{PG_j}{\sqrt{G_j^T PG_j}}$,where $\hat{\sigma}_j^2 = \left|\left|PY - \widetilde{G}_j\widehat{\beta_j}\right|\right|^2/(n-p-1)$. Note that generally the estimated variance $\hat{\sigma}_j^2$ are slightly different across SNPs. For null variants not associated with the outcome, $\hat{\sigma}_j^2$ are all unbiased estimates of the same outcome variance. Therefore for null variants, asymptotically $Var(z_j) = 1$ and $Cov(z_j, z_l) = r_{jl}$For a variant associated with the outcome, if it explains only a very small proportion of the total trait variation, which is true for most variants under polygenic model, the previous equations approximately hold.Denote the (raw or after adjusting for top ancestry PCs) LD correlation matrix $R = (r_{jl})$. For relatively homogeneous population (e.g., of European descent), we can use the LD matrix computed from public population samples.

Consider K separate GWAS each with n_k individuals, $k = 1,2,\ldots,K$. For the k-th GWAS, denote $\boldsymbol{Y}_k$ as the outcome, $\boldsymbol{P}_k$ the covariate projection matrix, $\boldsymbol{G}_{kj}$ the genotype scores for SNP j, and $\hat{\boldsymbol{\beta}}_{kj}$ the regression parameter estimates for the j-th SNP. With homogeneous population and no ancestry covariates, $\boldsymbol{P}_k\boldsymbol{G}_{kj}$ amounts to centering the genotypes. Without loss of generality, assume the genotype scores have been centered. We have $\boldsymbol{P}_k\boldsymbol{G}_{kj} = \boldsymbol{G}_{kj}$, and hence $\hat{\beta}_{kj} = G_{kj}^T Y_k / \left\{G_{kj}^T G_{kj}\right\}$. Denote $Var(Y_k) = \sigma^2 I_{n_k}$.Let θ_j denote the variance of genotype scores for SNP j. For null variants, we have asymptotically $Var(\hat{\beta}_{kj}) = \sigma^2/(n_k\theta_j)$, and $Cov\left(\hat{\beta}_{kj}, \hat{\beta}_{kl}\right) = r_{jl}\sigma^2/(n_k\sqrt{\theta_j\theta_l})$. They also approximately hold for those variants with small effects. The meta-analysis estimate is typically based on the weighted average, denoted as $\widehat{\beta_j} = \sum_{k=1}^K a_k\widehat{\beta_{kl}}$. Here the weight a_k is typically determined by the GWAS sample size. For example, one common choice is$a_k = n_k/(\sum_{k=1}^K n_k)$. Another popular choice is based on the inverse variance weighting, which amounts to setting $a_k = Var(\widehat{\beta_{kj}})^{-1}\left[\sum_{k=1}^K Var(\widehat{\beta_{kj}})^{-1}\right]$, which asymptotically equals to$n_k/(\sum_{k=1}^K n_k)$. Hence for null variants we have

asymptotically , $Var(\widehat{\beta_j}) = \left(\sum_{k=1}^{K} a_k^2/n_k\right)\sigma^2/\theta_j$, and $Cov(\widehat{\beta_j}, \widehat{\beta_l}) = \left(\sum_{k=1}^{K} a_k^2/n_k\right)r_{jl}\sigma^2/\sqrt{\theta_j\theta_l}$. Therefore for the GWAS meta-analysis Z-statistics, we also have asymptotically $Cov(z_j, z_l) = r_{jl}$. We note that with $a_k = n_k/(\sum_{k=1}^{K} n_k)$., we have $Var(\widehat{\beta_j}) = \sigma^2/\theta_j/\left(\sum_{k=1}^{K} n_k\right)$. Therefore the meta-analysis approach has the same asymptotic efficiency as the mega-analysis, which pools all GWAS individual samples for analysis. This agrees with the theoretical results of Lin and Zeng (2010).

References

Abecasis GR, Auton A, Brooks LD, and others from 1000 Genomes Project Consortium. (2012) An integrated map of genetic variation from 1,092 human genomes. Nature, 491 (7422), 56–65. [PubMed: 23128226]

Dupuis J, Langenberg C, Prokopenko I, and others. (2010) New genetic loci implicated in fasting glucose homeostasis and their impact on type 2 diabetes risk. Nature Genetics, 42 (2), 105–116. [PubMed: 20081858]

Evangelou E and Ioannidis JPA (2013) Meta-analysis methods for genome-wide association studies and beyond. Nature Reviews. Genetics, 14 (6), 379–389.

Gusev A, Ko A, Shi H, Bhatia G, and others. (2016) Integrative approaches for largescale transcriptome-wide association studies. Nature Genetics, 48 (3), 245–252. [PubMed: 26854917]

Hu YJ, Berndt SI, Gustafsson S, Ganna A, Genetic Investigation of ANthropometric Traits (GIANT) Consortium, Hirschhorn J, North KE., Ingelsson Eand Lin DY. (2013) Meta-analysis of gene-level associations for rare variants based on single-variant statistics. American Journal of Human Genetics, 93 (2), 236–248. [PubMed: 23891470]

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM and Haussler D (2002) The Human Genome Browser at UCSC. Genome Research, 12 (6), 996–1006. [PubMed: 12045153]

Lee S, Abecasis G, Boehnke M and Lin X (2014) Rare-Variant Association Analysis: Study Designs and Statistical Tests. The American Journal of Human Genetics, 95 (1), 5–23. [PubMed: 24995866]

Lee S, Wu MC and Lin X (2012) Optimal tests for rare variant effects in sequencing association studies. Biostatistics, 13 (4), 762–775. [PubMed: 22699862]

Lin DY and Zeng D (2010) Meta-analysis of genome-wide association studies: no efficiency gain in using individual participant data. Genetic Epidemiology, 34 (1), 60–66. [PubMed: 19847795]

Madsen BE and Browning SR (2009) A groupwise association test for rare mutations using a weighted sum statistic. PLOS Genetics, 5 (2), e1000384. [PubMed: 19214210]

Mancuso N, Shi H, Goddard P, Kichaev G, Gusev A and Pasaniuc B (2017) Integrating Gene Expression with Summary Association Statistics to Identify Genes Associated with 30 Complex Traits. The American Journal of Human Genetics, 100 (3), 473–487. [PubMed: 28238358]

Manolio TA, Collins FS, Cox NJ, and others. (2009) Finding the missing heritability of complex diseases. Nature, 461 (7265), 747–753. [PubMed: 19812666]

Pasaniuc B and Price AL (2017) Dissecting the genetics of complex traits using summary association statistics. Nature Reviews Genetics, 18 (2), 117–127.

Scott RA, Lagou V, Welch RP, and others. (2012) Large-scale association analyses identify new loci influencing glycemic traits and provide insight into the underlying biological pathways. Nature Genetics, 44 (9), 991–1005. [PubMed: 22885924]

Stephens M (2013) A unified framework for association analysis with multiple related phenotypes. PloS One, 8 (7), e65245. [PubMed: 23861737]

Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA and Yang J (2017) 10 Years of GWAS Discovery: Biology, Function, and Translation. American Journal of Human Genetics, 101 (1), 5–22. [PubMed: 28686856]

Wu B, Guan W, Pankow JS (2016) On efficient and accurate calculation of significance p-values for sequence kernel association test of variant set. Annals of Human Genetics, 80 (2), 123–135. [PubMed: 26757198]

Wu MC, Kraft P, Epstein MP, Taylor DM, Chanock SJ, Hunter DJ, Lin X, 2010 Powerful SNP-Set Analysis for Case-Control Genome-wide Association Studies. The American Journal of Human Genetics 86 (6), 929–942. [PubMed: 20560208]

Zhu X, Feng T, Tayo B, and others. (2015) Meta-analysis of Correlated Traits via Summary Statistics from GWASs with an Application in Hypertension. The American Journal of Human Genetics, 96 (1), 21–36. [PubMed: 25500260]

highlights.txt

1. 1. propose useful and efficient GWAS summary data based SNP-set association test methods to identify more novel variants
2. 2. provide efficient implementation of the proposed methods in a publicly available R package
3. 3. demonstrate the utility of proposed methods via application to analysis of fasting glucose GWAS meta-analysis results, and find several novel loci worth further study

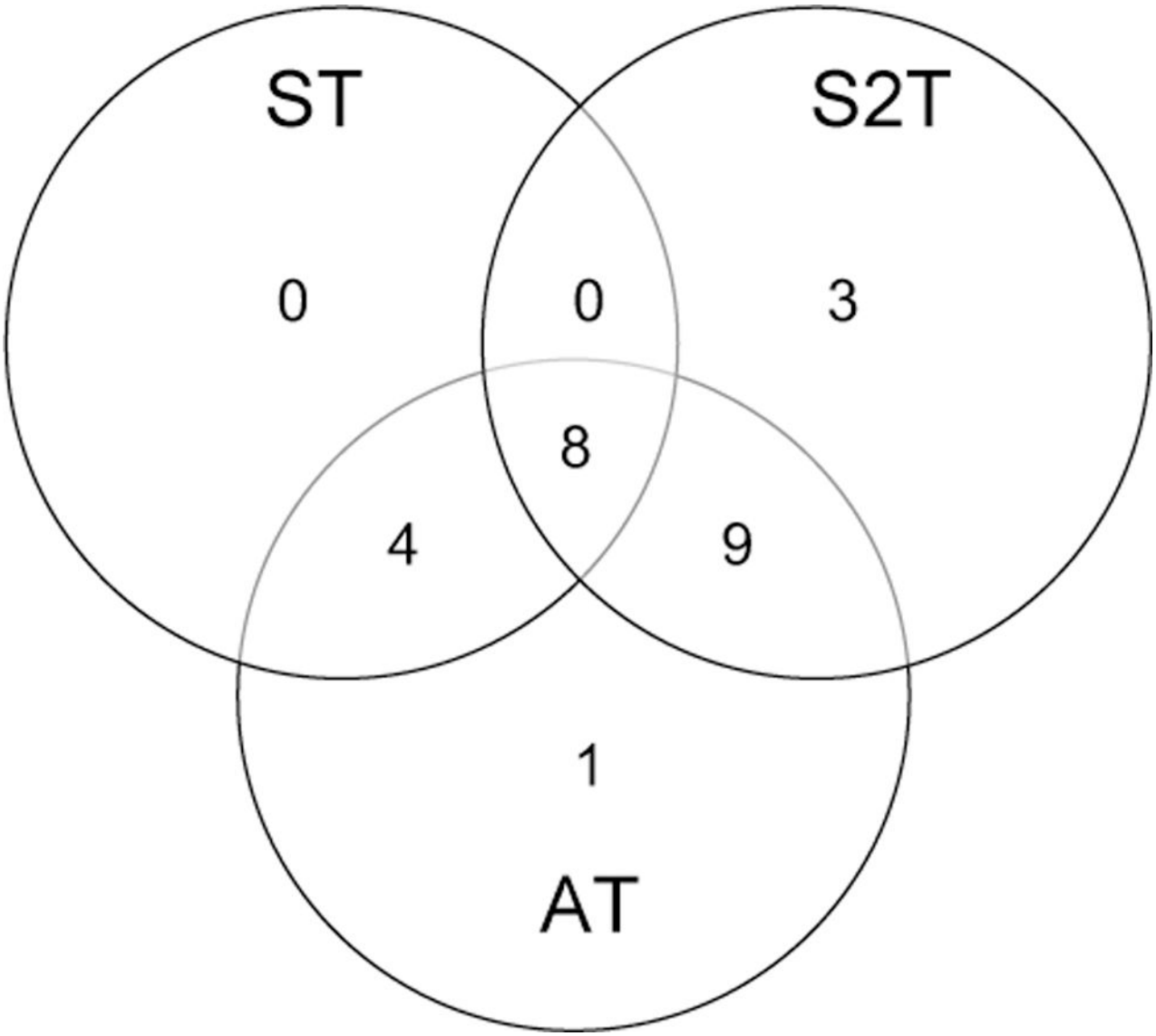

Figure 1:
Venn diagram of number of significant genes identified by three summary data based SNP-set tests.

Table 1:

Ratio of empirical type I errors divided by the significance level α estimated over 10^8 simulations: listed within parentheses are the standard errors. S2T is based on the sum of squared Z-statistics; ST is the sum of Z-statistics, and AT is the adaptive test.

	10^{-4}	10^{-5}	2.5×10^{-6}
S2T	1.01 (0.01)	0.99 (0.03)	0.98 (0.06)
ST	1.00 (0.01)	0.99 (0.03)	1.02 (0.06)
AT	0.99 (0.01)	0.88 (0.03)	0.81 (0.06)

Table 2:

Some novel genes found by the proposed SNP-set tests: we listed the test p-values for S2T, ST and AT and the minimum p-value across all SNPs in the gene for Dupuis *et al.* (2010) study (denoted as minP-2010) and Scott *et al.* (2012) study (denoted as minP-2012).

Gene	S2T	ST	AT	minP-2010	minP-2012
ZNF512	2.09e-06	5.49e-01	3.82e-06	1.40e-06	9.68e-20
GPN1	2.51e-06	5.38e-01	4.62e-06	2.96e-06	9.68e-20
SLC4A1AP	1.22e-06	6.06e-01	1.55e-06	2.96e-06	3.21e-17
PROX1	1.39e-06	8.17e-06	1.61e-06	7.08e-08	3.22e-12
SUPT7L	1.76e-06	4.16e-01	2.30e-06	2.96e-06	6.59e-12
C2CD4B	6.31e-07	5.64e-02	6.31e-07	5.89e-07	2.05e-08