# Supplementary Materials for "Efficient and powerful meta-analysis of variant-set association tests using MetaSAT"

Baolin Wu[1] and Hongyu Zhao[2]

[1]Division of Biostatistics, School of Public Health, University of Minnesota

[2]Department of Biostatistics, School of Public Health, Yale University

## 1 Meta-analysis of variant-set association test: technical details

### 1.1 Single study variant-set association test (SAT)

First let us consider a single set of $m$ variants with weighted test statistics $\boldsymbol{S}$ and asymptotic covariance matrix $\boldsymbol{V}$. Define a general burden test (BT) $B = (\eta^T \boldsymbol{S})^2$, where $\eta$ is a pre-specified vector, the variance component test (VT) $Q = \boldsymbol{S}^T \boldsymbol{S}$ in the same vein as the SKAT statistic (Wu *et al.*, 2011), and the adaptive test (AT) based on the minimum p-values of $Q_\rho = (1-\rho)(\boldsymbol{S}^T \boldsymbol{S}) + \rho(\eta^T \boldsymbol{S})^2$ over $\rho \in [0,1]$, in the same vein as the SKAT-O test (Lee *et al.*, 2012).

One can readily check that the orthogonal projection of $\boldsymbol{S}$ onto $\eta^T \boldsymbol{S}$ is $\beta \eta^T \boldsymbol{S}$, where $\beta = Cov(\boldsymbol{S}, \eta^T \boldsymbol{S})/Var(\eta^T \boldsymbol{S}) = \boldsymbol{V}\eta/(\eta^T \boldsymbol{V}\eta)$. Hence we can decompose $Q$ into two approximately independent components $B_1 = (\eta^T \boldsymbol{S})^2(\beta^T \beta) = \nu(\eta^T \boldsymbol{S})^2$ and $Q_e = Q - B_1$, where $\nu = (\eta^T \boldsymbol{V}^2 \eta)/(\eta^T \boldsymbol{V}\eta)^2$. Note that $Q_e = \boldsymbol{S}^T(\boldsymbol{I}_m - \nu\eta\eta^T)\boldsymbol{S}$. Denote eigen decomposition $\boldsymbol{V} = UD^2U^T$. Assume $\boldsymbol{Z} \sim \mathcal{N}(0, \boldsymbol{I}_m)$. Under the null $\boldsymbol{S}$ is distributed as $UD\boldsymbol{Z}$. Therefore we can write

$$Q_e = \boldsymbol{Z}^T[D^2 - \nu(DU^T\eta)(DU^T\eta)^T]\boldsymbol{Z},$$

and

$$Q_\rho = \boldsymbol{Z}^T[(1-\rho)D^2 + \rho(DU^T\eta)(DU^T\eta)^T]\boldsymbol{Z} = (1-\rho)Q_e + [(1-\rho)\nu + \rho](\eta^T\boldsymbol{S})^2,$$

and we can then efficiently compute the p-value for the adaptive test as follows.

Given a grid of $\rho_1 < \cdots < \rho_L = 1$, denote the p-value of $Q_{\rho_j}$ as $p_j$. Denote $P_0 = \min_{j\le L} p_j$, and let $q_j$ be the $(1-P_0)$th quantile of the null distribution of $Q_{\rho_j}$. We compute the p-value of $P_0$ as

$$P = \Pr(\min_{j\le L} p_j \le P_0) = \Pr(p_L \le P_0) + \Pr(p_L > P_0, \min_{j<L} p_j \le P_0) = P_0 + \Pr(p_L > P_0, \min_{j<L} p_j \le P_0).$$

Note that $\min_{j<L} p_j \le P_0$ is equivalent to $Q_{\rho_j} \ge q_j$ for some $j < L$, which can be expressed as

$$Q_e \ge \min_{j<L} \frac{q_j - \psi_j(\eta^T\boldsymbol{S})^2}{1-\rho_j}, \quad \psi_j = (1-\rho_j)\nu + \rho_j.$$

So we can analytically calculate

$$P = P_0 + 2\int_0^{q_h} G_e\Big[\min_{j<L} \frac{q_j - \psi_j(\eta^T\boldsymbol{V}\eta)x^2}{1-\rho_j}\Big]\phi(x)dx,$$

where $q_h = \Phi^{-1}(1-P_0/2)$, $(\phi, \Phi)$ are the standard normal density and distribution functions, and $G_e$ calculates the tail probability of $Q_e$.

### 1.2 Meta-analysis of variant-set association test

Consider $K$ studies with weighted test statistics $\boldsymbol{S}_k$ and asymptotic covariance matrix $\boldsymbol{V}_k$ for the $k$th study, $k = 1, \cdots, K$. Under the fixed-effects (FE) model, we assume all studies have similar effect sizes, and consider summarized test statistics $\boldsymbol{S} = \sum_{k=1}^K \boldsymbol{S}_k$ with asymptotic covariance matrix $\boldsymbol{V} = \sum_{k=1}^K \boldsymbol{V}_k$, which reduces to the form of a single-study SAT.

Under the heterogeneous-effects (HE) model, we allow studies to have different individual variant effects, and consider a vector of stacked summary statistics $\boldsymbol{S} = (\boldsymbol{S}_1^T, \cdots, \boldsymbol{S}_K^T)^T$ with

asymptotic covariance matrix $\boldsymbol{V} = \text{diag}\{\boldsymbol{V}_1, \cdots, \boldsymbol{V}_K\}$.

For the adaptive test (AT), the FE model considers the weighted test $(1-\rho)(\sum_{k=1}^K \boldsymbol{S}_k)^T(\sum_{k=1}^K \boldsymbol{S}_k)+\rho(\sum_{k=1}^K \eta^T\boldsymbol{S}_k)^2$, and the HE model considers the weighted test $(1-\rho)\sum_{k=1}^K \boldsymbol{S}_k^T\boldsymbol{S}_k+\rho(\sum_{k=1}^K \eta_k^T\boldsymbol{S}_k)^2$, $\rho \in [0,1]$. Here $\eta$ and $\eta_k$ are all given $m$-vectors.

### 1.3 A robust heterogeneous-effects (RHE) meta-analysis model

Alternatively we can consider the following heterogeneous-effects model

$$\sum_{k=1}^K \Big\{(1-\rho_k)\boldsymbol{S}_k^T\boldsymbol{S}_k + \rho_k(\eta_k^T\boldsymbol{S}_k)^2\Big\}, \quad \rho_k \in [0,1],$$

which essentially takes the sum of weighted tests across all studies, and can model variant-set level heterogeneous effects. One strategy is to search for the optimal combinations of all $\rho_k$'s, which however is too computationally intensive. As a compromise, we fix all $\rho_k$'s to be functions of a common parameter $\rho$ as follows. Note that for the $k$th study, we can decompose

$$(1-\rho_k)\boldsymbol{S}_k^T\boldsymbol{S}_k + \rho_k(\eta_k^T\boldsymbol{S}_k)^2 = (1-\rho_k)R_{ek} + [(1-\rho_k)\nu_k+\rho_k](\eta_k^T\boldsymbol{S}_k)^2, \quad \nu_k = \frac{\eta_k^T\boldsymbol{V}_k^2\eta_k}{(\eta_k^T\boldsymbol{V}_k\eta_k)^2},$$

where $R_{ek} = \boldsymbol{S}_k^T\boldsymbol{S}_k - \nu_k(\eta_k^T\boldsymbol{S}_k)^2$. Therefore we can set $[(1-\rho_k)\nu_k+\rho_k]/[(1-\rho_k)\nu_k+1]$ as the same across studies, and develop an efficient algorithm to compute the AT p-value analytically. Operationally this is equivalent to considering the following test

$$R_\rho = (1-\rho)\sum_{k=1}^K \boldsymbol{S}_k^T\boldsymbol{S}_k + \rho\sum_{k=1}^K \nu_k(\eta_k^T\boldsymbol{S}_k)^2 = (1-\rho)\sum_{k=1}^K R_{ek} + \sum_{k=1}^K \nu_k(\eta_k^T\boldsymbol{S}_k)^2.$$

Under the null, $R_\rho$ is distributed as the weighted sum of independent $\chi_1^2$ random variables with weights being the eigenvalues of $(1-\rho)\boldsymbol{V} + \rho\text{diag}\{\nu_1\boldsymbol{V}_1^{1/2}\eta_1\eta_1^T\boldsymbol{V}_1^{1/2}, \cdots, \nu_K\boldsymbol{V}_K^{1/2}\eta_K\eta_K^T\boldsymbol{V}_K^{1/2}\}$. We can accurately compute its p-value using the Davies method (Davies, 1980). Denote $R_e = \sum_{k=1}^K R_{ek}$ and $R_b = \sum_{k=1}^K \nu_k(\eta_k^T\boldsymbol{S}_k)^2$. Under the null, $R_e$ and $R_b$ are approximately independent and distributed as weighted sum of independent $\chi_1^2$ random variables. Furthermore, $(\eta_k^T\boldsymbol{S}_k)^2/(\eta_k^T\boldsymbol{V}_k\eta_k) \sim \chi_1^2$. Given $\rho_1 \le \rho_2 \cdots \le \rho_L = 1$, denote the p-value of $R_{\rho_j}$ as $p_j$ and define

$P_0 = \min_{j \le L} p_j$. Let $q_j$ denote the $(1 - P_0)$th quantile of the null distribution of $R_{\rho_j}$. The AT p-value can be analytically computed as

$$\begin{aligned} P &= \Pr(\min_{j \le L} p_j \le P_0) = \Pr(p_L \le P_0) + \Pr(p_L > P_0, \min_{j<L} p_j \le P_0) \\ &= P_0 + \Pr\Big(R_b < q_L, R_e \ge \min_{j<L} \frac{q_j - R_b}{1 - \rho_j}\Big) \\ &= P_0 + \int_0^{q_L} G_e\Big(\min_{j<L} \frac{q_j - x}{1 - \rho_j}\Big) d[-G_b(x)], \end{aligned}$$

which can be computed efficiently. Here $G_e$ and $G_b$ compute the tail probabilities of $R_e$ and $R_b$ respectively.

The RHE BT test $R_b$ is essentially the sum of squared burden tests across studies. It assumes similar effects across variants within one study while allows for different variant-set level effects across studies, which is potentially relevant for cross-ancestry meta-analysis. In contrast, the FE BT performs well only when all variants have similar effects across studies.

In addition, we also implement an AT based on adaptively weighting the RHE BT and FE BT, denoted as BAT. BAT works well when we have similar variant effects within each study and potentially heterogeneous effects across studies.

Due to low frequency, some rare variants may not be observed in some studies. We set their test statistics and corresponding covariance matrix terms as zero.

## 2 Implemented tests in MetaSAT

When studying the joint association of a variant-set with a trait across multiple studies, we could potentially have heterogeneous effects across both variants and studies. With homogeneous effects, the burden type tests (BT) that directly sum over the variant test statistics perform well. With heterogeneous effects (e.g., a mix of positive and negative effects), the variance component tests (VT) that sum over the squared variant test statistics tend to perform better.

Table 1 summarizes the four general tests implemented in MetaSAT that address varying levels of heterogeneity of variant effects both within and across studies. In summary, we imple-

Table 1: MetaSAT test methods for meta-analysis of variant-set associations.

| | | Effects across studies | |
|---|---|---|---|
| | | Heterogeneous | Homogeneous |
| Effects within a study | Heterogeneous | HE VT: $\sum_{k=1}^{K} \boldsymbol{S}_k^T \boldsymbol{S}_k$ | FE VT: $(\sum_{k=1}^{K} \boldsymbol{S}_k)^T (\sum_{k=1}^{K} \boldsymbol{S}_k)$ |
| | Homogeneous | RHE BT: $\sum_{k=1}^{K} \nu_k (\eta_k^T \boldsymbol{S}_k)^2$ | FE BT: $\eta^T (\sum_{k=1}^{K} \boldsymbol{S}_k)$ |

ment (1) FE BT: which assumes similar variant effects both across and within studies; (2) FE VT: which assumes similar variant effects across studies but heterogeneous effects across variants within each study; (3) RHE BT: which assumes heterogeneous effects across studies but similar effects across variants within each study; and (4) HE VT: which assumes heterogeneous variant effects both across and within studies.

In addition, MetaSAT implements four adaptive tests (AT) that optimally combine various BT and VT to achieve more robust performance. In summary, we implement (1) FE AT (FAT): which combines FE BT and FE VT; (2) HE AT (HAT): which combines FE BT and HE VT; (3) RHE AT (RAT): which combines RHE BT and HE VT; and (4) BT based AT (BAT): which combines RHE BT and FE BT. In MetaSAT, the BT assumes equal weights as default, and the AT searches over $\rho \in \{0, 0.1^2, 0.2^2, \cdots, 0.9^2, 1\}$ as default.

## 3 Simulation study

As shown previously, the FE and HE MA of SAT are essentially equivalent to a single study SAT. Previously Wu *et al.* (2016) have shown that in the context of a single study rare variant-set association test, the commonly used R packages, SKAT (Lee *et al.*, 2017) and hence MetaSKAT (Lee, 2015), and skatMeta/seqMeta (Voorman *et al.*, 2013, 2017), have liberal type I errors at the small genome-wide significance levels, especially for the adaptive tests.

We have proposed new algorithms to more accurately and efficiently compute the analytical p-values for all tests implemented in MetaSAT. Here we conduct simulation studies to show that they properly control the type I errors.

We consider three separate studies with 2000, 3000, and 5000 unrelated individuals, respectively. We simulate a binary covariate $X_1$ from a Bernoulli distribution with 0.5 event

Table 2: Empirical type I errors (divided by the significance level $\alpha$) of meta-analysis methods in MetaSAT.

| $\alpha$ | FE BT | FE VT | HE VT | RHE BT | FAT | HAT | RAT | BAT |
|---|---|---|---|---|---|---|---|---|
| $10^{-4}$ | 1.00 | 0.99 | 0.97 | 0.99 | 1.08 | 0.89 | 1.05 | 1.16 |
| $10^{-5}$ | 0.98 | 0.98 | 0.96 | 0.99 | 1.00 | 0.70 | 0.92 | 1.22 |
| $10^{-6}$ | 0.99 | 0.98 | 0.97 | 0.96 | 0.86 | 0.51 | 0.77 | 1.15 |
| $10^{-7}$ | 0.96 | 0.94 | 0.92 | 1.02 | 0.73 | 0.35 | 0.58 | 1.09 |
| $5 \times 10^{-8}$ | 0.95 | 0.98 | 0.94 | 0.96 | 0.74 | 0.33 | 0.49 | 1.05 |

probability, and a continuous covariate $X_2$ from the standard normal distribution. We first simulate a panel of 10,000 haplotypes over a 200Kb region from a calibrated coalescent model with a LD structure mimicking the European ancestry (Schaffner *et al.*, 2005). We randomly pair the haplotypes to simulate genotypes for individuals. We consider those variants with MAF $\leq 0.02$ and total minor allele counts $\geq 10$ in a randomly selected region of 10Kb. The outcomes are simulated from $Y = 0.5X_1 + 0.5X_2 + \epsilon$, where $\epsilon \sim \mathcal{N}(0, 1)$. We meta-analyze the score statistics for all variants computed for each study.

We conduct $2 \times 10^9$ null simulations to estimate the type I errors at significance level $\alpha = 10^{-4}, 10^{-5}, 10^{-6}, 10^{-7}, 5 \times 10^{-8}$, respectively. Table 2 summarizes the results. Overall we can see that all test methods can properly controll the type I errors.

# 4 Application

For illustration, we consider a rare variant set association test example studied in Wu *et al.* (2016). Specifically they analyzed the association of fasting glucose levels with a set of nine rare variants in the G6PC2 gene measured in 5866 non-diabetic white ARIC participants (Wessel *et al.*, 2015). These subjects have been recruited from three study sites. For illustration of meta-analysis, we analyze the rare variant set association within each site and meta-analyze the results across the three sites. The variant score test statistics are multiplied by (1-MAF)$^{24}$ to upweight rarer variants. We obtain their weighted association test statistics $\boldsymbol{U}_k$ together with the asymptotic covariance matrix $\boldsymbol{V}_k$ for site $k = 1, 2, 3$.

Table 3 summarizes the MetaSAT rare variant set association test meta-analysis p-values.

Table 3: P-values for MA of G6PC2 gene rare variant set association with fasting glucose level: shown are the four general BT and VT together with the AT that adaptively combine various BT and VT.

| BT and VT | | | |
|---|---|---|---|
| FE BT | FE VT | HE VT | RHE BT |
| $3.08 \times 10^{-7}$ | $5.32 \times 10^{-5}$ | $6.05 \times 10^{-4}$ | $3.03 \times 10^{-7}$ |
| AT | | | |
| FAT | HAT | RAT | BAT |
| $6.36 \times 10^{-7}$ | $1.30 \times 10^{-6}$ | $6.79 \times 10^{-7}$ | $2.10 \times 10^{-7}$ |

Overall, the FE model performs better than the HE model, reflecting similar association pattern of fasting glucose levels across three sites. The BT performs better than the VT, reflecting similar association signals across rare variants. Compared to the FE BT and HE VT, the RHE BT reports the smallest p-value, suggesting some heterogeneous variant-set effects across sites. Among the adaptive tests, combining the RHE BT and FE BT performs the best.

# 5 R package

The following are some sample codes to install and use the 'MetaSAT' R package.

```
## install the package
devtools::install_github('baolinwu/MetaSAT')
library(MetaSAT)
## 3 studies with 10 variants and r=0.05 corr
K = 3; m=10; r = 0.05
Vs = array(0, dim=c(m,m,K)); Us = matrix(0, m,K)
for(k in 1:K){
  ak = matrix(rnorm(100*m),100,m)*sqrt(1-r)+rnorm(100)*sqrt(r)
  Vs[,,k] = cor(ak)
  Rh = chol(Vs[,,k])
  Us[,k] = colSums(Rh*rnorm(m))
}
```

```
## FE model
FESAT(Us,Vs)
## HE model
HESAT(Us,Vs)
## robust HE model
RESAT(Us,Vs)
RBAT(Us,Vs) ## AT combining RHE BT and FE BT
U1 = Us + rnorm(m*K,1,1)
FESAT(U1,Vs)
HESAT(U1,Vs)
RESAT(U1,Vs)
RBAT(U1,Vs)
```

The 'MetaSAT' R package also contains the summary statistics and their associated asymptotic covariance matrices for nine variants across three studies as discussed in Section 4. They can be used to reproduce the results in Table 3.

```
data(WZda)
Us = WZda$Us; Vs = WZda$Vs
FESAT(Us,Vs)
HESAT(Us,Vs)
RESAT(Us,Vs)
RBAT(Us,Vs)
```

## References


Davies,R.B. (1980) Algorithm AS 155: The Distribution of a Linear Combination of $\chi^2$ Random Variables. *Journal of the Royal Statistical Society. Series C (Applied Statistics),* **29** (3), 323–333.

Lee,S., Wu,M.C. and Lin,X. (2012) Optimal tests for rare variant effects in sequencing association studies. *Biostatistics,* **13** (4), 762–775.

Lee,S. (2015). MetaSKAT: Meta Analysis for SNP-Set (Sequence) Kernel Association Test. R package version 0.6.0. `http://cran.r-project.org/web/packages/MetaSKAT`

Lee,S., Miropolsky,L. and Wu,M.C. (2017). SKAT: SNP-Set (Sequence) Kernel Association Test. R package version 1.3.2.1. `http://cran.r-project.org/web/packages/SKAT`

Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D. (2005). Calibrating a coalescent simulation of human genome sequence variation. *Genome Research,* **15** (11), 1576–1583.

Voorman,A., Brody,J. and Lumley,T. (2013). skatMeta: Efficient meta analysis for the SKAT test. R package version 1.4.3. `http://cran.r-project.org/src/contrib/Archive/skatMeta`

Voorman,A., Brody,J., Chen,H., Lumley,T. and Davis,B. (2017). seqMeta: Meta-Analysis of Region-Based Tests of Rare DNA Variants. R package version 1.6.7. `http://cran.r-project.org/web/packages/seqMeta`

Wessel,J., Chu,A.Y., Willems,S.M., and others. (2015) Low-frequency and rare exome chip variants associate with fasting glucose and type 2 diabetes susceptibility. *Nature Communications,* **6**, 5897.

Wu,B., Guan,W. and Pankow,J.S. (2016) On Efficient and Accurate Calculation of Significance P-Values for Sequence Kernel Association Testing of Variant Set. *Ann. Hum. Genet.,* **80** (2), 123–135.

Wu,M.C., Lee,S., Cai,T., Li,Y., Boehnke,M. and Lin,X. (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.,* **89** (1), 82–93.