Supplementary material 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]Division 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 lets consider a single variant-set with weighted test statistics $\boldsymbol{S}$ with asymptotic covariance matrix $\boldsymbol{V}$. Define a general burden test (BT) $B = (\eta^T \boldsymbol{S})^2$, where η is a pre-given 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 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_{ρ_j} as p_j. Denote $P_0 = \min_{j\le L} p_j$, and let q_j be the $(1-T)$th quantile of the null distribution of Q_{ρ_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 $q_h = \Phi^{-1}(1-P_0/2)$, (ϕ, Φ) 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

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 $\boldsymbol{V} = \sum_{k=1}^K \boldsymbol{V}_k$. Therefore it 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 $\boldsymbol{V} = \mathrm{diag}\{\boldsymbol{V}_1, \cdots, \boldsymbol{V}_K\}$.

In summary, the FE model considers the adaptive test based on $(1-\rho)(\sum_{k=1}^{K}\boldsymbol{S}_k)^T(\sum_{k=1}^{K}\boldsymbol{S}_k)+\rho(\sum_{k=1}^{K}\eta^T\boldsymbol{S}_k)^2$, and the HE model considers the adaptive test based on $(1-\rho)\sum_{k=1}^{K}\boldsymbol{S}_k^T\boldsymbol{S}_k+\rho(\sum_{k=1}^{K}\eta_k^T\boldsymbol{S}_k)^2$.

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

Alternatively we can consider the following heterogeneous-effects model

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

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 ρ_k's, which however is too computationally intensive. As a compromise, we fix all ρ_k's to be a function of one common parameter ρ as follows. Note that for the kth 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 null, R_ρ is distributed as the weighted sum of independent χ_1^2 random variables with weights being the eigenvalues of $(1-\rho)\boldsymbol{V}+\rho\mathrm{diag}\{\nu_1\boldsymbol{V}_1^{1/2}\eta_1\eta_1^T\boldsymbol{V}_1^{1/2},\cdots,\nu_K\boldsymbol{V}_K^{1/2}\eta_K\eta_K^T\boldsymbol{V}_K^{1/2}\}$. 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 null, R_e and R_b are approximately independent and distributed as weighted sum of independent χ_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_{ρ_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_{ρ_j}. The

AT p-value can be analytically computed as

$$\begin{aligned} P &= \Pr(\min_{j\leq L} p_j \leq P_0) = \Pr(p_L \leq P_0) + \Pr(p_L > P_0, \min_{j<L} p_j \leq P_0) \\ &= P_0 + \Pr\left(R_b < q_L, R_e \geq \min_{j<L} \frac{q_j - R_b}{1-\rho_j}\right) \\ &= P_0 - \int_0^{q_L} G_e\Big(\min_{j<L} \frac{q_j - x}{1-\rho_j}\Big) dG_b(x), \end{aligned}$$

which can be computed efficiently. Here G_e, 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 perform well.

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 implement (1) FE BT: assume similar variant effects both across and within studies; (2) FE VT:

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: $\eta^T (\sum_{k=1}^{K} \boldsymbol{S}_k)$

assume similar variant effects across studies but heterogeneous effects across variants within each study; (3) RHE BT: assume heterogeneous effects across studies but similar effects across variants within each study; (4) HE VT: assume 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): combine FE BT and FE VT; (2) HE AT (HAT): combine FE BT and HE VT; (3) RHE AT (RAT): combine RHE BT and HE VT; (4) BT based AT (BAT): combine 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 ATs.

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.

!!!TABLE TBD HERE!!!

4 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 0.1 corr
K = 3; m=10
Rs = 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(0.9)+rnorm(100)*sqrt(0.1)
  Rs[,,k] = cor(ak)
  Rh = chol(Rs[,,k])
  Us[,k] = colSums(Rh*rnorm(m))
}
## FE model
FMSAT(Us,Rs)
## HE model
HMSAT(Us,Rs)
## robust HE model
RMSAT(Us,Rs)
RBAT(Us,Rs) ## AT combining RHE BT and FE BT
U1 = Us + rnorm(m*K,1,1)
FMSAT(U1,Rs)
HMSAT(U1,Rs)
RMSAT(U1,Rs)
RBAT(U1,Rs)
```

References

Davies,R.B. (1980) Algorithm AS 155: The Distribution of a Linear Combination of χ^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`

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`

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.