

A Gene Based Combination Test Using GWAS Summary Data

Jianjun Zhang, Xiaoyu Liang, Samantha Gonzales, Jianguo Liu, Xiaoyi
Raymond Gao and Xuexia Wang*

*Correspondence: Xuexia Wang, xuexwang@fiu.edu, Professor of Biostatistics,
Department of Biostatistics, Florida International University, 11200 SW 8th
Street, Miami, FL 33174, USA.

Derivation of the score test

Use the notation in the Method section. Consider a raw data set of a sample including n individuals, where each individual has been genotyped at M variants in a genomic region (gene or pathway). Denote y_i as the trait value of the i^{th} individual for either a quantitative or qualitative trait (1 for cases and 0 for controls for a qualitative trait) and denote $X_i = (x_{i1}, \dots, x_{iM})'$ as the genotypic score of the i^{th} individual, where $x_{im} \in \{0, 1, 2\}$ is the number of minor alleles that the i^{th} individual has at the m^{th} variant.

We use the generalized linear model to model the relationship between the trait (quantitative or qualitative trait) and the genetic variants in the considered region:

$$f(\mathbb{E}(y_i|X_i)) = \beta_0 + \beta_c' X_i$$

where $f(\cdot)$ is a monotone “link” function and β_c is the parameter of interest. Testing the association of the genetic variants in the considered region is equivalent to testing the effect of the weighted combination of variants $g_i = \sum_{m=1}^M w_m^0 x_{im}$. Under the generalized linear model, we can use the score test statistic to test the null hypothesis $H_0 : \beta_c = 0$, which is given by

$$\mathbf{S}(w_1^0, \dots, w_M^0) = n \frac{(\sum_{i=1}^n (y_i - \bar{y})(g_i - \bar{g}))^2}{\sum_{i=1}^n (y_i - \bar{y})^2 \sum_{i=1}^n (g_i - \bar{g})^2}$$

where the score test statistic $\mathbf{S}$ can be viewed as a function of weight $\mathbf{W}_0 = (w_1^0, \dots, w_M^0)'$. Let $\mathbf{X} = (X_1, \dots, X_n)'$, $\mathbf{Y} = (y_1, \dots, y_n)'$ and $P = I_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n'$ where $\mathbf{1}_n$ represents a column vector containing all ones. Then, we have $x_i = X_i' \mathbf{W}_0$. We can rewrite the score test as:

$$\begin{aligned} \mathbf{S}(w_1^0, \dots, w_M^0) &= n \frac{(PY)' P \mathbf{X} \mathbf{W}_0 (PY)' P \mathbf{X} \mathbf{W}_0}{(PY)' PY (P \mathbf{X} \mathbf{W}_0)' P \mathbf{X} \mathbf{W}_0} \\ &= n \frac{\mathbf{W}_0' \mathbf{X}' P \mathbf{Y} \mathbf{Y}' P \mathbf{X} \mathbf{W}_0}{\mathbf{Y}' P \mathbf{Y} \mathbf{W}_0' \mathbf{X}' P \mathbf{X} \mathbf{W}_0} \\ &= n \frac{\mathbf{W}_0' \mathbf{X}' P \mathbf{Y} \mathbf{Y}' P \mathbf{X} \mathbf{W}_0}{\mathbf{W}_0' \mathbf{X}' P \mathbf{X} \mathbf{Y}' P \mathbf{Y} \mathbf{W}_0} \end{aligned}$$

where the second equation holds because we have $P = P'$ and $PP' = P$, and the third equation holds because $\mathbf{Y}' P \mathbf{Y}$ is a constant.

To test the association between a single trait and a single variant, which we refer to as single-nucleotide polymorphism (SNP), we usually employ a Z test. To test the main effect of the m^{th} variant in the considered region, we use the Z test: $Z_m = \frac{Y' P X_m}{\sigma \sqrt{X_m' P X_m}}$ where $\sigma = \sqrt{\frac{1}{n} Y' P Y}$

, $X_{m\cdot} = (x_{m1}, \dots, x_{mn})$. Let LD matrix $\mathbf{R} = \text{diag}(\mathbf{D})^{-1/2} \mathbf{D} \text{diag}(\mathbf{D})^{-1/2}$ where $\mathbf{D} = \mathbf{X}' \mathbf{P} \mathbf{X}$ and $\text{diag}(\mathbf{D})$ denote the diagonal matrix of $\mathbf{D}$. When GWAS summary statistics such as the Z-statistics and the LD matrix for SNP-SNP correlations are available, the score test can be written as:

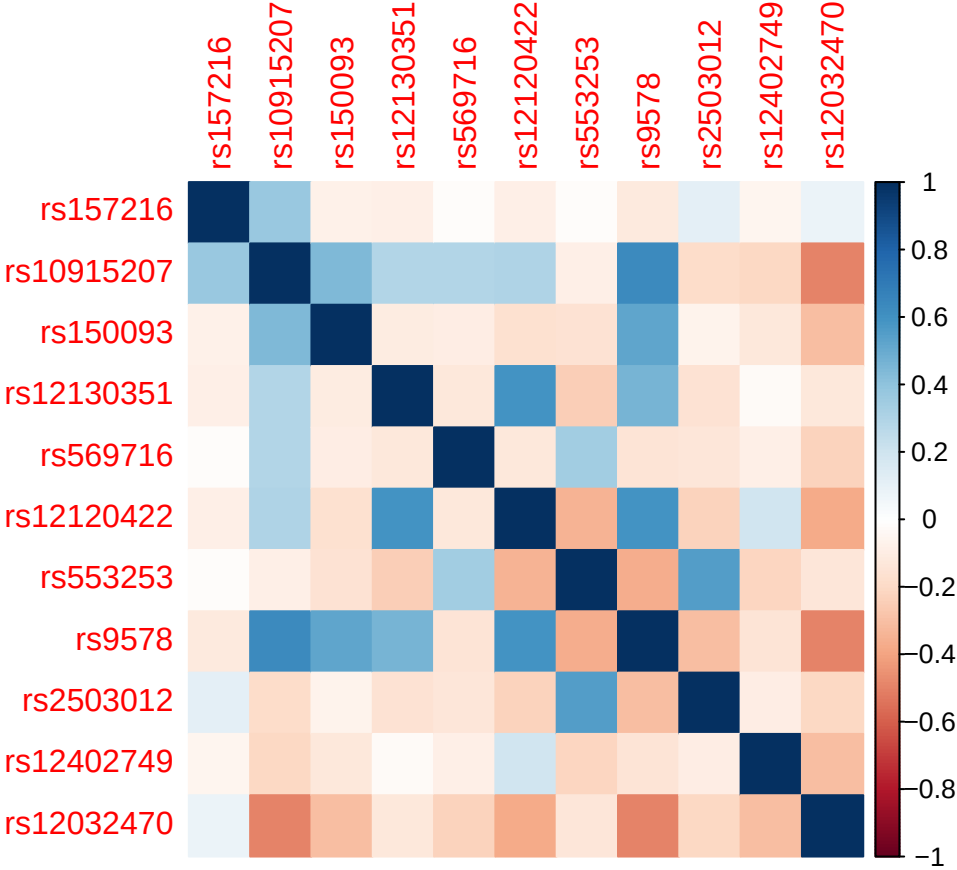
$$\begin{aligned}
& \mathbf{S}(w_1, \dots, w_M) \\
&= \frac{\mathbf{W}_0' \text{diag}(\mathbf{D})^{1/2} \text{diag}(\mathbf{D})^{-1/2} \sqrt{n} \mathbf{X}' \mathbf{P} \mathbf{Y}}{\sqrt{\mathbf{Y}' \mathbf{P} \mathbf{Y}}} \\
&= \frac{\mathbf{W}_0' \text{diag}(\mathbf{D})^{1/2} \text{diag}(\mathbf{D})^{-1/2} \mathbf{X}' \mathbf{P} \mathbf{X} \text{diag}(\mathbf{D})^{-1/2} \text{diag}(\mathbf{D})^{1/2} \mathbf{W}_0}{\sqrt{\mathbf{Y}' \mathbf{P} \mathbf{Y}}} \\
&\quad \times \frac{\sqrt{n} \mathbf{Y}' \mathbf{P} \mathbf{X} \text{diag}(\mathbf{D})^{-1/2}}{\sqrt{\mathbf{Y}' \mathbf{P} \mathbf{Y}}} \text{diag}(\mathbf{D})^{1/2} \mathbf{W}_0 \\
&= \frac{\mathbf{W}_0' \text{diag}(\mathbf{D})^{1/2} \mathbf{Z} \mathbf{Z}' \text{diag}(\mathbf{D})^{1/2} \mathbf{W}_0}{\mathbf{W}_0' \text{diag}(\mathbf{D})^{1/2} \mathbf{R} \text{diag}(\mathbf{D})^{1/2} \mathbf{W}_0} \\
&= \frac{\mathbf{W}' \mathbf{Z} \mathbf{Z}' \mathbf{W}}{\mathbf{W}' \mathbf{R} \mathbf{W}}
\end{aligned}$$

where $\mathbf{Z} = (Z_1, \dots, Z_M)'$ and $\mathbf{W} = (w_1, \dots, w_M)' = \text{diag}(\mathbf{D})^{1/2} \mathbf{W}_0$. From Equation (1), the score test statistic $\mathbf{S}$ is equivalent to a linear weighted test statistic based on Z-scores:

$$\mathbf{L}(w_1, \dots, w_M) = \sum_{m=1}^M w_m Z_m = \mathbf{W}' \mathbf{Z}$$

where $\mathbf{Z}$ follows multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\mathbf{R}$ under null hypothesis. This conclusion clearly demonstrates that testing the weighted combination of variants in a considered region using the score test is the same as using the weighted combination of Z-scores for the variants in the considered region.

Supplementary Figure S1. The linkage disequilibrium (LD) matrix with pairwise correlations for the 11 SNPs on gene EPB41.



Legends

Supplementary Figure S1. linkage disequilibrium (LD) matrix with pairwise correlations for the 11 SNPs on gene *EPB41*.

Supplementary Table S1. Significant genes identified by OWC, aSPU, GATES, sumSTAAR and GW in SCZ1 data.

Supplementary Table S2. Significant genes identified by OWC, aSPU, GATES, sumSTAAR and GW in SCZ2 data.

Supplementary Table S3. Significant genes identified by OWC, aSPU, GATES, sumSTAAR and GW in UKB data.