

# GeneGeneInteR vignette

## Statistical analysis of the interaction between a pair of genes.

Mathieu Emily

May 24, 2016

In this vignette we consider statistical procedures to test for the interaction between two genes in susceptibility with a binary phenotype, typically a case/control disease status. Let  $Y \in \{0, 1\}$  be the phenotype, where  $Y = 0$  stands for a control and  $Y = 1$  a case, and  $X_1$  and  $X_2$  be the two genes for which the interaction is tested.

Let consider a sample of  $n$  individuals with  $n_c$  controls and  $n_d$  cases ( $n_c + n_d = n$ ) and  $\mathbf{Y} = [y_1, \dots, y_n]'$  the vector of the observed binary phenotypes. Each gene is a collection of respectively  $m_1$  and  $m_2$  SNPs. The observed genotypes for gene  $X_1$  can be represented by a  $n \times m_1$  matrix:  $\mathbf{X}_1 = [x_{ij}^1]_{i \in 1 \dots n; j \in 1 \dots m_1}$  where  $x_{ij}^1 \in \{0; 1; 2\}$  is the number of copies of the minor allele for SNP  $j$  carried by individual  $i$ . A similar representation is used for gene  $X_2$  where  $\mathbf{X}_2$  is a  $n \times m_2$  matrix. Let us further introduce  $\mathbf{X}_1^c$  and  $\mathbf{X}_2^c$  the matrices of observed genotypes among controls for gene 1 and 2 and  $\mathbf{X}_1^d$  and  $\mathbf{X}_2^d$  among cases for both genes. Thus  $\mathbf{X}_1^c$  is a  $n_c \times m_1$  matrix,  $\mathbf{X}_1^d$  a  $n_d \times m_1$  matrix,  $\mathbf{X}_2^c$  a  $n_c \times m_2$  matrix and  $\mathbf{X}_2^d$  a  $n_d \times m_2$  matrix. A general setup of the observed values can be presented as follows:

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ \vdots \\ y_{n_c} \\ y_{n_c+1} \\ \vdots \\ y_{n_c+n_d} \end{bmatrix} \quad \mathbf{X}_1 = \begin{bmatrix} \mathbf{X}_1^c \\ \mathbf{X}_1^d \end{bmatrix} = \begin{bmatrix} x_{11}^1 & \dots & x_{1m_1}^1 \\ \vdots & \ddots & \vdots \\ x_{n_c1}^1 & \dots & x_{n_cm_1}^1 \\ x_{(n_c+1)1}^1 & \dots & x_{(n_c+1)m_1}^1 \\ \vdots & \ddots & \vdots \\ x_{(n_c+n_d)1}^1 & \dots & x_{(n_c+n_d)m_1}^1 \end{bmatrix} \quad \mathbf{X}_2 = \begin{bmatrix} \mathbf{X}_2^c \\ \mathbf{X}_2^d \end{bmatrix} = \begin{bmatrix} x_{11}^2 & \dots & x_{1m_2}^2 \\ \vdots & \ddots & \vdots \\ x_{n_c1}^2 & \dots & x_{n_cm_2}^2 \\ x_{(n_c+1)1}^2 & \dots & x_{(n_c+1)m_2}^2 \\ \vdots & \ddots & \vdots \\ x_{(n_c+n_d)1}^2 & \dots & x_{(n_c+n_d)m_2}^2 \end{bmatrix}$$

In our package we proposed 10 methods for testing interaction at the gene level. These 10 methods are all based on three main parameters: **Y**, a **numeric** or **factor** vector with exactly two distinct values, **G1** and **G2** two **SnpMatrix** objects as proposed by the R **Bioconductor** package **snpStats**. Our implementation is illustrated by the dataset **gene.pair** provided with the **GeneGeneInteR** package and summarized in the following command lines:

```
R> library(GeneGeneInteR)

> data(gene.pair)
> head(gene.pair$Y)

[1] HealthControl HealthControl HealthControl HealthControl HealthControl HealthControl
Levels: HealthControl RheumatoidArthritis

R> gene.pair$G1
A SnpMatrix with 453 rows and 8 columns
Row names: Id1 ... Id453
Col names: rs1491710 ... rs2298849

R> gene.pair$G2
A SnpMatrix with 453 rows and 4 columns
Row names: Id1 ... Id453
Col names: rs2057094 ... rs1005753
```

The 10 methods implemented in our package can be divided into two main families: 6 methods based on a multidimensional modeling of the interaction at the gene level and 4 methods that combine interaction tests at the SNP level into a single test at the gene level.

# 1 Multidimensional methods at the gene level

In the **GeneGeneInterR** package, 6 multidimensional methods have been implemented that are based on:

- Principal Components Analysis - `PCA.test` function,
- Canonical Correlation Analysis - `CCA.test` function,
- Kernel Canonical Correlation Analysis - `KCCA.test` function,
- Composite Linkage Disequilibrium - `CLD.test` function,
- Partial Least Square Path Modeling - `PLSPM.test` function,
- Gene-Based Information Gain Method - `GBIGM.test` function.

In the remainder of this section, technical and practical details are given regarding these 6 methods.

## 1.1 PCA-based

In the PCA-based method, a likelihood ratio test is performed to compare the model  $\mathcal{M}_{Inter}$  to the model  $\mathcal{M}_{No}$ , where  $\mathcal{M}_{Inter}$  is defined by:

$$\text{logit}(\mathbb{P}[Y = 1 | PC_{X_1}^1 \dots PC_{X_1}^{n_1}, PC_{X_2}^1 \dots PC_{X_2}^{n_2}]) = \beta_0 + \sum_{i=1}^{n_1} PC_{X_1}^i + \sum_{j=1}^{n_2} PC_{X_2}^j + \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} PC_{X_1}^i PC_{X_2}^j$$

and  $\mathcal{M}_{No}$  by:

$$\text{logit}(\mathbb{P}[Y = 1 | PC_{X_1}^1 \dots PC_{X_1}^{n_1}, PC_{X_2}^1 \dots PC_{X_2}^{n_2}]) = \beta_0 + \sum_{i=1}^{n_1} PC_{X_1}^i + \sum_{j=1}^{n_2} PC_{X_2}^j$$

In models  $\mathcal{M}_{Inter}$  and  $\mathcal{M}_{No}$ ,  $PC_{X_1}^i$  and  $PC_{X_2}^j$  are the  $i^{th}$  principal component of  $\mathbf{X}_1$  and the  $j^{th}$  principal component of  $\mathbf{X}_2$ . The number of principal components,  $n_1$  and  $n_2$ , kept in the interaction test is determined by the percentage of inertia retrieved by the PCA. Such a percentage is defined by the user and corresponds to the `threshold` parameter.

In our package, two distinct Principal Component decomposition are provided by the functions `PCA.test` via the argument `method`. With `method="Std"`, dataset is standardized using variables' standard deviation while with `method="GenFreq"`, dataset is standardized using standard deviation under Hardy-Weinberg equilibrium, as proposed in the **snpStats** package.

If model fitting fails then a PC is removed from the pool and a new attempt to fitting is made. The process loops until either the model is fitted or there are only two PC for each genes remaining. Principal components are removed according to lowest inertia percentage.

The following lines provide an example of the `PCA.test` function:

```
R> PCA.test(Y=gene.pair$Y, G1=gene.pair$G1, G2=gene.pair$G2,threshold=0.7,
method="GenFreq")
Gene-Gene Interaction method performed with:
Principal Component Analysis
Deviance = 8.215662, df = 6, p-value = 0.2227253
R> PCA.test(Y=gene.pair$Y, G1=gene.pair$G1, G2=gene.pair$G2,threshold=0.7,
method="Std")
```

Gene-Gene Interaction method performed with:  
Principal Component Analysis  
Deviance = 8.507404, df = 6, p-value = 0.2032347

## 1.2 Canonical Correlation Analysis (CCA)

The CCA based test is based on a Wald-type statistic defined as follows (see [PZX10] for details):

$$U_{CCA} = \frac{z_d - z_c}{\sqrt{\mathbb{V}(z_d) + \mathbb{V}(z_c)}}$$

where  $z_d = \frac{1}{2}(\log(1 + r_d) - \log(1 - r_d))$  and  $z_c = \frac{1}{2}(\log(1 + r_c) - \log(1 - r_c))$  with  $r_d$  the maximum canonical correlation coefficient between  $\mathbf{X}_1^d$  and  $\mathbf{X}_2^d$  and  $r_c$  the maximum canonical correlation coefficient between  $\mathbf{X}_1^c$  and  $\mathbf{X}_2^c$  computed for controls ( $Y = 0$ ). As suggested by [PZX10], the sampled variances  $\mathbb{V}(z_d)$  and  $\mathbb{V}(z_c)$  were evaluated by applying a bootstrapping method. The number of bootstrap sample used to estimate  $\mathbb{V}(z_d)$  and  $\mathbb{V}(z_c)$  is determined by the `n.boot` argument. P-value is then obtained by noting that under the null hypothesis  $U_{CCA} \sim \mathcal{N}(0, 1)$ .

CCA based gene-gene interaction is implemented in the `CCA.test` function and mainly depends on the `cancor` function from the **Stats** package [R C16].

```
R> set.seed(1234)

R> CCA.test(Y=gene.pair$Y, G1=gene.pair$G1, G2=gene.pair$G2,n.boot=500)
Gene-Gene Interaction method performed with:
      Canonical Correlation Analysis
CCU = 0.6030414, p-value = 0.5464811
```

## 1.3 Kernel Canonical Correlation Analysis (KCCA)

The KCCA based test provides a generalization of CCA test to detect non-linear co-association between  $\mathbf{X}_1$  and  $\mathbf{X}_2$  [YGH<sup>+</sup>12, LS13] and is based on the following Wald-type statistic:

$$U_{KCCA} = \frac{kz_d - kz_c}{\sqrt{\mathbb{V}(kz_d) + \mathbb{V}(kz_c)}}$$

where  $kz_d = \frac{1}{2}(\log(1 + kr_d) - \log(1 - kr_d))$  and  $kz_c = \frac{1}{2}(\log(1 + kr_c) - \log(1 - kr_c))$  with  $kr_d$  the maximum kernel canonical correlation coefficient between  $\mathbf{X}_1^d$  and  $\mathbf{X}_2^d$  and  $kr_c$  the maximum kernel canonical correlation coefficient between  $\mathbf{X}_1^c$  and  $\mathbf{X}_2^c$ .

Similar to the CCA test,  $\mathbb{V}(kz_d)$  and  $\mathbb{V}(kz_c)$  are estimated using bootstrap techniques [YGH<sup>+</sup>12, LS13] and the p-value is obtained using the standard gaussian distribution of  $U_{KCCA}$  under the null hypothesis. Since the performance of kernel methods strongly relates to the choice of kernel functions, the default is the Radial Basis kernel Function (RBF) owing to its flexibility in parameter specification. However, other kernel functions, such as linear, polynomial or spline kernels, can be used. Thus, in addition to the three arguments `Y`, `G1` and `G2`, our implementation of the KCCA test proposes two optional arguments: `n.boot` that determines the number of bootstrap samples and `kernel` that provides the kernel function to be used. This `kernel` parameter can be set to any function, of class `kernel` as provided by the **kernlab** package [KSHZ04], which computes an inner product in feature space between two vector arguments.

KCCA based gene-gene interaction test is implemented in the `KCCA.test` function and mainly depends on the `kcca` function from the **kernlab** package [KSHZ04].

```
R> set.seed(1234)

R> KCCA.test(Y=gene.pair$Y, G1=gene.pair$G1,G2=gene.pair$G2,
      kernel=rbfdot(sigma = 0.05),n.boot=500)
Gene-Gene Interaction method performed with:
      Kernel Canonical Correlation Analysis
KCCU = 1.407369, p-value = 0.1593179
```

```
R> set.seed(1234)
R> KCCA.test(Y=gene.pair$Y, G1=gene.pair$G1,G2=gene.pair$G2,
             kernel=polydot(degree = 1, scale = 1, offset = 1))
Gene-Gene Interaction method performed with:
Kernel Canonical Correlation Analysis
KCCU = -1.413508, p-value = 0.1575063
```

## 1.4 Partial Least Square Path Modeling (PLSPM)

The PLSPM testing has been introduced by [ZYY<sup>+</sup>13] and is based on the Wald-like statistic:

$$U_{PLSPM} = \frac{\beta_d - \beta_c}{\sqrt{\mathbb{V}(\beta_d - \beta_c)}}$$

where  $\beta_d$  (resp.  $\beta_c$ ) is the path coefficient between  $\mathbf{X}_1^d$  and  $\mathbf{X}_2^d$  (resp.  $\mathbf{X}_1^c$  and  $\mathbf{X}_2^c$ ). As quoted by [ZYY<sup>+</sup>13], the distribution of  $U_{PLSPM}$  is unknown and significance can be tested with bootstrapping method.

PLSPM based gene-gene interaction test is implemented in the `PLSPM.test` function and mainly depends on the `plspm` function from the `plspm` package [STR15].

```
R> set.seed(1234)
R> PLSPM.test(Y=gene.pair$Y, G1=gene.pair$G1,G2=gene.pair$G2,n.perm=1000)
Gene-Gene Interaction method performed with:
Partial Least Squares Path Modeling
U = 4.093781, p-value = 0.18
```

## 1.5 Composite Linkage Disequilibrium (CLD)

The CLD method, proposed in [RPM<sup>+</sup>12] is based on the normalized quadratic distance (NQD) and is defined as

$$\delta^2 = \text{tr.} \left( (\tilde{D} - \tilde{C})W^{-1}(\tilde{D} - \tilde{C})W^{-1} \right)$$

where  $\tilde{D}$ ,  $\tilde{C}$  and  $W$  are three  $(m_1 + m_2) \times (m_1 + m_2)$  matrices of the covariance between the whole set of SNPs that combines SNPs from both genes. More precisely,  $\tilde{D}$  and  $\tilde{C}$  are defined as follows:

$$\tilde{D} = \begin{bmatrix} W_{11} & D_{12} \\ D_{21} & W_{22} \end{bmatrix} \quad \tilde{C} = \begin{bmatrix} W_{11} & C_{12} \\ C_{21} & W_{22} \end{bmatrix}$$

where  $W_{11}$  (resp.  $W_{22}$ ) is the pooled estimate of the covariance matrix for  $\mathbf{X}_1$  (resp.  $\mathbf{X}_2$ ,  $D_{12}(= D'_{21})$  and  $C_{12}(= C'_{21})$  are the sample covariance matrix between the two genes estimated from  $(\mathbf{X}_1^d, \mathbf{X}_2^d)$  and  $(\mathbf{X}_1^c, \mathbf{X}_2^c)$  respectively. In more details, the sample covariance matrices in cases, denoted by  $D$ , and in controls, denoted by  $C$ , can be partitioned in 4 blocks as follows:

$$D = \text{Cov}(\mathbf{X}_1^d, \mathbf{X}_2^d) = \begin{bmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{bmatrix} \quad C = \text{Cov}(\mathbf{X}_1^c, \mathbf{X}_2^c) = \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix}$$

The pooled estimate of the covariance matrix,  $W$ , can thus been obtained by:

$$W = \frac{n_c C + n_d D}{n_c + n_d} = \begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix}$$

Since the distribution of  $\delta^2$  is not known under the null hypothesis, the significance levels of this statistics is established using permutation tests, as proposed by [RPM<sup>+</sup>12]. Such a test has been implemented in our package in the `CLD.test` function where the number of permutations is determined by the argument `n.perm`.

```
R> set.seed(1234)
R> CLD.test(Y=gene.pair$Y, G1=gene.pair$G1,G2=gene.pair$G2,n.perm=2000)
Gene-Gene Interaction method performed with:
Composite Linkage Disequilibrium
CLD = 0.4925654, p-value = 0.8865
```

## 1.6 Gene-Based Information Gain Method (GBIGM)

Introduced by [LHG<sup>+</sup>15], the GBIGM method is based on the information gain rate  $\Delta R_{1,2}$ .  $\Delta R_{1,2}$  is defined as follows:

$$\Delta R_{1,2} = \frac{\min(H_1, H_2) - H_{1,2}}{\min(H_1, H_2)}$$

where  $H_1$ ,  $H_2$ ,  $H_{1,2}$  are the conditional entropies, given the  $\mathbf{Y}$ , of  $\mathbf{X}_1$ ,  $\mathbf{X}_2$  and the pooled SNP set  $(\mathbf{X}_1, \mathbf{X}_2)$  respectively. Assuming that  $H(\cdot)$  is the classical entropy function, we have:

$$\begin{aligned} H_1 &= H(\mathbf{Y}, \mathbf{X}_1) - H(\mathbf{X}_1) \\ H_2 &= H(\mathbf{Y}, \mathbf{X}_2) - H(\mathbf{X}_2) \\ H_{1,2} &= H(\mathbf{Y}, \mathbf{X}_1, \mathbf{X}_2) - H(\mathbf{X}_1, \mathbf{X}_2) \end{aligned}$$

Since the distribution of  $\Delta R_{1,2}$  is unknown, the significance testing is performed by permutations as suggested by [LHG<sup>+</sup>15]. The GBIGM method has been implemented in the `GBIGM.test` function and the number of permutations is defined by the argument `n.perm`.

```
R> set.seed(1234)

R> GBIGM.test(Y=gene.pair$Y, G1=gene.pair$G1, G2=gene.pair$G2, n.perm=2000)
Gene-Gene Interaction method performed with:
  Gene-based Information Gain Method
DeltaR1,2 = 0.4644093, p-value = 0.441
```

## 2 From SNP-SNP interaction to Gene-Gene interaction testing

This section provides details of the four statistical methods that proposes a gene-based test from SNP-based tests [Emi16]. Rather than considering multiple SNPs in both gene as part of a joint model, these methods aim at aggregating p-values obtained at the SNP level into a single p-value at a gene level.

### Interaction testing at the SNP level

Let consider a pair of SNPs,  $(X_{1,j}, X_{2,k})$  where  $X_{1,j}$  is the  $j^{\text{th}}$  SNP of gene  $X_1$  and  $X_{2,k}$  the  $k^{\text{th}}$  SNP of gene  $X_2$  ( $1 \leq j \leq m_1$  and  $1 \leq k \leq m_2$ ). To test for interaction at the SNP level, we used the following Wald statistic:

$$W_{jk} = \frac{\widehat{\beta_3^{j,k}}}{\sigma\left(\widehat{\beta_3^{j,k}}\right)}$$

where  $\widehat{\beta_3^{j,k}}$  is an estimate of the interaction coefficient  $\beta_3^{j,k}$  of the following logistic model:

$$\log\left(\frac{\mathbb{P}[Y = 1 | X_{1,j} = x_1, X_{2,k} = x_2]}{1 - \mathbb{P}[Y = 1 | X_{1,j} = x_1, X_{2,k} = x_2]}\right) = \beta_0^{j,k} + \beta_1^{j,k} x_1 + \beta_2^{j,k} x_2 + \beta_3^{j,k} x_1 x_2$$

$\widehat{\beta_3^{j,k}}$  is obtained by maximizing the likelihood function on the observed data  $\mathbf{Y}$ ,  $\mathbf{X}_1$  and  $\mathbf{X}_2$  while  $\sigma\left(\widehat{\beta_3^{j,k}}\right)$  is calculating by inverting the Hessian of the likelihood. Since the solution of the maximization of the likelihood function does not have a closed form, we compute  $W_{jk}$  according to the iteratively reweighted least squares algorithm proposed in the `glm` function of the **stats** package [R C16].

To combine the statistics  $W_{jk}$  into a single test, [MCK13] proposed four methods that all account for covariance matrix  $\Sigma = [\sigma_{(j,k),(j',k')}]_{\substack{j=1\dots m_1; k=1\dots m_2 \\ j'=1\dots m_1; k'=1\dots m_2}}$ , a  $(m_1 \times m_2) \times (m_1 \times m_2)$  symmetric matrix where  $\sigma_{(j,k),(j',k')} = \text{Cov}(W_{jk}, W_{j',k'})$ . As proposed by [Emi16], the covariance between  $W_{jk}$  and  $W_{j',k'}$  is estimated by:

$$\sigma_{(j,k),(j',k')} = r_{j,j'} r_{k,k'}$$

where  $r_{j,j'} = \frac{p_{jj'} - p_j p_{j'}}{\sqrt{p_j(1-p_j)p_{j'}(1-p_{j'})}}$  is the widely used correlation measure between SNP  $j$  and SNP  $j'$ , given that  $p_j$  and  $p_{j'}$  are the respective allelic frequencies and  $p_{jj'}$  is the joint allelic frequency [HR68].

In the remainder of this section, the four methods: minP (function `minP.test`), GATES (function `gates.test`), tTS (function `tTS.test`) and tProd (function `tProd.test`) are detailed.

## 2.1 minP

The minP test is based on the minimum p-value that is often used to combine p-values of association [CB07]. Let  $W_{\max} = \max(|W_{11}|, \dots, |W_{m_1, m_2}|)$  be the maximum of the absolute observed statistics. The minP is then defined by:

$$\text{minP} = 1 - \mathbb{P}\left[\max(|Z_1|, |Z_2|, \dots, |Z_{m_1 m_2}|) < W_{\max}\right]. \quad (1)$$

where  $\mathbb{Z} = (Z_1, Z_2, \dots, Z_{m_1 m_2})$  is a random vector that follows a multivariate normal distribution  $\mathbb{Z} \sim \mathcal{N}(\mathbf{0}, \Sigma)$ .

The computation of Equation (1) requires the calculation of the probability distribution of a multivariate normal random variable. For that purpose, we used the `pmvnorm` function from the R package `mvtnorm` [GB09].

```
R> set.seed(1234)

R> minP.test(Y=gene.pair$Y, G1=gene.pair$G1, G2=gene.pair$G2)
Gene-Gene Interaction method performed with:
  minP
minP = 0.009924148, p-value = 0.1795672
```

## 2.2 GATES

The GATES procedure, proposed by [LGKS11], is an extension of the Simes procedure used to assess the gene level association significance. Let  $p_{(1)}, \dots, p_{(m_1 m_2)}$  be the ascending SNP-SNP interaction  $m_1 \times m_2$  p-values, GATES p-value is then defined by

$$p_{GATES} = \min\left(\frac{mep_{(1)}}{me_{(1)}}, \frac{mep_{(2)}}{me_{(2)}}, \dots, \frac{mep_{(m_1 m_2)}}{me_{(m_1 m_2)}}\right)$$

where  $m_e$  is the number of effective tests among the  $m_1 \times m_2$  tests and  $me_{(i)}$  the number of effective tests among the  $i$  most significant tests associated with the lowest order p-values  $p_{(1)}, \dots, p_{(i)}$ . The number of effective ought to characterize the number of independent tests equivalent to the correlated tests that really performed and is often used to account for dependence in a multiple testing correction.

Although no formal definition of the number of effective tests has been formulated in the literature, several procedures have been proposed to estimate such number. All methods are based on a transformation of the set of eigenvalues of the SNP covariance matrix assuming that (1) if the SNPs are independent, the number of effective tests is the number of performed, (2) if the absolute value of the correlation between any pair of SNPs is equal to 1, the number of effective tests is 1. In the **GeneGeneInteR** package, four main methods have been implemented and can be chosen by the user with the argument `merest`: Cheverud-Nyholt method - `me.est="ChevNy"` [Che01, Nyh04], Keff method - `me.est="Keff"` [MS08], Li and Ji method - `me.est="LiJi"` [LJ05] and Galwey - `me.est="Galwey"` [Gal09].

```
R> set.seed(1234)

R> gates.test(Y=gene.pair$Y, G1=gene.pair$G1, G2=gene.pair$G2, me.est="ChevNy")
Gene-Gene Interaction method performed with:
  GATES
GATES = 0.009924148, p-value = 0.293932

R> set.seed(1234)
```

```

R> gates.test(Y=gene.pair$Y, G1=gene.pair$G1,G2=gene.pair$G2,alpha=0.05,me.est="Keff")
Gene-Gene Interaction method performed with:
      GATES
GATES = 0.01394543, p-value = 0.1899414

R> set.seed(1234)

R> gates.test(Y=gene.pair$Y, G1=gene.pair$G1,G2=gene.pair$G2,me.est="LiJi")
Gene-Gene Interaction method performed with:
      GATES
GATES = 0.01394543, p-value = 0.1255088

R> set.seed(1234)

R> gates.test(Y=gene.pair$Y, G1=gene.pair$G1,G2=gene.pair$G2,me.est="Galwey")
Gene-Gene Interaction method performed with:
      GATES
GATES = 0.01394543, p-value = 0.1596024

```

## 2.3 tTS and tProd

tTS and tProd procedures are two truncated tail strength methods that aim at combining signals from all single-SNP p-values less than a predefined cutoff value [JZZK11]. Denoting by  $\tau$  the cutoff value, the two truncated p-values are defined as follows [ZZWW02]:

$$\begin{aligned}
tTS &= \frac{1}{m_1 m_2} \sum_{i=1}^{m_1 m_2} \mathbb{I}(p_{(i)} < \tau) \left( 1 - p_{(i)} \frac{m_1 m_2 + 1}{i} \right) \\
tProd &= \prod_{i=1}^{m_1 m_2} p_i^{\mathbb{I}(p_i < \tau)}
\end{aligned}$$

where  $\mathbb{I}$  is the indicator function.

When p-values are correlated, the null distribution of  $tTS$  and  $tProd$  are unknown. Following the approach proposed by [ZZWW02], a p-value is obtained in the **GeneGeneInteR** package by computing an empirical null distribution using Monte-Carlo (MC) simulations. For each MC iteration, an empirical value for  $tTS$  (or  $tProd$ ) is obtained by simulating a vector of  $W_{jk}$  with respect to a multivariate normal distribution with a vector of 0 means and  $\hat{\Sigma}$  as covariance matrix. The empirical p-value is calculated as the proportion of simulated statistics larger than the observed statistic on the “true” set of  $W_{jk}$ .

tTS and tProd methods have been implemented in the functions `tTS.test` and `tProd.test` of the **GeneGeneInteR** package. Additional to the mandatory `Y`, `G1` and `G2` arguments, these two functions have two optional arguments: `tau` and `n.sim` that control the cutoff value and the number of simulations used to estimate the empirical value respectively. The following coding lines give an example of the `tTS.test` and `tProd.test`:

```

R> set.seed=1234

R> tTS.test(Y=gene.pair$Y, G1=gene.pair$G1,G2=gene.pair$G2,tau=0.5,n.sim=10000)
Gene-Gene Interaction method performed with:
      Truncated Tail Strength
tTS = -0.009912706, p-value = 0.5104

R> set.seed(1234)

R> tProd.test(Y=gene.pair$Y, G1=gene.pair$G1,G2=gene.pair$G2,tau=0.05,n.sim=1000)
Gene-Gene Interaction method performed with:
      Truncated Product
tProd = 3.69877e-12, p-value = 0.485

```

## References

- [CB07] Karen N. Conneely and Michael Boehnke. So many correlated tests, so little time! rapid adjustment of p values for multiple correlated tests. *The American Journal of Human Genetics*, 81(6):1158–1168, 2007.
- [Che01] James M Cheverud. A simple correction for multiple comparisons in interval mapping genome scans. *Heredity*, 87(1):52–58, 07 2001.
- [Emi16] Mathieu Emily. Aggregator: A gene-based gene-gene interaction test for case-control association studies. *Statistical Applications in Genetics and Molecular Biology*, 15(2):151–171, 2016.
- [Gal09] Nicholas W. Galwey. A new measure of the effective number of tests, a practical tool for comparing families of non-independent significance tests. *Genetic Epidemiology*, 33(7):559–568, 2009.
- [GB09] Alan Genz and Frank Bretz. *Computation of Multivariate Normal and T Probabilities*. Springer Publishing Company, Incorporated, 1st edition, 2009.
- [HR68] W. G. Hill and A. Robertson. Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics*, 38:226–231, 1968.
- [JZZK11] Bo Jiang, Xiao Zhang, Yijun Zuo, and Guolian Kang. A powerful truncated tail strength method for testing multiple null hypotheses in one dataset. *Journal of Theoretical Biology*, 277(1):67 – 73, 2011.
- [KSHZ04] Alexandros Karatzoglou, Alex Smola, Kurt Hornik, and Achim Zeileis. kernlab – an S4 package for kernel methods in R. *Journal of Statistical Software*, 11(9):1–20, 2004.
- [LGKS11] Miao-Xin Li, Hong-Sheng Gui, JohnnyÊS.H. Kwan, and PakÊC. Sham. Gates: A rapid and powerful gene-based association test using extended simes procedure. *The American Journal of Human Genetics*, 88(3):283–293, 2011.
- [LHG<sup>+</sup>15] J. Li, D. Huang, M. Guo, X. Liu, C. Wang, Z. Teng, R. Zhang, Y. Jiang, H. Lv, and L. Wang. A gene-based information gain method for detecting gene-gene interactions in case-control studies. *European Journal of Human Genetics*, Online:Online, 03 2015.
- [LJ05] J. Li and L. Ji. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity*, 95:221–227, 2005.
- [LS13] Nicholas B. Larson and Daniel J. Schaid. A kernel regression approach to gene-gene interaction detection for case-control studies. *Genetic Epidemiology*, 37(7):695–703, 2013.
- [MCK13] Li Ma, Andrew G. Clark, and Alon Keinan. Gene-based testing of interactions in association studies of quantitative traits. *PLoS Genet*, 9(2):e1003321, 02 2013.
- [MS08] Valentina Moskvina and Karl Michael Schmidt. On multiple-testing correction in genome-wide association studies. *Genetic Epidemiology*, 32(6):567–573, 2008.
- [Nyh04] Dale R. Nyholt. A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. *The American Journal of Human Genetics*, 74(4):765 – 769, 2004.
- [PZX10] Qianqian Peng, Jinghua Zhao, and Fuzhong Xue. A gene-based method for detecting gene-gene co-association in a case-control association study. *European Journal of Human Genetics*, 18(5):582–587, 05 2010.
- [R C16] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2016.
- [RPM<sup>+</sup>12] Indika Rajapakse, Michael D. Perlman, Paul J Martin, John A. Hansen, and Charles Kooperberg. Multivariate detection of gene-gene interactions. *Genetic Epidemiology*, 36(6):622–630, 2012.



- [STR15] Gaston Sanchez, Laura Trinchera, and Giorgio Russolillo. *plspm: Tools for Partial Least Squares Path Modeling (PLS-PM)*, 2015. R package version 0.4.7.
- [YGH<sup>+</sup>12] Zhongshang Yuan, Qingsong Gao, Yungang He, Xiaoshuai Zhang, Fangyu Li, Jinghua Zhao, and Fuzhong Xue. Detection for gene-gene co-association via kernel canonical correlation analysis. *BMC Genetics*, 13(1):83, 2012.
- [ZYY<sup>+</sup>13] Xiaoshuai Zhang, Xiaowei Yang, Zhongshang Yuan, Yanxun Liu, Fangyu Li, Bin Peng, Dianwen Zhu, Jinghua Zhao, and Fuzhong Xue. A plspm-based test statistic for detecting gene-gene co-association in genome-wide association study with case-control design. *PLoS ONE*, 8(4):e62129, 04 2013.
- [ZZWW02] D.V. Zaykin, Lev A. Zhivotovsky, P.H. Westfall, and B.S. Weir. Truncated product method for combining p-values. *Genetic Epidemiology*, 22(2):170–185, 2002.