Supplement of “Efficiently Finding Genome-wide Three-way Gene Interactions from Transcript- and Genotype-Data”

Mitsunori Kayano 1, Ichigaku Takigawa 1, Motoki Shiga 1, Koji Tsuda 2 and Hiroshi Mamitsuka 1*

1Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji 611-0011, Japan.
2Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology (AIST), 2-42 Aomi, Koto-ku, Tokyo 135-0064 Japan.

Fig. 1. Synthetic examples: Expressions of two genes under the three genotypes of another gene.

1 INTRODUCTION

In this supplement, we describe the following points in detail: 1) The proposed method, particularly the pseudocode of each part of our method, and 2) Experimental results, particularly the result obtained by using the GEO (Gene Expression Omnibus) database (Barrett et al., 2007) extensively.

2 METHODS

2.1 Notations and Preliminaries

Let \( X \) be an input matrix, in which each row is an individual and each column is a numerical vector of gene expressions or a categorical vector of SNPs (in genes). Let \( E \) be the set of genes for which expressions are measured in \( X \) and \( Q \) be the set of SNPs in \( X \), indicating that \(|E| + |Q|\) is the total number of columns of \( X \). To test the three-way interaction, we choose one combination, i.e., two genes \((c_1, e_1)\) and one SNP \((q)\) out of \( E \) and \( Q \), respectively, and we write \( X(c_1, e_1, q) \) which has only three columns of \( X \), corresponding to \( e_1, e_2 \) and \( q \) (we write \( X(c, q) \) when we choose only one gene \( e \) out of \( E \) and \( q \) out of \( Q \)). Hereafter until Section 2.6, we assume that we already choose one combination.

For gene expressions, let \( X = (X_1, \ldots, X_K)' \in \mathbb{R}^K \) be a \( K \)-dimensional numerical variable, taking value \( x = (x_1, \ldots, x_K)' \). We note that using two genes in expressions does not necessarily means \( K = 2 \). For example, for two genes, we can set \( K = 3 \), where \( X_1, X_2 \) and \( X_3 \) correspond to one gene, the other gene and the interaction between these two genes, respectively. For genotypes, let \( C \) be the number of groups (or classes), and in fact, \( C = 3 \). We denote three genotypes by \( G_1, G_2 \) and \( G_3 \), into one of which each individual falls. Let \( Y \) be the class variable, taking value \( y \), where \( Y = (Y_1, Y_2)' \in \{0, 1\} \times \{0, 1\}. \) Here we note that \( y \) takes the following values: \( y = (1, 0)' \) if \( x \in G_1 \), \( y = (0, 1)' \) if \( x \in G_2 \) and \( y = (0, 0)' \) if \( x \in G_3 \). We denote \( N \) inputs (individuals) by \( X = (x_1, \ldots, x_N)' \) and \( Y = (y_1, \ldots, y_N)' = (y(1), y(2)) \), which can be classified into \( N_1, N_2 \) and \( N_3 \) inputs for \( G_1, G_2 \) and \( G_3 \), respectively. The average expression values can be defined for each class \( c \) and all classes: \( x_c = \frac{1}{N_c} \sum_{j|y_j = c} x_j \), \( \bar{x} = \frac{1}{N} \sum_{j=1}^{N} x_j \), respectively, where \( \bar{x} = 1/N \sum_{c=1}^{C} N_c x_c \). \( I_K \) is the identity matrix of size \( K \), and \( I \) is an \( n \)-dimensional vector in which all elements are 1.

We incorporate some basic statistics \( T \) and \( B \) by:

\[
T = \sum_{j=1}^{N} (x_j - \bar{x})(x_j - \bar{x})',
\]

\[
B = \frac{C}{N} \sum_{c=1}^{C} N_c (x_c - \bar{x})(x_c - \bar{x})',
\]

\[
W = \frac{1}{N} \sum_{j=1}^{N} \sum_{j|y_j = c} (x_j - \bar{x}_c)(x_j - \bar{x}_c)',
\]

where \( T = B + W \). We can further define covariance matrix \( S_c \) for class \( c \) and total covariance matrices \( S \) and \( S_T \) as follows:

\[
S_c = \frac{1}{N_c} \sum_{j=1|y_j = c}^{N_c} (x_j - \bar{x}_c)(x_j - \bar{x}_c)' (c = 1, \ldots, C),
\]

\[
S = \frac{1}{N} \sum_{c=1}^{C} N_c \sum_{j=1|y_j = c}^{N_c} (x_j - \bar{x}_c)(x_j - \bar{x}_c)' (= \frac{1}{N} W),
\]

\[
S_T = \frac{1}{N} \sum_{c=1}^{C} \sum_{j=1|y_j = c}^{N_c} (x_j - \bar{x}_c)(x_j - \bar{x}_c)' (= \frac{1}{N} T).
\]

We note that \( W = \sum_{c=1}^{C} N_c S_c \) and \( S = \frac{1}{N} \sum_{c=1}^{C} N_c S_c. \)
We explain the multivariate normal distribution, which will be used in our approach. This distribution has two parameters, \( \mu \) and \( \Sigma \), which are the means and the covariance matrix of class \( c \), and the density function of this distribution can be given as follows:

\[
f(X|\mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)}
\]

which is also the likelihood function, and the log-likelihood function \( \ell(\mu, \Sigma|X) \) is given as follows:

\[
\ell(\mu, \Sigma|X) = -\frac{N}{2} \log(2\pi) - \frac{1}{2} (x-\mu)^T \Sigma^{-1} (x-\mu) - \frac{1}{2} \log|\Sigma|
\]

From this equation, we can see that \( x \), covariance matrix \( \Sigma \), and covariance matrix \( S \) can be the maximum likelihood estimators of \( \mu \), \( \Sigma \), and \( \Sigma = \Sigma_1 = \cdots = \Sigma_c \), respectively.

We briefly describe likelihood ratio test, which will be used. We first assume that \( x_1, x_2, \ldots, x_n \) are generated according to parameter vector \( \theta \). Let \( H_0 : \theta \in \Theta_0 \) be a null hypothesis and \( H_1 : \theta \in \Theta_1 \) be the alternative hypothesis. The likelihood ratio statistic \( \lambda \) for testing \( H_0 \) against \( H_1 \) can be defined as follows:

\[
\lambda = \frac{L_0}{L_1},
\]

where \( L_0 \) and \( L_1 \) are the maximum likelihoods under \( \theta \in \Theta_0 \) and \( \theta \in \Theta_1 \), respectively. We note that the following can be used instead of Eq.(3):

\[
-2 \log \lambda = 2(\ell_1^0 - \ell_0^0),
\]

where \( \ell_0^0 \) and \( \ell_1^0 \) are the maximum log-likelihoods under \( \theta \in \Theta_0 \) and \( \theta \in \Theta_1 \), respectively. We note that the following theorem holds regarding the asymptotic distribution of the likelihood ratio statistic.

**Theorem 2.1** (Mardia et al. (1979)). If \( \Theta_0 \) is a region in \( \mathbb{R}^d \), and if \( \Theta_0 \) is an r-dimensional subregion of \( \Theta_1 \), then under suitable regularity conditions, for each \( \theta \in \Theta_0 \), \(-2 \log \lambda\) has an asymptotic \( \chi^2_{d-r} \) distribution as \( N \to \infty \).

Here \( q - r \) is the degree of freedom (df) of the \( \chi^2 \) distribution.

### 2.2 Finding Three-way Interactions: Interaction Test (Likelihood Ratio Test of Logistic Regression)

A standard and exact approach for our problem is likelihood ratio test of logistic regression (McCullagh and Nelder, 1989), which we simply call interaction test.

#### 2.2.1 Logistic Regression

We first denote the parameter that \( \alpha \) is in \( G_1 \) by \( p_1(\alpha) \), and similarly the probability that \( \alpha \) is in \( G_2 \) by \( p_2(\alpha) \), by which the probability that \( \alpha \) is in \( G_1 \) is \( p_3(\alpha) = 1 - p_1(\alpha) - p_2(\alpha) \). We use logistic regression to link these probabilities to \( R \)-dimensional input \( \alpha \) by using weight parameters (or coefficients) \( \omega = (\omega_1, \omega_2, \cdots, \omega_R)^T \), where \( \omega_1 = (\omega_{10}, \omega_{11}, \cdots, \omega_{1M}) \), \( \omega_2 = (\omega_{20}, \omega_{21}, \cdots, \omega_{2M}) \) as follows:

\[
\begin{align*}
p_1(\alpha) &= \frac{1}{1 + \exp(\omega_1^T \alpha)} \\
p_2(\alpha) &= \frac{1}{1 + \exp(\omega_2^T \alpha)}
\end{align*}
\]

Here we denote \( p_1(\alpha), p_2(\alpha) \) and \( p_3(\alpha) \) by \( p_1(\alpha; \omega), p_2(\alpha; \omega) \) and \( p_3(\alpha; \omega) = (1 - p_1(\alpha; \omega) - p_2(\alpha; \omega)) \), respectively, because they can be functions of \( \alpha \). We can then write the likelihood of logistic regression for

| Table 1. Log-likelihoods by LDA and Newton-Raphson for logistic regression and \( 2(\ell(w) - \ell(\hat{w}_0)) \) (LLR) with \( p \)-value in parentheses |
|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| \( \ell(\hat{w}_0) \) | \( \ell(w) \) | \( \ell(\hat{w}_0) \) | \( \ell(w) \) | \( \ell(\hat{w}_0) \) | \( \ell(w) \) | (LLR) (p-value)         |
|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|-----------------|
| a)              | -567.2          | -196.4          | -195.5          | -195.5          | -194.4          | 2.23 (0.45)     |
| b)              | -21.2           | -1.86           | -1.56           | -0.42           | -2.36           | -3.87 (1.00)    |
| c)              | -305.1          | -83.5           | -18.3           | -1.52           | -6.00           | -8.97 (1.00)    |
| d)              | -197.8          | -197.4          | -197.4          | -126.4          |                | 142.12 (0.00)   |

where \( \gamma = (\gamma_1, \gamma_2) \).

#### 2.2.2 Parameter Estimation

We can obtain the maximum likelihood estimator for \( \omega \) by maximizing the log-likelihood \( \ell(w) = \log L(w) \). A standard approach for this purpose is the Newton-Raphson method, which is an iterative gradient descent, having the following updating rule by which we can have \( \omega^{(t+1)} \) at the \( (t+1) \)-th iteration, using \( \omega^{(t)} \) of the \( t \)-th iteration:

\[
\omega^{(t+1)} = \omega^{(t)} - \left( H(\omega^{(t)}) \right)^{-1} U(\omega^{(t)})
\]

where Hessian matrix \( H(w) = \frac{\partial^2 L(w)}{\partial \omega \partial \omega^T} \) and gradient vector \( U(\omega) = \partial L(\omega)/\partial \omega \) can be given in the following:

\[
U(\omega) = X^T a(w),
\]

where \( X = \text{diag}(XX) \) (diagonal matrix of \( X^T X \)), \( a(w) = (a_1(w)^T, a_2(w)^T)^T \) where \( a_i(w) = y_j^2 - p_j(w) \) and \( p_j(w) = p_j(x_1; w, \cdots, p_j(x_n; w))^T \) \((j = 1, 2)\).

\[
H(\omega) = (X' R_{11}(w) X X' R_{12}(w) X X' R_{22}(w) X X' R_{21}(w) X X' R_{12}(w) X X' R_{11}(w) X)^T = X_s^T \hat{R}(w) X_s,
\]

where \( N \times N \) matrix \( R_{jk}(w) (j, k = 1, 2) \) is given by \( R_{jk}(w) = \text{diag}(p_j(w) \cup p_k(w) - 1) \) and \( \hat{R}(w) = \text{diag}(p_j(w) \cup p_k(w)) \) \((j \neq k)\).

Finally, the updating rule of the Newton-Raphson method for logistic regression can be rewritten in the following:

\[
\omega^{(t+1)} = \omega^{(t)} - \{X_s(R(\omega^{(t)})) X_s^T\}^{-1} X_s \hat{R}(\omega^{(t)}),
\]

In practice, we start with some initial values \( \omega^{(0)} \) and update \( \omega^{(t+1)} \) according to Eq.(7) until the following equation is satisfied:

\[
\|\omega^{(t+1)} - \omega^{(t)}\|^2 < 2K\delta,
\]

where \( \delta \) is set at a certain value.
Input: $X(e_1, e_2, q)$: Input three vectors of genes $e_1, e_2$ and SNP $q$.

Output: One if $e_1$ and $e_2$ are interacting with each other under $q$; otherwise zero.

Interaction Test($e_1$, $e_2$, $q$, $\alpha$)
1: $\omega_0$ ← some initial value.
2: repeat
3: Update $\omega_0$, according to the iterative rule of Eq.(7)
4: until Eq.(8) is satisfied
5: $\omega$ ← some initial value.
6: repeat
7: Update $\omega$, according to the iterative rule of Eq.(7)
8: until Eq.(8) is satisfied
9: if $-2(\ell(\omega) - \ell(\omega_0)) > \chi^2_2(\alpha_i)$ then
10: return 1
11: else
12: return 0
13: end if

Fig. 2. Pseudocode of interaction test.

Fig. 3. The likelihood ratio between likelihoods with and without the interaction term.

Then the test statistic of the likelihood ratio test and its asymptotic distribution can be given as follows:

$$-2 \log \lambda = 2(\ell(\omega) - \ell(\omega_0)) \sim \chi^2_2(\alpha_i).$$

where $\chi^2_2(\alpha_i)$ is the $\chi^2$ distribution with the df of two, meaning that interacting genes can be obtained as those which have lower $p$-values under this distribution than the input significance level $\alpha_i$. We run interaction test 100 times over four examples in Fig. 1, and the last three columns of Table 1 show the average results over the 100 runs. This table clearly shows that the $p$-value is very large for each of (a)-(c) of Fig. 1, while that is zero for (d), indicating that this test can detect our target sample correctly.

Fig. 2 shows a pseudocode of interaction test. A significant drawback of this approach is practical computation time. First, Eq. (9) shows $K = 8$, meaning that Newton-Raphson needs to compute an $8 \times 8$ inverse-matrix at each iterative step. We then have to conduct two iterative procedures until convergence, as shown in Fig. 2. In fact, as will be shown in our experiments, it took more than 24 hours to finish interaction test over only $10^5$ combinations ($=1,000$ SNPs $\times$ 100 genes $\times$ 100 genes). Thus we need devices to avoid running interaction test over all given combinations.

2.3 Key Idea for Speeding-up Interaction Finding

A basic idea for accelerating finding three-way interactions is to prune some combinations, to which interaction test do not have to be applied. From Eq. (10), we can see that the interacting genes should have a larger log-likelihood ratio. Fig. 3 shows a schematic figure, in which we plot the log-likelihood without the interaction term in the left-hand side and that with the interaction term in the right-hand side. We note that the range of the log-likelihood can be limited, because the maximum log-likelihood is zero and the minimum log-likelihood can be given by the case of the uniform distribution for $p_i(x)$. The log-likelihood ratio in question can be then given by the distance between these two plots which is parallel to the vertical axis, being shown by a dotted line in the figure. Thus two interacting genes should have a long dotted line, meaning that the point in the left-hand side should be lower and that in the right-hand side should be higher. This observation indicates that we can prune the following two cases: I) We already have a large likelihood without the interaction term, and II) We have only a small likelihood even if we use the interaction term. These I) and II) correspond to Areas I and II, respectively, of Fig. 3. We then attempt to efficiently detect examples in Areas I and II by assuming the normality on data distribution.

2.4 Linear Discriminant Analysis (LDA)

Area I of Fig. 3 contains examples in which expressions can be easily separated into three genotypes without the interaction term, as shown in Fig. 1 (b)-(d). Thus this case, we can consider a simpler, easily-computable estimation method for parameters of the same logistic regression model without an interaction term, and if the likelihood for a given three-way combination is zero, then we can consider a simpler, easily-computable estimation method, by the distance between these two plots which is parallel to the vertical axis, being shown by a dotted line in the figure.

Then the test statistic of the likelihood ratio test and its asymptotic distribution can be given as follows:

$$-2 \log \lambda = 2(\ell(\omega) - \ell(\omega_0)) \sim \chi^2_2(\alpha_i).$$

where $\chi^2_2(\alpha_i)$ is the $\chi^2$ distribution with the df of two, meaning that interacting genes can be obtained as those which have lower $p$-values under this distribution than the input significance level $\alpha_i$. We run interaction test 100 times over four examples in Fig. 1, and the last three columns of Table 1 show the average results over the 100 runs. This table clearly shows that the $p$-value is very large for each of (a)-(c) of Fig. 1, while that is zero for (d), indicating that this test can detect our target sample correctly.

Fig. 2 shows a pseudocode of interaction test. A significant drawback of this approach is practical computation time. First, Eq. (9) shows $K = 8$, meaning that Newton-Raphson needs to compute an $8 \times 8$ inverse-matrix at each iterative step. We then have to conduct two iterative procedures until convergence, as shown in Fig. 2. In fact, as will be shown in our experiments, it took more than 24 hours to finish interaction test over only $10^5$ combinations ($=1,000$ SNPs $\times$ 100 genes $\times$ 100 genes). Thus we need devices to avoid running interaction test over all given combinations.
Input: \( X(e_1, e_2, q) \): Input three vectors of genes \( e_1, e_2 \) and SNP \( q \). \( \alpha_i \): Significance level for interaction test

Output: One if the likelihood by LDA is high enough; otherwise zero.

LDA\( (e_1, e_2, q, \alpha_i) \):
1: Estimate means and covariances according to Eqs.\( (12) \)
2: Compute \( l(\hat{\theta}_0) \) using Eq.\( (11) \)
3: if \( l(\hat{\theta}_0) > -\frac{1}{2} \chi^2_{K}(\alpha_i) \) then
4: return 1
5: else
6: return 0
7: end if

Fig. 4. Pseudocode of LDA. For LDA\( (e, q, \alpha_i) \) with two inputs, \( e \) and \( q \), \( l(\hat{\theta}_0) \) is computed instead of \( l(\hat{\theta}_0) \).

Table 1 shows the average log-likelihoods over 100 trials by using the parameters estimated by LDA for examples in Fig. 1. We can see that \( \hat{\theta}_0 \) achieves a very high log-likelihood for each of (b) and (c) (especially (b)), implying that they can be pruned by LDA.

2.5 Randomness Test

Area II of Fig. 3 contains examples for which the maximum likelihoods with the interaction term are very low, implying that expression values are almost uniformly distributed in terms of genotypes, as shown in Fig. 1 (a). To detect the randomness of expression values, we can use hypothesis tests for randomness. If we can compute the randomness test with a significantly smaller amount of computation time than that of Newton-Raphson, we can speed up the procedure for finding interacting genes. We use an assumption that expression values follow the \( K \)-dimensional normal distribution for all class of genotypes, and under this assumption, we first show the two most typical random tests, multivariate analysis of variance (MANOVA) and Box’s M test (Mardia et al., 1979), and then present our approach, which combines these two tests. We can set \( K = 2 \) for our test, meaning that the largest matrix size is \( 2 \times 2 \), making the computation very efficient.

2.5.1 Multivariate Analysis of Variance (MANOVA)

MANOVA considers the following null hypotheses over the means:
\[ H_0: \mu_1 = \cdots = \mu_C, \ H_1: \mu_i \neq \mu_j \text{ for some pair of } i \text{ and } j \]

For testing \( H_0 \) against \( H_1 \), we use the statistic 
\[ -2 \log \lambda = 2(\ell^*_1 - \ell^*_0), \]
which follows the \( \chi^2 \) distribution. By replacing \( \Sigma \) in Eq.\( (2) \) with \( \Sigma_i \) and using the maximum likelihood estimators \( \bar{x}_i \) and \( S \), we can have the following:
\[ \ell^*_1 = \frac{N}{2} \log \det \left( \frac{2\pi W}{N} \right) - \frac{NK}{2}, \]
(14)

On the other hand, for the log-likelihood under null hypothesis, we can use the maximum likelihood estimators \( S \) for \( \mu_i \) and \( \Sigma \), respectively, and have the following:
\[ \ell^*_0 = \frac{N}{2} \log \det \left( \frac{2\pi W}{N} \right) - \frac{NK}{2}. \]
(15)

Thus we can finally have the following statistic:
\[ -2 \log \lambda = -N \log \frac{\det(W)}{\det(B + W)} \]

We can see that \( \lambda \) is \( KC + \frac{K(C+1)}{2} \) and \( r \) is \( K + \frac{K(C+1)}{2} \). In practice, we can follow Johnson and Wichern (2002), which introduces an approximation (to the \( \chi^2 \) distribution) in which \( N \) is replaced with \( N - 1 - \frac{1}{2}(K + C)/2 \).

We conducted MANOVA over four examples in Fig. 1. Table 2 shows the average p-value over 100 runs for each case with the standard deviation in parentheses. The p-value of MANOVA for (a) was high (0.53) while that for (b) and (c) was zero, meaning that MANOVA can discriminate (a) from (b) and (c).

Table 2. Average p-values over 100 runs with standard deviations in parentheses for four examples in Fig. 1 by using MANOVA, Box’s M test and Means-Covariance (MC) test

| Examples in Fig. 1 | MANOVA | Box’s M test | MC test |
|--------------------|--------|-------------|---------|
| (a)                | 0.53 (0.28) | 0.70 (0.25) | 0.60 (0.30) |
| (b)                | 0.00 (0.00) | 0.68 (0.25) | 0.00 (0.00) |
| (c)                | 0.00 (0.00) | 0.71 (0.25) | 0.00 (0.00) |
| (d)                | 0.94 (0.09) | 0.00 (0.00) | 0.00 (0.00) |

Input: \( X(e_1, e_2, q) \): Input three vectors of genes \( e_1, e_2 \) and SNP \( q \). \( \alpha_i \): Significance level for interaction test

Output: One if two genes \( e_1 \) and \( e_2 \) are randomly generated in terms of SNP \( q \); otherwise zero.

MC test\( (e_1, e_2, q, \alpha_m) \):
1: Compute \( \ell^*_i \) according to Eq. \( (15) \).
2: Compute \( \ell^*_j \) according to Eq. \( (16) \).
3: Compute \( -2 \log \lambda < \chi^2_{i}(\alpha_m) \) then
4: return 1
5: else
6: return 0
7: end if

Fig. 5. Pseudocode of Means-covariance (MC) test.

2.5.2 Box’s M Test

We then consider the following hypotheses over the covariance:
\[ H_0: \Sigma_1 = \Sigma_2 = \cdots = \Sigma_C, \ H_1: \Sigma_i \neq \Sigma_j \text{ for some pair of } i \text{ and } j \]

Here \( \ell^*_i \) can be given by \( \ell^*_i \) of MANOVA (i.e. Eq.\( (14) \)), and \( \ell^*_j \) can be obtained by using maximum likelihood estimators \( \bar{x}_i \) and \( S \) for \( \mu_i \) and \( \Sigma_i \), respectively, in Eq.\( (2) \):
\[ \ell^*_j = \sum_{c=1}^{C} N_c \log \det(2\pi S) - \frac{NK}{2}. \]
(16)

Thus we can have the following statistic:
\[ -2 \log \lambda = -\sum_{c=1}^{C} N_c \log \det(S_c^{-1} S). \]

Here \( q \) is \( KC + \frac{K(C+1)}{2} \) and \( r \) is \( K + \frac{K(C+1)}{2} \). We run Box’s M test over four examples in Fig. 1. Table 2 shows the average p-values over 100 runs for each of four examples with the standard deviation in parenthesis. This result shows that the p-value of (a) was high (0.70) while that of (d) was zero, meaning that M-test separated (a) from (d). However, this, the p-value for (b) and (c) was also high (0.68), meaning that this test could not discriminate (a) from (b) and (c). Thus we need another hypothesis test, but this result showed that Box’s M test can be a complement of MANOVA, implying that we can combine these two tests for our purpose of detecting random distributions such as Fig. 1 (a).

2.5.3 MC Test (MANOVA + M Test)

We finally consider the following hypotheses over both the means and covariance:
\[ H_0: \mu_1 = \cdots = \mu_C \text{ and } \Sigma_1 = \cdots = \Sigma_C, \ H_1: \mu_i \neq \mu_j \text{ or } \Sigma_i \neq \Sigma_j \text{ for some pair of } i \text{ and } j \]

We emphasize that this test suits our purpose the most, although this is an unpopular statistic and not named. We then call this test Mean-Covariance (MC) test. The statistic \(-2 \log \lambda = 2(\ell^*_1 - \ell^*_0)\) of this test is easily
Proposed Procedure

1: for each pair of gene $e \in E$ and SNP $q \in Q$ do
2: if $\text{LDA}(e, q, \alpha_i) = 1$ then
3: \( F \leftarrow F \cup (e, q) \)
4: end if
5: end for
6: \( \text{Pruned} \)
7: for each combination of genes $e_1 \in E, e_2 \in E$ and SNP $q \in Q$ do
8: if \( (e_1, q) \notin F \) and \( (e_2, q) \notin F \) then
9: \( I \leftarrow I \cup (e_1, e_2, q) \)
10: if $\text{MC}_\text{test}(e_1, e_2, q, \alpha_m) = 1$ then
11: This combination should be in Area II. go to Pruned
12: end if
13: if $\text{Pruning by LDA: Two genes and a SNP}$
14: if $\text{LDA}(e_1, e_2, q, \alpha_i) = 1$ then
15: This combination should be in Area I. go to Pruned
16: end if
17: if $\text{Interaction test for unpruned combinations}$
18: if $\text{Interaction}_\text{test}(e_1, e_2, q, \alpha_i) = 1$ then
19: $I \leftarrow I \cup (e_1, e_2, q)$
20: end if
21: end if
22: Pruned
23: end for

Fig. 6. Pseudocode of our entire procedure: FTGI.

Obtained from MANOVA and M test. That is, \( \ell^2 \) of this test is given by \( \ell^2 \) of MANOVA, i.e. Eq. (15) and \( \ell^2 \) is given by \( \ell^2 \) of M test, i.e. Eq. (16). Thus, the statistic of this test is given as follows:

\[
-2 \log \lambda = \sum_{c=1}^{C} N_c \log \det(S_c^{-1} S_T),
\]

since $T = S_T$. Here $q = KC + KK^+(K+1)$ and $r = K + K^+(K+1)$, meaning that $df$ is 10 in our case. Fig. 5 shows a pseudocode of MC test, in which we can set significance level $\alpha_m$, to remove given combination \((e_1, e_2, q)\) if its $p$-value is larger than $\alpha_m$, meaning that a larger number of combinations can be removed if $\alpha_m$ is smaller.

We checked the performance of this test using synthetic four examples of Fig. 1. Table 2 shows that all $p$-values are zero, except (a), which has the $p$-value of 0.60, indicating that MC test can successfully detect (a) out of other three examples and is expected to work on real data as well.

2.6 Proposed Procedure

Fig. 6 shows a pseudocode of our entire procedure. We can first check each pair of a gene in expressions and a SNP by using LDA, and if expressions can be categorized into three genotypes, this pair is saved in $F$ to be pruned (Lines 1-6). We then generate all possible combinations of two genes and a SNP out of given data (Line 7). For each combination, we apply three pruning conditions one by one: The first is LDA, and if any gene-SNP pair in the given combination is in $F$, it is pruned (Line 8). The next is MC test, and if expression values are randomly distributed, this combination is pruned (Lines 9-12). The last is LDA again, and if expressions can be separated without the interaction term, this combination is pruned (Lines 13-16). Finally we run interaction test over the unpruned combination to find the three-way gene interaction (Lines 17-23). Hereafter we call our proposed procedure FTGI, standing for Fast finding Three-way Gene Interactions, while we call the approach of running Interaction Test Only over all possible combinations ITO.

3 EXPERIMENTS

In this section, $p$-values are shown by $\log_{10}(p$-values).

3.1 Annotating Genes in Detected Three-way Interactions

For the top ten three-way interactions in Table 4 in the main text, Table 3 shows the identifiers used in Reactome (Vastrik et al., 2007). Then Table 4 shows the corresponding annotations of the identifiers shown in Table 3.

3.2 Validating Top Ten Three-way Interactions Detected by FTGI

For each gene pair of Table 4 in the main text, we explored the possibility that there exists a switching mechanism we addressed under the alteration of experimental conditions for gene expressions. To do this, we used the entire GEO (Gene Expression Omnibus) database to generate datasets with binary classes, and measured $p$-values of interaction test over them. The detail procedure is described in the main text. In this supplement we show the resultant list of datasets with the (top ten) smallest $p$-values for each gene pair.

Tables 5 and 6 show the GEO datasets (GDs) with ten smallest $p$-values of interaction test for each of ten interactions in Table 4 in the main text. All these $p$-values are significantly small, indicating
Table 4. Annotations and their identifier in Reactome.

| Identifier   | Pathway                                                                 |
|--------------|-------------------------------------------------------------------------|
| REACT-107.4  | Apoptotic cleavage of cellular proteins                                 |
| REACT-1105.1 | Rho GTPase cycle                                                        |
| REACT-1248.4 | EGFR downregulation                                                     |
| REACT-1255.1 | Downstream TCR signaling                                                |
| REACT-1341.5 | p75NTR recruits signalling complexes                                    |
| REACT-1344.3 | Regulated proteolysis of p75NTR                                         |
| REACT-1352.7 | Further platelet releasate                                              |
| REACT-1354.1 | Caspase-mediated cleavage of cytoskeletal proteins                      |
| REACT-1364.3 | NIF signals cell death from the nucleus                                 |
| REACT-1369.6 | NF-kB is activated and signals survival                                 |
| REACT-1432.1 | TNF signaling                                                           |
| REACT-1503.2 | Caspase-8 is formed from procaspase-8                                   |
| REACT-1827.7 | PERK regulated gene expression                                           |
| REACT-1834.8 | Activation of Chaperones by ATF6-alpha                                 |
| REACT-1836.8 | Activation of Chaperones by IRE1 alpha                                  |
| REACT-402.1  | TRAIL signaling                                                         |
| REACT-5213.2 | Influenza Virus Induced Apoptosis                                       |
| REACT-6305.1 | Electron Transport Chain                                                |
| REACT-6759.1 | Formation of ATP by chemiosmotic coupling                               |
| REACT-6784.3 | Glucuronidation                                                          |
| REACT-6809.2 | TRAM Cascade                                                            |
| REACT-6898.2 | Viral dsRNA:TLR3:TRIF Complex Activates TBK1                            |
| REACT-6976.2 | Viral dsRNA:TLR3:TRIF Complex Activates RIP1                             |
| REACT-701.2  | Activation, myristylation of BID and translocation to mitochondria     |
| REACT-7016.1 | Vpr-mediated induction of apoptosis by mitochondrial outer membrane permeabilization |
| REACT-832.2  | Activation of Pro-Caspase 8                                             |
| REACT-900.1  | FasL/CD95L signaling                                                    |

that there exist a switching mechanism under the (alteration of) conditions which were used to measure the expression values in GDSs. This result directly implies that there might exist the switching mechanism under the alteration of genotypes for each interaction. Thus this result supports the reliability of the three-way interactions which were detected by our method, FTGI.

REFERENCES

Barrett, T. et al. (2007). NCBI GEO: mining tens of millions of expression profiles database and tools update. NAR, 35, D760–D765.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer-Verlag.

Johnson, R. and Wichern, D. (2002). Applied Multivariate Statistical Analysis. Prentice Hall, New York.

Mardia, K. V. et al. (1979). Multivariate Analysis. Academic Press, New York.

McCullagh, P. and Nelder, J. (1989). Generalized Linear Models. Chapman & Hall CRC.

Vastrik, I. et al. (2007). Reactome: a knowledge base of biologic pathways and processes. Genome Biol., 8, R39.
Table 5. List of top ten GDSs for each gene pair of top 1 to 5 of three-way interactions we detected.

| GDS   | p-value  | #ex. in class 1 | #ex. in class 2 | Annotation                                                                 |
|-------|----------|----------------|----------------|-----------------------------------------------------------------------------|
| Rank 1 (COX6C and UBA1)  |
| 1 GDS2960 | -3.9532  | 60             | 41             | Marfan syndrome: cultured skin fibroblasts                                  |
| 2 GDS2960 | -3.1890  | 60             | 41             | Marfan syndrome: cultured skin fibroblasts                                  |
| 3 GDS2733 | -3.1814  | 18             | 17             | Cytosine arabinoside effect on Ewing’s sarcoma cell line: time course and dose response |
| 4 GDS2960 | -2.9117  | 60             | 41             | Marfan syndrome: cultured skin fibroblasts                                  |
| 5 GDS1615 | -2.7542  | 42             | 26             | Ulcerative colitis and Crohn’s disease comparison: peripheral blood mononuclear cells |
| 6 GDS2545 | -2.6416  | 63             | 18             | Metastatic prostate cancer (HG-U95A)                                       |
| 7 GDS724  | -2.6162  | 31             | 42             | Kidney transplant rejection expression profiling                           |
| 8 GDS2733 | -1.6415  | 18             | 16             | Cytosine arabinoside effect on Ewing’s sarcoma cell line: time course and dose response |
| 9 GDS2733 | -1.3510  | 17             | 16             | Cytosine arabinoside effect on Ewing’s sarcoma cell line: time course and dose response |
| 10 GDS1615 | -1.3415 | 59             | 42             | Ulcerative colitis and Crohn’s disease comparison: peripheral blood mononuclear cells |

| Rank 2 (RERE and TNFRSF1A)  |
| 1 GDS2736 | -5.9049  | 19             | 15             | Malignant fibrous histiocytoma and various soft tissue sarcomas              |
| 2 GDS2643 | -4.6912  | 13             | 11             | Waldenstrom’s macroglobulinemia: B lymphocytes and plasma cells            |
| 3 GDS2255 | -4.5097  | 17             | 14             | Transmigrated neutrophils in the alveolar space of endotoxin-exposed lung  |
| 4 GDS2255 | -4.3663  | 17             | 14             | Transmigrated neutrophils in the alveolar space of endotoxin-exposed lung  |
| 5 GDS2736 | -4.1722  | 19             | 15             | Malignant fibrous histiocytoma and various soft tissue sarcomas              |
| 6 GDS2255 | -4.0081  | 17             | 17             | Transmigrated neutrophils in the alveolar space of endotoxin-exposed lung  |
| 7 GDS2736 | -3.7308  | 19             | 15             | Malignant fibrous histiocytoma and various soft tissue sarcomas              |
| 8 GDS2255 | -3.5448  | 17             | 17             | Transmigrated neutrophils in the alveolar space of endotoxin-exposed lung  |
| 9 GDS2643 | -3.3449  | 13             | 11             | Waldenstrom’s macroglobulinemia: B lymphocytes and plasma cells            |
| 10 GDS2644 | -3.1556 | 20             | 11             | Waldenstrom’s macroglobulinemia: B lymphocytes and plasma cells            |

| Rank 3 (ATPS3 and ITCH)  |
| 1 GDS1865 | -5.1235  | 27             | 24             | Host cell response to HIV-1 Vpr-induced cell cycle arrest: time course      |
| 2 GDS534 | -4.2667  | 34             | 18             | Smoking-induced changes in airway transcriptome                             |
| 3 GDS2255 | -3.9027  | 20             | 11             | Waldenstrom’s macroglobulinemia: B lymphocytes and plasma cells            |
| 4 GDS534 | -3.7176  | 34             | 23             | Smoking-induced changes in airway transcriptome                             |
| 5 GDS2255 | -3.7175  | 17             | 14             | Transmigrated neutrophils in the alveolar space of endotoxin-exposed lung  |
| 6 GDS1956 | -3.6001  | 20             | 14             | Various muscle diseases (HG-U133A)                                        |
| 7 GDS2736 | -3.2145  | 16             | 15             | Malignant fibrous histiocytoma and various soft tissue sarcomas              |
| 8 GDS2643 | -3.1770  | 20             | 11             | Waldenstrom’s macroglobulinemia: B lymphocytes and plasma cells            |
| 9 GDS2767 | -2.6176  | 29             | 28             | Blood response to various beverages: time course                           |
| 10 GDS2362 | -2.5896 | 22             | 22             | Presymptomatic and symptomatic malaria: peripheral blood mononuclear cells  |

| Rank 4 (ATPSG1 and ATP5H)  |
| 1 GDS2733 | -7.9996  | 17             | 17             | Cytosine arabinoside effect on Ewing’s sarcoma cell line: time course and dose response |
| 2 GDS2733 | -7.9996  | 17             | 17             | Cytosine arabinoside effect on Ewing’s sarcoma cell line: time course and dose response |
| 3 GDS2733 | -3.5411  | 18             | 17             | Cytosine arabinoside effect on Ewing’s sarcoma cell line: time course and dose response |
| 6 GDS2733 | -3.5411  | 18             | 17             | Cytosine arabinoside effect on Ewing’s sarcoma cell line: time course and dose response |
| 7 GDS2733 | -3.5411  | 18             | 17             | Cytosine arabinoside effect on Ewing’s sarcoma cell line: time course and dose response |
| 8 GDS2733 | -3.5411  | 18             | 17             | Cytosine arabinoside effect on Ewing’s sarcoma cell line: time course and dose response |
| 9 GDS274 | -2.4366  | 29             | 20             | Hepatocellular carcinoma metastasis                                        |
| 10 GDS274 | -2.4366  | 29             | 20             | Hepatocellular carcinoma metastasis                                        |

| Rank 5 (NCSTN and HSPA5)  |
| 1 GDS2545 | -6.4398  | 63             | 25             | Metastatic prostate cancer (HG-U95A)                                       |
| 2 GDS2545 | -4.1952  | 25             | 18             | Metastatic prostate cancer (HG-U95A)                                       |
| 3 GDS2545 | -3.2969  | 65             | 25             | Metastatic prostate cancer (HG-U95A)                                       |
| 4 GDS274 | -3.1166  | 29             | 20             | Hepatocellular carcinoma metastasis                                        |
| 5 GDS274 | -2.9108  | 28             | 20             | Hepatocellular carcinoma metastasis                                        |
| 6 GDS274 | -2.9019  | 29             | 20             | Hepatocellular carcinoma metastasis                                        |
| 7 GDS274 | -2.7325  | 28             | 20             | Hepatocellular carcinoma metastasis                                        |
| 8 GDS2733 | -1.9764  | 18             | 17             | Cytosine arabinoside effect on Ewing’s sarcoma cell line: time course and dose response |
| 9 GDS2733 | -1.9324  | 18             | 17             | Cytosine arabinoside effect on Ewing’s sarcoma cell line: time course and dose response |
| 10 GDS1875 | -1.7328 | 24             | 16             | Host cell response to HIV-1 Vpr-induced cell cycle arrest: time course      |
| Rank | GDS          | p-value | #ex of class 1 | #ex of class 2 | Annotation                                                                 |
|------|--------------|---------|---------------|---------------|-----------------------------------------------------------------------------|
| 6    | GDS2733_4    | -4.7027 | 17            | 16            | Cytosine arabinoside effect on Ewing’s sarcoma cell line: time course and dose response |
|      | GDS2736_12   | -3.6633 | 16            | 15            | Malignant fibrous histiocytoma and various soft tissue sarcomas             |
| 8    | GDS2733_10   | -3.4264 | 17            | 16            | Cytosine arabinoside effect on Ewing’s sarcoma cell line: time course and dose response |
| 5    | GDS2733_8    | -3.3663 | 17            | 16            | Cytosine arabinoside effect on Ewing’s sarcoma cell line: time course and dose response |
| 6    | GDS2733_6    | -3.3190 | 18            | 16            | Cytosine arabinoside effect on Ewing’s sarcoma cell line: time course and dose response |
| 7    | GDS1449_11   | -2.9389 | 10            | 10            | HIV-1 infection effect on peripheral blood mononuclear cells                |
| 8    | GDS2733_12   | -2.7464 | 18            | 16            | Cytosine arabinoside effect on Ewing’s sarcoma cell line: time course and dose response |
| 9    | GDS2736_5    | -2.6202 | 34            | 18            | Smoking-induced changes in airway transcriptome                             |
| 10   | GDS2736_19   | -2.5776 | 21            | 15            | Malignant fibrous histiocytoma and various soft tissue sarcomas             |
| 6    | GDS1627_2    | -3.2808 | 16            | 15            | Breast cancer cell lines response to chemotherapeutic drugs: time course    |
| 2    | GDS1627_5    | -2.5799 | 20            | 18            | Breast cancer cell lines response to chemotherapeutic drugs: time course    |
| 3    | GDS1627_3    | -2.3132 | 18            | 15            | Breast cancer cell lines response to chemotherapeutic drugs: time course    |
| 4    | GDS1627_9    | -1.7245 | 18            | 15            | Breast cancer cell lines response to chemotherapeutic drugs: time course    |
| 5    | GDS1627_4    | -1.4653 | 20            | 16            | Breast cancer cell lines response to chemotherapeutic drugs: time course    |
| 6    | GDS1627_7    | -1.3858 | 20            | 15            | Breast cancer cell lines response to chemotherapeutic drugs: time course    |
| 7    | GDS1627_8    | -1.1312 | 16            | 15            | Breast cancer cell lines response to chemotherapeutic drugs: time course    |
| 8    | GDS1962_3    | -0.8665 | 153           | 23            | Glioma-derived stem cell factor effect on angiogenesis in the brain        |
| 9    | GDS2819_3    | -0.5034 | 33            | 33            | Leukemic white blood cells and various RNA preparation protocols            |
| 10   | GDS1615_6    | -0.4423 | 153           | 23            | Glioma-derived stem cell factor effect on angiogenesis in the brain        |
| 6    | GDS2960_1    | -3.1628 | 60            | 41            | Marfan syndrome: cultured skin fibroblasts                                 |
| 2    | GDS2960_2    | -3.1628 | 60            | 41            | Marfan syndrome: cultured skin fibroblasts                                 |
| 3    | GDS2767_5    | -3.0667 | 21            | 21            | Breast cancer cell lines response to chemotherapeutic drugs: time course    |
| 4    | GDS1627_11   | -3.0667 | 21            | 21            | Breast cancer cell lines response to chemotherapeutic drugs: time course    |
| 5    | GDS1627_17   | -3.0667 | 21            | 21            | Breast cancer cell lines response to chemotherapeutic drugs: time course    |
| 6    | GDS1627_23   | -3.0667 | 21            | 21            | Breast cancer cell lines response to chemotherapeutic drugs: time course    |
| 7    | GDS353_3     | -2.2410 | 23            | 18            | Smoking-induced changes in airway transcriptome                             |
| 8    | GDS353_4     | -2.2410 | 23            | 18            | Smoking-induced changes in airway transcriptome                             |
| 9    | GDS353_9     | -2.2410 | 23            | 18            | Smoking-induced changes in airway transcriptome                             |
| 10   | GDS353_12    | -2.2410 | 23            | 18            | Smoking-induced changes in airway transcriptome                             |
| 6    | GDS2960_1    | -3.1628 | 60            | 41            | Marfan syndrome: cultured skin fibroblasts                                 |
| 2    | GDS2960_2    | -3.1628 | 60            | 41            | Marfan syndrome: cultured skin fibroblasts                                 |
| 3    | GDS2733_5    | -3.1814 | 18            | 17            | Cytosine arabinoside effect on Ewing’s sarcoma cell line: time course and dose response |
| 4    | GDS2960_3    | -2.9117 | 60            | 41            | Marfan syndrome: cultured skin fibroblasts                                 |
| 5    | GDS1615_1    | -2.7340 | 42            | 26            | Ulcerative colitis and Crohn’s disease comparison: peripheral blood mononuclear cells |
| 6    | GDS2545_1    | -2.6416 | 63            | 18            | Metastatic prostate cancer (HG-U95A)                                      |
| 7    | GDS724_1     | -2.6162 | 31            | 31            | Kidney transplant rejection expression profiling                           |
| 8    | GDS2733_6    | -1.6415 | 18            | 16            | Cytosine arabinoside effect on Ewing’s sarcoma cell line: time course and dose response |
| 9    | GDS2733_2    | -1.3510 | 17            | 16            | Cytosine arabinoside effect on Ewing’s sarcoma cell line: time course and dose response |
| 10   | GDS1615_2    | -1.3415 | 59            | 42            | Ulcerative colitis and Crohn’s disease comparison: peripheral blood mononuclear cells |
| 6    | GDS2643_9    | -6.2133 | 13            | 12            | Waldenstrom’s macroglobulinemia: B lymphocytes and plasma cells            |
| 2    | GDS2643_7    | -5.3750 | 20            | 13            | Waldenstrom’s macroglobulinemia: B lymphocytes and plasma cells            |
| 3    | GDS2767_13   | -4.0154 | 26            | 25            | Blood response to various beverages: time course                           |
| 4    | GDS2643_1    | -3.5855 | 20            | 13            | Waldenstrom’s macroglobulinemia: B lymphocytes and plasma cells            |
| 5    | GDS2771_3    | -3.0461 | 97            | 90            | Large airway epithelial cells from cigarette smokers with suspect lung cancer |
| 6    | GDS2643_2    | -2.6757 | 13            | 12            | Waldenstrom’s macroglobulinemia: B lymphocytes and plasma cells            |
| 7    | GDS2767_16   | -2.5530 | 29            | 26            | Blood response to various beverages: time course                           |
| 8    | GDS2926_1    | -2.4070 | 27            | 17            | Megakaryocytic differentiation: time course                                |
| 9    | GDS2771_2    | -2.3254 | 97            | 90            | Large airway epithelial cells from cigarette smokers with suspect lung cancer |
| 10   | GDS2643_12   | -2.2933 | 12            | 11            | Waldenstrom’s macroglobulinemia: B lymphocytes and plasma cells            |

Table 6. List of top ten GDSs for each gene pair of top 6 to 10 of three-way interactions we detected.