BIOINFORMATICS ORIGINAL PAPER

Vol. 25 no. 21 2009, pages 2735–2743
doi:10.1093/bioinformatics/btp531

Genome analysis

Efficiently finding genome-wide three-way gene interactions from transcript- and genotype-data

Mitsunori Kayano[1,2], Ichigaku Takigawa[1,2], Motoki Shiga[1,2], Koji Tsuda[2,3] and Hiroshi Mamitsuka[1,2,*]

[1]Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji 611-0011, [2]Institute for Bioinformatics Research and Development (BIRD), Japan Science and Technology Agency (JST) and [3]Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology (AIST), 2-42 Aomi, Koto-ku, Tokyo 135-0064, Japan

Received on March 17, 2009; revised on August 6, 2009; accepted on August 25, 2009

Advance Access publication September 7, 2009

Associate Editor: Alfonso Valencia

ABSTRACT

Motivation: We address the issue of finding a three-way gene interaction, i.e. two interacting genes in expression under the genotypes of another gene, given a dataset in which expressions and genotypes are measured at once for each individual. This issue can be a general, switching mechanism in expression of two genes, being controlled by categories of another gene, and finding this type of interaction can be a key to elucidating complex biological systems. The most suitable method for this issue is likelihood ratio test using logistic regressions, which we call *interaction test*, but a serious problem of this test is computational intractability at a genome-wide level.

Results: We developed a fast method for this issue which improves the speed of interaction test by around 10 times for any size of datasets, keeping highly interacting genes with an accuracy of ~85%. We applied our method to $\sim 3\times 10^8$ three-way combinations generated from a dataset on human brain samples and detected three-way gene interactions with small P-values. To check the reliability of our results, we first conducted permutations by which we can show that the obtained P-values are significantly smaller than those obtained from permuted null examples. We then used GEO (Gene Expression Omnibus) to generate gene expression datasets with binary classes to confirm the detected three-way interactions by using these datasets and interaction tests. The result showed us some datasets with significantly small P-values, strongly supporting the reliability of the detected three-way interactions.

Availability: Software is available from http://www.bic.kyoto-u.ac.jp/pathway/kayano/bioinfo_three-way.html

Contact: kayano@kuicr.kyoto-u.ac.jp

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

We address the issue of efficiently finding a three-way gene interaction, precisely two interacting genes in expression under the genotypes of a different gene, given a dataset in which both gene expressions and genotypes are measured for each individual. We illustrate our problem setting by using synthetic 2D diagrams in Figure 1, where expression values of two genes are plotted with three classes (genotypes): +, * and △. In this figure, panel (a) shows expression values being just randomly distributed; (b) shows expression values being easily categorized into three classes; and (c) shows that classes can be categorized by expressions without using two genes at the same time. We are not interested in (a–c) but in (d), which shows that the correlation in expression between two genes differs for each class. More concretely, two genes are positively correlated for one class, whereas they are negatively correlated for another. This is exactly a switching mechanism in expression between correlation and inverse-correlation of two genes, controlled by another gene. Also this is the three-way gene interaction which we attempt to find in this article. We note that this can be categorized into a general switch in biology. A simple, well-known example is Max, a transcription factor, which plays a role of an activator or a suppressor, depending on whether it binds to Myc (i.e. Myc-Max)

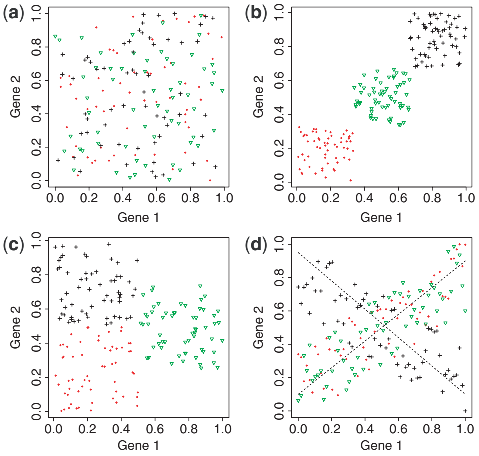

Fig. 1. Synthetic examples: expressions of two genes under the three classes of another gene. (**a**) randomly distributed, (**b**,**c**) easily categorized into three classes and (**d**) a switching mechanism.

*To whom correspondence should be addressed.

or Mad (i.e. Mad-Max) (Ayer and Eisenman, 1993). We emphasize that this type of interaction must be a key to elucidating complex biological systems.

A reasonable approach to detect such three-way interactions is the likelihood ratio test for regression (LRTR). Particularly, logistic regression must be suitable the most, because of categorical responses (genotypes) in our setting (McCullagh and Nelder, 1989). The first item of note is that parameter estimation for logistic regression is based on the maximum likelihood, for which a time-consuming iterative gradient descent, Newton–Raphson, is usually used. Secondly, in our case, classes are genotypes, causing a problem of an explosive number of combinations of one SNP (genotypes) and two genes (expressions). For example, for 50 000 SNPs and 1000 genes, we have roughly 5×10^{10} $(=50\,000\times1000\times1000)$ combinations, making scanning over all possible combinations intractable. In fact, >24 h are needed to run Newton–Raphson over only 10^7 combinations in our experiments. Thus, the main focus of this article is to speed up the procedure of finding the three-way interactions. Our strategy for this issue is to prune irrelevant combinations, such as those in which the expression values of two genes are randomly distributed as in Figure 1a, by using a hypothesis test assuming the normality of given examples.

The contribution of this article can be summarized into three folds: (i) We present a problem setting of finding a three-way gene interaction of two numerical variables and one categorical, corresponding to a biological switch in expression. (ii) LRTR and LRT of logistic regression (LRTLR) are the standard approaches for this problem, but these are computationally inefficient, particularly for a huge number of combinations that we can have. We then propose an efficient method for pruning large part of input combinations. (iii) Our experiment with a huge dataset of human brain samples showed that our method run 10 times faster than LRTLR for any data size, keeping the accuracy of detecting three-way interactions at ∼85%.

2 RELATED WORK

Three-way interactions in expression have not been considered except only a few cases of using simple methods (Li *et al.*, 2004; Zhang *et al.*, 2007). There are two reasons for this: (i) dealing with more than two-way correlations is intractable at a genome-wide level, because of the explosive number of combinations and (ii) three-way interactions along this line can be inferred from two-way co-expression. We emphasize that our three-way interaction is different from them, in terms that correlation or inverse-correlation in expression of two genes is controlled by genotypes of another gene.

Genome-wide association studies (GWA) using genotypes, especially single nucleotide polymorphisms (SNPs), have been highlighted in these few years (McCarthy and Hirschhorn, 2008), whereas cDNA microarrays have been a standard tool for understanding gene/protein behaviors in a cell. Thus, currently a large number of studies use both gene expressions and genotypes, showing the importance of combining these two information sources (Nica and Dermitzakis, 2008). Consequently, we now have a unique dataset, in which both gene expressions and genotypes are measured at once for each individual, and this type of dataset, which we use in this article, is increasing in these few years, which makes our approach very promising (Dixon *et al.*, 2007; Myers *et al.*, 2007; Schadt *et al.*, 2008).

A standard analysis in GWA is conducted between a single SNP (i.e. genotypes at a locus) and a categorical or continuous outcome (phenotype). For this analysis, the two most typical approaches are ANOVA (Analysis of Variance) and LRTR (Balding, 2006). Usually more complex analysis is multiple (usually two) SNPs with a single phenotype where two-way ANOVA or LRTR with two explanatory variables can be considered. This situation is closely related with *epistasis*, a general concept in modern quantitative genetics (Aylor and Zeng, 2008; Cordell, 2002), meaning the interaction between multiple loci and phenotype (Marchini *et al.*, 2005). Our problem setting looks similar to this but interestingly in the reverse direction. That is, we consider the interaction between two expression phenotypes under categorical genotypes which thus have not been examined in GWA. We note that ANOVA cannot be applied to this issue,[1] whereas LRTR can be applied as a standard manner for our setting. Another item of note is that finding three-way interactions in only SNPs exists (Lo *et al.*, 2008), but their problem setting is straightforward and totally different from our setting.

3 METHODS

3.1 Notations and preliminaries

Let $\mathcal{X}$ 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 $\mathcal{X}$ and Q be the set of SNPs in $\mathcal{X}$, indicating that $|E|+|Q|$ is the total number of columns of $\mathcal{X}$. To test the three-way interaction, we choose one *combination*, i.e. two genes (e_1 and e_2) and one SNP (q) out of E and Q, respectively, and we write $\mathcal{X}(e_1,e_2,q)$ which has only three columns of $\mathcal{X}$, corresponding to e_1, e_2 and q [we write $\mathcal{X}(e,q)$ when we choose only one gene e out of E and q out of Q]. Hereafter, until Section 3.6, we assume that we already choose one combination.

For gene expressions, let $X=(X_1,\dots,X_K)'\in\mathbb{R}^K$ be a K-dimensional numerical variable, taking value $\boldsymbol{x}=(x_1,\dots,x_K)'$. We note that using two genes in expression does not necessarily mean $K=2$. 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 $\boldsymbol{y}$, where $Y=(Y_1,Y_2)'\in\{0,1\}\times\{0,1\}$. Here, we note that $\boldsymbol{y}$ takes the following values: $\boldsymbol{y}=(1,0)'$ if $\boldsymbol{x}\in G_1$, $\boldsymbol{y}=(0,1)'$ if $\boldsymbol{x}\in G_2$ and $\boldsymbol{y}=(0,0)'$ if $\boldsymbol{x}\in G_3$. We denote N inputs (individuals) by $\boldsymbol{X}=(\boldsymbol{x}_1,\dots,\boldsymbol{x}_N)'$ and $\boldsymbol{Y}=(\boldsymbol{y}_1,\dots,\boldsymbol{y}_N)'=(\boldsymbol{y}^{(1)},\boldsymbol{y}^{(2)})$, 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: $\bar{\boldsymbol{x}}_c=\frac{1}{N_c}\sum_{j=1|j\in G_c}^{N}\boldsymbol{x}_j$ and $\bar{\boldsymbol{x}}=\frac{1}{N}\sum_{j=1}^{N}\boldsymbol{x}_j$, respectively, where $\bar{\boldsymbol{x}}=1/N\sum_{c=1}^{K}N_c\bar{\boldsymbol{x}}_c$. I_K is the identity matrix of size K, and $\mathbf{1}$ is an n-dimensional vector in which all elements are 1.

We incorporate some basic statistics: $T=\sum_{j=1}^{N}(\boldsymbol{x}_j-\bar{\boldsymbol{x}})(\boldsymbol{x}_j-\bar{\boldsymbol{x}})'$, $B=\sum_{c=1}^{C}N_c(\bar{\boldsymbol{x}}_c-\bar{\boldsymbol{x}})(\bar{\boldsymbol{x}}_c-\bar{\boldsymbol{x}})'$ and $W=\sum_{c=1}^{C}\sum_{j=1|j\in G_c}^{N}(\boldsymbol{x}_j-\bar{\boldsymbol{x}}_c)(\boldsymbol{x}_j-\bar{\boldsymbol{x}}_c)'$, where $T=B+W$. We can further define covariance matrix S_c for class c, $S_c=\frac{1}{N_c}\sum_{j=1|j\in G_c}^{N_c}(\boldsymbol{x}_j-\bar{\boldsymbol{x}}_c)(\boldsymbol{x}_j-\bar{\boldsymbol{x}}_c)'$ $(c=1,\dots,C)$, and total covariance matrices S and S_T, $S=\frac{1}{N}\sum_{c=1}^{C}\sum_{j=1|j\in G_c}^{N_c}(\boldsymbol{x}_j-\bar{\boldsymbol{x}}_c)(\boldsymbol{x}_j-\bar{\boldsymbol{x}}_c)'$ $(=\frac{1}{N}W)$ and

[1] ANOVA can be applied only to the case with a single continuous response (phenotype) and one or more discrete explanatory variables (genotypes).

$S_T = \frac{1}{N}\sum_{c=1}^{C}\sum_{j=1|j\in G_c}^{N_c}(\boldsymbol{x}_j-\bar{\boldsymbol{x}})(\boldsymbol{x}_j-\bar{\boldsymbol{x}})'\ (=\frac{1}{N}T)$. We note that $W=\sum_{c=1}^{C}N_cS_c$ and $S=\frac{1}{N}\sum_{c=1}^{C}N_cS_c$.

We show the multivariate normal distribution, having two parameters, $\boldsymbol{\mu}_c$ and Σ_c (the mean and the covariance matrix of class c), and the log-density (log-likelihood) function of this distribution can be given as follows:

$$\ell(\boldsymbol{\mu},\boldsymbol{\Sigma}|\mathbf{X}) = -\sum_{c=1}^{C}\frac{N_c}{2}\log\,\det(2\pi\Sigma_c) - \sum_{c=1}^{C}\frac{N_c}{2}\left\{\mathrm{tr}\,\Sigma_c^{-1}S_c+(\bar{\boldsymbol{x}}_c-\boldsymbol{\mu}_c)'\Sigma_c^{-1}(\bar{\boldsymbol{x}}_c-\boldsymbol{\mu}_c)\right\}. \qquad (1)$$

From this equation, we can see that $\bar{\boldsymbol{x}}_c$, covariance matrix S_c and covariance matrix S can be the maximum likelihood estimators of $\boldsymbol{\mu}_c$, Σ_c and Σ ($=\Sigma_1=\cdots=\Sigma_C$), respectively.

We briefly describe likelihood ratio test (LRT), which will be used. We first assume that examples $\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_n$ are generated according to parameter vector $\boldsymbol{\theta}$. Let $H_0:\boldsymbol{\theta}\in\Omega_0$ be a null hypothesis and $H_1:\boldsymbol{\theta}\in\Omega_1$ be the alternative hypothesis. The statistic λ for testing H_0 against H_1 can be defined as $\lambda=L_0^*/L_1^*$, where L_0^* and L_1^* are the maximum likelihoods under $\boldsymbol{\theta}\in\Omega_0$ and $\boldsymbol{\theta}\in\Omega_1$, respectively. Usually we can use the log-likelihood ratio (LLR), $-2\log\lambda=2(\ell_1^*-\ell_0^*)$, where $\ell_1^*=\log L_1^*$ and $\ell_0^*=\log L_0^*$. We note that this statistic follows χ^2_{q-r} distribution as $N\to\infty$, where $q-r$ is the degree of freedom (*df*) of the χ^2 distribution.

3.2 Finding three-way interactions: interaction test (Likelihood Ratio Test of Logistic Regression, LRTLR)

A standard and exact approach for our problem is LRTLR (McCullagh and Nelder, 1989), which we simply call *interaction test* in this article.

3.2.1 Logistic regression We first denote the probability that $\boldsymbol{x}$ is in G_1 by $p_1(\boldsymbol{x})$, and similarly the probability that $\boldsymbol{x}$ is in G_2 by $p_2(\boldsymbol{x})$, by which the probability that $\boldsymbol{x}$ is in G_3 is $p_3(\boldsymbol{x})$ $(=1-p_1(\boldsymbol{x})-p_2(\boldsymbol{x}))$. We use logistic regression to link these probabilities to K-dimensional input $\boldsymbol{x}$ by using weight parameters (or coefficients) $\boldsymbol{w}=(\boldsymbol{w}_1',\boldsymbol{w}_2')'$, where $\boldsymbol{w}_1=(w_{10},w_{11},\ldots,w_{1(K-1)})'$, $\boldsymbol{w}_2=(w_{20},w_{21},\ldots,w_{2(K-1)})'$ as follows:

$$\begin{cases} p_1(\boldsymbol{x})=\dfrac{\exp(\boldsymbol{w}_1'\boldsymbol{x})}{1+\exp(\boldsymbol{w}_1'\boldsymbol{x})+\exp(\boldsymbol{w}_2'\boldsymbol{x})} \\ p_2(\boldsymbol{x})=\dfrac{\exp(\boldsymbol{w}_2'\boldsymbol{x})}{1+\exp(\boldsymbol{w}_1'\boldsymbol{x})+\exp(\boldsymbol{w}_2'\boldsymbol{x})} \end{cases} \Leftrightarrow \begin{cases} \log\dfrac{p_1(\boldsymbol{x})}{p_3(\boldsymbol{x})}=\boldsymbol{w}_1'\boldsymbol{x} \\ \log\dfrac{p_2(\boldsymbol{x})}{p_3(\boldsymbol{x})}=\boldsymbol{w}_2'\boldsymbol{x} \end{cases} \qquad (2)$$

Here, we denote $p_1(\boldsymbol{x})$, $p_2(\boldsymbol{x})$ and $p_3(\boldsymbol{x})$ by $p_1(\boldsymbol{x};\boldsymbol{w})$, $p_2(\boldsymbol{x};\boldsymbol{w})$ and $p_3(\boldsymbol{x};\boldsymbol{w})$ $(=1-p_1(\boldsymbol{x};\boldsymbol{w})-p_2(\boldsymbol{x};\boldsymbol{w}))$, respectively, because they can be functions of $\boldsymbol{w}$. We can then write the likelihood of logistic regression for given N examples and parameters $\boldsymbol{w}$, as follows:

$$L(\boldsymbol{w})=\prod_{i=1}^{N}p_1(\boldsymbol{x}_i;\boldsymbol{w})^{y_{i1}}p_2(\boldsymbol{x}_i;\boldsymbol{w})^{y_{i2}}p_3(\boldsymbol{x}_i;\boldsymbol{w})^{1-y_{i1}-y_{i2}},$$

where $\boldsymbol{y}_i=(y_{i1},y_{i2})'$.

3.2.2 Parameter estimation We can obtain the maximum likelihood estimator $\hat{\boldsymbol{w}}$ for $\boldsymbol{w}$ by maximizing the log-likelihood $l(\boldsymbol{w})=\log L(\boldsymbol{w})$. 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 $\hat{\boldsymbol{w}}^{(t+1)}$ at the $(t+1)$-th iteration, using $\hat{\boldsymbol{w}}^{(t)}$ of the t-th iteration:

$$\hat{\boldsymbol{w}}^{(t+1)}=\hat{\boldsymbol{w}}^{(t)}-\left(H(\boldsymbol{w})|_{\boldsymbol{w}=\hat{\boldsymbol{w}}^{(t)}}\right)^{-1}U(\boldsymbol{w})|_{\boldsymbol{w}=\hat{\boldsymbol{w}}^{(t)}}, \qquad (3)$$

where Hessian matrix $H(\boldsymbol{w})$ $(=\partial^2 l/\partial\boldsymbol{w}\partial\boldsymbol{w}')$ and gradient vector $U(\boldsymbol{w})$ $(=\partial l/\partial\boldsymbol{w})$ can be given in the following:

$$U(\boldsymbol{w})=\boldsymbol{X}_*'\boldsymbol{a}(\boldsymbol{w}),$$

Table 1. Log-likelihoods and LLR by Newton–Raphson

Figure 1	$l(\hat{\boldsymbol{w}}_{01})$	$l(\hat{\boldsymbol{w}}_0)$	$l(\hat{\boldsymbol{w}})$	LLR (*P*-value)
(a)	−196.4	−195.5	−194.4	2.23 (0.45)
(b)	−1.86	−0.42	−2.36	−3.87 (1.00)
(c)	−83.5	−1.52	−6.00	−8.97 (1.00)
(d)	−197.8	−197.4	−126.4	142.12 (0.00)

where $\boldsymbol{X}_*=\mathrm{diag}(\boldsymbol{X},\boldsymbol{X})$ (block diagonal matrix of $\boldsymbol{X}$), $\boldsymbol{a}(\boldsymbol{w})=(\boldsymbol{a}_1(\boldsymbol{w})',\boldsymbol{a}_2(\boldsymbol{w})')'$ where $\boldsymbol{a}_j(\boldsymbol{w})=\boldsymbol{y}^{(j)}-\boldsymbol{p}_j(\boldsymbol{w})$ and $\boldsymbol{p}_j(\boldsymbol{w})=(p_j(\boldsymbol{x}_1;\boldsymbol{w}),\ldots,p_j(\boldsymbol{x}_N;\boldsymbol{w}))'$ $(j=1,2)$.

$$H(\boldsymbol{w})=\begin{pmatrix}\boldsymbol{X}'R_{11}(\boldsymbol{w})\boldsymbol{X} & \boldsymbol{X}'R_{12}(\boldsymbol{w})\boldsymbol{X}\\ \boldsymbol{X}'R_{21}(\boldsymbol{w})\boldsymbol{X} & \boldsymbol{X}'R_{22}(\boldsymbol{w})\boldsymbol{X}\end{pmatrix}=\boldsymbol{X}_*'R(\boldsymbol{w})\boldsymbol{X}_*,$$

where $N\times N$ matrix $R_{jk}(\boldsymbol{w})(j,k=1,2)$ is given by $R_{jj}(\boldsymbol{w})=\mathrm{diag}\{\boldsymbol{p}_j(\boldsymbol{w})\odot(\boldsymbol{p}_j(\boldsymbol{w})-\mathbf{1})\}$ and $R_{jk}(\boldsymbol{w})=\mathrm{diag}\{\boldsymbol{p}_j(\boldsymbol{w})\odot\boldsymbol{p}_k(\boldsymbol{w})\}\,(j\neq k)$.

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

$$\hat{\boldsymbol{w}}^{(t+1)}=\hat{\boldsymbol{w}}^{(t)}-\{\boldsymbol{X}_*'R(\hat{\boldsymbol{w}}^{(t)})\boldsymbol{X}_*\}^{-1}\boldsymbol{X}_*'\boldsymbol{a}(\hat{\boldsymbol{w}}^{(t)}). \qquad (4)$$

In practise, we start with some initial values $\hat{\boldsymbol{w}}^{(0)}$ and update $\hat{\boldsymbol{w}}^{(t+1)}$ according to Equation (4) until the following equation is satisfied:

$$\|\hat{\boldsymbol{w}}^{(t+1)}-\hat{\boldsymbol{w}}^{(t)}\|^2<2K\delta, \qquad (5)$$

where δ is set at a certain value.

3.2.3 Interaction test We then examine the significance of the interaction in expression between two genes in terms of classes of another gene. Let x_{i1} and x_{i2} be expression values of the corresponding two genes for input i. The interaction term is $x_{i1}x_{i2}$, meaning that our purpose is to find the case that the logistic model is well fitted to the data when this term is added. We then let $\boldsymbol{x}_i=(1,x_{i1},x_{i2},x_{i1}x_{i2})'$ and $\boldsymbol{w}=(w_{10},w_{11},w_{12},w_{13},w_{20},w_{21},w_{22},w_{23})'$, and the logistic model with the interaction term is given as follows:

$$\log\frac{p_c(\boldsymbol{x};\boldsymbol{w})}{p_3(\boldsymbol{x};\boldsymbol{w})} = w_{c0}+w_{c1}x_{i1}+w_{c2}x_{i2}+w_{c3}x_{i1}x_{i2}\quad (c=1,2)$$

If $w_{c3}=0$, the model does not have the interaction term, meaning that the null hypothesis and $\boldsymbol{w}_0$ are given as follows:

$$\begin{aligned} H_0 &: w_{13}=0 \text{ and } w_{23}=0 \\ \boldsymbol{w}_0 &= (w_{10},w_{11},w_{12},0,w_{20},w_{21},w_{22},0)'. \end{aligned} \qquad (6)$$

Then the test statistic, LLR and its asymptotic distribution can be given:

$$-2\log\lambda=2\{l(\hat{\boldsymbol{w}})-l(\hat{\boldsymbol{w}}_0)\}\sim\chi_2^2(\alpha_i), \qquad (7)$$

where $\chi_2^2(\alpha_i)$ is the χ^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 α_i. We run interaction test 100 times over four examples in Figure 1, and Table 1 shows the average results over the 100 runs. This table clearly shows that the *P*-value is very large for Figure 1a–c, while that is zero for Figure 1d, indicating that intraction test can detect our target sample correctly.

Figure 2 shows a pseudocode of interaction test. We can write interaction test by function Interaction_test(e_1, e_2, q, α_i), which outputs one if given example (e_1,e_2,q) has the three-way interaction; otherwise zero. A significant drawback of interaction test is computational inefficiency. In fact, Equation (6) shows $K=8$, meaning that Newton–Raphson needs to compute an 8×8 inverse-matrix at each of its iteration procedure.

3.3 Key idea for speeding-up interaction finding

A basic idea for accelerating the finding of a three-way interactions is to prune some combinations, to which interaction test does not have to be applied.

```
Input: $\mathcal{X}(e_1,e_2,q)$: Input three vectors of genes $e_1,e_2$ and SNP $q$.
    $\alpha_i$: Significance level for interaction test
Output: One if $e_1$ and $e_2$ are interacting with each other under $q$; otherwise
    zero.
Interaction_test($e_1$, $e_2$, $q$, $\alpha_i$)
w0 ← some initial value.
repeat
    Update $\boldsymbol{w}_0$, according to the iterative rule of Eq.(4)
until Eq.(5) is satisfied
w ← some initial value.
repeat
    Update $\boldsymbol{w}$, according to the iterative rule of Eq.(4)
until Eq.(5) is satisfied
if $-2\{l(\hat{\boldsymbol{w}})-l(\hat{\boldsymbol{w}_0})\}>\chi^2_2(\alpha_i)$ then
    return 1
else
    return 0
end if
```

Fig. 2. Pseudocode of interaction test.

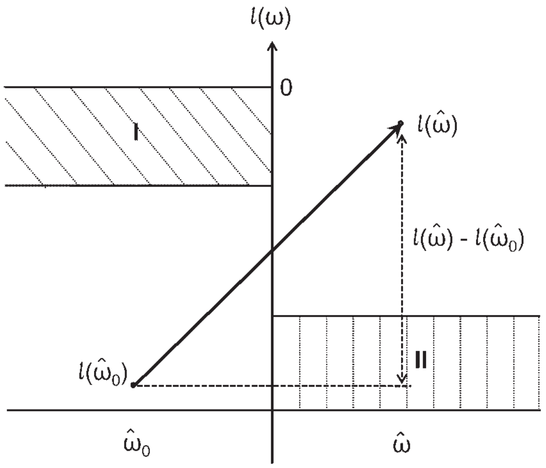

Fig. 3. LLR and its components.

From Equation (7), we can see that the interacting genes should have a larger LLR. Figure 3 shows a schematic figure, in which we plot the log-likelihood without the interaction term in the left-hand side and 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 $p_i(\boldsymbol{x})$. The LLR in question can be then given by the distance being shown by a dotted line in Figure 3. 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) a large likelihood can be obtained without the interaction term, and (II) only a small likelihood can be obtained even if we use the interaction term. These (I) and (II) correspond to areas I and II, respectively, in Figure 3. We then attempt to efficiently detect examples in areas I and II by assuming the normality on data distribution.

3.4 Linear discriminant analysis

Area I in Figure 3 contains examples in which expressions can be easily separated into three classes without the interaction term, as shown in Figure 1b and c. Thus, in this case, we can consider a simpler, easily computable estimation method for parameters of the logistic regression model without the interaction term, and if the likelihood for a given combination by that model is high enough, this combination can be pruned. For the simpler estimation method, we use linear discriminant analysis (LDA), which assumes that $\boldsymbol{x}$ follows the normal distribution $N(\boldsymbol{\mu},\Sigma)$ with the same covariance Σ for all three classes (Hastie *et al.*, 2001). We skip the detail of this method due to space limitations because in our experiment only a small part of all given examples can be pruned by LDA. Interested readers should refer the Supplementary Material. We can write LDA by function LDA(e_1, e_2, q, α_i) [or LDA(e, q, α_i)], which outputs one if given example (e_1,e_2,q) [or (e,q)] should be pruned; otherwise zero.

Table 2. MANOVA, Box's M test and Means-Covariances (MC) test on four examples in Figure 1

Examples in Figure 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)

3.5 Randomness test

Area II in Figure 3 contains an example for which the maximum likelihood with the interaction term is very low, implying that expression values are almost randomly distributed in terms of classes, as shown in Figure 1a. To detect the randomness of expression values, if we use a faster hypothesis test for randomness than Newton–Raphson, we can speed up the procedure for finding the three-way interaction. We assume that expression values follow the K-dimensional normal distribution for each class of genotypes, and under this assumption, we present our approach, which combines multivariate ANOVA (MANOVA) and Box's M test (Mardia *et al.*, 1979). We can set $K=2$ for our test, meaning that the largest matrix size is 2×2, making the computation very efficient.

3.5.1 MANOVA MANOVA considers the following null hypotheses over the means:

$$H_0:\boldsymbol{\mu}_1=\boldsymbol{\mu}_2=\cdots=\boldsymbol{\mu}_C,\quad H_1:\boldsymbol{\mu}_i\neq\boldsymbol{\mu}_j \text{ for some pair of } i \text{ and } j$$

For testing H_0 against H_1, we use LLR, $-2\log\lambda$ $(=2(\ell_1^*-\ell_0^*))$. By replacing Σ_c in Equation (1) with Σ and using the maximum likelihood estimators $\bar{\boldsymbol{x}}_c$ and S for $\boldsymbol{\mu}_k$ and Σ, respectively, we have the following:

$$\ell_1^*=-\frac{N}{2}\log\,\det\left(\frac{2\pi}{N}W\right)-\frac{NK}{2}. \tag{8}$$

On the other hand, for the log-likelihood under null hypothesis, we can use the maximum likelihood estimators $\bar{\boldsymbol{x}}$ and S_T for $\boldsymbol{\mu}_k$ and Σ, respectively, and we have the following:

$$\ell_0^*=-\frac{N}{2}\log\,\det\left(\frac{2\pi}{N}T\right)-\frac{NK}{2}. \tag{9}$$

Thus, the statistic can be given as follows: $-2\log\lambda=-N\log\frac{\det(W)}{\det(T)}$. We can further see that q is $KC+\frac{K(K+1)}{2}$ and r is $K+\frac{K(K+1)}{2}$.

We conducted MANOVA over four samples in Figure 1, and Table 2 shows the resultant average over 100 runs with SDs in parentheses. The P-value of MANOVA for (a) was high (0.53), whereas that for (b) [and (c)] was zero, meaning that MANOVA can discriminate (a) from (b) [and (c)]. However, the P-value of (d) was also high (0.94), meaning that MANOVA could not separate (a) from (d). Thus, we need another hypothesis test, which can distinguish (a) from (d).

```
Input: $\mathcal{X}(e_1, e_2, q)$: Input three vectors of genes $e_1, e_2$ and SNP $q$.
    $\alpha_m$: Significance level for MC test
Output: One if two genes $e_1$ and $e_2$ are randomly generated in terms of
    SNP $q$; otherwise zero.
MC_test($e_1$, $e_2$, $q$, $\alpha_m$)
Compute $\ell_0^*$ according to Eq. (9).
Compute $\ell_1^*$ according to Eq. (10).
Compute $-2\log\lambda$ according to Eq. (11).
if $-2\log\lambda < \chi_2^2(\alpha_m)$ then
    return 1
else
    return 0
end if
```

Fig. 4. Pseudocode of MC test.

3.5.2 Box's M test We then consider the following hypotheses over the covariance:

$$H_0 : \Sigma_1 = \Sigma_2 = \cdots = \Sigma_C, \quad H_1 : \Sigma_i \neq \Sigma_j \text{ for some pair of } i \text{ and } j$$

Here, ℓ_0^* can be given by ℓ_1^* of MANOVA [i.e. Equation (8)], and ℓ_1^* can be obtained by using maximum likelihood estimators $\bar{\boldsymbol{x}}_k$ and S_k for μ_k and Σ_k, respectively, in Equation (1).

$$\ell_1^* = -\sum_{c=1}^{C} \frac{N_c}{2} \log \det(2\pi S_c) - \frac{NK}{2} \qquad (10)$$

Thus, the statistic is $-2\log\lambda = \sum_{c=1}^{C} N_c \log \det(S_c^{-1} S)$. Here, q is $KC + \frac{KC(K+1)}{2}$ and r is $KC + \frac{K(K+1)}{2}$.

We run Box's M test over four samples in Figure 1, and Table 2 shows the results. This result shows that the P-value of (a) was high (0.70), whereas that of (d) was zero, meaning that M-test separated (a) from (d). However, this time, this test could not discriminate (a) from (b) [and (c)], since the P-value of (b) [and (c)] was also high. Thus, this result showed that Box's M test can be a complement of MANOVA, implying that we can combine these two tests for detecting random distributions such as Figure 1a.

3.5.3 MC test (MANOVA + M Test) We finally consider the following hypotheses over both the means and covariances:

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \cdots = \boldsymbol{\mu}_C \text{ and } \Sigma_1 = \Sigma_2 = \cdots = \Sigma_C$$
$$H_1 : \boldsymbol{\mu}_i \neq \boldsymbol{\mu}_j \text{ or } \Sigma_i \neq \Sigma_j \text{ for some pair of } i \text{ and } j$$

We emphasize that this test suits our purpose the most, although this is an unpopular statistic and not named. We then call this test as *MC* test. Interestingly, ℓ_0^* of this test is given by ℓ_0^* of MANOVA, i.e. Equation (9) and ℓ_1^* is given by ℓ_1^* of M test, i.e. Equation (10). Thus, the statistic of MC test is given as follows:

$$-2\log\lambda = \sum_{c=1}^{C} N_c \log \det(S_c^{-1} S_T), \qquad (11)$$

since $\frac{1}{N} T = S_T$. Here, $q = KC + \frac{KC(K+1)}{2}$ and $r = K + \frac{K(K+1)}{2}$, meaning that *df* is 10 in our case. Figure 4 shows a pseudocode of MC test. We can write MC test by function MC_test(e_1, e_2, q, α_m), having significance level α_m as an input which removes given combination (e_1, e_2, q) if its P-value is larger than α_m, meaning that a larger number of combinations can be removed if α_m is smaller. This function outputs one if (e_1, e_2, q) should be pruned; otherwise zero.

We checked the performance of MC test using synthetic four samples of Figure 1. Table 2 shows that all P-values are zero, except (a) with the P-value of 0.60, indicating that MC test can successfully detect (a) out of the four examples and is expected to work on real data as well.

3.6 Proposed procedure

Figure 5 shows a pseudocode of our entire procedure. We can first check each pair of a gene and a SNP by LDA, and if the log-likelihood is high, this pair is stored to be pruned. We then generate all possible combinations of two genes and a SNP out of given data. For each of these combinations, it is first pruned if it contains the stored gene–SNP pair. Then, LDA and MC test are run in sequence for pruning, and finally interaction test is applied to the remaining. Hereafter, we call our proposed procedure FTGI, standing for Fast finding Three-way Gene Interactions, whereas we call the approach of running Interaction Test Only over all possible combinations as ITO. More details of our proposed method is shown in the Supplementary Material.

```
Input: $\mathcal{X}$: Input dataset
    $\alpha_i$: Significance level for interaction test,
    $\alpha_m$: Significance level for MC test
Output: $\mathcal{I}$: Interacting gene pairs with the corresponding SNPs
Proposed_procedure($\mathcal{X}$, $\alpha_i$, $\alpha_m$)
for each pair of gene $e \in E$ and SNP $q \in Q$ do
    // Pruning by LDA: One gene and a SNP
    if LDA($e$, $q$, $\alpha_i$) == 1 then
        $\mathcal{F} \leftarrow \mathcal{F} \cup (e, q)$
    end if
end for
for each combination of genes $e_1 \in E, e_2 \in E$ and SNP $q \in Q$ do
    if $(e_1, q) \notin \mathcal{F}$ and $(e_2, q) \notin \mathcal{F}$ then
        // Pruning by MC test
        if MC_test($e_1$,$e_2$,$q$,$\alpha_m$) == 1 then
            This combination should be in Area II. go to Pruned
        end if
        // Pruning by LDA: Two genes and a SNP
        if LDA($e_1$,$e_2$,$q$, $\alpha_i$) == 1 then
            This combination should be in Area I. go to Pruned
        end if
        // Interaction test for unpruned combinations
        if Interaction_test($e_1$, $e_2$, $q$, $\alpha_i$) == 1 then
            $\mathcal{I} \leftarrow \mathcal{I} \cup (e1, e2, q)$
        end if
    end if
    Pruned
end for
```

Fig. 5. Pseudocode of our entire procedure: FTGI.

4 EXPERIMENTS

4.1 Data

We used the human brain-derived dataset of Myers *et al.* (2007), which originally has 193 rows (individuals) and 14 078 numerical columns (corresponding to gene expressions) and 366 140 categorical columns (corresponding to SNPs). We first removed the columns containing missing values and the columns which have a genotype to which only less than 10 individuals are assigned. Our purpose is to find three-way gene interactions, and so we further removed SNPs which are neither in coding regions nor in introns, by specifying genes on sequences using the FTP site of NCBI Mapviewer for *Homo sapiens*. Finally, we obtained 5269 numerical vectors (in expression of genes) and 13 411 categorical vectors (in genotypes of SNPs) for 193 individuals, which we call the *Source* dataset. Myers *et al.* (2007) collected the original dataset from human brains, and so we focused on neurodegenerative diseases [including Alzheimer's disease (AD) and Parkinson's disease, etc.] out of five disease pathways in the KEGG disease database (Kanehisa *et al.*, 2008), resulting in 142 genes which we call *Neuro*. All experiments were run on a machine with Dual-Core

AMD Opteron 2222 SE (3.0 GHz) and 18 GB RAM. Throughout Section 4, each P-value is shown by $\log_{10}(P\text{-value})$.

4.2 Results and discussion

4.2.1 Speeding-up finding three-way interactions and pruning accuracy We examined the improvement in time efficiency by FTGI over ITO. Figure 6 shows the real computation time of ITO and FTGI, when we changed the number of combinations randomly chosen from the source dataset. We here focused on Area II of Figure 3 only, since we found that in the Source dataset of Area I had only a small number of examples, which do not affect the efficiency greatly. This figure clearly shows that as α_m decreased, the amount of running time of FTGI became smaller for any size of inputs, by pruning a larger number of them. In particular, at α_m of 0.001, FTGI runs approximately 10 times faster than ITO, resulting in only ~2 h for 10^7 combinations, being a sizable improvement. This means that for 5×10^{10} (= 50 000 SNPs × 1000 genes × 1000 genes) combinations, FTGI just needs only a couple of days with 100 CPUs, while ITO needs more than a month.

The α_m controls the number of pruned combinations, and Table 3 shows the *pruning rate*, i.e. the ratio of pruned combinations to all input combinations, with varying α_m for 10^7 input combinations. We further checked the *pruning accuracy*, which can be defined as the overlap between the resultant top $\mathcal{K}$ (set at 100) combinations by P-values of ITO and those of FTGI. Table 3 shows that for α_m of 0.05, FTGI can prune around ~70% of input combinations with pruning accuracy of almost 100%. If α_m is reduced to 0.001, ~94% inputs can be pruned, keeping the pruning accuracy of ~85%. This high pruning rate effects the time efficiency of FTGI.

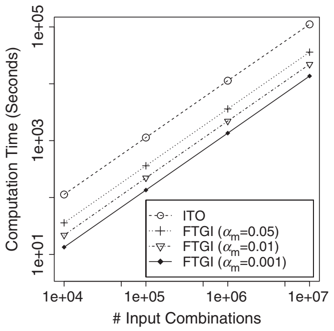

Fig. 6. Computation time improvement by reducing α_m.

We note that all results in this section were averaged over three runs at each corresponding setting.

4.2.2 Detecting three-way interactions We then generated all combinations from the Source dataset, focusing on the genes in Neuro, meaning that we had totally ~3×10^8 combinations (= 13 411 SNPs × 142 genes × 142 genes). We then run FTGI with α_m of 0.001 over these combinations. Figure 7 shows the gene expressions of the resultant top 10 combinations in terms of P-values. We note that these P-values of interaction test were computed by the procedure in Section 3.2. Each of Figure 7 is a 2D diagram on which expression values of the corresponding two genes are plotted with Contour lines for each genotype. This figure shows that the topographical distribution of different genotypes are clearly crossed in all cases, meaning that in each of all the top 10 combinations, genes are interacting in expression, being controlled by genotypes, as shown in Figure 1d.

Table 4 shows the detail (Gene name for one SNP and the name with GeneID, the definition and the pathway for each of two

Table 3. Pruning rates and pruning accuracies (top 100) at three α_m values of FTGI for 10^7 combinations

α_m	0.05	0.01	0.001
Pruning rate	0.7095	0.8611	0.9354
Pruning accuracy (top 100)	0.9967	0.9567	0.8467

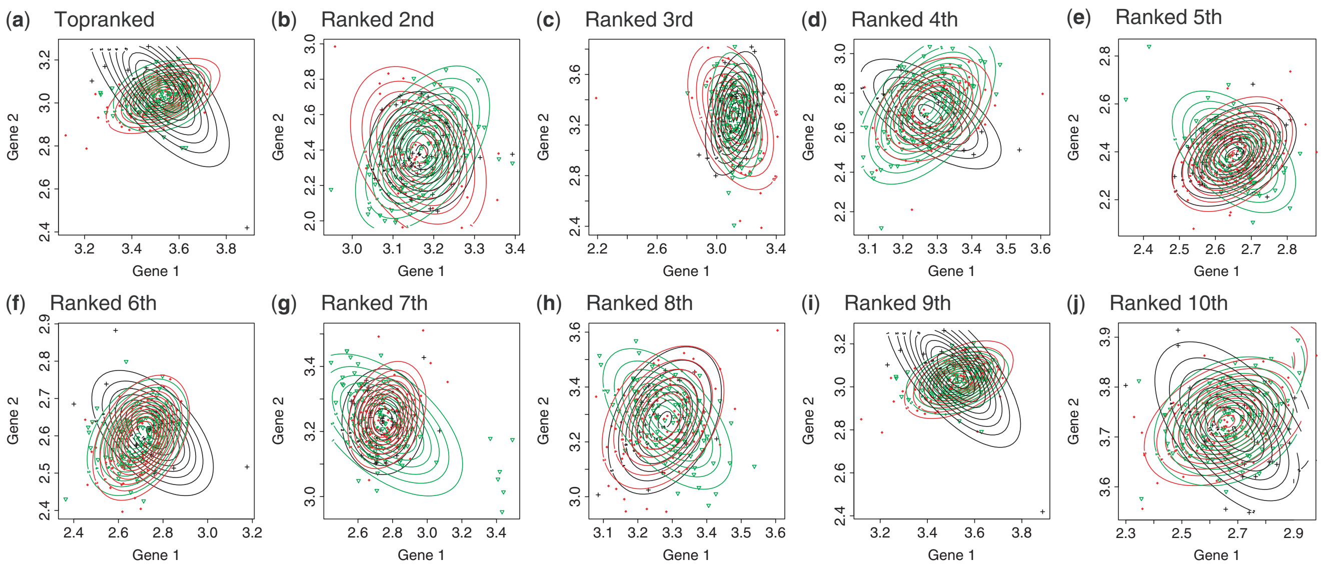

Fig. 7. Expressions of two genes under three genotypes of another gene for top 10 (**a–j**) ranked three-way interactions out of 3×10^8 combinations.

Table 4. Details of the top 10 three-way interactions in Figure 7

	P-value	SNP (GeneID and name)	Gene 1		Gene 2	
			Name (GeneID)	Definition	Name (GeneID)	Definition
1	−8.91108	rs7487429 (113251, LARP4)	COX6C (1345)	Cytochrome c oxidase subunit VIc (EC:1.9.3.1)	UBA1 (7317)	Ubiquitin-like modifier activating enzyme 1 (EC:6.3.2.19)
2	−8.4901	rs13086670 (80163, FLJ11827)	RERE (473)	Arginine-glutamic acid dipeptide (RE) repeats	TNFRSF1A (7132)	Tumor necrosis factor receptor superfamily, member 1A
3	−8.10611	rs2175200 (439992, RPS3AP5)	ATP5D (513)	ATP synthase, H+ transporting, mitochondrial F1 complex, δ subunit (EC:3.6.1.14)	ITCH (83737)	ITCHY E3 ubiquitin protein ligase homolog (mouse)
4	−8.06076	rs2797425 (55227, LRRC1)	ATP5G1 (516)	ATP synthase, H+ transporting, mitochondrial F0 complex, subunit C1 (subunit 9)	ATP5H (10476)	ATP synthase, H+ transporting, mitochondrial F0 complex, subunit d (EC:3.6.1.14)
5	−8.02645	rs7116710 (440031, LOC440031)	NCSTN (23385)	Nicastrin	HSPA5 (3309)	Heat shock 70kDa protein 5 (glucose-regulated protein, 78kDa)
6	−8.02495	rs2058619 (728730, LOC728730)	NDUFA8 (4702)	NADH dehydrogenase (ubiquinone) 1 α subcomplex, 8, 19 kDa (EC:1.6.5.3 1.6.99.3)	NDUFA6 (4700)	NADH dehydrogenase (ubiquinone) 1 α subcomplex, 6, 14kDa
7	−8.0149	rs1893261 (25833, POU2F3)	ALS2 (57679)	Amyotrophic lateral sclerosis 2 (juvenile)	SLC25A6 (293)	Solute carrier family 25 (mitochondrial carrier; adenine nucleotide translocator), member 6
8	−7.86801	rs1571176 (9044, BTAF1)	ATP5G1 (516)	ATP synthase, H+ transporting, mitochondrial F0 complex, subunit C1 (subunit 9)	ATP5J (522)	ATP synthase, H+ transporting, mitochondrial F0 complex, subunit F6 (EC:3.6.1.14)
9	−7.84081	rs12425705 (91012, LASS5)	COX6C (1345)	Cytochrome c oxidase subunit VIc (EC:1.9.3.1)	UBA1 (7317)	Ubiquitin-like modifier activating enzyme 1 (EC:6.3.2.19)
10	−7.73205	rs12698191 (393078, tcag7.1023)	NDUFA10 (4705)	NADH dehydrogenase (ubiquinone) 1 α subcomplex, 10, 42kDa (EC:1.6.5.3 1.6.99.3)	COX4 (1327)	Cytochrome c oxidase subunit IV isoform 1 (EC:1.9.3.1)

interacting genes in expression) of the 10 three-way interactions in Figure 7, all information in this table being retrieved from KEGG.[2] For example, the first interaction of Table 4 shows the switching mechanism of two genes, COX6C and UBA1, being controlled by a SNP in LARP4.

4.2.3 Validating detected interactions with permutations To confirm the statistical significance of the detected three-way interactions, we conducted permutations by measuring P-values of 'null data', generated in the following three manners, and comparing them with those of the interactions we detected.

- Null data 1: we randomly chose 10 000 combinations out of all combinations using the Source dataset (13 411 SNPs × 5269 genes × 5269 genes) and randomly permuted the genotypes of these combinations 100 times. Totally, we had one million null examples.
- Null data 2: we randomly chose 10 000 combinations out of all combinations using the Neuro dataset (13 411 SNPs × 142 genes × 142 genes) and randomly permuted the genotypes of these combinations 100 times. Totally, we had one million null examples.
- Null data 3: we permuted the genotypes of each of the detected top 10 interactions in Figure 7 one million times, resulting in one million null examples for each combination.

We first show the results of permutation tests when we use Null data 1 and 2. Figure 8 shows the distribution of P-values of null examples, being located in the right side, for Null data 1 and 2. In this figure, the distribution of P-values for the top 10 000 interactions detected by FTGI is located in the left side. This figure shows that the red-colored distribution is clearly separated from the black-colored one, meaning that the detected three-way interactions have significantly small P-values. For Null data 3, we show the result, focusing on two cases (the top and the 10th interactions), since the trend of results was kept the same for all 10 interactions in Table 4. Figure 9 shows the distribution of P-values of null examples generated from the top interaction (or the 10th), with the P-value of the top (or the 10th) interaction by an arrow. This figure indicates that the P-value of the top (or the 10th) interaction is clearly distant from the P-value distribution of null examples, implying that P-values of the detected interactions are statistically significant.

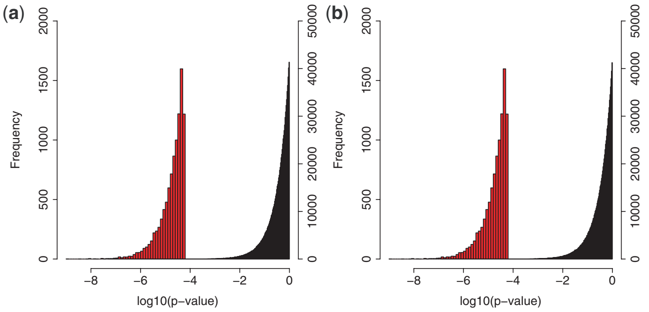

Fig. 8. Distributions (left side) of P-values of the top 10 000 interactions detected by FTGI, with those (right side) of Null data (**a**) 1 and (**b**) 2.

4.2.4 Validating detected interactions with GEO To confirm the reliability of the interactions in Table 4, we tried to find, for each gene pair, the switching mechanism in expression which can be controlled by some experimental condition of gene expression. This is because

[2] The Supplementary Material shows annotations by Reactome (Vastrik *et al.*, 2007) for interacting genes.

Table 5. Results of interaction test over the datasets from GEO

Rank	Gene pair	#datasets from GEO	GDS	*P*- value	#ex. class 1	#ex. class 2	Annotation
1	{COX6C,UBA1}	117	GDS2960_1	−3.9532	60	41	Marfan syndrome: cultured skin fibroblasts
2	{RERE,TNFRSF1A}	284	GDS2736_25	−5.9049	19	15	Malignant fibrous histiocytoma and various soft tissue sarcomas
3	{ATP5D,ITCH}	324	GDS1875_3	−5.1235	27	24	Host cell response to HIV-1 Vpr-induced cell cycle arrest
4	{ATP5G1,ATP5H}	392	GDS2733_1	−7.9996	17	17	Cytosine arabinoside effect on Ewing’s sarcoma cell line
5	{NCSTN,HSPA5}	102	GDS2545_5	−6.4398	63	25	Metastatic prostate cancer (HG-U95A)
6	{NDUFA8,NDUFA6}	142	GDS2733_4	−4.7027	17	16	Cytosine arabinoside effect on Ewing’s sarcoma cell line
7	{ALS2,SLC25A6}	108	GDS1627_2	−3.2808	16	15	Breast cancer cell lines response to chemotherapeutic drugs
8	{ATP5G1,ATP5J}	418	GDS2960_1	−3.1628	60	41	Marfan syndrome: cultured skin fibroblasts
9	{COX6C,UBA1}	117	GDS2960_1	−3.9532	60	41	Marfan syndrome: cultured skin fibroblasts
10	{NDUFA10,COX4}	232	GDS2643_9	−6.2133	13	12	Waldenstrom’s macroglobulinemia: B lymphocytes and plasma cells

For each gene pair of 10 interactions in Table 4, the number of datasets obtained from GEO, the GDS which gave the smallest *P*-value, the *P*-value, the number of examples (ex.) in two classes of the GDS and the annotation of the GDS are shown.

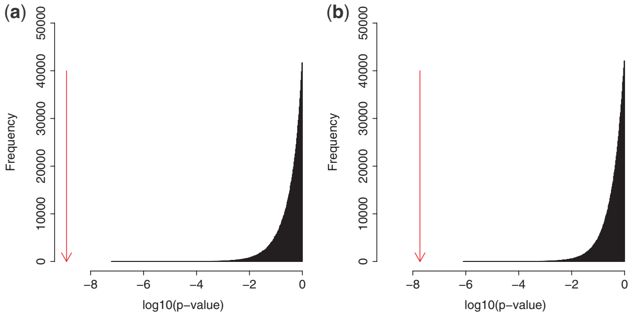

Fig. 9. The *P*-values of the (**a**) top and (**b**) 10th interactions (shown by arrows) and the distributions of *P*-values of the corresponding null examples generated.

if found, this directly means that the corresponding gene pair can be controlled by another categorical factor, such as genotypes of another gene.

For this purpose, we used GEO (version of June 1, 2009; Barrett *et al.*, 2007), from which we found 2089 GDSs (gene datasets) which are annotated. Out of the 2089 datasets, we selected datasets which satisfy all the following four conditions for each gene pair in Table 4: (i) expression values of the corresponding gene pair are contained; (ii) the total number of experiments is ≥50; (iii) experimental conditions can be divided into two or more classes; and (iv) each class has 10 or more experiments. We then obtained 36 datasets.[3] For each gene pair of the top 10 list, we conducted interaction test by using pairwise (binary) classes in each dataset and ranked them according to *P*-values of interaction test. Table 5 shows a list of datasets, each giving the lowest *P*-value for each gene pair of the 10 interactions in Table 4. For example, for COX6C and UBA1, the gene pair of the first interaction of Table 4, we found a switching mechanism in GDS2960_1 with the *P*-value of −3.9532, showing the statistical significance of this mechanism. This directly indicates that there must exist a switching mechanism in expression between these two genes under the alteration of experimental conditions which is specified by the annotation of GDS2960_1. In fact, Table 5 indicates that the switching mechanism happens between patients of Marfan syndrome and controls. This type of explanation is possible for all 10 GDSs in Table 5 by using annotations in this table. As well all *P*-values shown in Table 5 are small enough,[4] supporting the reliability of the three-way interactions in Table 4 which our method detected. Furthermore, Figure 10 shows the real expression values of two genes, being categorized into two classes, for each GDS of Table 5. These orthogonal Contour plots also assist the reliability of three-way interactions that we detected in Table 4.

We further briefly checked the genes having SNPs in the first and the third interactions in Table 4: (i) the first interaction in Table 4 has two genes, COX6C and UBA1, which is controlled by a SNP in LARP4, i.e. La ribonucleoprotein domain family member 4. This gene was already known as an important gene in both AD and aging, being already pointed out that LARP4 increases expression with increasing AD progression and normal aging (Miller *et al.*, 2008). As our focus was on 142 genes on neurodegenerative diseases including AD, the known function on LAPR4 is consistent enough with the interaction with COX6C and UBA1, being possibly in the switching mechanism. (ii) The gene with the SNP in the second interactions in Table 4 was a hypothetical one, but the third interaction has two genes, ATP5D and ITCH, being controlled by a SNP in RPS3AP5, which is a pseudogene of RPS3A, i.e. ribosomal protein S3A. This gene is known to be downregulated in the same manner as some genes in oxidative phosphorylation pathway (Welle *et al.*, 2003), which includes ATP5D. Thus, these observations reveal the possibility that the third interaction also may exist as the switching mechanism in expression of two genes, i.e. ATP5D and ITCH.

Overall our extensive analysis has implied that the detected three-way interactions can exist. These results show the potential of our approach to explicate complex biological systems appearing in modern biology and medical sciences.

[3] In each GDS, if it has more than two classes or replicated experiments, we consider all possible pairwise combinations of them. We then name generated multiple datasets from one GDS (e.g. GDS2960) those like GDS2960_1, GDS2960_2, etc. This results in that the number of datasets we used could be >36. The actual number of datasets for each gene pair is shown in Table 5.

[4] For each gene pair, not only the dataset giving the top *P*-value but also 10 datasets providing the top 10 *P*-values are shown in the Supplementary Material. All *P*-values in the Supplement Material are small, showing the statistical significance of the switching mechanism of each gene pair.

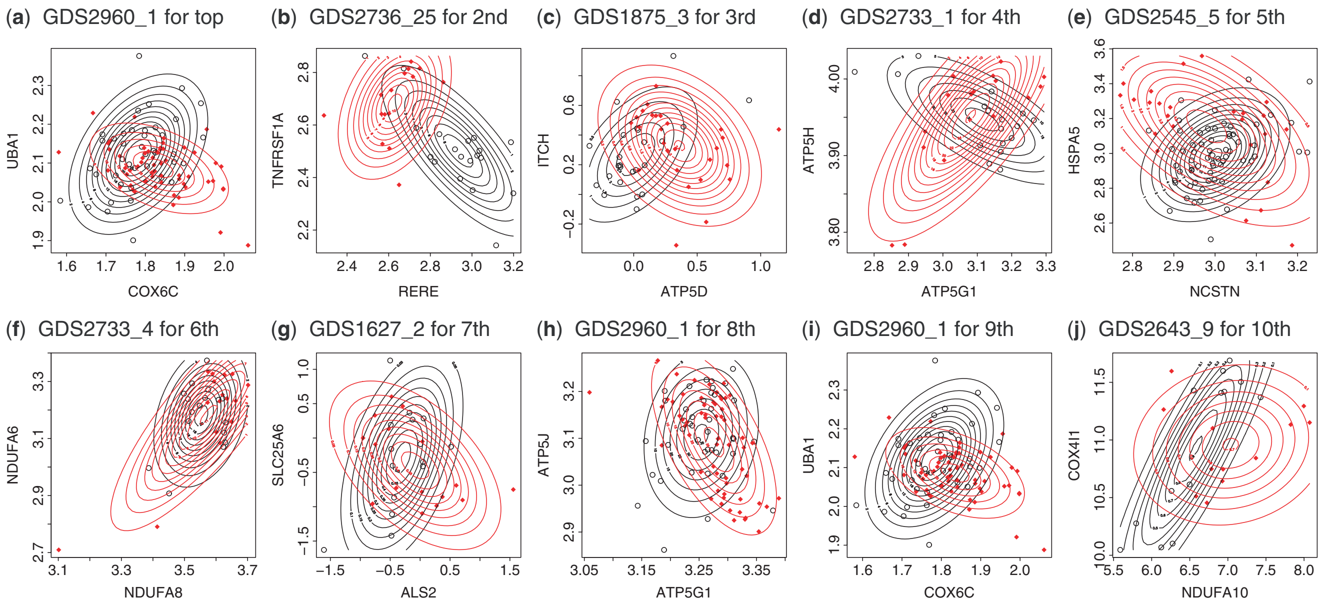

Fig. 10. (**a–j**) Expressions of two genes which give the smallest *P*-value of interaction test in the corresponding GDS of GEO.

5 CONCLUDING REMARKS

We have presented a fast method for finding three-way gene interactions from transcript-and genotype-data and showed experimental results obtained by applying this method to $\sim 3 \times 10^8$ human brain samples. In our experiments, we confirmed the three-way interactions that we found in various manners. Possible future work would be to apply our approach to various types of transcript- and genotype-data further to uncover three-way gene interactions, i.e. biological switches by genotypes.

ACKNOWLEDGEMENTS

The authors would like to thank anonymous reviewers for their helpful comments and advice.

Funding: Grant-in-Aid for Young Scientists 20700134, 20700269, 21680025 from Ministry of Education, Culture, Sports, Science and Technology (MEXT, in parts); the Functional RNA Project of New Energy and Industrial Technology Development Organization (NEDO, in parts).

Conflict of Interest: none declared.

REFERENCES

Ayer,D.E. and Eisenman,R.N. (1993) A switch from myc:max to mad:max heterocomplexes accompanies monocyte/macrophage differentiation. *Genes Dev.*, **7**, 2110–2119.

Aylor,D.L. and Zeng,Z.-B. (2008) From classical genetics to quantitative genetics to systems biolog: modern epistasis. *PLoS Genet.*, **4**, e1000029.

Balding,D.J. (2006) A tutorial on statistical methods for population association studies. *Nat. Genet.*, **7**, 781–791.

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

Cordell,H.J. (2002) Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Hum. Mol. Genet.*, **11**, 2463–2468.

Dixon,A.L. *et al.* (2007) A genome-wide association study of global gene expression. *Nat. Genet.*, **39**, 1202–1207.

Hastie,T. *et al.* (2001) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York.

Kanehisa,M. *et al.* (2008) KEGG for linking genomes to life and the environment. *Nucleic Acids Res.*, **36**, D480–D484.

Li,K.-C. *et al.* (2004) A system for enhancing genome-wide coexpression dynamics study. *Proc. Natl Acad. Sci. USA*, **101**, 15561–15566.

Lo,S.-H. *et al.* (2008) Discovering interactions among brca1 and other candidate genes associated with sporadic breast cancer. *Proc. Natl Acad. Sci. USA*, **105**, 12387–12392.

Marchini,J. *et al.* (2005) Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat. Genet.*, **37**, 413–417.

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

McCarthy,M.I. and Hirschhorn,J.N. (2008) Genome-wide association studies: Past, present and future. *Hum. Mol. Genet.*, **17** (Review Issue 2), R100–R101.

McCullagh,P. and Nelder,J. (1989) *Generalized Linear Models*. 2nd edn. Chapman & Hall CRC, Boca, Raton.

Miller,J.A. *et al.* (2008) A systems level analysis of transcriptional changes in Alzheimer's disease and normal aging. *J. Neurosci.*, **28**, 1410–1420.

Myers,A.J. *et al.* (2007) A survey of genetic human cortical gene expression. *Nat. Genet.*, **39**, 1494–1499.

Nica,A.C. and Dermitzakis,E.T. (2008) Using gene expression to investigate the genetic basis of complex disorders. *Hum. Mol. Genet.*, **17** (Review Issue 2), R129–R134.

Schadt,E.E. *et al.* (2008) Mapping the genetic architecture of gene expression in human liver. *PLoS Biol.*, **6**, e107.

Vastrik,I. *et al.* (2007) Reactome: a knowledge base of biologic pathways and processes. *Genome Biol.*, **8**, R39.

Welle,S. *et al.* (2003) Gene expression profile of aging in human muscle. *Physiol. Genomics*, **14**, 149–159.

Zhang,J. *et al.* (2007) Extracting three-way gene interactions from microarray data. *Bioinformatics*, **23**, 2903–2909.