Bioinformatics, 31(5), 2015, 656–664
doi: 10.1093/bioinformatics/btu727
Advance Access Publication Date: 30 October 2014
Original Paper

Genome analysis

eQTL epistasis: detecting epistatic effects and inferring hierarchical relationships of genes in biological pathways

Mingon Kang[1], Chunling Zhang[2], Hyung-Wook Chun[3], Chris Ding[1], Chunyu Liu[2] and Jean Gao[1,*]

[1]Department of Computer Science and Engineering, University of Texas at Arlington, Arlington, TX 76019, USA, [2]Department of Psychiatry, University of Illinois at Chicago, Chicago, IL 66012, USA and [3]Department of Mathematics, University of Texas at Arlington, Arlington, TX 76019, USA

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on August 8, 2014; revised on September 28, 2014; accepted on October 25, 2014

Abstract

Motivation: Epistasis is the interactions among multiple genetic variants. It has emerged to explain the 'missing heritability' that a marginal genetic effect does not account for by genome-wide association studies, and also to understand the hierarchical relationships between genes in the genetic pathways. The Fisher's geometric model is common in detecting the epistatic effects. However, despite the substantial successes of many studies with the model, it often fails to discover the functional dependence between genes in an epistasis study, which is an important role in inferring hierarchical relationships of genes in the biological pathway.

Results: We justify the imperfectness of Fisher's model in the simulation study and its application to the biological data. Then, we propose a novel generic epistasis model that provides a flexible solution for various biological putative epistatic models in practice. The proposed method enables one to efficiently characterize the functional dependence between genes. Moreover, we suggest a statistical strategy for determining a recessive or dominant link among epistatic expression quantitative trait locus to enable the ability to infer the hierarchical relationships. The proposed method is assessed by simulation experiments of various settings and is applied to human brain data regarding schizophrenia.

Availability and implementation: The MATLAB source codes are publicly available at: http://biomecis.uta.edu/epistasis.

Contact: gao@uta.edu

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

1 Introduction

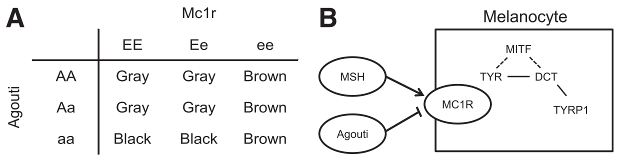
Epistasis, the interaction effect among unlinked loci or between genes, makes a considerable contribution to phenotypic variation in the polygenic mechanism of complex human diseases such as psychiatric disorder, diabetes and cancer. Nonetheless, many studies that identify the genetic susceptible factors tend to ignore the interaction effects between loci (Carlborg and Haley, 2004; Pandey *et al.*, 2012). It is not surprising at all that the investigations of the genetic variability of individual genes by genome-wide association study give the explanation only up to 40% in psychiatric disorder (Pandey *et al.*, 2012). This fact implies that there should indeed be a 'missing heritability'. The missing heritability may include the interaction between multiple genes and environmental factors (Eichler *et al.*, 2010; Pandey *et al.*, 2012).

Although epistasis studies traditionally take into account qualitative categories of phenotypes, the recent rapid development of high-throughput technologies, such as gene microarray, enables one to characterize the quantitative interaction, e.g. gene expression. Gene expression is an intermediate molecular phenotype between genotype and the higher level phenotypes such as human diseases. For the quantitative analysis, the expression quantitative trait loci (eQTL) mapping studies have prevailed to identify susceptible genetic loci and capture the insight of the genetic architecture of gene expression. Furthermore, the importance of detecting the quantitative epistasis analysis is recently being emphasized in many studies (Huang *et al.*, 2013; Mackay, 2014).

When William Bateson created the terminology, epistasis, it described the non-independent prediction of segregation ratios by the action of individual gene (Bateson, 1909). However, epistasis has been interpreted in various ways to describe different phenomena for the last 100 years (Cordell, 2002; Phillips, 2008). The uses of the term, epistasis, were viewed with three different perspectives: (1) compositional epistasis, (2) statistical epistasis and (3) functional epistasis (Phillips, 2008). To be short, compositional epistasis represents the biological phenomena that one allelic effect masks an allele at another locus, which describes the traditional meaning of epistasis. The presence of the compositional epistasis can be interpreted in the biological hierarchy between genes (Aylor and Zeng, 2008; Cordell, 2009; Phenix *et al.*, 2011). Statistical epistasis takes into account Fisher's model and its deviations, which describe the combinatory genetic effect within a population. The functional epistasis, which is also called protein–protein interaction, is not dealt with in this article since the functional epistasis considers the molecular interactions between protein without genetic description.

Compositional epistasis emphasizes the functional dependence between genes. It aims to infer gene regulatory networks or signaling pathways. Epistasis occurs when one allelic effect is modified or blocked by an allele at another locus, or when the combined multiple genetic variants produce non-Mendelian segregation ratios against the individual genetic effect. Since genes must be interacting with others, at least on the same pathway, the interactive effect to phenotype is clear. When a genetic mutation has the 'stopping' or 'standing above' effect to another mutation, the mutation is said to be epistatic and can be interpreted as the gene is downstream of the other. The epistatic relationship between the Melanocortin 1 receptor (Mc1r) and agouti on the coat color genetic pathway of a mouse is a good example to show the epistatic relationship (Fig. 1) (Phillips, 2008). As shown in the 3 × 3 genotype contingency table in Figure 1A, the body colors of the offspring were investigated with all the combinations of knockout mutations of the genes. Mc1r seems to regulate the functionality of the genetic effect of agouti. That is, mc1r is the downstream gene of agouti in the melanocyte pathway. The melanocyte pathway can elucidate the epistatic relationship between the genes in Figure 1B.

The quantitative epistasis analysis, eQTL epistasis, holds significant promise in inferring the hierarchical relationships between genes in biological pathways and its enrichment as well as the qualitative epistasis study (Aylor and Zeng, 2008; Cordell, 2009; Phenix *et al.*, 2011). The hierarchical interpretation of the epistatic gene pairs enables one to construct or enrich the biological pathways. A number of regression-based methods have been suggested for inferring the gene regulatory network and their hierarchical relationships. As the classical quantitative epistasis analysis, the regression-based approaches investigated the data that include reciprocal effects on the triplet combination of deletion mutant of a gene (Aylor and Zeng, 2008; Phenix *et al.*, 2011). The studies have succeeded in inferring the hierarchical pathways of the genes as well as a rule between the interaction, for example, activating or repressing the downstream gene. However, the experiments were required tremendously expensive experiments for gene deletions, which is unfeasible to measure the effects of all possible gene deletions in practice.



Agouti	Mc1r EE	Mc1r Ee	Mc1r ee
AA	Gray	Gray	Brown
Aa	Gray	Gray	Brown
aa	Black	Black	Brown

Fig. 1. (**A**) The 3 × 3 genotype contingency table of all the combinations of knockout mutations of the genes. Mc1r is the downstream gene of agouti in the melanocyte pathway (**B**) The functional relationship of Mc1r and agouti genes in the melanocyte pathway

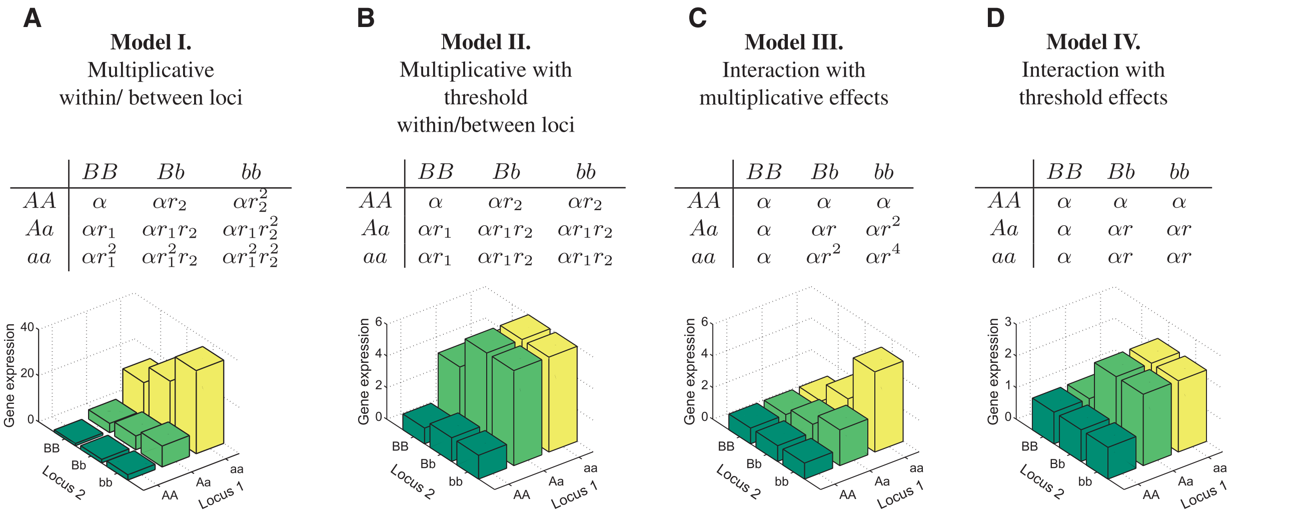
Most epistasis studies have been performed by using the statistical approach of Fisher's. Recently, the limitations of the statistical epistasis have been argued, and the development of a new approach has consistently been demanded (Cordell, 2002, 2009; VanderWeele, 2010; Wan *et al.*, 2013). In this article, we develop a novel method to solve two problems: (i) detecting the compositional epistasis and (ii) identifying the genetic hierarchical relationship between genes based using a regression model. We first present the biologically putative epistasis models and Fisher's model in Sections 2.1 and 2.2. Then, a generic epistasis model (GEM) is suggested to represent the biological models in Section 2.3. The optimal solutions for the parameter estimation of GEM are derived in Section 2.4. For the second problem, we propose a statistical approach to determine the hierarchical relationships when an epistasis occurs in the multiple genes in Section 2.5. In the simulation studies in Section 3.1, we show the performance of GEM, comparing Fisher's model on the various designed settings. We also apply GEM to the human brain data of schizophrenia, where the epistasis findings are described and a gene hierarchy on schizophrenia is inferred.

2 Methods

2.1 Biological models for eQTL epistasis

Three explicit models for the interaction between multiple loci were elucidated (Marchini *et al.*, 2005) for genome-wide epistasis, which are: (i) multiplicative within and between loci (corresponding to Model I in our work), (ii) two-locus interaction multiplicative effects (corresponding to Model III) and (iii) two-locus interaction threshold effects (corresponding to Model IV of our work). Model I specifies the minor allelic effects increasing in a multiplicative fashion both within and between loci with the different genetic effect size (r_1 and r_2) of each locus. The baseline effect α describes the background effect when no mutation exists. Model III is equivalent to Fisher's model, where it has equal genetic effect size at the two loci ($r_1 = r_2 = r$). Model IV is a deviation of Model III with a threshold, where the equal effect of the mutation (no matter what the minor allele number) is considered.

Although the above models were designed for the odds of diseases (qualitative analysis), the models can be converted for the eQTL epistasis analysis (quantitative analysis) by replacing the odds with the gene expression levels. The new models are illustrated in Figure 2, in which the models are modified for the eQTL epistasis, and an additory model (multiplicative with threshold within and between loci, Model II) is added.



A

Model I.
Multiplicative
within/ between loci

	BB	Bb	bb
AA	α	αr_2	αr_2^2
Aa	αr_1	$\alpha r_1 r_2$	$\alpha r_1 r_2^2$
aa	αr_1^2	$\alpha r_1^2 r_2$	$\alpha r_1^2 r_2^2$

B

Model II.
Multiplicative with
threshold
within/between loci

	BB	Bb	bb
AA	α	αr_2	αr_2
Aa	αr_1	$\alpha r_1 r_2$	$\alpha r_1 r_2$
aa	αr_1	$\alpha r_1 r_2$	$\alpha r_1 r_2$

C

Model III.
Interaction with
multiplicative effects

	BB	Bb	bb
AA	α	α	α
Aa	α	αr	αr^2
aa	α	αr^2	αr^4

D

Model IV.
Interaction with
threshold effects

	BB	Bb	bb
AA	α	α	α
Aa	α	αr	αr
aa	α	αr	αr

Fig. 2. Epistasis models for two loci. The four scenarios of the epistatic effects of two loci (A and B) to the gene expression levels are illustrated, where 'A' and 'a' represent the major allele and the minor allele, respectively, in a single locus, α is a baseline effect, r_1 and r_2 are the genotypic effects of A and B. The illustrations are depicted with $\alpha = 1$, $r_1 = 1.5$, $r_2 = 4$, $r = 1.5$

2.2 Deviation of Fisher's model

The deviation in the multiplicative model from Fisher's geometric model is common in detecting the epistatic effects. Fisher defined epistasis by using the interaction term in an additive fashion (Fisher, 1918). However, since the late 1960s, due to the dependency issue of the additive manner, a deviation of a multiplicative model for the interaction term has begun to be used, and this regression-based model is still most widely applied in many studies (Cordell, 2002; Schupbach *et al.*, 2010; Lee and Xing, 2012; Huang *et al.*, 2013). The statistical interaction model (called Fisher's model hereafter) is formed as,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \varepsilon, \tag{1}$$

where genotypes at a single locus are represented by the number of minor alleles and $\varepsilon \sim N(0, \sigma^2)$. Testing the significance of the interaction term β_{12} is performed by comparing with the following marginal model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon. \tag{2}$$

A significant interaction term represents that Equation (1) explains a significantly larger proportion of phenotypic variance in quantitative trait than Equation (2) which involves only marginal effects of the two individual genetic loci. Therefore, the significant interaction term determines the presence of the epistatic effects between the two loci. The significance is assessed by *t*-statistic.

Despite the substantial success by Fisher's model in epistasis studies, it is often imperfect to represent all putative epistatic phenomena in biology. The limitations of Fisher's model for the compositional epistasis have been investigated by researchers (Cordell, 2002, 2009; VanderWeele, 2010; Wan *et al.*, 2013).

2.3 GEM for eQTL epistasis

We propose a GEM for detecting complex eQTL epistatic effects with adapting kernel function $\varphi()$, where the kernel function is redefined for the targeting epistatic effects.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} r_1^{\varphi(x_1)} r_2^{\varphi(x_2)} + \varepsilon, \tag{3}$$

where y is the continuous values of the observed phenotype (e.g. gene expression level), and x_1 and x_2 are the genotypic data of the two loci, where single-nucleotide polymorphisms (SNPs) are considered in diploid organisms. β_1 and β_2 are the coefficients of the marginal effect on x_1 and x_2, respectively. β_{12} is the coefficient of the epistatic genotypic effect between the two loci. r_1 and r_2 show the genotypic effect sizes of the two loci. Note that $r_i = 1$ describes no epistatic effect at locus i. $0 < r_i < 1$ represents the degrading effect, whereas $r_i > 1$ shows the propagating effect by the allele at the ith locus.

Using the kernel function, we can consider various epistasis models in the single analytical expression form. For instance, the kernel function for Model I can be defined as

$$\varphi(x_i)_{\text{Model I}} = \begin{cases} 2 & \text{if } x_i = aa. \\ 1 & \text{if } x_i = \text{Aa}. \\ 0 & \text{if } x_i = AA. \end{cases} \tag{4}$$

In the same sense, the kernel function for Model II is

$$\varphi(x_i)_{\text{Model II}} = \begin{cases} 1 & \text{if } x_i = aa. \\ 1 & \text{if } x_i = \text{Aa}. \\ 0 & \text{if } x_i = AA. \end{cases} \tag{5}$$

More specifically, GEM for the instance of the two loci of 'Aa' and 'bb' in Model I becomes (see the table in Fig. 2A),

$$y = \beta_0 + \beta_1 + 2\beta_2 + \beta_{12} r_1^1 r_2^2. \tag{6}$$

Models III and IV can be considered as the specialized models of Models I and II respectively, where $r_1 = r_2 = r$. Therefore, the kernel functions for Models I and II can also easily deal with Models III and IV instead of using additional kernel definition. We will consider the two kernel functions throughout this study.

2.4 Optimization

Given a number of samples, we can characterize the global genetic effects by optimizing the parameters to the sample. Let $\mathbf{x}_i$ be a column vector of SNP at the ith locus, and $\mathbf{y}$ be a column vector of gene expression. The learning function is obtained by using least squares with a ridge regularization term,

$$\begin{aligned}\text{argmin}\quad & F(\mathbf{X},\mathbf{y};\beta,\mathbf{r})\\ & = \|\mathbf{y}-\frac{1}{w}(\beta_0+\beta_1\mathbf{x}_1+\beta_2\mathbf{x}_2+\beta_{12}r_1^{\varphi(\mathbf{x}_1)}r_2^{\varphi(\mathbf{x}_2)})\|^2\\ & +\lambda(\beta_0^2+\beta_1^2+\beta_2^2+\beta_{12}^2)\\ & \text{s.t. } r_1,r_2\geq 0,\end{aligned}\tag{7}$$

where $\mathbf{X}=\{\mathbf{x}_1,\mathbf{x}_2\}$, $\beta=\{\beta_0,\beta_1,\beta_2,\beta_{12}\}$, and $\mathbf{r}=\{r_1,r_2\}$. w gives a weight to compromise the different samples for the contingency table, and λ is a hyper-parameter for the ridge regularization. In this study, w is computed as the sum of the sample numbers of 3 × 3 different genetic combinations for two loci.

The learning function (7) is non-convex. Nevertheless, the gradient of the learning function with respect to each parameter can be easily computed. Since there is no analytic closed form for the optimal solution, the local optimum is obtained by the coordinate descent algorithm. The solution for the parameters will be presented here.

In the coordinate descent algorithm, the parameters are estimated iteratively: $\mathbf{r}=\{r_1,r_2\}$ are estimated by Proposition 2.1 with given β_1, β_2 and β_{12}, and then the coefficients β_{12},β_1 and β_2, are alternatively estimated by Proposition 2.2. The iterative optimization is performed until they converge. Note that the estimation sequence of the parameters is important in the proposed method, i.e. in the order of $\mathbf{r}=\{r_1,r_2\},\beta_{12},\beta_1$ and β_2. Consideration of the coefficients of the main effects (β_1 and β_2) prior to the epistatic effects ($\mathbf{r}=\{r_1,r_2\}$ and β_{12}) makes the epistatic effects to be neglected, but main effects to be emphasized. In other words, the estimation of the epistatic effects in advance of the main effects enables us to detect the epistatic effects even if there are no main effects.

The genotypic effect sizes r_1 and r_2 play an important role in determining the gene hierarchy. Fisher's model takes into account only the presence or absence of the interaction effects, i.e. epistasis. Therefore, it lacks to infer the hierarchy of the genes involved in the epistasis, which it consequently has limits to biologically interpret the epistatic phenomena. Introduction of the genotypic effect sizes r_1 and r_2 into the GEM makes it possible to infer the hierarchical orders of the genes. The detailed description for the identification approach of the hierarchical relationships in the epistasis will be shown in Section 2.5. The proofs of the propositions and the algorithm are provided in the Supplementary Material.

Testing the significance of the interaction term β_{12} is performed in the same way as Fisher's model presented in Section 2.2. The significant interaction term shows the presence of epistasis between two loci to the gene expression of a target gene.

Proposition 2.1 Let $\gamma_i=y_i-\beta_1x_{i1}-\beta_2x_{i2}$. Given β_1, β_2, β_{12} and r_2, the optimal genotypic effect size r_1 is given by,

$$r_1=\underset{s_i\in\mathrm{S}}{\text{argmin}}\,F(\mathbf{X},\mathbf{y};\beta,s_i,r_2),\tag{8}$$

where a set S is defined as,

$$\begin{aligned}\mathbf{S}=\Bigg\{&\{0,s_1,\ldots,s_k\}|d_hs^h+d_{h-1}s^{h-1}+\cdots+d_1s^1+d_0=0,\\ & d_I=\left[\sum_{i=1}^{n}-\gamma_i\beta_{12}r_2^{\varphi(x_{i2})}\varphi(x_{i1})\right]_{\varphi(x_{i1})-1=I}\\ & +\left[\sum_{i=1}^{n}\beta_{12}^2r_2^{2\varphi(x_{i2})}\varphi(x_{i1})\right]_{2\varphi(x_{i1})-1=I},\\ & s\geq 0,\ \ s\in\Re,\ \ 0\leq I\leq h,\ \ 1\leq k\leq h\Bigg\}.\end{aligned}\tag{9}$$

r_2 is given by,

$$r_2=\underset{t_j\in\mathrm{T}}{\text{argmin}}\,F(\mathbf{X},\mathbf{y};\beta,r_1,t_j),\tag{10}$$

where a set $\mathbf{T}$ is defined as,

$$\begin{aligned}\mathbf{T}=\Bigg\{&\{0,t_1,\ldots,t_k\}|d_ht^h+d_{h-1}t^{h-1}+\cdots+d_1t^1+d_0=0,\\ & d_I=\left[\sum_{i=1}^{n}-\gamma_i\beta_{12}r_1^{\varphi(x_{i1})}\varphi(x_{i2})\right]_{\varphi(x_{i2})-1=I}\\ & +\left[\sum_{i=1}^{n}\beta_{12}^2r_1^{2\varphi(x_{i1})}\varphi(x_{i2})\right]_{2\varphi(x_{i2})-1=I},\\ & t\geq 0,\ \ t\in\Re,\ \ 0\leq I\leq h,\ \ 1\leq k\leq h\Bigg\}.\end{aligned}\tag{11}$$

Proposition 2.2 Given r_1 and r_2, the optimal parameters, β_{12},β_1 and β_2, are given by,

$$\beta_{12}=\sum_{i=1}^{n}\frac{\gamma_ir_1^{\varphi(x_{i1})}r_2^{\varphi(x_{i2})}}{r_1^{2\varphi(x_{i1})}r_2^{2\varphi(x_{i2})}+\frac{\lambda}{w_i}},\tag{12}$$

$$\beta_1=\sum_{i=1}^{n}\frac{y_ix_{i1}-\beta_2x_{i1}x_{i2}-\beta_{12}x_{i1}r_1^{\varphi(x_{i1})}r_2^{\varphi(x_{i2})}}{x_{i1}^2+\frac{\lambda}{w_i}},\tag{13}$$

$$\beta_2=\sum_{i=1}^{n}\frac{y_ix_{i2}-\beta_2x_{i1}x_{i2}-\beta_{12}x_{i2}r_1^{\varphi(x_{i1})}r_2^{\varphi(x_{i2})}}{x_{i2}^2+\frac{\lambda}{w_i}}.\tag{14}$$

2.5 Inference of hierarchical relationships between genes

In this section, we propose a statistical strategy in order to infer the hierarchical relationships between genes, when their epistatic effect is detected. The proposed GEM detects complex epistatic effects underlying a targeting biological interaction model. However, it is still challenging to infer hierarchical relationships between the genes in epistatic effects. The significant interaction term (β_{12}) shows the presence or absence of epistasis, but it does not explain the hierarchical order. In the GEM, the genetic effect size at the individual locus, r_1 and r_2, enables one to compare their effect size to the phenotype variation. For instance, the genetic effect size of Locus 1 ($r_1=1.5$) is much less than of Locus 2 ($r_2=4$) in Model I in Figure 2. It shows that the genetic variants of Locus 2 may repress the genetic effects of Locus 1 on the activation of the gene. Also, it can be interpreted as the dominance of the effect size in the epistasis. Therefore, the statistical testing of each genetic effect size may provide the hierarchical relationships between the genes.

Given the detected paired genetic loci of epistasis, the tests of significance of the genetic effect size are individually performed by comparing with a model which consists of only marginal effects (null hypothesis). That is, the individual genetic effect size terms are considered separately as,

$$y=\beta_0+\beta_1x_1+\beta_2x_2+\beta_{r_1}r_1^{\varphi(x_1)}+\beta_{r_2}r_2^{\varphi(x_2)}+\varepsilon.\tag{15}$$

The significant β_{r_1} and β_{r_2} show the dominance of the effect size in the epistasis. All possible determinations of the hierarchical relationship between the genes are illustrated in Table 1. $L1^+$ represents the first locus whose coefficient β_{r_1} is rejected to the null hypothesis, i.e. $p(\beta_{r_1})<0.05$, whereas $L1^-$ shows the insignificant genetic

Table 1. Determination of hierarchical relationship when epistasis occurs

Locus 1	Locus 2	Hierarchical relationship
$L1^+$	$L2^+$	$L1^+ \rightarrow G,\ L2^+ \rightarrow G$
$L1^+$	$L2^-$	$L2^- \dashv L1^+ \rightarrow G$
$L1^-$	$L2^+$	$L1^- \dashv L2^+ \rightarrow G$
$L1^-$	$L2^-$	$L1^-, L2^- \rightarrow G$

Notes: $L1^+$ represents the first locus whose coefficient β_{r_1} is rejected to the null hypothesis, i.e. $p(\beta_{r_1}) < 0.05$, whereas $L1^-$ shows the insignificant genetic effect, i.e. $p(\beta_{r_1}) \geq 0.05$. G is the gene that the two genetic effects contribute to.

effect, i.e. $p(\beta_{r_1}) \geq 0.05$. $p(\cdot)$ represents a P-value of a coefficient. G is the gene that the two genetic effects contribute to.

Since either detecting the gene order from the hierarchical relationships or inferring exact gene-regulatory/signaling pathways are beyond the scope of this study, we would not go further in depth. However, we briefly address how to derive the gene hierarchy from the proposed method. From the epistasis study, a series of a set of genetic variations manifesting the epistatic relationship and the corresponding gene expression are provided. By performing the proposed statistical strategy for the identification of the hierarchical relationships, the four possible situations in Table 1 can be observed. When the genetic effect sizes at the two loci are both significant, it infers to the substantial effects to the phenotype variation simultaneously. In this setting, the genetic effect size on each individual locus takes significance within the interaction. On the other hand, insignificance of the genetic effect size at both loci represents a closer cooperative interaction effect. Even though no individual genetic effect size is significant, the interaction between the two loci makes a significant contribution to the variation of the target gene expression. Finally, the situation, one significant genetic effect size with an insignificant one, is of interest, since it clearly shows the hierarchical relationships between the genes. When 'A' is epistatic to 'B' (also means that 'A' is a downstream gene of 'B'), it can be notated as 'B' ⊣ 'A'. The notation describes the hierarchical relationships that 'B' represses 'A' not to active 'A' in the wild type. However, when 'A' is activated, the effect of 'B' is masked by 'A'. The interpretation gives an important information to infer the hierarchy for functional dependencies between genes of genes. The procedure to infer the gene hierarchy order is briefly illustrated in Figure 3.

3 Experiment results

3.1 Simulation studies

The simulation studies were conducted to assess the performances of GEM and Fisher's model. A number of types of simulation data were generated, based on the biological epistasis models in Figure 2. We considered the 14 case studies in various models and parameters to represent all potential epistatic situations as described in Table 2. The case studies 1 and 2 describe the hierarchical relationship with different extents in Model I, where locus 2 is epistatic to locus 1. On the other hand, case studies 3 and 4 have significant genetic effect sizes at both loci, but in reverse (degrading for locus 1/propagating for locus 2). Case study 5 is equivalent to case studies 11 and 12 with different genetic effect sizes. The same settings were given in Models II and IV. Note that the parameters, r_1 and r_2, are identical in Models III and IV.

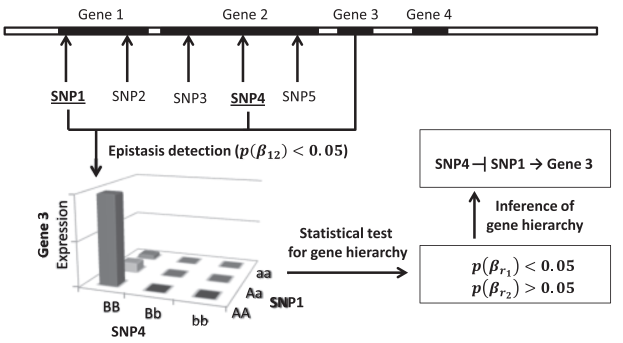


Fig. 3. Inference of hierarchical relationships. The illustration shows the epistatic effect of SNP1 and SNP4 to the gene expression of Gene 3. The proposed statistical test determines the gene hierarchy

Table 2. Simulation case designs for the possible biological epistatic models

Case No.	Model	Parameters		Case No.	Model	Parameters	
		r_1	r_2			r_1	r_2
Case 1	Model I	1	2	Case 6	Model II	1	2
Case 2		1	5	Case 7		1	5
Case 3		0	2	Case 8		0	2
Case 4		0	5	Case 9		0	5
Case 5		5	5	Case 10		5	5
Case 11	Model III	0	0	Case 13	Model IV	0	0
Case 12		2	2	Case 14		2	2

Given the genetic effect sizes of two loci and the epistatic model (Model I–Model IV) at each case study, SNP data at two loci $(\mathbf{x}_1, \mathbf{x}_2)$ of a hundred samples were randomly generated, where r_1 and r_2 were varied by adding the positively normally distributed random variables $(|N(0, 0.1)|)$ to the given values in Table 2. In a similar way, the parameters of the marginal effects β_1 and β_2 were randomly chosen with the positively normally distributed random variables $(|N(0, 1)|)$. The intercept parameter β_0 was ignored. The corresponding gene expressions were derived from the models with the noise by normal distributed noise. The random noise was generated by various signal-to-noise ratios (SNRs)—5 and 1.5. The genetic effect size, which is close to 1, shows that the locus has no effect, while a size that is close to 0 or bigger than 1 shows a large effect size of degradation or propagation to gene expression, respectively (Fig. 2). Note that either r_1 or r_2 of zero value in Table 2 does not mean zero but a fraction less than one due to the addition of the random variables.

We first compared P-values of the interaction term β_{12} on GEM and Fisher's model. A thousand replications were conducted for all experiments in this article unless otherwise specifically stated. The negative logarithmic P-values, $-\log_{10}(p(\beta_{12}))$, and the reject rates to the null hypothesis of the interaction term when SNR = 5 are depicted in Figure 4 (Supplementary Fig. S1 for SNR = 1.5 in the Supplementary Material). The significantly higher negative logarithmic P-values of the interaction term on GEM than on Fisher's model were observed for all case studies (Fig. 4A and Supplementary Fig. S1C). The result shows that GEM has higher chance to detect epistasis than Fisher's model. Moreover, we examined the statistical assessment of the epistasis findings. A null hypothesis, which consists of only marginal effects without interaction effect, is rejected when $p(\beta_{12}) < 0.05$. The reject rates on Fisher's model were 0.29, 0.16,

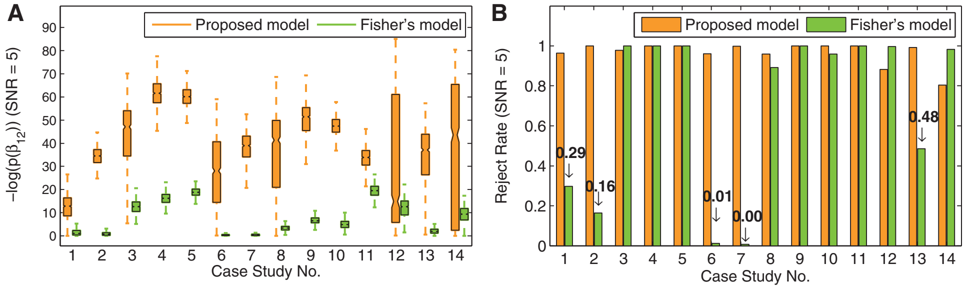


Fig. 4. The comparison of the performance between GEM and Fisher's model. (**A**) Negative logarithmic *P*-values of the interaction term, $-\log_{10}(p(\beta_{12}))$, and (**B**) reject rates to the null hypothesis for SNR = 5

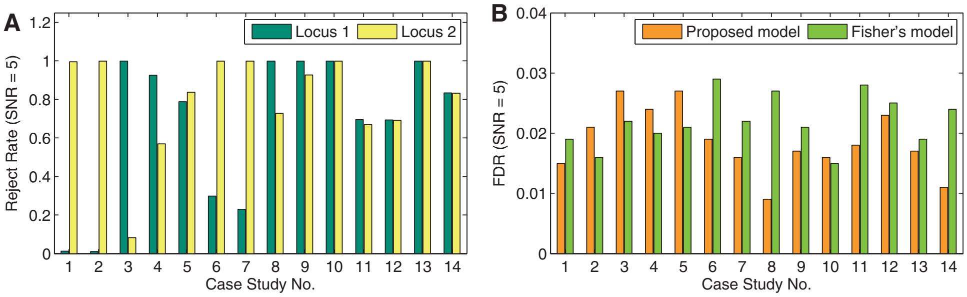


Fig. 5. (**A**) The evaluation of the proposed statistical strategy for the determination of hierarchical relationship (**B**) FDRs of GEM and Fisher's model. The average FDR over the case studies are 0.018 ± 0.0054 and 0.022 ± 0.0042, respectively. The incorrect rejection rate to the null hypothesis less than 3% is expected on the proposed model when SNR = 5

0.01, 0.00, 0.89 and 0.48 in case studies 1, 2, 6, 7, 8 and 13, respectively, when SNR = 5 (Fig. 4B), and are 0.06, 0.06, 0.01, 0.01, 0.46 and 0.18 when SNR = 1.5 (Supplementary Fig. S1D). On the other hand, GEM consistently detected almost all interactions across the case studies when SNR = 5. The slightly low reject rates in case studies 1, 6, 12 and 14 were observed on GEM when SNR = 1.5. However, they were still significantly higher than Fisher's model (Supplementary Fig. S1D). The observations clearly describe the substantial lack of epistasis detection power on Fisher's model, especially when the epistasis is involved in hierarchy relationship of genes, and support the argument for the limited use of Fisher's model to represent biological model in practice.

The proposed statistical strategy for the determination of hierarchical relationships were evaluated. When epistasis is detected in the simulation study, the *P*-values of β_{r_1} and β_{r_2} in Equation (15) were computed. Then, the significance of the individual genetic effect size of the two loci was determined with the significance level of 0.05. The reject rates of each locus are depicted in Figure 5A. Since the effect size of locus 1 was set by around one in case studies 1 and 2, the low reject rates (~0) at locus 1 were observed in Figure 5A. The reject rates of case studies 5, 10, and 11–14, which is set by the same effect size in, are shown as fairly even.

False discovery rate (FDR) analysis was conducted to measure the expected proportion of incorrectly rejected null hypotheses. The simulation data for the FDR analysis was generated by taking into account only marginal effects and noise (SNR = 5) without interaction, i.e. $\beta_{12} = 0$. The incorrectly rejected interaction term on GEM and Fisher's model was considered as a false positive, and FDR was computed by false positive/(true positive + false positive). The FDRs of GEM (0.018 ± 0.0054) and Fisher's model (0.022 ± 0.0042) on average were observed throughout the case studies (Fig. 5B). A FDR of less than 3% at most is expected on GEM.

3.2 Human brain data on schizophrenia

We applied the proposed GEM to human brain data of schizophrenia. Both SNP and gene expression data used in the preparation of this experiment were acquired from the human prefrontal cortex of 39 patients with schizophrenia and 44 controls (Liu *et al.*, 2010). We removed SNPs with minor allele frequency less than 1% or in linkage disequilibrium ($r^2 > 0.2$ within 1 Mb).

The exhaustive search approach for a whole large-scale genome study is very expensive. Moreover, this work rather focused on proposing the GEM to detect complex eQTL epistasis and its optimizing solution. Thus, we finally selected 76 genes by biological literature in this study. The 76 genes are widely reported major susceptible components of the pathways on schizophrenia such as PIK3K/AKT, growth factors, adhesion/junctions, NMDA receptor, glutamate-related, dopaminergic, serotonergic, GABA, circadian, cytokines and oxidative stress (Carter, 2006). The genes are listed in Supplementary Table S1. We also considered 35 698 SNPs located within the regions of the selected 76 genes. Note that although upstream SNPs of the genes known as enhancers or promoters were not considered in this study, the consideration of the enhancer and promoter regions may provide a great benefit to discover epistasis in regulatory regions.

A significance test (α = 0.05) for detecting the epistasis effect was performed. For each gene, the pairs of the SNPs located in the corresponding gene and the gene expression were examined by the proposed method. Among a total of 1 549 788 scans, 15 epistatic effects were found by the proposed method with a threshold

Table 3. Detected epistatic eQTLs and the genes on schizophrenia

Model	SNP 1–SNP 2	Gene	r_1 ($p(\beta_{r_1})$)	r_2 ($p(\beta_{r_2})$)	$p(\beta_1)$	$p(\beta_2)$	$p(\beta_{12})^{\dagger}$	$p(\beta_{12})^{\ddagger}$
II	**rs2324046 (BMP6)–rs2876117 (GRM7)**[a]	**GRID1**	**0.03 (0.47)**	**0.09 (0.04)**	**0.939**	**0.465**	**0.037**	**0.328**
I	**rs363223 (SNAP25)–rs6108464 (SLC18A2)**[b]	**HTR5A**	**0.51 (0.78)**	**0.16 (0.05)**	**0.263**	**0.437**	**0.014**	**0.276**
I	rs1566368 (HTR2A)–rs1928042 (GRM7)	GRIN2A	1.92 (0.92)	2.48 (0.06)	0.002	0.028	0.001	0.027
II	rs7341537 (SLC18A2)–rs363223 (CNTNAP2)	ANK3	0.46 (0.13)	0.07 (0.00)	0.447	0.159	0.002	0.848
I	rs10230882 (CNTNAP2)–rs7792210 (CNTNAP2)	PIK3C3	1.84 (0.53)	1.70 (0.09)	0.025	0.008	0.026	0.099
I	rs11977660 (ANK3)–rs10509123 (EGFR)	GABRA1	0.72 (0.84)	2.09 (0.04)	0.817	0.095	0.035	0.658
II	rs2452801 (TIMELESS)–rs812279 (GRIA1)	IL10	2.51 (0.20)	0.19 (0.08)	0.262	0.651	0.013	0.111
II	rs4686101 (BMP6)–rs2876117 (GRM7)	AKT1	1.69 (0.51)	1.88 (0.12)	0.003	0.002	0.039	0.517
I	rs779710 (PCLO)–rs10954712 (GRM7)	JAM3	0.44 (0.89)	0.08 (0.12)	0.566	0.595	0.041	0.136
II	rs917880 (SLC18A2)–rs363223 (EGFR)	SNAP25	2.09 (0.00)	2.15 (0.09)	0.000	0.000	0.001	0.894
I	rs12354209 (NRXN1)–rs991566 (MTR)	HTR5A	0.16 (0.74)	0.12 (0.21)	0.231	0.839	0.050	0.365
I	rs7341537 (SLC18A2)–rs363223 (CNTNAP2)	HTR5A	0.26 (0.67)	0.28 (0.91)	0.998	0.214	0.043	0.073
I	rs779710 (CNTNAP2)–rs7341537 (GRM7)	GABRA1	0.31 (0.57)	11.58 (0.27)	0.353	0.472	0.049	0.119
I	rs851814 (ANK3)–rs10509123 (CNTNAP2)	NPAS2	0.21 (0.84)	10.96 (0.36)	0.348	0.729	0.039	0.215
I	rs11776959 (SLC18A2)–rs363223 (NRG1)	TIMELESS	0.28 (0.28)	0.01 (0.38)	0.814	0.855	0.050	0.162

Notes: Model: the used target model for the kernel function, $p(*)$: P-value of $*$, $p(\beta_{12})^{\dagger}$: P-value of β_{12} on GEM. $p(\beta_{12})^{\ddagger}$: P-value of β_{12} on Fisher's model.

[a] The bold is described in detail in Figure 6A.

[b] The bold is described in detail in Figure 6B.

($r^2 > 0.7$), which are listed in Table 3. In the table, the model indicates the biological epistatic model used to define the kernel function. The latent epistatic properties of the genetic effects to the quantitative phenotype can be deduced from the optimally estimated model. It is noticeable that only one in the findings, rs1566368, rs1928042 → GRIN2A, appears as significant by Fisher's model (see the third row in Table 3, where $p(\beta_{12})^{\ddagger} < 0.05$). It represents that compositional epistasis is much pervasive in genetic systems and it has overlooked the influence with the statistical epistasis studies.

The detailed epistatic effects of the two representative findings are illustrated in Figure 6 (also see the first two rows in Table 3), where the gene expressions of the contingency table at the genetic mutants at two loci are described. The putative gene expressions on epistasis estimated by the proposed method is colored in red, and the sample mean and median values of the gene expressions are indicated with blue and black lines, respectively, in the left figures. The right figures show the latent epistatic effects of the model by removing the marginal effects for elucidating the genetic effects of the mutation at the loci. As the first significant observation, the gene expressions of GRID1 on double minor alleles of rs2324046 and rs2876117 are depicted in Figure 6A. In the observation, no significant marginal effect of both rs2324046 and rs2876117 appears, (i.e. $p(\beta_1) = 0.939 > 0.05$ and $p(\beta_2) = 0.465 > 0.05$), but significant interaction effect (i.e. $p(\beta_{12})^{\dagger} = 0.037$). The fractional genetic effect sizes ($r_1 = 0.03, r_2 = 0.09$) indicate that the genetic mutants repress the gene expression of GRID1 in a multiplicative fashion with threshold. The epistatic phenomena are clearly observed in the right figure in Figure 6A, where the gene is highly activated when no mutant exists at both loci. This discovery reflects the two variants may work together to regulate GRID1 expression. By the hierarchy test, the effect of rs2876117 is observed as highly significant ($\mathrm{p}(\beta_{r_2}) = 0.04$) in the interaction. It shows the hierarchical relationship between the two genetic mutants to the gene, which consequently infers the gene hierarchy: BMP6 ⊣ GRM7 → GRID1. As the second significant observation, the gene expressions of HTR5A on double mutants of rs2324046 and rs2876117 are depicted in Figure 6B. It also does not have marginal effects at both loci, but a significant interaction effect ($\mathrm{p}(\beta_{12}) = 0.014$). The significantly repressed gene expression on the mutants on rs2876117 is described with the fractional genetic effect size, $r_2 = 0.16$ and $\mathrm{p}(\beta_{r_2}) = 0.05$. The gene hierarchy of SNAP25 ⊣ SLC18A2 → HTR5A is inferred. SNAP25 inhibits (but not completely) the effect of SLC18A2 at the wide type. However, the mutant of SNAP25 activates SLC18A2, which significantly affects the gene activation of HTR5A.

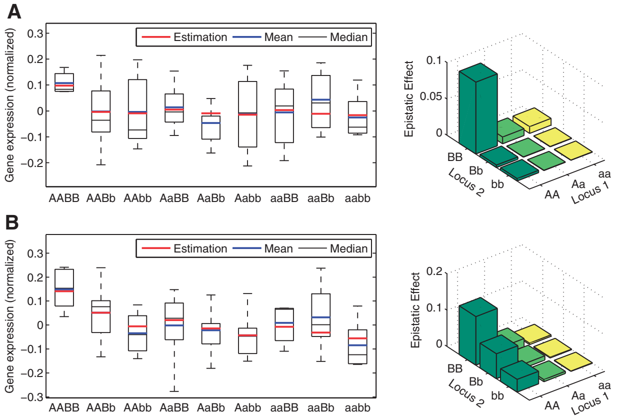


Fig. 6. Epistasis effect detection on schizophrenia. (**A**) Gene expressions and the fitting on the GEM for rs2324046–rs2876117 with GRID1 (**B**) rs363223–rs6108464 with HTR5A. The right figures show the genetic effect to the gene expression taking only interaction term without the marginal effects

We constructed a biological network from the set of the gene hierarchies, as illustrated in Figure 7. The epistatic effects, in which only one genetic effect size is significant to another by Equation (15), were used for the construction, since the epistasis relationship clearly provides a clue to infer the hierarchy between genes. The genes are listed in the seven rows on top in Table 3. In Figure 7, round rectangles represent epistatic eQTLs, circle nodes show the genes which are regularized by the genetic effects. Hexagon nodes

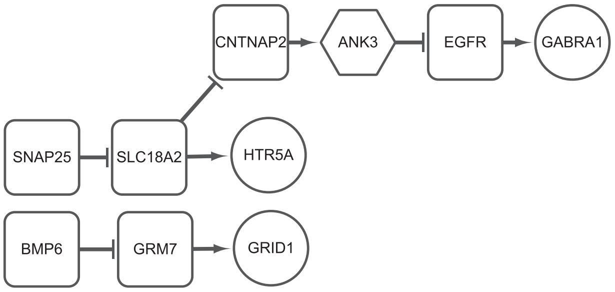


Fig. 7. The inference of the gene hierarchy on schizophrenia. Round rectangles represent epistatic eQTLs, and circle nodes show the genes which are regularized by the genetic effects. Hexagon nodes represent the genes that have both genotypic epistatic effects and phenotype variations of gene expression

represent the genes that have both genotypic epistatic effects and phenotype variations of gene expression. For instance, ANK3 gene is regularized by the two genotypic mutants of SLC18A2 and CNTNAP2, and also ANK3 controls the gene expression of GABRA1 with the genotypic variation of SLC18A2. Although the network does not provide the full description of the gene regulatory network or signaling pathway, the inferred gene hierarchies play an important role in the enrichment of the biological network inference.

To find the biological support of the constructed network, extensive resources have been checked. There are only a few eQTL-epistasis studies have done. One study addressed the relationship between BMP6 and glutamine (Yabe *et al.*, 2002). According to the study, BMP6 plays an inductive role in the generation of cerebellar granule cells (CGCs), and glutamate causes a reduction in the ability of CGCs. In our findings, it appears that the genetic mutant of GRM7 masks the effect of BMP6. Although we could not make a direct connection between the findings at this time, we found the apparent epistasis phenomena between the genes in the experiment.

4 Conclusion

In this article, we proposed a novel GEM to detect the quantitatively compositional epistasis as well as the statistical epistasis. Fisher's model has been commonly used to detect epistasis as a tool for the statistical epistasis in many studies. However, many argue for the limited usage of Fisher's model due to the lack of the biological interpretation. The proposed method is designed to provide a flexible tool that detects epistasis on various biological models. The individual genetic effect size on the interactive effect of the GEM enables one to infer the hierarchical relationships between the genes. The optimal solutions for best-fitted parameters to the data and the statistical approach for identifying the hierarchical relationship are provided.

To justify the lack of Fisher's model and to assess the proposed model, we conducted simulation studies designed with the various epistatic settings. In the simulation studies, we showed the epistatic situations that Fisher's model lacks the ability to detect, and the outstanding performance of the proposed model comparing Fisher's model. Furthermore, we applied the model to human brain data on schizophrenia. The assessment with the real biological data is not easy due to the lack of a well-known grounded truth. However, the clear epistatic evidences are shown by the significance test.

The expansion of the proposed method to the epistasis study with more than two genes can be easily achieved in terms of a theory by adding more terms to the model. However, the genetic interaction effects with more than two would make the epistasis very complicated, and would need more biological putative scenarios. Furthermore, the interpretation of the hierarchical relationship would be challenging. Intensive future research is necessary for this issue.

This exhaustive search approach is extremely expensive on a whole large-scale genome study. However, it gives descriptional information of the epistatic model on multiple genes when certain designed targeting genes are given. In this article, the total scanning of the human brain data of schizophrenia between 35 698 SNPs and 76 genes (only susceptible genes) in the schizophrenia human brain data experiment was completed in several minutes with the program implemented by MATLAB on a personal computer. For the whole large-scale genome study, the total search space will be (number of SNPs) × (number of SNPs−1) × (number of genes)/2, i.e. 852 963(SNP) × 852 962(SNP−1) × 25 833(Gene)/2, which is intractable to investigate whole eQTL epistasis interactions. The improvements of its performance can be possible with parallel computing and distributed systems as well as extending it to a semi-exhaustive approach for future research.

Conflict of interest: none declared.

References

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

Bateson,W. (1909) *Mendel's Principles of Heredity*. Cambridge University Press, Cambridge.

Carlborg,O. and Haley,C.S. (2004) Epistasis: too often neglected in complex trait studies? *Nat. Rev. Genet.*, **5**, 618–625.

Carter,C. (2006) Schizophrenia susceptibility genes converge on interlinked pathways related to glutamatergic transmission and long-term potentiation, oxidative stress and oligodendrocyte viability. *Schizophr. Res.*, **86**, 1–14.

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

Cordell,H.J. (2009) Detecting gene–gene interactions that underlie human diseases. *Nat. Rev. Genet.*, **10**, 392–404.

Eichler,E.E. *et al.* (2010) Missing heritability and strategies for finding the underlying causes of complex disease. *Nat. Rev. Genet.*, **11**, 446–450.

Fisher,R. (1918) The correlations between relatives on the supposition of Mendelian inheritance. *Trans. R. Soc. Edinb.*, **52**, 399–433.

Huang,Y. *et al.* (2013) eQTL epistasis—challenges and computational approaches. *Front. Genet.*, **4**, 51.

Lee,S. and Xing,E.P. (2012) Leveraging input and output structures for joint mapping of epistatic and marginal eQTLs. *Bioinformatics*, **28**, i137–i146.

Liu,C. *et al.* (2010) Whole-genome association mapping of gene expression in the human prefrontal cortex. *Mol. Psychiatry*, **15**, 779–784.

Mackay,T.F.C. (2014) Epistasis and quantitative traits: using model organisms to study gene-gene interactions. *Nat. Rev. Genet.*, **15**, 22–33.

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

Pandey,A. *et al.* (2012) Epistasis network centrality analysis yields pathway replication across two GWAS cohorts for bipolar disorder. *Transl. Psychiatry*, **2**, e154.

Phenix,H. *et al.* (2011) Quantitative epistasis analysis and pathway inference from genetic interaction data. *PLoS Comput. Biol.*, **7**, e1002048.

Phillips,P.C. (2008) Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems. *Nat. Rev. Genet.*, **9**, 855–867.

Schupbach,T. *et al*. (2010) Fastepistasis: a high performance computing solution for quantitative trait epistasis. *Bioinformatics*, **26**, 1468–1469.

VanderWeele,T.J. (2010) Empirical tests for compositional epistasis. *Nat. Rev. Genet.*, **11**, 166.

Wan,X. *et al*. (2013) The complete compositional epistasis detection in genome-wide association studies. *BMC Genet.*, **14**, 7.

Yabe,T. *et al*. (2002) Bone morphogenetic proteins bmp-6 and bmp-7 have differential effects on survival and neurite outgrowth of cerebellar granule cell neurons. *J. Neurosci. Res.*, **68**, 161–168.