Briefings in Bioinformatics, 18(2), 2017, 250–259

doi: 10.1093/bib/bbw014
Advance Access Publication Date: 4 March 2016
Paper

iFORM/eQTL: an ultrahigh-dimensional platform for inferring the global genetic architecture of gene transcripts

Kirk Gosik, Lan Kong, Vernon M. Chinchilli and Rongling Wu*

*Corresponding author. Rongling Wu, Division of Biostatistics and Bioinformatics, Department of Public Health Sciences, Penn State College of Medicine, Hershey, PA 17033, USA. Tel: (717)531-2037; Fax: (717)531-0480; E-mail: rwu@phs.psu.edu

Abstract

Knowledge about how changes in gene expression are encoded by expression quantitative trait loci (eQTLs) is a key to construct the genotype–phenotype map for complex traits or diseases. Traditional eQTL mapping is to associate one transcript with a single marker at a time, thereby limiting our inference about a complete picture of the genetic architecture of gene expression. Here, we implemented an ultrahigh-dimensional variable selection model to build a computing platform that can systematically scan main effects and interaction effects among all possible loci and identify a set of significant eQTLs modulating differentiation and function of gene expression. This platform, named iFORM/eQTL, was assembled by forward-selection-based procedures to tackle complex covariance structures of gene–gene interactions. iFORM/eQTL can particularly discern the role of *cis*-QTLs, *trans*-QTLs and their epistatic interactions in gene expression. Results from the reanalysis of a published genetic and genomic data set through iFORM/eQTL gain new discoveries on the genetic origin of gene expression differentiation in *Caenorhabditis elegans*, which could not be detected by a traditional one-locus/one-transcript analysis approach.

Key words: gene expression; eQTL; variable selection; genetic architecture

Introduction

As activation or inhibition of gene expression causes change in phenotypic formation, the identification of expression quantitative trait loci (eQTLs) that regulate the pattern of gene expression is essential for constructing a precise genotype–phenotype map [1–3]. With the advent and development of various biotechnologies, it has become possible that genome-scale marker and expression data can be generated, providing an important fuel to systematically study the biological function of all types of cellular components in an organism [4–6]. Several genome-wide association studies (GWAS) have been initiated to map a complete set of eQTLs for the abundance of genome-wide transcripts whose expression levels are related to biological or clinical traits [3, 7, 8]. Statistical analysis and modeling are playing an increasing role in mapping and identifying the underlying eQTLs from massive amounts of observed data [9–11].

A typical eQTL mapping approach is to associate a gene transcript with a single marker such as single nucleotide polymorphism (SNP). By analyzing the significance of all these markers one by one and adjusting for multiple testing, a researcher can identify significant loci that contribute to variation of expression by the gene. This marginal approach based on a simple regression model has been instrumental for the identification of eQTLs in a variety of organisms [4, 12]. However, there are two major limitations for the results by such a marginal analysis. First, it does not take into account the

Kirk Gosik is a doctoral student in statistical genetics in the Department of Public Health Sciences at The Pennsylvania State University.
Lan Kong is Associate Professor of Biostatistics and Bioinformatics in the Department of Public Health Sciences at The Pennsylvania State University.
Vernon M. Chinchilli is Distinguished Professor of Biostatistics and Bioinformatics and Chair of the Department of Public Health Sciences at The Pennsylvania State University.
Rongling Wu is Distinguished Professor of Biostatistics and Bioinformatics and Director of the Center of Statistical Genetics in the Departments of Public Health Sciences and Statistics at The Pennsylvania State University.
Submitted: 13 December 2015; **Received (in revised form):** 16 January 2016

dependence of different markers, so that a significant association detected by one marker may be owing to the other markers that are linked with it. The marginal marker analysis cannot separate the confounding effect of eQTLs owing to marker–marker dependence or linkage [13]. Second, an eQTL may act through its interaction with other eQTLs and environmental factors. Because of their paramount importance in affecting complex diseases and traits, gene–gene interactions, or epistatic effects, and gene–environment interactions have been studied intensively in modern biological and medical research [14–17].

These two limitations can be overcome by analyzing all markers and their pairwise interactions simultaneously through formulating a high-dimensional regression model. Although it can infer a complete picture of the genetic architecture of gene expression, this endeavor is highly challenged by the curse of dimensionality, i.e. the number of predictors far exceeds the number of observations. The past two decades have witnessed the tremendous development of variable selection models via penalized least squares or likelihood for high-dimensional data analysis. The basic principle used to develop these models is 'sparsity', i.e. only a small set of predictors explain variation in the response. Tibshirani [18] pioneered the least absolute shrinkage and selection operator (LASSO), which can select significant predictors and estimate their regression coefficients at the same time within a high-dimensional setting. Fan and Li [19] improved this approach in terms of solution sparsity, model stability and estimation accuracy by proposing a so-called smoothly clipped absolute deviation approach. Zou and Hastie [20] further developed the elastic net that resolves an issue of high pairwise correlations among different variables. A different approach, named minimax concave penalty by Zhang [21], was shown to considerably increase the accuracy of penalized variable selection. When the number of predictors is considerably larger than the number of samples, Candes and Tao [22] proposed the Dantzig selector to select the best subset of variables by solving a simple convex problem. Many methods possess favorable theoretical properties such as model selection consistency [23] and oracle properties [24]. However, when the number of predictors is extremely high, compared with the number of observations, sure screening is a more realistic goal to achieve than oracle properties or selection consistency [24, 25]. Sure screening assures that all important variables are identified with a probability tending to 1, hence achieving effective dimension reduction without information loss and providing a reasonable starting point for low-dimensional methods to be applied.

More recently, Hao and Zhang [26] extended variable selection approaches to jointly model main and interaction effects from high-dimensional data. Based on a greedy forward approach, their model can identify all possible interaction effects through two algorithms, iFORT and iFORM, which have been proved to possess sure screening property in an ultrahigh-dimensional setting. iFORT first searches for main effects, followed by interaction searching, whereas iFORM models main and interaction effects jointly in a high-dimensional setting. In this article, we implemented and reformed Hao and Zhang's model to build a computational platform for mapping the genetic architecture of eQTL actions and interactions for gene expression profiles. This so-called iFORM/eQTL platform was modified to cope with the feature of a genetic mapping or GWAS design in which molecular markers as genetic predictors are discrete although some additional continuous predictors can also be considered. At each marker, there are three distinct genotypes (i.e. categories), thus a pair of markers form nine genotype combinations. Classic quantitative genetic theory partitions nine genotypic values into the overall mean, and eight genetic effects, i.e. additive and dominant main effects at each marker, and additive × additive, additive × dominant, dominant × additive and dominant × dominant interaction effects between the two markers [27]. Therefore, if the number of markers is p, a total number of predictors including all main and two-way interaction terms is $2p^2$. For a typical moderate-sized mapping study, in which several thousands of markers are genotyped on a few hundred individuals, consideration of pair-wise genetic interactions will quickly make the dimension of predictors a high one. The dimensionality encountered in a GWAS study with thousands of thousands of markers on thousands of subjects becomes ultra-high.

By modeling all markers jointly at one time under an organizing framework, iFORM/eQTL can detect all possible significant eQTLs and their epistasis. An eQTL can be either a *cis*-QTL, coming from the same physical location as the gene expression, or a *trans*-QTL, coming from other areas of the genome. iFORM/eQTL can more precisely discern these two different types of eQTLs and their interactions than traditional marginal analysis. By reanalyzing a published data set collected in a mapping population of *Caenorhabditis elegans* [12], iFORM/eQTL has validated previous results by the marginal approach, while yielding new discoveries on the genetic origin of gene expression differentiation, which could not be detected in a traditional way.

iFORM/eQTL platform

Experimental design

An experimental population for genetic mapping includes the backcross, the F_2, both initiated two inbred lines, and a full-sib family derived from two outcrossing parents. These types of populations are used specifically for different species. Although they have different levels of complexities for statistical modeling, the genetic dissection of different populations underlies a similar principle. For the purpose of simplicity, we consider a backcross design in which there are only two genotypes at each marker.

Suppose the backcross contains n progeny, each of which is genotyped by p markers, such as SNPs, distributed over different chromosomes. The number of SNPs, p, should be large enough to completely cover the entire genome at an adequate depth so that we are likely to capture all possible genetic variants. An increasing body of evidence suggests that significant SNPs associated with complex traits or diseases are more likely to be eQTLs [7]. Hence, the identification of eQTLs is an important first step toward the genetic dissection of end-point phenotypes. For this reason, we assume that genome-wide gene transcripts are available for the assumed study population. We also assume that all progenies are recorded for the same organ by microarray, leading to expression abundance data of m gene transcripts. We purport to identify all possible genetic variants including main effects and interaction effects of SNPs that contribute to each gene transcript.

Adaptation of iFORM procedure

Hao and Zhang [26] formulated an interaction forward selecting procedure under the marginality principle (iFORM). The marker and gene transcript data of the study population can be denoted as $(\mathbf{X}_i, \mathbf{Y}_i)(i = 1, \ldots, n)$, which are independent and identically distributed copies of (X,Y), where $X = (X_1, ., X_p)^T$ is a p-dimensional predictor vector and Y is the response, expressed by a linear regression model:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon \tag{1}$$

where βs are the coefficients for the genetic effects of each marker. Like most genome-wide data sets, the number of markers here grossly outnumbers the number of observations, i.e. $p \gg n$. Therefore, selection procedures would need to be implemented to fit a linear regression model such as (1). We are already at the point of high-dimensional data, but if we want to include epistatic effects between different markers as predictors as well, then it would increase the amount of predictors by $(p + p^2)/2$. The resulting linear model would grow to be

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \gamma_{11} X_1^2 + \gamma_{12} X_1 X_2 + \cdots + \gamma_{pp} X_p^2 + \epsilon, \tag{2}$$

where γs are the coefficients for the epistatic effects for all the quadratic and two-way interactions between the markers. For convenience, we will assume that the markers and the transcripts are standardized before running the selection procedure. Therefore, $E(X_{ij}) = 0, \mathrm{Var}(X_{ij}) = 1, E(Y_i) = 0$ **and** $\mathrm{Var}(Y_i) = 1$ (**for** $i = 1, \ldots, n; j = 1, \ldots, p$). Also, the quadratic and two-way interaction effects will be centered, which we will write as $Z_i = (\ldots, X_{ik} X_{il} - E(X_{ik} X_{il}), \ldots)^T$. By doing so, we eliminate the need for an intercept in regression model (2). This reduces the model to the form,

$$Y = X^T \beta + Z^T \gamma + \epsilon \tag{3}$$

Some notation that will be used to define the elements of Han and Zhang's (2014) iFORM procedure are as follows:

$$\mathcal{P}_1 = \{1, 2, \ldots, p\}, \mathcal{P}_2 = \{(j, k) : 1 \le j \le k \le p\}$$

which are the index sets for the linear and two-way interaction terms, respectively. The significant main effects for the markers and their interaction effects are

$$\mathcal{T}_1 = \{j : \beta_j \ne 0, j \in P_1\}, \mathcal{T}_2 = \{(j, k) : \beta_{jk} \ne 0, (j, k) \in \mathcal{P}_2\}$$

For any model $\mathcal{M}$, $|\mathcal{M}|$ is used to denote the number of predictors contained in the model. The true model size is indicated by $|\mathcal{T}_1| = p_0$ and $|\mathcal{T}_2| = q_0$, or together is denoted by $|\mathcal{T}| = d_0 = p_0 + q_0$. For the procedure, three sets will be used throughout, which are $\mathcal{M}$ for the model set, $\mathcal{C}$ for the candidate set of predictors and $\mathcal{S}$ for the solution set of predictors currently selected in the model.

Two principles are used in the selection procedure when considering interactions as candidates for selection into the final model. The first principle important to the procedure is the heredity principle. The strong case of the heredity principle states that an interaction effect should not be considered unless both the contributing main effects are in the model [23]. The second is the principle of marginality. This principle states that it is inappropriate to model interaction terms when the main effects contributing to the interaction have been deleted because their effects become marginal by the inclusion of the interaction effect. This translates into

$$\gamma_{jk} \ne 0 \textbf{ only if } \beta_j, \beta_k \ne 0 \; \forall 1 \le j, k \le p$$

for model (2). Including both principles during the selection process allows for dynamically including both main effects and the quadratic and interaction effects. The quadratic and interaction effects can only be considered between the main effects currently selected into the solution set of the model according to the discussed principles. A more formal description of the procedure is given below.

Model selection

The initial step of Hao and Zhang's [26] iFORM procedure starts with the empty set for both the initial solution set ($\mathcal{S}_0$) and the initial model set ($\mathcal{M}_0$), $S_0 = \varnothing$ and $M_0 = \varnothing$. The candidate set contains all main effects at the beginning, $C_0 = P_1$, for each of the markers as a possible eQTL. Typical forward selection procedures are carried out to start the selection. Each marker is tested individually using a marker regression. The marker that results in the lowest residual sum of squares is the marker selected from the candidate set into the solution set as an eQTL. This is then iterated again for a selection of another marker into the model set. Once there are at least two main effects selected into the solution set, under the strong heredity principle, the quadratic and two-way interactions are then created and placed into the candidate set as possible eQTLs for selection in the next step. This process continues selecting main effects or the newly created quadratic and interaction effects into the solution set. If another main effect is selected into the solution set, then the candidate set grows with the creation of all possible quadratic and two-way interactions of the main effects that are currently in the solution set. This is continued until a designated stopping value, say *d*, is reached. For the number of predictors placed into the model set from the solution set, the Bayesian information Criterion was used, i.e.

$$\mathbf{BIC}_2(\hat{\mathcal{M}}) = \mathbf{log}(\hat{\sigma}\hat{\mathcal{M}}^2) + n^{-1}|\hat{\mathcal{M}}|(\mathbf{log}(n) + 2\mathbf{log}(d^*))$$

where $\hat{\sigma}\hat{\mathcal{M}}^2$ is the sample variance for the given model, $|\hat{\mathcal{M}}|$ is the size of the model or the number of predictors selected into the given model and *n* is the sample size. The $d^* = p + q$ term is the number of predictors in the full model, where $|\mathcal{P}_1| = p$ and $|\mathcal{P}_2| = q$. This was proposed as $\mathbf{BIC}_2$ by Chen and Chen [29], which they derived to help control the false discovery rate in high-dimensional data situations. They also showed that it was selection consistent if $d^* = O(n^\xi)$ for some $\xi > 0$. The only difference between the traditional Bayesian information criterion (BIC) calculation and the BIC_2 is the additional term involving $2\mathbf{log}(d^*)$. Ignoring the BIC, the maximum number of steps in the solution path is of size *n*. The parameter *d* controls the overall length of the solution path. In practice, the exact number of predictors to include, say d_0, in the true model is unknown. We want to make *d* large enough to include d_0 but not so large as to fit the model to the point where it becomes oversaturated. Using the BIC_2 should help avoid oversaturation. It is reasonable to assume that d_0 is much smaller than n in high-dimensional sparse regression problems [24]. As this is the case, for the purposes of our model, *d* is set to be no larger than $n/\mathbf{log}(n)$. Generally, the BIC_2 should reach a minimum, indicating the optimal stopping point, before the designated stopping value is reached.

Some consideration

A consideration is about the type of coding used for the genotypes. At any given eQTL, say the *j*th eQTL, there are two possible genotypes: $Q_j Q_j$ and $Q_j q_j$, making the total number of possible QTL genotypes in the population 2^m. The goal of a genetic model is to relate the 2^m possible genotypic values to a set of genetic parameters, such that these parameters are interpretable in terms of main and epistatic effects of the m eQTL. A genetic model is to use orthogonal contrast scales because it is consistent in the sense that the effect of an eQTL is consistently defined whether the genetic model includes one, two, three or more eQTL [28]. The orthogonal contrasts for the genetic model can be expressed by

$$x_{ij} = \left[-\frac{1}{2}\textbf{if homozygote } \mathrm{Q}_j\mathrm{Q}_j, \frac{1}{2}\textbf{if heterozygote } \mathrm{Q}_j\boldsymbol{q}_j\right]$$

Typically, in an inbred-line backcross population, a given genotype is coded with a 0 and 1. However, there are two drawbacks to this coding when considering the selection procedures discussed above. The first issue comes with not including an intercept in model (2). If this is the case, then each of the predictors would need to be centered, which yields a coding of −1/2 and 1/2 instead of 0 and 1. Besides meeting the assumptions of the model that the predictors are centered, it is also beneficial for the interaction effects. If the coding would remain at 0's and 1's, the interaction coding would also consist of 0's and 1's. This could lead to a problem because three of the four scenarios of epistasis between markers would result in a coding of 0 for the level in the interaction effect. This has the potential to falsely skew the data of no additive effect for interactions terms because of the sparseness of coding. Centering the coding to (−1/2,1/2) results in an interaction effect being coded as

(−1/4,1/4). This coding occurs for different scenarios for each of the levels. The coding of −1/4 could arise when the interaction is composed of a homozygote interacting with a heterozygote genotype. A coding of 1/4 would arise by either a homozygote interacting with another homozygote genotype, or when a heterozygote interacts with another heterozygote genotype.

Results

Simulation

We conducted simulation studies to test the theoretical properties of iFORM/eQTL. The response was generated from model (2) involving 500 genes with a sample size of $n = 200$ under three possible scenarios: (i) main and epistasis scenario in which both main and epistatic effects are simulated with the true $\boldsymbol{\beta} = (3, 0, 0, 3, 0, 3, 3, 0_{493})$, yielding $\mathcal{T}_1 = \{1, 4, 6, 7\}$ and $\boldsymbol{p}_0 = 4$, and the relevant interactions set to the pairs $\mathcal{T}_2 = \{(1, 6), (1, 7), (4, 7), (6, 7)\}$ and $\boldsymbol{q}_0 = 4$ all with $\gamma_{jk} = 3$ where $(\boldsymbol{j}, \boldsymbol{k}) \in \mathcal{T}_2$, (ii) main scenario in which only main effects were simulated with the true $\boldsymbol{\beta} = (3, 0, 0, 3, 0, 3, 3, 0_{493})$ and (iii) weak-main and -epistasis scenario in which both main and epistatic effects are simulated with the true $\boldsymbol{\beta} = (1.1, 0, 0, 1.1, 0, 1.1, 1.1, 0_{493})$, yielding $\mathcal{T}_1 = \{1, 4, 6, 7\}$ and $\boldsymbol{p}_0 = 4$, and the relevant interactions set to the pairs $\mathcal{T}_2 = \{(1, 6), (1, 7), (4, 7), (6, 7)\}$ and $\boldsymbol{q}_0 = 4$ all with $\gamma_{jk} = 3$ where $(\boldsymbol{j}, \boldsymbol{k}) \in \mathcal{T}_2$. In each scenario, we studied the influence of different effect values on parameter estimation by using different sizes of random error, $\sigma = 1$, 2 and 3. Note that scenario (iii) assumes epistatic effects that are larger than main effects so that this scenario can test how well iFORM functions when its underlying heredity principle is partly isolated. We did not assume that epistasis occurs between two loci of zero main effects because this case may not be pervasive (Mackay 2014).

In each of the scenarios, $\mathbf{X}_i^{'}\mathbf{s}$ were all independently and identically distributed realizations generated from Binomial (0.5), and then orthogonal contrasts were used to make each $\mathbf{x}_{ij} \in (-1/2, 1/2)$. The results were compared with several other commonly used methods for eQTL mapping. The methods that were used to model the data were single-marker analysis, forward selection involving only main effects (FS), forward selection involving all main effects and interaction (FS2) and the iFORM procedure. Several outcomes were evaluated to compare across each of the models. The outcomes are separated into three parts. The first part focuses on the selection of main effects, the second part focuses on the selection of interaction effects and the third part is the overall model performance. Simulations with $\mathcal{M} = 100$ replicates were run and the outcomes considered include

- Convergence Probability (Cov) $\sum_{m=1} \frac{I(T\hat{T})}{M}$
- Percentage of correct zeros (Cor0) $\sum_{m=1}^{M}\sum_{j=1}^{p} \frac{I\left(\widehat{(\beta_j)}=0, \beta_j=0\right)}{\left[M\left(p-p_0\right)\right]}$
- Percentage of incorrect zeros (Inc0) $\sum_{m=1}^{M}\sum_{j=1}^{p} \frac{I\left(\widehat{(\beta_j)}=0, \beta_j\neq 0\right)}{\left[M\left(p_0\right)\right]}$
- Exact Selection probability (Exact) $\sum_{m=1}^{M} \frac{I(\mathcal{T}=\hat{\mathcal{T}})}{M}$
- The average model size
- Mean Square Error (MSE)
- Adjusted R-square
- Computation Time in seconds

where j is the index of the β coefficient selected in the model.

In simulation scenario (i), iFORM/eQTL was closest to the simulated data (indicated as Oracle) (Table 1). Single-marker analysis was conducted on each of the main effects individually, and the significant markers were then designated as eQTLs. When comparing single-marker analysis, we can see it rarely identified the full set of main effects as significant from the simulated data. Also, no consideration for interactions could be assessed in single-marker analysis. iFORM/eQTL contains the identified main effects >90% of the time across all simulations. The procedure also includes interaction selection. The interaction screening shares a similar success rate where the interaction effects are correctly selected >90% of the time as well. Focusing on the computation time, we observed only a few seconds' increase, on average, than running single-marker analysis. The final models selected by iFORM/eQTL had similar adjusted R-square values as the Oracle results, on average. Examining the exact selection percentage, we can see that the vast majority of the time the correct predictors were selected and indicated as significant. To compare the interaction screening effectiveness, forward selection was implemented on both the main effects and interactions effects. The time it took to create the design matrix to implement forward selection was not included in the computation time.

As can be seen from the results, using forward selection on the full set of main effects and pair-wise interactions took substantially longer to run on average than any of the other methods, including iFORM/eQTL. Another drawback to implementing forward selection on such a large set seemed to come with overfitting the model. The selection included the maximum number of predictors allowed by the designated stopping value and did not use the BIC criteria for final model selection. This resulted in 19 additional predictors selected (Table 1). This increased the adjusted R-square value of the final model; however, this is suspected because of overfitting the data and not to be a true prediction of the response.

Scenario (ii) allows us to investigate false-positive rates (FPRs) of detecting epistasis because the simulated data contains no epistasis. First, it is not surprising that iFORM/eQTL can

Table 1. Results of simulation under scenario (i) with $\sigma = 1$ for the random error with independent predictors

$\boldsymbol{\sigma = 1}$	**Main effects**				**Interaction effects**				**Model**			
Method	**Cov**	**Cor0**	**Inc0**	**Exact**	**Cov**	**Cor0**	**Inc0**	**Exact**	**Size**	**MSE**	**Adj-R^2**	**Time (s)**
Single Marker	0 (0)	1 (0)	0.25 (0.06)	0 (0)	NA	NA	NA	NA	3 (0.24)	23.63 (0.09)	0.216 (0.028)	0.824
LASSO	1 (0)	0.933 (0.037)	0 (0)	0 (0)	NA	NA	NA	NA	36.7 (18.5)	1.33 (0.21)	0.851 (0.044)	1.24
FS	0.85 (0)	0.953 (0.001)	0.0625 (0.09)	0.85 (0.11)	NA	NA	NA	NA	27 (1.5)	10.23 (1.1)	0.660 (0.027)	3.47
FS2	1 (0)	0.996 (0.001)	0 (0)	1 (0)	0.95 (0.445)	0.981 (0.002)	0 (0)	0 (0)	27 (1.43)	0.302 (1.2)	0.989 (0.019)	72.31
iFORM	0.9 (0.002)	0.999 (0.001)	0.05 (0.0012)	0.9 (0.125)	0.9 (0.01)	1 (0)	0 (0)	0.9 (0.03)	7.55 (0.126)	2.93 (0.044)	0.894 (0.11)	4.08
Oracle	1	1	0	1	1	1	0	1	8	1.023 (0.0441)	0.965 (0.011)	NA

Note: The standard deviations of the estimates are given in parentheses.
Outcomes include the convergence Probability (Cov) $\sum_{m=1}^{M} I\left(\mathcal{T}\hat{\mathcal{T}}\right)/M$, percentage of correct zeros identified (Cor0) $\sum_{m=1}^{M}\sum_{j=1}^{p} I(\hat{\beta}_j = 0,\ \beta_j = 0)/[M(p - p_0)]$, percentage of incorrect zeros identified (Inc0) $\sum_{m=1}^{M}\sum_{j=1}^{p} I(\hat{\beta}_j = 0,\ \beta_j \neq 0)/[M(p_0)]$, the exact selection probability (Exact) $\sum_{m=1}^{M} I\left(\mathcal{T} = \hat{\mathcal{T}}\right)/M$, average model size, Mean Square Error for the model (MSE), the adjusted R-square of the model and the computational time in seconds.

Table 2. Results of simulation under scenario (ii) (testing FPRs for epistasis effects) with $\sigma = 1$ for the random error with independent predictors

$\boldsymbol{\sigma = 1}$	**Main effects**				**Interaction effects**				**Model**			
Method	**Cov**	**Cor0**	**Inc0**	**Exact**	**Cov**	**Cor0**	**Inc0**	**Exact**	**Size**	**MSE**	**Adj-R^2**	**Time (s)**
Single Marker	1.00 (0)	1.00 (0)	0 (0)	0.96 (0.196)	NA	NA	NA	NA	4.04 (0.196)	1.01 (0.048)	0.90 (0.0125)	0.824
LASSO	0 (0)	0.932 (0.032)	0.3175 (0.127)	0 (0)	NA	NA	NA	NA	34.7 (16.5)	2.61 (0.23)	0.291 (0.102)	1.24
FS	1.00 (0)	1.00 (0.003)	0 (0)	0.976 (0.153)	NA	NA	NA	NA	4.02 (0.153)	0.995 (0.053)	0.899 (0.0123)	3.52
FS2	1 (0)	1 (0)	0 (0)	1 (0)	0	0.999 (0.001)	0 (0)	0 (0)	4.01 (0.09)	0.995 (0.052)	0.89 (0.0122)	70.5
iFORM	1 (0)	1 (0.001)	0 (0)	0.976 (0.153)	0 (0)	1 (0)	0 (0)	0 (0)	4.032 (0.217)	1.002 (0.0534)	0.89 (0.0124)	4.08
Oracle	1	1	0	1	1	1	0	1	4	1.003 (0.0526)	0.898 (0.012)	NA

Note: The standard deviations of the estimates are given in parentheses.
Outcomes include the convergence Probability (Cov) $\sum_{m=1}^{M} I\left(\mathcal{T}\hat{\mathcal{T}}\right)/M$, percentage of correct zeros identified (Cor0) $\sum_{m=1}^{M}\sum_{j=1}^{p} I(\hat{\beta}_j = 0,\ \beta_j = 0)/[M(p - p_0)]$, percentage of incorrect zeros identified (Inc0) $\sum_{m=1}^{M}\sum_{j=1}^{p} I(\hat{\beta}_j = 0,\ \beta_j \neq 0)/[M(p_0)]$, the exact selection probability (Exact) $\sum_{m=1}^{M} I\left(\mathcal{T} = \hat{\mathcal{T}}\right)/M$, average model size, Mean Square Error for the model (MSE), the adjusted R-square of the model and the computational time in seconds.

estimate main effects as well as single-marker analysis (Table 2). Second, FPRs for epistatic detection were estimated as < 0.05, showing a low probability of identifying any epistasis if the data has no epistasis. Scenario (iii) shows the power of iFORM/eQTL for epistatic detection in a situation that deviates from the strong heredity principle. While single-marker analysis can only identify main effects, iFORM/eQTL can identify epistasis even if epistasis takes place between two loci that has weak main effect (Table 3). The power of epistatic detection in this case is good, ranging 0.80–0.95.

Similar results were obtained from simulation results with $\sigma = 2$ and 3 of the random error. All these simulation studies show that iFORM/eQTL can well be used as a tool to systematically search for epistasis in practice.

Real data analysis

Rockman *et al.* [12] reported an eQTL mapping study of *C. elegans* using 208 recombinant inbred advanced intercross lines (RIAIL) from a cross between the laboratory strain, N2, and a wild isolate from Hawaii, CB4856. Abundances of 20000 gene transcripts were measured by microarray in developmentally synchronized young adult hermaphrodites of these lines, providing a genome-wide coverage of *C. elegans* from WormBase, a public *C. elegans* genome database. The microarray data was preprocessed through a normal–exponential convolution background correction and normalized using quantile standardization. Although they are closely related, the two strains used for the cross are considered relatively divergent for *C. elegans*. The two strains differ roughly at approximately 1 base pair per 900. Their RIAILs were genotyped at 1454 ordered SNP markers that cover the whole genome of *C. elegans* including five autosomes (denoted as I–V) and one sex chromosome (denoted as X).

Rockman *et al.* [12] used a classic interval mapping approach to detect 2309 eQTLs by testing and scanning associations of each SNP with each gene transcript over the entire genome. Rockman *et al.*'s analysis allowed a rectangular map of eQTL positions × gene positions to be constructed (Figure 1), from

Table 3. Results of simulation under scenario (iii) (weak main effects and epistasis) with $\sigma = 1$ for the random error with independent predictors

$\sigma = 1$	Main effects				Interaction effects				Model			
Method	Cov	Cor0	Inc0	Exact	Cov	Cor0	Inc0	Exact	Size	MSE	Adj-R^2	Time (s)
Single Marker	0.75 (0.432)	1 (0.0012)	0.07 (0.138)	0.49 (0.50)	NA	NA	NA	NA	4.10 (0.86)	1.81 (0.88)	0.39 (0.79)	0.824
LASSO	0 (0)	0.93 (0.39)	0 (0)	0 (0)	NA	NA	NA	NA	33.55 (19.9)	2.01 (0.141)	0.075 (0.63)	1.24
FS	0.78 (0.42)	1.00 (0.0001)	0.07 (0.149)	0.74 (0.437)	NA	NA	NA	NA	3.67 (0.798)	1.82 (0.096)	0.38 (0.89)	3.47
FS2	0.44 (0.497)	1 (0)	0.15 (0.203)	0.44 (0.497)	0.01 (0.089)	1.00 (0.001)	0.60 (0.497)	0.00 (0)	3.11 (1.39)	1.85 (0.172)	0.36 (0.143)	72.31
iFORM	1 (0)	0.999 (0.001)	0 (0)	0.94 (0.23)	1 (0)	1 (0)	0 (0)	1 (0)	8.06 (0.23)	1.004 (0.049)	0.812 (0.029)	4.08
Oracle	1	1	0	1	1	1	0	1	8	1.006 (0.049)	0.811 (0.0292)	NA

Note: The standard deviations of the estimates are given in parentheses.

Outcomes include the convergence Probability (Cov) $\sum_{m=1}^{M} I\left(\mathcal{T}\hat{\mathcal{T}}\right)/M$, percentage of correct zeros identified (Cor0) $\sum_{m=1}^{M}\sum_{j=1}^{p} I(\hat{\beta}_j = 0,\ \beta_j = 0)/[M(p - p_0)]$, percentage of incorrect zeros identified (Inc0) $\sum_{m=1}^{M}\sum_{j=1}^{p} I(\hat{\beta}_j = 0,\ \beta_j \neq 0)/[M(p_0)]$, the exact selection probability (Exact) $\sum_{m=1}^{M} I\left(\mathcal{T} = \hat{\mathcal{T}}\right)/M$, average model size, Mean Square Error for the model (MSE), the adjusted R-square of the model and the computational time in seconds.

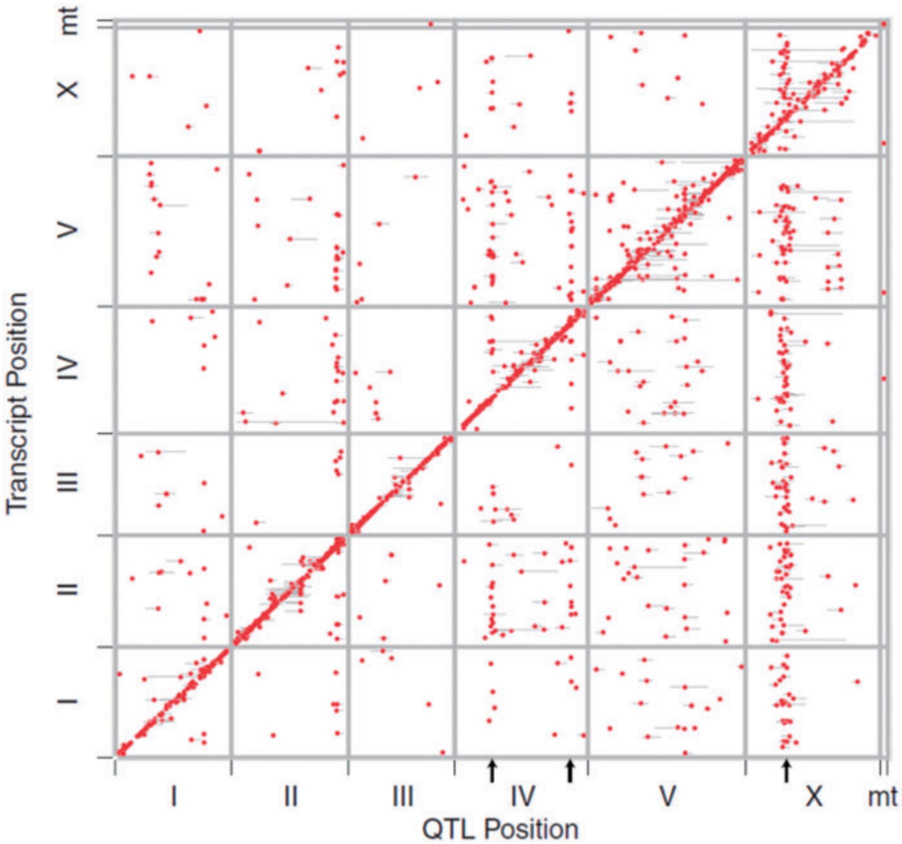

Figure 1. The distribution of eQTLs for each transcript abundance phenotype in *C. elegans*, located at the genomic positions of the transcripts. Those eQTLs on the diagonal are *cis*-eQTLs, whereas those off the diagonal are *trans*-eQTLs. Adapted from Rockman *et al.* [12].

which one can identify *cis*-eQTLs on the diagonal and *trans*-eQTLs off the diagonal. However, because their association analysis was conducted individually for each SNP, the detection of eQTLs was based on the marginal effects of individual eQTLs, which may lead to two issues being unsolved. First, of those eQTLs detected for the same gene transcript, some may include confounded effects by others. Second, the effects of genetic epistasis may take place but were not detected. By analyzing all SNPs simultaneously under a single framework, the high-dimensional model, iFORM, implemented in this study can more precisely characterize the genetic machineries underlying variation in each gene transcript. More specifically, we treat each transcript as a response with all SNP markers and their interactions as predictors by building a large regression model.

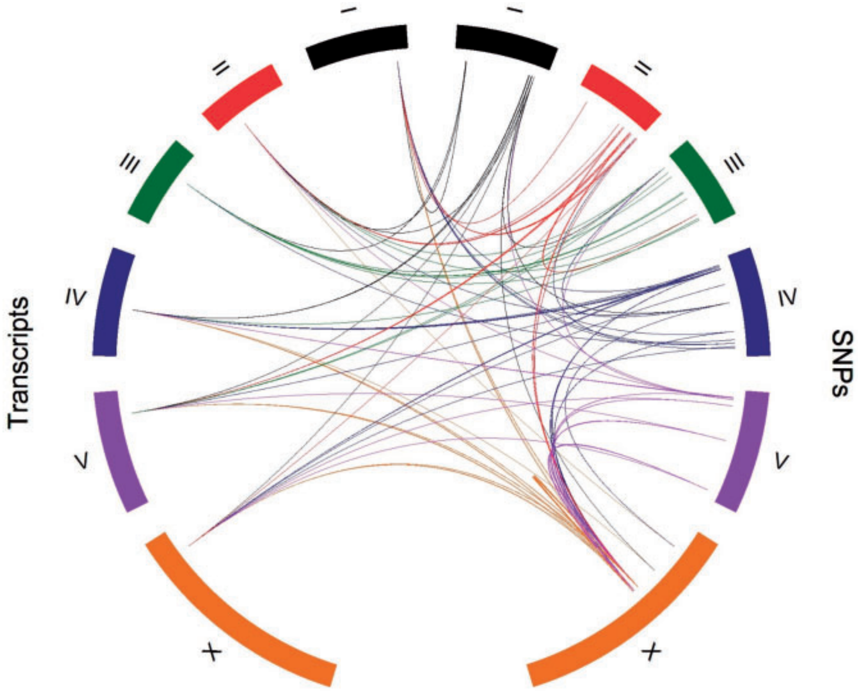

Figure 2. Circos plot illustrating the pattern of how a particular gene transcript is regulated by eQTLs on different chromosomes in C. *elegans*.

Significant predictors were then selected based on the iFORM procedure. A final model including both main and interaction effects can be evaluated by calculating adjusted R-square values.

Figure 2 illustrates the map of how a particular gene transcript is controlled by its eQTLs through main effects and interaction effects. For clarity of our presentation, we only chose one representative gene transcript from each chromosome. For example, gene transcript A_12_P103290 located at position 2069088–2069147 of chromosome I was detected to be controlled by main effects owing to X2_13516256 eQTLs on chromosomes II and X4_15632637 eQTLs on chromosome IV and X2_13516256:X4_15632637 interactions between some of these eQTLs on these two chromosomes.

iFORM/eQTL provides the estimates of each effect (either main effect or interaction effect), standard errors of each estimate and the significance tests of each effect. As an example, Table 4 gives the result of how gene transcript A_12_P103290 can be predicted by its eQTLs and their interactions. It can be seen that the final predictive model (adjusted $R^2 = 0.8964$) contains 10 markers, which exert their main effects and/or interaction effects on the transcript. The identification of these 10 markers should be well convinced because the previous simulation result (Table 2) suggests that iFORM/eQTL has reasonably low FPRs. The same data were also analyzed by single-marker analysis and FS, both of which can only identify a couple of significant main-effect eQTLs (Table 4), showing lower power for eQTL detection compared with iFORM/eQTLs. Of the 10 final markers, 7 show significant main effects ($P < 0.05$), with several (i.e. X_14636404, X4_15568674, X4_15632637 and X_14542103) explaining about 5% heritability (defined as a proportion of genetic variance owing to a predictor over the total phenotypic variance). Of these final markers, we identified eight significant epistatic interactions. Each epistasis accounts for 4.6–5.5% heritability (Table 4).

It is interesting to note that all predictors jointly contribute to 62.6% heritability for transcript A_12_P103290, of which main effects account for 26.7% and epistatic effects account for 35.9%. It is surprising that epistasis contributes to more than one-half of the heritability. Of the eight epistatic interactions, only one occurs owing to the interaction between two significant eQTLs, X_14542103 and X4_13532205 (Table 4). All the remaining ones are owing to interactions between one significant eQTL and one nonsignificant marker. Some eQTLs, such as X_14542103 and X_14636404, produce epistasis with a greater frequency than others. Despite their involvement in the final predictive model, some markers were tested to be insignificant in terms of both main and interaction effects, suggesting that they regulate a gene transcript in a subtle but important fashion. In summary, iFORM/eQTL not only provides an estimate of the overall heritability of gene transcript A_12_P103290 (i.e. the sum of individual heritability explained by each predictor), but also charts a detailed picture of how each genetic variant contributes to transcript variation. In particular, iFORM/eQTL characterizes epistasis and its role in trait control and is equipped with a capacity to retrieve so-called missing heritability [30], a significant issue arising from current GWAS.

Through analyzing associations between all markers and each transcript by iFORM/eQTL, we can identify the difference of *cis*- and *trans*-eQTLs for a particular transcript. For example, of the eQTLs affecting A_12_P103290, we detected that X1_2068168 is a *cis*-eQTL, whereas all others are *trans*-eQTLs (Table 4). We list the number and distribution of these two types of eQTLs and the pattern of how they interact with each other to determine gene transcripts (Table 5). By detecting *cis*-eQTLs and *trans*-eQTLs, iFORM/eQTL detected that genetic interactions take place mostly between *trans*-eQTLs.

Discussion

With the recent development of genotyping and sequencing techniques, the collection of genome-wide genetic and genomic data from any tissue of an organism has been rendered much easier and more efficient. The past decade has been a fertile one for genetic and genomic studies of complex diseases or traits,

Table 4. Estimated main and epistatic effects of eQTLs by iFORM/eQTL on gene transcript a_12_P103290 on chromosome I, in a comparison with the result by traditional single-marker analysis

	iForm/eQTL				Single-marker analysis			Forward select		
eQTL	Effect	SE	P-value	Heritability	Effect	SE	P-value	Effect	SE	P-value
X1_2068168 (cisQTL)	−0.197	0.035	0.000	0.060	−0.210	0.074	0.005	−0.145	0.027	0.000
X2_13516256	−0.069	0.039	0.080	0.007	−0.138	0.075	0.068	NA	NA	NA
X2_2482896	0.064	0.027	0.017	0.006	−0.057	0.076	0.453	NA	NA	NA
X_14636404	−1.768	0.092	0.000	4.794	0.024	0.076	0.751	0.0191	0.091	0.833
X4_15568674	−1.972	0.134	0.000	5.964	0.094	0.075	0.213	NA	NA	NA
X4_1873297	0.044	0.026	0.086	0.003	0.095	0.071	0.185	0.045	0.026	0.084
X4_15632637	1.960	0.143	0.000	5.892	0.104	0.075	0.169	NA	NA	NA
X4_13532205	0.064	0.028	0.024	0.006	0.111	0.075	0.143	NA	NA	NA
X_15820520	−0.014	0.055	**0.796**	0.000	0.146	0.075	0.054	0.032	0.054	0.552
X_14542103	1.786	0.087	0.000	4.892	0.162	0.075	0.031	0.012	0.084	0.883
X2_13516256.X4_15632637	−3.799	0.268	0.000	5.534	NA	NA	NA	NA	NA	NA
X2_13516256.X4_15568674	3.753	0.276	0.000	5.401	NA	NA	NA	NA	NA	NA
X_15820520.X_14636404	−3.771	0.172	0.000	5.453	NA	NA	NA	NA	NA	NA
X_15820520.X_14542103	3.691	0.172	0.000	5.224	NA	NA	NA	NA	NA	NA
X_14636404.X4_1873297	−3.534	0.163	0.000	4.789	NA	NA	NA	NA	NA	NA
X_14636404.X4_13532205	−3.567	0.166	0.000	4.879	NA	NA	NA	NA	NA	NA
X_14542103.X4_1873297	3.629	0.164	0.000	5.050	NA	NA	NA	NA	NA	NA
X_14542103.X4_13532205	3.469	0.167	0.000	4.614	NA	NA	NA	NA	NA	NA

Note: Multiple R-squared: 0.9074, adjusted R-squared: 0.8964.

Table 5. The distribution and proportion of *cis*- and *trans*-QTLs detected by iFORM/eQTL

eQTL type	Count	Proportion
cis-eQTL	14	0.0024
trans-eQTL	5509	0.9628
cis-eQTL × *cis*-eQTL	0	0.0000
trans-eQTL × *cis*-eQTL/*cis*-eQTL × *trans*-eQTL	2	0.0003
cis-eQTL × *trans*-eQTL	196	0.0340

A random sample of 1000 transcripts was used as a response, and iFORM/eQTL was implemented on each. The number of predictors selected during the procedure is shown, along with the eQTL type of each of the predictors in relation to the transcript used as the response.

which have now achieved a point at which we can draw a complete picture of genetic architecture for disease or trait formation and progression by GWAS [17].

Traditional marginal analysis based on simple regression has been instrumental for the detection of important genetic variants or quantitative trait loci in a variety of organisms, but a bottleneck has emerged quickly owing to its limitation in precisely and comprehensively charting genetic control landscapes. Many published GWAS studies are beset with missing heritability because of their incapacity to detect genome-wide epistasis, genotype × environment interactions and any possible other mechanisms [30]. Epistasis is a phenomenon by which the influence of a gene on the phenotype depends critically on the context provided by other genes [14]. It has been increasingly recognized that epistasis is an important source for trait variation [15, 31, 32], so that inclusion of epistasis would enhance the prediction accuracy of phenotypic performance and shed more light on the global genetic architecture of trait control [17]. However, epistasis is extremely hard to detect as an interaction term, whose inclusion may complicate the inference of the predictive model [17, 31]. Thanks to recent methodological progresses in high-dimensional data modeling, we have been able to implement several cutting-edge statistical models for systematical detection and characterization of genome-wide epistasis.

Hao and Zhang [26] proposed a new high-dimensional model, iFORM, that tackles an issue of interaction selection simultaneously from a large pool of continuous predictors. This model is based on forward-selection-based procedures, which are characteristic of computational feasibility and efficiency. The authors further proved that the detection of interactions by iFORM is consistent, even if the dimension increases exponentially for a sample size. As one of the first attempts to introduce high-dimensional models into genetic studies, we modified and implemented iFORM to accommodate the discrete nature of molecular markers. Our simulation studies indicate that the resulting iFORM/eQTL platform can provide reasonably accurate and precise estimates of genetic main effects and interaction effects. It shows greater power to detect significant genes and their interactions, which may not be detected by traditional single-marker analysis. Also, although its underlying assumption is the heredity principle, iFORM/eQTL was found to well detect epistasis taking place between any two loci that display weak main effects. This functionality remarkably increases the applicability of this high-dimensional model to large-scale genetic data. However, one drawback of iFORM/eQTL is its incapacity to detect the epistasis of two eQTLs whose main effects are insignificant. A modification to cope with this situation deserves further investigation.

We applied iFORM/eQTL to reanalyze gene expression data in an eQTL mapping study [12]. While our results confirmed those by the traditional approach, the new model provides some new findings including new eQTLs and epistasis, thus allowing a complete set of genetic variants to be characterized. As an important tool to understand the genetic mechanisms underlying both complex traits and diseases, eQTL mapping has been widely used to identify key regulatory pathways toward endophenotype and end-point phenotypes [1–3, 33]. The eQTLs displaying significant main effects detected by iFORM/eQTL were broadly in agreement with those reported in Rockman *et al.*'s [12] well-validated studies, showing the biological relevance of iFORM/eQTL. The new insight into epistasis gained by iFORM/eQTL provides molecular geneticists with next hypotheses to pursue subsequent experiments to understand biology. As a working example, we randomly chose one transcript involved in our iFORM/eQTL analysis. However, a complete analysis of all transcripts collected in Rockman *et al.* [12], from which new discoveries can be made, deserves a separate publication.

A typical eQTL study may not only include a large number of molecular markers like in a GWAS, but also record tens of thousands of gene transcripts throughout the entire genome. Our current version of iFORM/eQTL can only take into account one gene transcript as a response at a time, thus having a limitation to model the correlation and dependence among different genes. It is our next step to formulate a multivariate multiple regression model by which to test how an individual predictor, main effect or epistatic effect, pleiotropically affects correlated expression profiles of different genes.

Given the complexity of biological phenomena, pair-wise epistasis may be insufficient to explain phenotypic variation. Imielinski and Belta [34] argued that high-order interactions among more than two genes may provide a key pathway toward complex traits. Three-way interactions have been detected in trait control [35, 36]. A model for modeling three-way interactions has been developed in a case-control GWAS design [37] and a genetic mapping setting [38]. It is crucial to extend iFORM/eQTL to map main effects, two-way epistasis and three-way epistasis in an eQTL mapping study although no substantial change is needed in the computational algorithm, except for an enlarged test set and extra computing time. Our work is based on a backcross population in which there are only two genotypes at a locus. The backcross population can facilitate our estimation and test of genetic effects owing to a smaller number of parameters at each locus or locus pair, but its utility is limited in the F_2 design of model systems and natural populations of outcrossing species such as humans. A more general model of iFORM/eQTL should consider three genotypes at each locus, which provides estimates of additive and dominant effects at each locus and four types of epistasis, i.e. additive × additive, additive × dominant, dominant × additive and dominant × dominant, between each pair of loci [27]. Each of these epistatic types may affect a phenotype through a different pathway.

Although the current implementation of iFORM/eQTL focuses on SNPs as predictors, it is flexible to include other types of predictors, such as multi-nucleotide polymorphisms and indels, given their roles in regulating biological processes. With continuous falling of sequencing price, we will have desirable opportunities to study the dynamic behavior and pattern of gene expression profiles across time and space scales [39–41]. Many previous studies suggest that gene expression during cell and organ development may follow a particular form, which can be quantified by mathematical equations [42]. For example, abundance of gene expression may change periodically in human's brain during circadian clock. Several researchers used Fourier's series approximation to model the periodic changes of gene expression by estimating the period and amplitude of the cycles [43]. Statistical models for mapping QTLs that regulate dynamic change of phenotypic traits through mathematical equations, called functional mapping, have been developed [44, 45] and further have proven to be broadly useful for the genetic dissection of complex traits [46]. By integrating iFORM/eQTL into functional mapping, we will be able to map dynamic eQTLs for gene expression and make a quantitative prediction of temporal and spatial patterns of genetic control by eQTLs. We packed iFORM/eQTL in R with the source code available at http://statgen.psu.edu/software.html/ (after the manuscript is accepted).

Key Points

- The identification of expression quantitative trait loci (eQTLs) facilitates the precise reconstruction of the genotype–phenotype map for complex traits or diseases.
- We implement a variable selection model to map a comprehensive set of eQTLs and their interactions from an ultrahigh-dimensional array of genes throughout the entire genome.
- This implementation builds a computing platform to illustrate the genetic architecture of gene transcripts.

Supplementary data

Supplementary data are available online at http://bib.oxfordjournals.org/.

Acknowledgments

We thank Leonid Kruglyak and Mathew Rockman for providing their *C. elegans* data to test and validate the model. We also thank Helen Hao Zhang for useful discussion on model development and for sharing her computer code with us.

Funding

This research was funded by grants U01 HL119178 and UL1 TR000127.

References

1. Emilsson V, Thorleifsson G, Zhang B, *et al*. Genetics of gene expression and its effect on disease. *Nature* 2008;**452**:423–8.
2. Cookson W, Liang L, Abecasis G, *et al*. Mapping complex disease traits with global gene expression. *Nat Rev Genet 2009* 2009;**10**:184–94.
3. Nica AC, Dermitzakis ET. Expression quantitative trait loci: present and future. *Philos Trans R Soc Lond B Biol Sci* 2013;**368**:20120362.
4. Kim Y, Via K, Tao R, *et al*. A meta-analysis of gene expression quantitative trait loci in brain. *Transl Psychiatry* 2014;**4**:e459.
5. Fairfax BP, Humburg P, Makino S, *et al*. Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. *Science* 2014;**343**:1246949.
6. Lee MN, Ye C, Villani AC, *et al*. Common genetic variants modulate pathogen-sensing responses in human dendritic cells. *Science* 2014;**343**:1246980.
7. Li L, Kabesch M, Bouzigon E, *et al*. Using eQTL weights to improve power for genome-wide association studies: a genetic study of childhood asthma. *Front Genet* 2013;**4**:103.
8. Koopmann TT, Adriaens ME, Moerland PD, *et al*. Genome-wide identification of expression quantitative trait loci (eQTLs) in human heart. *PLoS One* 2014;**9**(5):e97380.
9. Chun H, Keles S. Expression quantitative trait loci mapping with multivariate sparse partial least squares regression. *Genetics* 2009;**182**:79–90.

10. Sun W. A statistical framework for eQTL mapping using RNA-seq data. *Biometrics* 2002;**68**:1–11.
11. Flutre T, Wen X, Pritchard J, *et al*. A statistical framework for joint eQTL analysis in multiple tissues. *PLoS Genet* 2013;**9**(5):e1003486.
12. Rockman MV, Skrovanek SS, Kruglyak L. Selection at linked sites shapes heritable phenotypic variation in *C. elegans*. *Science* 2010;**330**:372–6.
13. Wu R, Ma C, Casella G. *Statistical Genetics of Quantitative Traits: Linkage, Maps, and QTL*. New York: Springer-Verlag, 2007.
14. Cheverud JM, Routman EJ. Epistasis and its contribution to genetic variance components. *Genetics* 1995;**139**:1455–61.
15. Moore JH. The ubiquitous nature of epistasis in determining susceptibility to common human diseases. *Hum Hered* 2003;**56**:73–82.
16. van Eeuwijk FA, Bink MC, Chenu K, *et al*. Detection and use of QTL for complex traits in multiple environments. *Curr Opin Plant Biol* 2010;**13**:193–205.
17. Mackay TFC. Epistasis and quantitative traits: using model organisms to study gene–gene interactions. *Nat Rev Genet* 2014;**15**:22–33.
18. Tibshirani R. Regression shrinkage and selection via the Lasso. *J Roy Stat Soc Ser B* 1996;**58**:267–88.
19. Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J Am Stat Assoc* 2001;**96**: 1348–60.
20. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J Roy Stat Soc Ser B* 2005;**67**:301–20.
21. Zhang CH. Nearly unbiased variable selection under minimax concave penalty. *Ann Stat* 2010;**38**:894–942.
22. Candes E, Tao T. The dantzig selector: statistical estimation when p is much larger than n. *Ann Stat* 2007;**35**:2313–51.
23. Zhao P, Yu B. On model selection consistency of Lasso. *J Mach Learn Res* 2006;**7**:2541–63.
24. Fan J, Lv J. Sure independence screening for ultrahigh dimensional feature space. *J R Stat Soc Ser B* 2008;**70**:849–911.
25. Wang H. Forward regression for ultra-high dimensional variable screening. *J Am Stat Assoc* 2009;**104**:1512–24.
26. Hao N, Zhang HH. Interaction screening for ultrahigh-dimensional data. *J Am Stat Assoc* 2014;**507**:1285–301.
27. Fisher RA. The correlation between relatives on the supposition of Mendelian inheritance. *Trans R Soc (Edinburgh)* 1918;**52**:399–433.
28. Kao C-H, Zeng ZB. Modeling epistasis of quantitative trait loci using Cockerham's model. *Genetics* 2002;**160**:1243–61.
29. Chen J, Chen Z. Extended bayesian information criteria for model selection with large model spaces. *Biometrika* 2008;**95**: 759–71.
30. Manolio TA, *et al*. Finding the missing heritability of complex diseases. *Nature* 2009;**461**:747–53.
31. Carlborg O, Haley CS. Epistasis: too often neglected in complex trait studies? *Nat Rev Genet* 2004;**5**:618–25.
32. Cordell HJ. Detecting gene-gene interactions that underlie human diseases. *Nat Rev Genet* 2009;**10**:392–404.
33. Pickrell JK, Marioni JC, Pai AA, *et al*. Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* 2010;**464**:768–72.
34. Imielinski M, Belta C. Exploiting the pathway structure of metabolism to reveal high-order epistasis. *BMC Syst Biol* 2008;**2**:40.
35. McMullen MD, Byrne PF, Snook ME, *et al*. Quantitative trait loci and metabolic pathways. *Proc Natl Acad Sci USA* 1998;**95**:1996–2000.
36. Stich B, Yu J, Melchinger AE, *et al*. Power to detect higher-order epistatic interactions in a metabolic pathway using a new mapping strategy. *Genetics* 2007;**176**:563–70.
37. Wang Z, Liu T, Lin Z, *et al*. A general model for multilocus epistatic interactions in case-control studies. *PLoS One* 2010;**5**(8):e11384
38. Pang XM, Wang Z, Yap JS, *et al*. A statistical procedure to map high-order epistasis for complex traits. *Brief Bioinform* 2013;**14**:302–14.
39. Viñuela A, Snoek LB, Riksen JA, *et al*. Genome-wide gene expression regulation as a function of genotype and age in *C. elegans*. *Genome Res* 2010;**20**:929–37.
40. Ackermann M, Sikora-Wohlfeld W, Beyer A. Impact of natural genetic variation on gene expression dynamics. *PLoS Genet* 2013;**9**:e1003514.
41. Francesconi M, Lehner B. The effects of genetic variation on gene expression dynamics during development. *Nature* 2014;**505**:208–11.
42. Kim BR, McMurry T, Zhao W, *et al*. Wavelet-based functional clustering for patterns of high-dimensional dynamic gene expression. *J Comp Biol* 2010;**17**:1067–80.
43. Li N, McMurry T, Berg A, *et al*. Functional clustering of periodic transcriptional profiles through ARMA (p, q). *PloS One* 2010;**5**:e9894.
44. Ma CX, Casella G, Wu R. Functional mapping of quantitative trait loci underlying the character process: a theoretical framework. *Genetics* 2012;**161**:1751–62.
45. Wu R, Lin M. Functional mapping—how to map and study the genetic architecture of dynamic complex traits. *Nat Rev Genet* 2006;**7**:229–37.
46. Li Z, Sillanpaa MJ. Dynamic quantitative trait locus analysis of plant phenomic data. *Trends Plant Sci* 2015;**20**:822–33. doi:10.1016/j.tplants.2015.08.012.