**An ultra-high dimensional model for inferring the global genetic architecture of gene transcripts**

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

Department of Public Health Sciences, Penn State College of Medicine, Hershey, PA 17033

**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, I present an innovative application of variable selection approaches to systematically detect main effects and interaction effects among all possible loci on differentiation and function of gene expression. Forward-selection-based procedures were particularly implemented to tackle complex covariance structures of gene-gene interactions. We reanalyzed a published genetic and genomic dataset collected in a mapping population of *Caenorhabditis elegans*, gaining new discoveries on the genetic origin of gene expression differentiation, which could not be detected by a traditional one-locus/one-transcript analysis approach.

**Keywords:** eQTL, marker, variable selection, genetic effect, epistasis

**Introduction**

Since 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 (Emilsson et al. 2008; Cookson et al. 2009; Nica and Dermitzakis 2013). 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 (Kim et al. 2014; Fairfax et al. 2014; Lee et al. 2014). 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 (Nica and Dermitzakis 2013; Li et al. 2013; Koopmann et al. 2014). Statistical analysis and modeling are playing an increasing role in mapping and identifying the underlying eQTLs from massive amounts of observed data (Kendziorski et al. 2006; Chun and Keles 2009; Sun 2012; Flutre et al. 2013).

A typical eQTL mapping approach is to associate a gene transcript with a single marker such as single nucleotide polymorphism (SNP). By analzying 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 identifcation of eQTLs in a variety of organisms (Rockman et al. 2010; Kim et al. 2014). However, there are two major limitations for the results by such a marginal analysis. First, it does not take into account the dependence of different markers, so that a significant association detected by one marker may be due to the other markers that are linked with it. The marginal marker analysis cannot separate the confounding effect of eQTLs due to marker-marker dependence or linkage (Wu et al. 2007). 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 (Cheverud and Routman 1995; Moore 2003; van Eeuwijk et al. 2010; Mackay 2014)

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 decade has witnessed the tremendous development of variable selection models for high-dimensional data analysis, such as LASSO (Tibshirani 1996), SCAD (Fan and Li 2001), Dantzig selector (Candes and Tao 2007), elastic net (Zou and Hastie 2005), and minimax concave penalty (MCP) (Zhang 2010) among others. Many methods possess favorable theoretical properties such as model selection consistency (Zhao and Yu 2006) and oracle properties (Fan and Lv 2011). When the number of predictors is much larger than the number of observations, sure screening is a more realistic goal to achieve than oracle properties or selection consistency (Fan and Lv 2008; Wang 2009). Sure screening assures that all important variables are identified with a probability tending to one, 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 (2014) 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. In this article, we implement and reform Hao and Zhang’s model to map the genetic architecture of eQTL actions and interactions for gene expression profiles. This model is modified to accommodate 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. We expand Hao and Zhang’s regression model to include discrete components. Also, for an F2 or a natural population with three genotypes at each locus, we need to estimate a total of eight genetic effects for a pair of markers, which are additive and dominant effects at each locus, and additive ×additive, additive ×dominant, dominant ×additive and dominant ×dominant effects between the two loci (Fisher 1918). Thus, if the number of markers is *p*, a total number of predictors including all main and two-way interaction terms is 2*p*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 an ultrahigh one.

By modeling all markers jointly at one time under an organizing framework, the modified model 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. Our model can more precisely discern these two different types of eQTLs and their interactions than traditional marginal analysis. By reanalyzing a published dataset collected in a mapping population of *C. elegans* (Rockman et al. 2010), the new model validates 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.

**Method**

**Experimental design**

Consider an experimental population for genetic studies of complex traits, such as the backcross and F2 initiated from two inbred lines, full-sib family derived from two outcrossing parents, or random samples drawn from a natural population. 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 single nucleotide polymorphisms (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 (Li et al. 2013). 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 progeny 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**

The marker and gene transcript data of the study population can be denoted as (***x****i*,*yi*) (*i* = 1, …, *n*) which are independent and identically distributed copies of (***X***,*Y*), where is a *p*-dimensional predictor vector and *Y* is the response, expressed by a linear regression model:

(1)

The ***β***’sare the coefficients for the genetic effects of each marker. Like most genome-wide datasets, the number of markers here grossly outnumbers the number of observations, i.e., *p* >> *n*. Therefore, selection procedures would need to be implemented in order 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 . The resulting linear model would grow to be

(2)

where ***γ’s*** 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, . Also, the quadratic and two-way interaction effects will be centered which we will write as . By doing so we eliminate the need for an intercept in regression model (2). This reduces the model to the form,

(3)

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

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

For any model , is used to denote the number of predictors contained in the model. The true model size is indicated by , or together is denoted by . For the procedure, three sets will be used throughout. The sets are for the model set, for the candidate set of predictors and 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 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 either not been included in the model or are deleted because their effects become marginal by the inclusion of the interaction effect. The second 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 (Zhao et al. 2009; Yuan et al. 2009; Choi et al. 2010). This translates into

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.

**iFORM**

Hao and Zhang (2014) formulated an interaction forward selecting procedure under the marginality principle (iFORM). The procedure’s initial step starts with the empty set for both the solution set and the model set, . The candidate set contains all main effects at the beginning, , 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.,

where is the sample variance for the given model, is the size of the model or the number of predictors selected into the given model, and is the sample size. The term is the number of predictors in the full model. This was proposed as by Chen and Chen (2008), which they derived to help control the false discovery rate in high-dimensional data situations. They also showed that it was selection consistent if The only difference between the traditional calculation and the is the additional term involving 2log(*d*\*). Ignoring the , 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 , in the true model is unknown. We want to make large enough to include but not so large as to fit the model to the point where it becomes oversaturated. Using the should help avoid oversaturation. It is reasonable to assume that is much smaller than in high-dimensional sparse regression problems (Fan and Lv 2008). Since this is the case, for the purposes of our model, is set to be no larger than . Generally, the should reach a minimum, indicating the optimal stopping point, before the designated stopping value is reached.

**Some considerations**

There are some considerations and pre-processing steps to take before the iFORM procedure is implemented. The first consideration is to determine if there are any exact duplicate markers in the dataset. One drawback that could arise with marker datasets when attempting to run multiple linear regression is the possibility of duplicate markers in the dataset. If two different markers happen to have exactly the same genotypes for each subject and are included in the linear model, then it would show up as an exact linear combination of each other. Including redundant markers in a linear model would not add any additional information and therefore should not be included in the candidate set during the selection procedure. This also reduces the dimension slightly when there are duplicate markers are removed from the dataset.

Another consideration is the type of coding used for the genotypes. At any given eQTL, say the *j*th eQTL, there are two possible genotypes: and , making the total number of possible QTL genotypes in the population . The goal of a genetic model is to relate the possible genotypic values to a set of genetic parameters, such that these parameters are interpretable in terms of main and epistatic effects of the 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 (Kao and Zeng 2002; Zeng et al. 2005). The orthogonal contrasts for the genetic model can be expressed by

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 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 as well. 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 out 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 results in an interaction effect being coded as . This coding occurs for different scenarios for each of the levels. The coding of could arise when the interaction is comprised of a homozygote interacting with a heterozygote genotype. A coding of would arise by either a homozygote interacting with another homozygote genotype, or when a heterozygote interacts with another heterozygote genotype.

**Application**

**Simulation**

We conducted simulation studies to test the theoretical properties of the selection procedures and the results (Tables 1 -3). The results were compared to several other commonly used methods for eQTL mapping. In each of the examples the response was generated from model (2) with for the random error with a sample size of . The were all independently and identically distributed realizations generated from and then orthogonal contrasts were made with each . The true , yielding and . The relevant interactions were set to the pairs and all with . There were several methods compared during each of the simulations (Tables 1 -3). 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 replicates were run and the outcomes considered include

* Convergence Probability (Cov)
* Percentage of correct zeros (Cor0)
* Percentage of incorrect zeros (Inc0)
* Exact Selection probability (Exact)
* The average model size
* Mean Square Error (MSE)
* Adjusted R-square
* Computation Time in seconds

In each instance of the simulation, the iFORM procedure was closest to the simulated data, indicated as Oracle. Single marker analysis was conducted on each of the main effects individually and the significant markers were then designated as eQTLs. When comparing the 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. The iFORM procedure contains the identified main effects over 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 over 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 the iFORM procedure 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 in order 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 the iFORM procedure. Another drawback to implementing forward selection on such a large set seemed to come with over fitting 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 (Tables 1 -3). This increased the adjusted R-square value of the final model; however, this is suspected because of over-fitting the data and not to be a true prediction of the response.

**Real data analysis**

Rockman et al. (2010) 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 20,000 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 single-nucleotide polymorphism (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. (2010) 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 (Fig. 1), from 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. 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 due 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 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 = 0.896) contains 14 markers which exert their main effects and/or interaction effects on the transcript. Of the 14 final markers, one-half 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 due 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 very surprising that epistasis contributes to more than one-half of the heritability. Of the eight epistatic interactions, only one occurs due to the interaction between two significant eQTLs, X\_14542103 and X4\_13532205 (Table 4). All the remaining ones are due to interactions between one significant eQTL and one non-significant 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 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 characterizes epistasis and its role in trait control and is equipped with a capacity to retrieve so-called missing heritability (Manolio et al. 2009), a significant issue arising from current genome-wide association studies.

Through analyzing associations between all markers and each transcript by iFORM, 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 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. Because of this, genetic studies of complex diseases or traits have developed during the past decade to a point at which we can draw a complete picture of genetic architecture for disease or trait formation and progression by genome-wide association studies (GWAS) (Mackay et al. 2009). 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 due 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 and genotype × environment interactions (Manolio et al. 2009).

Epistasis is a phenomenon by which the influence of a gene on the phenotype depends critically upon the context provided by other genes (Cheverud and Routman 1995). It has been increasingly recognized that epistasis is an important source for trait variation (Moore 2003; Carlborg and Haley 2004; Cordell 2009), 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 (Mackay 2014). However, epistasis is extremely hard to detect as an interaction term, whose inclusion may complicate the inference of the predictive model (Carlborg and Haley 2004; Mackay 2014). Thanks to recent 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 (2014) 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 iFORM to accommodate the discrete nature of molecular markers. Our simulation studies indicate that iFORM can provide reasonably accurate and precise estimates of genetic main effects and interaction effects. Also, it shows greater power to detect significant genes and their interactions which may not be detected by traditional single marker analysis.

We applied iFORM to re-analyze gene expression data in an eQTL mapping study (Rockman et al. 2010). 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 (Schadt et al. 2005; Emilsson et al. 2008; Cookson et al. 2009; Pickrell et al. 2010; Nica and Dermitzakis 2013). A typical eQTL study may not only include a large number of molecular markers as like in a GWAS, but also record tens of thousands of gene transcripts throughout the entire genome. Our current version of iFORM 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 (2008) 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 (McMullen et al. 1998; Stich et al. 2007). A model for modeling three-way interactions has been developed in a case-control GWAS design (Wang et al. 2010) and a genetic mapping setting (Pang et al. 2012). It is crucial to extend iFORM 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 very limited in the F2 design of model systems and natural populations of outcrossing species such as humans. A more general model of iFORM 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 (Fisher 1918). Each of these epistatic types may affect a phenotype through a different pathway.

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 (Viñuela et al. 2010; Ackermann et al. 2013). Many previous studies suggest that gene expression during cell and organ development may follow a particular form, which can be quantified by mathematical equations (Kim et al. 2008). For example, abundance of gene expression may change periodically in human’s brain during circadian clock. Many researchers used Fourier’s series approximation to model the periodic changes of gene expression by estimating the period and amplitude of the cycles (Li et al. 2010). By integrating Fourier series into iFORM, 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.

**Aknowledgements**

We thank Leonid Kruglyak for helpful discussions on model development and further thank him and Mathew Rockman for providing their *C. elegans* data to test and validate the model. This research was funded by grants U01 HL119178 and UL1 TR000127.

**References**

Ackermann, M. et al. A. Impact of natural genetic variation on gene expression dynamics. PLoS Genet. 9, e1003514 (2013).

Candes, E., and Tao, T. (2007), “The Dantzig Selector: Statistical Estimation When p is Much Larger Than n,” The Annals of Statistics, 35, 2313–2351.

Carlborg O, Haley CS. Epistasis: Too often neglected in complex trait studies? Nat Rev Genet. 2004;5:618–625.

Chen, J. & Chen, Z. (2008). Extended bayesian information criteria for model selection with large model spaces. Biometrika 95, 759–771.

Chun, H. & Keles, S. Expression quantitative trait loci mapping with multivariate sparse partial least squares regression. Genetics 182, 79-90 (2009).

Cookson, W., Liang, L., Abecasis, G., Moffatt, M. & Lathrop, M. Mapping complex disease traits with global gene expression. Nat. Rev. Genet. 10, 184-194 (2009).

Cordell HJ. Detecting gene-gene interactions that underlie human diseases. Nat Rev Genet. 2009;10:392–404.

Emilsson V. et al. Genetics of gene expression and its effect on disease. Nature 452, 423-428 (2008).

Evans et al. 2006; Manolio and Collins 2007; Kooperberg and LeBlanc 2008; Cordell 2009).

Fairfax, B.P. et al. Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. Science 343, 1246949 (2014).

Fan, J., and Li, R. (2001), “Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties,” Journal of the American Statistical Association, 96, 1348–1360.

Fan, J. & Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **70**, 849–911.

Fan J, Lv J (2011), “Nonconcave Penalized Likelihood With np-Dimensionality, Information Theory, IEEE Transactions on, 57, 5467–5484.

Fisher RA. The correlation between relatives on the supposition of Mendelian inheritance. Trans Roy Soc (Edinburgh) 1918;52:399–433.

Flutre T, Wen X, Pritchard J, Stephens M (2013) A Statistical Framework for Joint eQTL Analysis in Multiple Tissues. PLoS Genet 9(5): e1003486.

Francesconi, M. & Lehner, B. The effects of genetic variation on gene expression dynamics during development. Nature 505, 208-211 (2014).

Hao N, Zhang HH (2014) Interaction screening for ultrahigh-dimensional data. J Am Stat Assoc 507: 1285-1301

Imielinski M, Belta C. Exploiting the pathway structure of metabolism to reveal high-order epistasis. BMC Syst Biol 2008;2:40.

J. H. Moore, “The ubiquitous nature of epistasis in determining susceptibility to common human diseases,” Hum. Hered., vol. 56, pp. 73–82, 2003.

J. M. Cheverud, E. J. Routman, “epistasis and its contribution to genetic variance components,” Genetics, vol 139, 1455–1461, 1995.

Kao, C.-H., and Z-B. Zeng, 2002 Modeling epistasis of quantitative trait loci using Cockerham’s model. Genetics 160: 1243–1261.

Kendziorski, C.M. et al. Statistical methods for expression quantitative trait loci (eQTL) mapping. Biometrics 62, 19-27 (2006).

Kim, B.R. et al. Wavelet-based functional clustering for patterns of high-dimensional dynamic gene expression. J. Comp. Biol. 17, 1067-1080 (2010).

Kim, Y. et al. A meta-analysis of gene expression quantitative trait loci in brain. Transl. psychiatry 4, e459 (2014).

Koopmann TT, Adriaens ME, Moerland PD, Marsman RF, Westerveld ML, et al. (2014) Genome-Wide Identification of Expression Quantitative Trait Loci (eQTLs) in Human Heart. PLoS ONE 9(5): e97380

Lee, M.N. et al. Common genetic variants modulate pathogen-sensing responses in human dendritic cells. Science 343, 1246980 (2014).

[Li L](http://www.ncbi.nlm.nih.gov/pubmed/?term=Li%20L%5BAuthor%5D&cauthor=true&cauthor_uid=23755072), [Kabesch M](http://www.ncbi.nlm.nih.gov/pubmed/?term=Kabesch%20M%5BAuthor%5D&cauthor=true&cauthor_uid=23755072), [Bouzigon E](http://www.ncbi.nlm.nih.gov/pubmed/?term=Bouzigon%20E%5BAuthor%5D&cauthor=true&cauthor_uid=23755072), [Demenais F](http://www.ncbi.nlm.nih.gov/pubmed/?term=Demenais%20F%5BAuthor%5D&cauthor=true&cauthor_uid=23755072), [Farrall M](http://www.ncbi.nlm.nih.gov/pubmed/?term=Farrall%20M%5BAuthor%5D&cauthor=true&cauthor_uid=23755072), [Moffatt MF](http://www.ncbi.nlm.nih.gov/pubmed/?term=Moffatt%20MF%5BAuthor%5D&cauthor=true&cauthor_uid=23755072), [Lin X](http://www.ncbi.nlm.nih.gov/pubmed/?term=Lin%20X%5BAuthor%5D&cauthor=true&cauthor_uid=23755072), Liang L (2013) Using eQTL weights to improve power for genome-wide association studies: a genetic study of childhood asthma. Front Genet 4:103.

Li, N. et al. Functional clustering of periodic transcriptional profiles through ARMA (p, q). PloS One 5, e9894 (2010).

Mackay TFC, Stone EA, and Ayroles JF. (2009). The genetics of quantitative traits: Challenges and prospects. Nat. Rev. Genet. 10: 565–577.

## Mackay TFC (2014) Epistasis and quantitative traits: using model organisms to study gene–gene interactions. Nature Reviews Genetics 15, 22–33

Manolio TA, et al. (2009) Finding the missing heritability of complex diseases. Nature 461: 747–753.

McMullen MD, Byrne PF, Snook ME, Wiseman BR, Lee EA, Wildstrom NW, Coe EH. Quantitative trait loci and metabolic pathways. Proc Natl Acad Sci U S A 1998;95:1996-2000.

Nica AC, Dermitzakis ET (2013) Expression quantitative trait loci: present and future. Philos Trans R Soc Lond B Biol Sci 368(1620): 20120362.

Pang XM, Wang Z, Yap JS, Bo WH, Lv YF, Xu F, Zhou T, Peng SF, Shen DF, Wu RL (2013) A statistical procedure to map high-order epistasis for complex traits. Brief Bioinform 14: 302-314.

**Pickrell JK**, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, Veyrieras J-B, Stephens M, Gilad Y, Pritchard JK. Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature. 2010 Apr 1;464(7289):768-72

Rockman, M.V., Skrovanek, S.S. & Kruglyak, L. Selection at linked sites shapes heritable phenotypic variation in C. elegans. Science 330, 372-376 (2010).

Schadt et al. 2005; Emilsson et al. 2008; Cookson et al. 2009; Nica and Dermitzakis 2013

Schadt E.E. et al. An integrative genomics approach to infer causal associations between gene expression and disease. Nat. Genet. 37, 710-717 (2005).

Stich B, Yu J, Melchinger AE, Piepho H-P, Utz HF, Maurer HP, Buckler ES. Power to detect higher-order epistatic interactions in a metabolic pathway using a new mapping strategy. Genetics 2007;176:563-570.

Sun W (2012) A statistical framework for eQTL mapping using RNA-seq data. Biometrics 68: 1–11.

Tibshirani, R. (1996), “Regression Shrinkage and Selection via the Lasso,” Journal of the Royal Statistical Society, Series B, 58, 267–288.

van Eeuwijk FA, Bink MC, Chenu K, Chapman SC (2010) Detection and use of QTL for complex traits in multiple environments. Curr Opin Plant Biol 13: 193-205.

Viñuela, A. et al. Genome-wide gene expression regulation as a function of genotype and age in C. elegans. Genome Res. 20, 929–937 (2010).

Wang, H. (2009), “Forward Regression for Ultra-High Dimensional Variable Screening,” Journal of the American Statistical Association, 104, 1512–1524.

Wang Z, Liu T, Lin ZW, Hegarty J, Koltun WA, Wu RL, 2010 A general model for multilocus epistatic interactions in case-control studies. PLoS ONE 5(8): e11384.

Wu RL, Ma C-X, Casella G (2007) Statistical Genetics of Quantitative Traits: Linkage, Maps, and QTL. Springer-Verlag, New York.

Zeng, Z-B., T. Wang and W. Zou, 2005 Modeling quantitative trait loci and interpretation of models. Genetics 169: 1711–1725

Zhao, P., and Yu, B. (2006), “On Model Selection Consistency of Lasso,” Journal of Machine Learning Research, 7, 2541–2563.

Zhang, C.-H. (2010), “Nearly Unbiased Variable Selection Under Minimax Concave Penalty,” The Annals of Statistics, 38, 894–942.

Zou, H., and Hastie, T. (2005), “Regularization and Variable Selection via the Elastic Net,” Journal of the Royal Statistical Society, Series B, 67, 301–320.

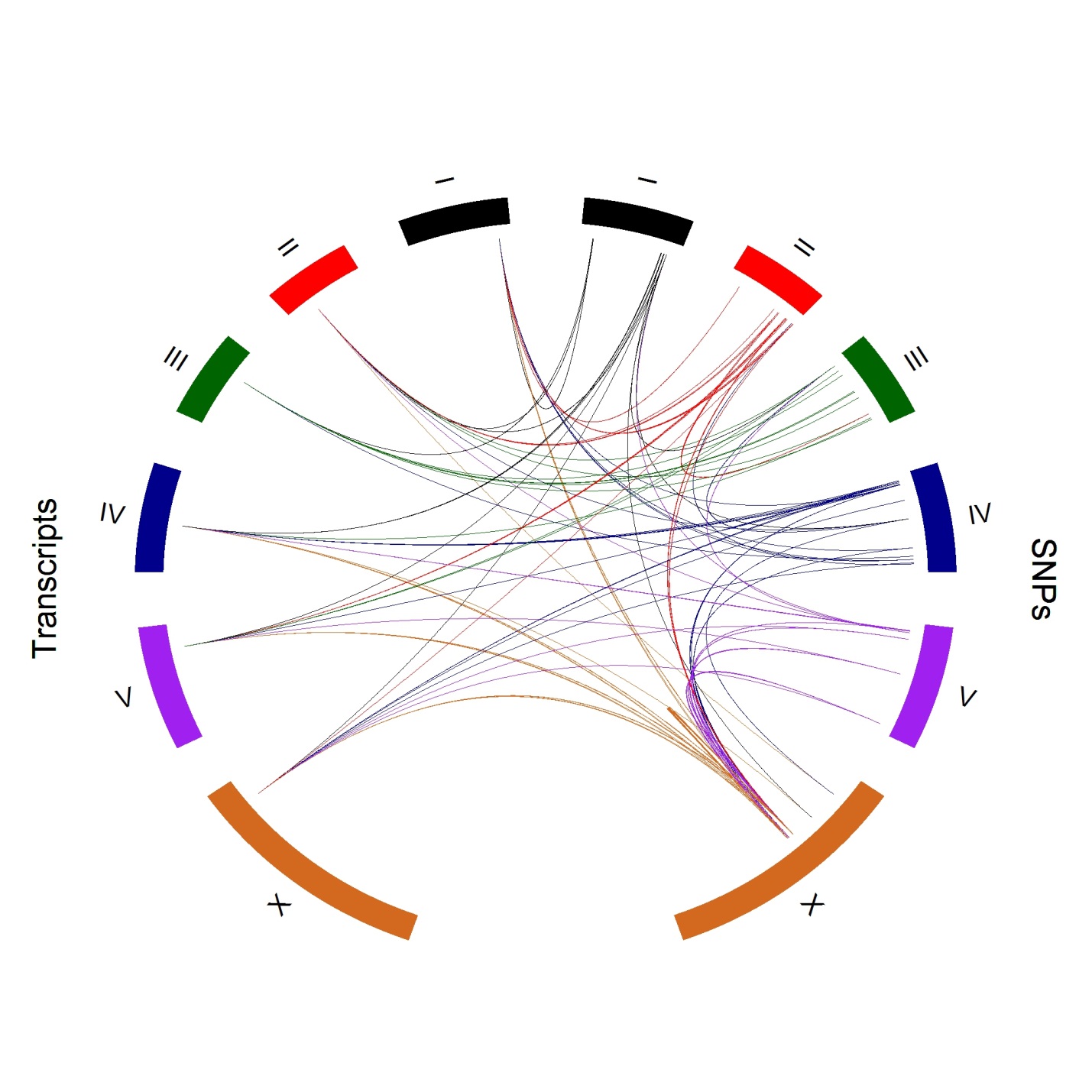
**Figure Legends**

**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. (2010).

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



**Figure 1**



**Figure 2**

Table 1 Results of simulation 1 with for the random error with independent predictors.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Main Effects** | | | | **Interactions Effects** | | | | **Model** | | | |
| **Method** | **Cov** | **Cor0** | **Inc0** | **Exact** | **Cov** | **Cor0** | **Inc0** | **Exact** | **Size** | **MSE** |  | **Time** |
| **Single Marker** | 0 | 1 | 0.25 | 0 | NA | NA | NA | NA | 3 | 23.63 | 0.216 | 0.824 |
| **FS** | 0.85 | 0.953 | 0.0625 | 0.85 | NA | NA | NA | NA | 27 | 10.23 | 0.660 | 3.47 |
| **FS2** | 1 | 0.996 | 0 | 1 | 0.95 | 0.981 | 0 | 0 | 27 | 0.302 | 0.989 | 72.31 |
| **iFORM** | 0.9 | 0.999 | 0.05 | 0.9 | 0.9 | 1 | 0 | 0.9 | 7.55 | 2.93 | 0.894 | 4.08 |
| **Oracle** | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 8 | 1.023 | 0.965 | NA |

Outcomes include the convergence Probability (Cov) , percentage of correct zeros identified (Cor0) , percentage of incorrect zeros identified (Inc0) , the exact selection probability (Exact) , 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 2 with for the random error with independent predictors.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Main Effects** | | | | **Interactions Effects** | | | | **Model** | | | |
| **Method** | **Cov** | **Cor0** | **Inc0** | **Exact** | **Cov** | **Cor0** | **Inc0** | **Exact** | **Size** | **MSE** |  | **Time** |
| **Single Marker** | 0.02 | 0.999 | 0.529 | 0.029 | NA | NA | NA | NA | 1.97 | 27.02 | 0.178 | 0.69 |
| **FS** | 0.8 | 0.953 | 0.05 | 0.8 | NA | NA | NA | NA | 27 | 11.49 | 0.651 | 3.22 |
| **FS2** | 1 | 0.996 | 0 | 1 | 0.98 | 0.98 | 0 | 0 | 27 | 1.17 | 0.964 | 68.2 |
| **iFORM** | 0.97 | 0.998 | 0.007 | 0.97 | 0.95 | 1 | 0 | 0.93 | 8.7 | 4.41 | 0.865 | 3.84 |
| **Oracle** | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 8 | 3.92 | 0.880 | NA |

Outcomes include the convergence Probability (Cov) , percentage of correct zeros identified (Cor0) , percentage of incorrect zeros identified (Inc0) , the exact selection probability (Exact) , average model size, Mean Square Error for the model (MSE), the adjusted R-square of the model, and the computational time in seconds.

Table 3 Results of simulation 3 with for the random error with independent predictors.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Main Effects** | | | | **Interactions Effects** | | | | **Model** | | | |
| **Method** | **Cov** | **Cor0** | **Inc0** | **Exact** | **Cov** | **Cor0** | **Inc0** | **Exact** | **Size** | **MSE** |  | **Time** |
| **Single Marker** | 0 | 0.999 | 0.612 | 0 | NA | NA | NA | NA | 1.65 | 33.90 | 0.138 | 0.69 |
| **FS** | 0.82 | 0.953 | 0.043 | 0.827 | NA | NA | NA | NA | 27 | 14.44 | 0.633 | 3.22 |
| **FS2** | 1 | 0.997 | 0 | 1 | 0.98 | 0.97 | 0 | 0 | 27 | 2.69 | 0.931 | 68.2 |
| **iFORM** | 0.89 | 0.995 | 0.060 | 0.896 | 0.96 | 1 | 0 | 0.94 | 7.93 | 11.10 | 0.713 | 3.83 |
| **Oracle** | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 8 | 8.98 | 0.771 | NA |

Outcomes include the convergence Probability (Cov) , percentage of correct zeros identified (Cor0) , percentage of incorrect zeros identified (Inc0) , the exact selection probability (Exact) , average model size, Mean Square Error for the model (MSE), the adjusted R-square of the model, and the computational time in seconds.

Table 4 Estimated main and epistatic effects of eQTLs by iFORM on gene transcript A\_12\_P103290 on chromosome I , in a comparison with the result by traditional single marker analysis.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **iFORM** | | | | | **Single Marker Analysis** | | |
| **eQTL** | **Effect** | | **SE** | ***p*-value** | **Herit-ability(%)** | **Effect** | **SE** | ***p*-**  **value** |
| X1\_2068168 (cis-eQTL) | | -0.197 | 0.035 | 0.000 | 0.060 | -0.210 | 0.074 | 0.005 |
| X2\_13516256 | | -0.069 | 0.039 | 0.080 | 0.007 | -0.138 | 0.075 | 0.068 |
| X2\_13813025 | | -0.023 | 0.052 | 0.665 | 0.001 | -0.095 | 0.074 | 0.201 |
| X2\_13694563 | | 0.074 | 0.062 | 0.238 | 0.008 | -0.089 | 0.075 | 0.242 |
| X2\_2482896 | | 0.064 | 0.027 | 0.017 | 0.006 | -0.057 | 0.076 | 0.453 |
| X\_15500580 | | 0.073 | 0.064 | 0.253 | 0.008 | 0.010 | 0.076 | 0.899 |
| X\_14636404 | | -1.768 | 0.092 | 0.000 | 4.794 | 0.024 | 0.076 | 0.751 |
| X4\_16403215 | | 0.028 | 0.040 | 0.489 | 0.001 | 0.087 | 0.075 | 0.252 |
| X4\_15568674 | | -1.972 | 0.134 | 0.000 | 5.964 | 0.094 | 0.075 | 0.213 |
| X4\_1873297 | | 0.044 | 0.026 | 0.086 | 0.003 | 0.095 | 0.071 | 0.185 |
| X4\_15632637 | | 1.960 | 0.143 | 0.000 | 5.892 | 0.104 | 0.075 | 0.169 |
| X4\_13532205 | | 0.064 | 0.028 | 0.024 | 0.006 | 0.111 | 0.075 | 0.143 |
| X\_15820520 | | -0.014 | 0.055 | 0.796 | 0.000 | 0.146 | 0.075 | 0.054 |
| X\_14542103 | | 1.786 | 0.087 | 0.000 | 4.892 | 0.162 | 0.075 | 0.031 |
| X2\_13516256.X4\_15632637 | | -3.799 | 0.268 | 0.000 | 5.534 | NA | NA | NA |
| X2\_13516256.X4\_15568674 | | 3.753 | 0.276 | 0.000 | 5.401 | NA | NA | NA |
| X\_15820520.X\_14636404 | | -3.771 | 0.172 | 0.000 | 5.453 | NA | NA | NA |
| X\_15820520.X\_14542103 | | 3.691 | 0.172 | 0.000 | 5.224 | NA | NA | NA |
| X\_14636404.X4\_1873297 | | -3.534 | 0.163 | 0.000 | 4.789 | NA | NA | NA |
| X\_14636404.X4\_13532205 | | -3.567 | 0.166 | 0.000 | 4.879 | NA | NA | NA |
| X\_14542103.X4\_1873297 | | 3.629 | 0.164 | 0.000 | 5.050 | NA | NA | NA |
| X\_14542103.X4\_13532205 | | 3.469 | 0.167 | 0.000 | 4.614 | NA | NA | NA |

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

|  |  |  |
| --- | --- | --- |
| **eQTL Type** | **Count** | **Proportion** |
| **Cis-eQTL** | 14 | 0.0024 |
| **Trans-eQTL** | 5509 | 0.9628 |
| **Cis-eQTL x cis-eQTL** | 0 | 0.0000 |
| **Trans-eQTL x cis-eQTL or cis-eQTL x trans-eQTL** | 2 | 0.0003 |
| **Cis-eQTL x trans-eQTL** | 196 | 0.0340 |

A random sample of 1000 transcripts was used as a response and the iFORM selection procedure was implemented on each. Table 5 shows the number of predictors selected during the procedure and the eQTL type of each of the predictors in relation to the transcript used as the response.