Detection of Quantitative Trait Associated Genes Using Cluster Analysis

Zhenyu Jia[1,⋆], Sha Tang[2], Dan Mercola[1], and Shizhong Xu[3]

[1] Department of Pathology and Laboratory Medicine, University of California, Irvine, CA 92697, USA
{zjia,dmercola}@uci.edu
[2] Department of Pediatrics, University of California, Irvine, CA 92697, USA
shat@uci.edu
[3] Department of Botany and Plant Sciences, University of California, Riverside, Riverside, CA 92521, USA
shxu@ucr.edu

Abstract. Many efforts have been involved in association study of quantitative phenotypes and expressed genes. The key issue is how to efficiently identify phenotype-associated genes using appropriate methods. The limitations for the existing approaches are discussed. We propose a hierarchical mixture model in which the relationship between gene expressions and phenotypic values is described using orthogonal polynomials. Gene specific coefficient, which reflects the strength of association, is assumed to be sampled from a mixture of two normal distributions. The association status for a gene is determined based on which distribution the gene specific coefficient is sampled from. The statistical inferences are made via the posterior mean drawn from a Markov Chain Monte Carlo sample. The new method outperforms the existing methods in simulated study as well as the analysis of a mice data generated for obesity research.

Keywords: Gibbs sampler, Microarray.

1 Introduction

Microarray technology allows us to measure the expression levels of many thousands of genes simultaneously. The objective of microarray experiments is to closely examine the changes of gene expression under different experimental conditions. These conditions may simply be control and various treatments [1], or may represent different time slots after a certain treatment is applied to the experimental subjects [2], or may refer to measurements of quantitative phenotype for different subjects [3]. Many statistical methodologies have been proposed to analyze data generated from microarray experiments. Fundamental microarray data analyses aim to identify a list of genes as being differentially expressed across experimental conditions [4,5,6]. Recent methods, such as various cluster analyses, have been devised not to find individual genes but to search for groups

⋆ To whom all correspondence should be addressed.

E. Marchiori and J.H. Moore (Eds.): EvoBIO 2008, LNCS 4973, pp. 83–94, 2008.

of genes that are functionally related [7,8,9,10]. However, many existing cluster analyses only use expression data, which requires extra steps to infer the functions of genes that form a cluster. Incorporating biological information [11] or phenotypic information [12] into cluster analyses seems to be more efficient and reasonable.

Efforts have been provided to uncover genes that affect the phenotype of interest. For example, [3] conducted an experiment to study the relationship of gene expression and Alzheimer's disease. The Pearson's correlation was calculated for each gene with the phenotypic values separately. Genes were declared to be disease-associated if their correlation coefficients are statistically significant. Similarly, [13] used the Pearson's correlation analysis to study the relationship of gene expression and mouse weight. First, the highly correlated genes were clustered into the same group which was called "module". Next, they assessed the physiological relevance of each module by examining the overall correlation of the module genes with the phenotype. The genes within a significant module were claimed to be associated with the phenotype.

As suggested by [12], the Pearson's correlation analysis may not be optimal for two reasons: (1) Genes are not jointly analyzed leading to a poor information sharing across genes. (2) A significant correlation is not always biologically meaningful unless the regression is also high. They proposed a mixed model in which the gene expression levels are linearly regressed on the phenotypic values. The regression coefficients, which reflect the affiliation of the genes with the phenotype, are used to cluster genes into a number of functional groups. A cluster is claimed to be significant if the regression coefficient of the mean expression profile is not equal to zero; otherwise, the cluster is claimed as neutral. Genes that have been assigned into non-neutral clusters are target genes. Because the mixed model of [12] is solely built upon the assumption of linear association, it is limited to pick up genes that are associated with the phenotype in a non-linear manner. In order to solve this problem, [14] developed a mixed model to cluster genes based on the non-linear association using orthogonal polynomials. For these two model-based analyses, the optimal number of clusters is not known and needs to be determined by comparing the BIC [15] values for different models. To our experience, this often requires credible evaluations for at least 10 models with distinct dimensionality of parameters which is defined by the number of clusters. It would be more computationally intensive if non-linear association is considered due to the complex nature of microarray data.

In current study, we developed a Bayesian hierarchical model to cluster genes with fixed number of clusters. The non-linear relationship between gene and the phenotype is also described using orthogonal polynomials. The orthogonal polynomials can be constructed as described by [16]. For each gene, the coefficient of each polynomial is assumed to be sampled from a two-components mixture Normal distribution. Both Normal components have mean zero but different variances, i.e., one has a very small variance while the other has a larger variance. If the corresponding coefficient is sampled from the component with small variance, the coefficient is enforced to be zero; otherwise, the coefficient is non-

trivial and its magnitude should be estimated from data. That is to say each gene may be assigned into one of two clusters based on whether it is associated with the polynomial. Suppose that there are p polynomials in the expression model. Therefore, there are a total of 2^p patterns or clusters to illustrate genes under study. Once p is chosen, the number of clusters is immediately determined, which circumvents the model evaluations required by the aforementioned methods.

2 Methods

2.1 Hierarchical Linear Model

Let m and N be the number of genes and the number of subjects under study, respectively. Let Z be the measurements of a quantitative phenotype collected from N subjects. The expression levels of gene i across N subjects can be described in the following model:

$$Y_i(Z) = \alpha_i + \beta_i(Z) + \epsilon_i, \tag{1}$$

where $i = 1, \ldots, m$. In the model 1, $Y_i(Z)$ is a $N \times 1$ matrix, α_i represents the gene specific intercept, $\beta_i(Z)$ is an arbitrary function chosen to describe the relationship between the gene expressions and the phenotypic values, and ϵ_i is used to model the random error with assumed $\mathrm{N}(\mathbf{0}, I\sigma^2)$ distribution.

There are different ways to choose function $\beta_i(Z)$. In current study, we only consider the orthogonal polynomials [16], such that, $\beta_i(Z)$ can be expressed as:

$$\beta_i(Z) = X\beta_i = \sum_{j=1}^{p} X_j \beta_{ij},$$

where p is the degree of orthogonal polynomials after transformation, Z is transformed into a $N \times p$ matrix which is denoted by $X = (X_1, \ldots, X_p)$, and $\beta_i = (\beta_{i1}, \ldots, \beta_{ip})$ represents the corresponding coefficients for gene i. Then, model 1 can be rewritten as:

$$Y_i = \alpha_i + \sum_{j=1}^{p} X_j \beta_{ij} + \epsilon_i. \tag{2}$$

Using a linear contrasting scheme (see [14]), model 2 can be further written as

$$y_i = \sum_{j=1}^{p} x_j \beta_{ij} + \varepsilon_i, \tag{3}$$

where $\sum_{k=1}^{N} y_{ik} = 0$ and $\sum_{k=1}^{N} x_{jk} = 0$. In fact, we do not have N pieces of independent information for each gene after linear contrasting. Therefore, the last element of vector y_i should be removed and y_i becomes an $n \times 1$ vector for $n = N - 1$. Accordingly, x_j becomes an $n \times 1$ vector, and ε_i is now $\mathrm{N}(\mathbf{0}, R\sigma^2)$ distributed with a known $n \times n$ positive definite matrix R (see [14]).

For the sake of convenience for presentation, we use the following notation to express different probability densities throughout current study:

$$p(\text{variable}|\text{parameter list}) = \text{DensityName}(\text{variable}; \text{parameter list}).$$

For example, the probability of y_i given all the β_{ij} variables and σ^2 is described as

$$p\left(y_i \left| \sum_{j=1}^{p} x_j \beta_{ij}, R\sigma^2 \right.\right) = \text{Normal}\left(y_i; \sum_{j=1}^{p} x_j \beta_{ij}, R\sigma^2\right). \tag{4}$$

The model 3 is the lowest level in the hierarchical structure, which is governed by higher parameters, such as regression coefficients (β_{ij}) and the residual variance (σ^2). These parameters themselves are controlled by assumed higher distributions. In this study, we assign a mixture distribution to β_{ij} as originally suggested by [17],

$$p(\beta_{ij}|\eta_{ij}, \sigma_j^2) = (1 - \eta_{ij})\text{Normal}(\beta_{ij}; 0, \delta) + \eta_{ij}\text{Normal}(\beta_{ij}; 0, \sigma_j^2) \tag{5}$$

where $\delta = 10^{-4}$ (a small positive number) and σ_j^2 is an unknown variance assigned to the jth polynomial. Variable $\eta_{ij} = \{0, 1\}$ is used to indicate whether β_{ij} is sampled from a $\text{N}(0, \delta)$ or a $\text{N}(0, \sigma_j^2)$ distribution. If it comes from the first normal distribution, β_{ij} is virtually fixed at zero; otherwise, β_{ij} has a non-trivial value and should be estimated from the data. Therefore, $\eta_{ij} = 1$ means that $\beta_{ij} \neq 0$ and gene i is associated with the jth polynomial. The hierarchical level of density 5 is regulated by η_{ij} and σ_j^2. We further describe η_{ij} by a Bernoulii distribution with probability ρ_j, denoted by

$$p(\eta_{ij}|\rho_j) = \text{Bernoulii}(\eta_{ij}; \rho_j). \tag{6}$$

The parameter ρ_j will control the proportion of the genes that are associated with the jth polynomial. Because of the hierarchical nature, we may further describe ρ_j by a Dirichlet distribution, denoted by $\text{Dirichlet}(\rho_j; 1, 1)$. The variance components of the hierarchical model are assigned scaled inverse chi-square distributions, denoted by $\text{Inv} - \chi^2(\sigma_j^2; d_0, \omega_0)$. We choose $d_0 = 5$ and $\omega_0 = 50$ for σ_j^2, and choose $d_0 = 0$ and $\omega_0 = 0$ for σ^2.

2.2 Markov Chain Monte Carlo

The typical technique for inferring the posterior distributions of the parameters is to use MCMC sampling since the posterior distributions are intractable. We draw a posterior sample from which empirical posterior means of interested parameters can be found. First, we choose initial values for parameter θ, where $\theta = (\sigma^2, \sigma_1^2, \ldots, \sigma_p^2, \rho)$. We then derive the distribution of one parameter conditional on the data and values of all other variables, i.e., $p(\theta_k|\text{data}, \theta_{-k})$, where θ_k is current parameter of interest and θ_{-k} is the list of remaining variables. This distribution usually has a simple form from which a value for θ_k can be sampled.

The parameter θ_k is then updated using the realized value, and it will be used as known parameter to update all other parameters in the same manner. The detailed sampling scheme for each variable is described as follows.

(1) Variable η_{ij} is simulated from Bernoulli$(\eta_{ij}; \pi_{ij})$, where

$$\pi_{ij} = \frac{\rho_j \mathrm{N}(\gamma_{ij}; 0, \sigma_j^2)}{\rho_j \mathrm{N}(\gamma_{ij}; 0, \sigma_j^2) + (1-\rho_j)\mathrm{N}(\gamma_{ij}; 0, \delta)} \tag{7}$$

(2) Variable β_{ij} is simulated from $\mathrm{N}(\beta_{ij}; \mu_\beta, \sigma_\beta^2)$, where

$$\mu_\beta = \left[x_j^T R^{-1} x_j + \frac{\sigma^2}{\eta_{ij}\sigma_j^2 + (1-\eta_{ij})\delta} \right]^{-1} x_j^T R^{-1} \triangle y_i, \tag{8}$$

$$\sigma_\beta^2 = \left[x_j^T R^{-1} x_j + \frac{\sigma^2}{\eta_{ij}\sigma_j^2 + (1-\eta_{ij})\delta} \right]^{-1} \sigma^2 \tag{9}$$

and

$$\triangle y_i = y_i - \sum_{j' \neq j}^{p} x_{j'} \beta_{ij'} \tag{10}$$

which is called the offset of y_i adjusted for the jth polynomial effect.

(3) Sample σ_j^2 from

$$\mathrm{Inv} - \chi^2 \left(\sigma_j^2; \sum_{i=1}^{m} \eta_{ij} + 5, \sum_{i=1}^{m} \eta_{ij}\beta_{ij}^2 + 50 \right).$$

(4) Sample σ^2 from

$$\mathrm{Inv} - \chi^2 \left(\sigma^2; mn, \sum_{i=1}^{m} (y_i - \sum_{j=1}^{p} x_j \beta_{ij})^T R^{-1} (y_i - \sum_{j=1}^{p} x_j \beta_{ij}) \right).$$

(5) Simulate ρ_j from

$$\mathrm{Dirichlet} \left(\rho_j; \sum_{i=1}^{m} \eta_{ij} + 1, m - \sum_{i=1}^{m} \eta_{ij} + 1 \right).$$

So far, every variable has been updated. Once every variable is updated, we complete one iteration or sweep. The sampling process continues until the Markov chain reaches its stationary distribution. The length of the chain required for convergence can be determined by the R package "coda" [18]. We discard a number of iterations from the beginning of the chain, which is so-called burn-in period. For the remaining portion of the chain, we save one observation in every 10 sweeps to form a posterior sample until the sample is sufficiently large to allow an accurate estimate of the posterior mean for each variable. Let

$\bar{\eta}_{ij} = N_p^{-1} \sum_{l=1}^{N_p} \eta_{ij}^{(l)}$ be the posterior mean of variable η_{ij}, where N_p is the posterior sample size. Gene i is said to be associated with the jth polynomial if $\bar{\eta}_{ij}$ is greater than some pre-specified threshold. We use 0.8 as such cutoff point throughout the current study since [19] showed that 0.8 was quite sufficient to achieve the false discovery rate (FDR) control at $\leq 1\%$ level in the similar analysis.

3 Implementation

3.1 Simulation Study

In the simulation study, a total of 20 datasets were simulated independently. For each dataset, expression levels of 1000 genes were simulated for 50 subjects. The phenotypic value for each subject was randomly selected from $\mathrm{U}(0, 10)$. The 50×1 phenotype matrix was then transformed into 50×3 orthogonal polynomials matrix with degree 3. The corresponding 3×1 regression coefficient matrix for each gene was generated as follows. For genes 1 to 5 and genes 21 to 35, the coefficients for the polynomial of the first order were simulated from $\mathrm{N}(0, 3^2)$. For genes 6 to 10, genes 16 to 25, and genes 31 to 35, the coefficients for the polynomial of the second order were simulated from $\mathrm{N}(0, 1^2)$. For genes 11 to 20 and genes 26 to 35, the coefficients for the polynomial of the third order were simulated from $\mathrm{N}(0, 0.5^2)$. In current study, we define a gene-polynomial association as a linkage. Thus, a total of 60 linkages were generated in the simulation study. Such set up made the 1000 genes fall into $2^3 = 8$ binary-based categories, which were represented by (0 0 0), (1 0 0), (0 1 0), (0 0 1), (0 1 1), (1 1 0), (1 0 1) and (1 1 1), respectively. For example, gene 1 can be regarded as a member of the cluster (1 0 0) since it was only associated with the polynomial of the first order; while gene 35 belonged to the cluster (1 1 1) because the coefficients for all three polynomials are non-trivial. Only the first 35 genes were associated with phenotype while the majority of the genes were placed in the neutral cluster represented by (0 0 0). The residual error for each gene was sampled from $\mathrm{N}(0, 0.4^2)$. The aim of our analysis is to detect genes represented by significant linkages with the phenotypic polynomials.

We used the new method to analyze the 20 simulated datasets separately. The results summarized from 20 analyses are presented in Table 1. The estimated parameters agreed with the true values very well. Due to the small sample size

Table 1. True and estimated values by the new method for the parameters used in simulation study

	Parameter						
	ρ_1	ρ_2	ρ_3	σ_1^2	σ_1^2	σ_1^2	σ^2
True	0.020	0.020	0.020	9.00	1.00	0.25	0.160
Estimate	0.021	0.019	0.017	10.72	3.18	2.93	0.160

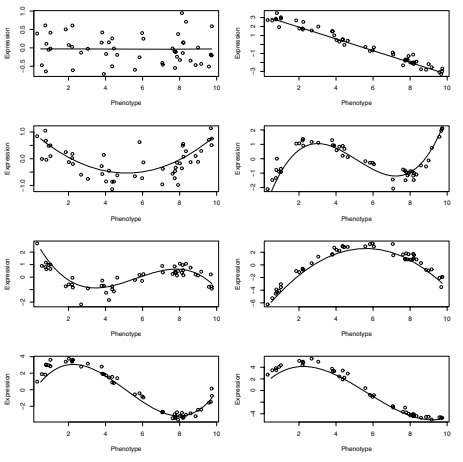

Fig. 1. Plots of the typical genes selected from each of 8 clusters for the analysis of one simulated dataset

Table 2. Comparison between two methods based on the percentages of true genes identified by each of them. Dataset one is the one simulated in current study and dataset two is that simulated in [14]. Method I is the proposed method and method II is the method of [14].

	Dataset	
Method	One	Two
I	88.57%	100.0%
II	62.86%	98.38%

(≤ 30) for each polynomial, the estimated σ_j^2 showed some deviations from the true values. Figure 1 gives the plots of the typical genes selected from each of 8 clusters for the analysis of one simulated dataset. The expression pattern of each gene across phenotypic values is satisfactorily depicted with a regression curve approximated by the new method. We also used the method of [14] to analyze the same dataset. The optimal BIC occurred when the number of clusters was set to 7. Because two methods use different criterions to cluster genes, that is the method of [14] classifies genes based on their mean expression pattern across the phenotypic polynomials; while the proposed method clusters genes based on whether they are significantly associated with the phenotypic polynomials. Thus, we compared two methods by checking the proportions of true associated genes that have been successfully identified by each of them. Note that the numbers of falsely identified genes were zero for both methods. The results are listed in Table 2, from where we can see that the new method identified more true genes than the method of [14]. We examined all 7 true linkages missed by our analysis. The average of the absolute effects for these 7 linkages was 0.08, which is too

small to be detectable with reasonable analysis methods. We understand that the better performance of the new method may result from simulation scheme which could be biased to the new method. To eliminate this nuisance factor, we also analyzed the dataset simulated in [14] using two methods. From Table 2, we can see that the new method outperformed the method of [14] again.

3.2 Analysis of Mice Data

To demonstrate the new method, we analyzed a mice data collected for obesity study by [20]. The data are publicly available at gene expression omnibus (GEO)

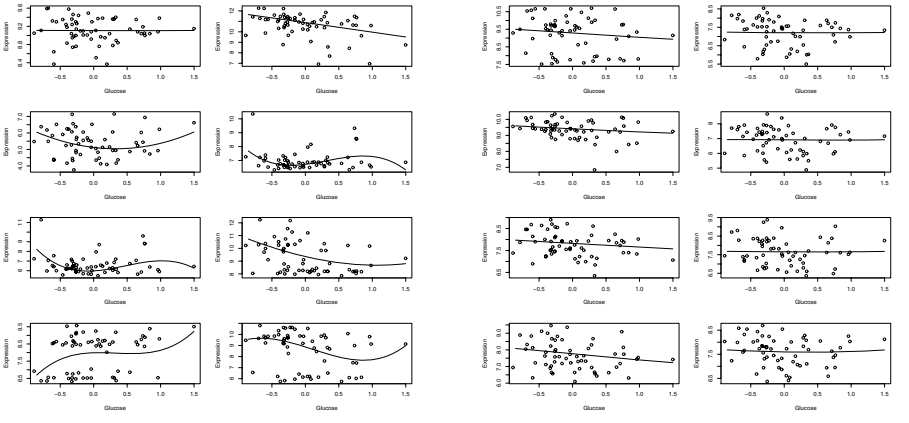

Fig. 2. (a) Plots of the typical transcripts selected from each of the 8 clusters for the analysis of glucose-expression associations for mice data. (b) Scatter plots of selected transcripts identified by one method by missed by the other method for the analysis of glucose-expression associations for mice data. Four transcripts on the left panel are those only detected by the proposed method; while the other four on the right panel are transcripts solely detected by the method of [14].

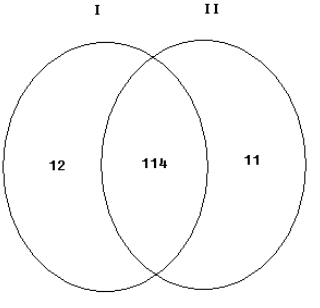

Fig. 3. Transcripts detected by the two methods for the analysis of glucose-expression associations for mice data. Method I is the proposed method and method II is the method of [14].

with accession no. GSE3330. In their experiment, a total of 60 ob/ob mice were examined. For each mouse, the expression levels of over 40,000 transcripts and 25 obesity related phenotypes were measured. Since the expression levels of most transcripts are constant across 60 mice and they do not provide any information, we eliminated those non-variant transcripts prior to the analysis to lessen the computation burden. We sorted all transcripts by their variances across 60 individuals and deleted the transcripts with variances less than 0.05, leaving 5185 most varying transcripts for further analysis. Similar pre-screening scheme has been used for array data analyses [13,19]. In current study, we only investigated the association between gene expression and plasma glucose level (mg/dl). The phenotypic data were collected at eight weeks of age. A total of 126 transcripts were detected to be associated with the glucose level. The typical transcripts selected from each of the 8 clusters are presented in Figure 2(a). We also used the method of [14] to analyze the expression-glucose data. The optimal BIC occurred when the number of clusters was 2, which might not be sound. More distinct clusters were expected due to the complexity of array data. The BIC value kept going down as the number of clusters increased, though the differences of analytic results from different models were trivial. In this case we chose the number of clusters as 8 to achieve the parallel between two methods. A total of 125 transcripts were identified. From Figure 3, we can see that both methods detected 114 common transcripts. We checked all the transcripts that have been detected by one method but missed by the other one. The left four transcripts in Figure 2(b) were detected by the new method but missed by the method of [14]. These four transcripts had slight slopes which should be accounted. The four transcripts on the right panel were only detected by the method of [14]. We could not see any regressions between the expression levels and the phenotypic values. It seemed that the new method had more power than the method of [14]; on the other hand, the new method was subject to lower type I error than the method of [14].

4 Discussion

The purpose of current study is to introduce a more sensitive and convenient approach for association study on gene expressions and quantitative traits. Similar to the existing method of [14], the new method is also based on the non-linear relationship assumption and is realized via orthogonal polynomial transformation. The differences are: (1) the method of [14] organize genes based on their mean expression patterns across phenotypic values; while the new method clusters genes by examine their associations with the polynomials of phenotype. (2) in method of [14], extra model evaluations are needed to find the optimal number of clusters; however, this is not necessary for new method, where the number of clusters is always fixed. In the analysis of [14], the significant tests are performed on the coefficients of the mean expressions for genes. In such the

case, the coefficients for all polynomials are jointly considered, which sometimes leads to loss of power. Suppose that, for a gene, the coefficients for the second and the third order are trivial, while the coefficient for the first order is somewhat significant. This gene may not be detected when the overall significance is considered. In the new method, this would not happen since we test the association of gene expression with each polynomial individually. We can sharpen the prior (δ) to make the analysis sensitive to a satisfactory extent. A gene that is significantly linked to any of the polynomials will be picked up. We also noticed that all the missed significant genes by the method of [14] are all linearly associated with the phenotype, which means that the genes linked with phenotype with higher orders are relatively easier to be seen. That makes senses because the high-order association tends to show a more obvious pattern than the linear association does. This explains why the genes with slight first order regressions have been overlooked by the method of [14]. For the analysis of [14], we need to compare the BIC values for different models to find the optimal number of clusters, which requires considerable extra effort. Such evaluation failed probably due to the complexity of the array data. We consider this extra computation can be avoided by fixing the number of clusters through a meaningful way. In current study, we classify genes into one of two clusters for each polynomial, that is, cluster contains non-associated genes and cluster contains associated genes. Thus, the association of a gene with the phenotype may be describe by one of 2^p patterns, which makes the new method more efficient in implementation.

As aforementioned in Methods section, different functions can be adopted for $\beta_i(Z)$, which is used to describe the relationship between the gene expression and the phenotype in current study. For example, we may use B-spline transformation instead. Simulation studies indicated that B-spline version is equivalent to orthogonal polynomials version (data not shown). B-spline is a alternative way of constructing a basis for piecewise polynomial; however, it is not a natural method of describing spline. Thus, we prefer using orthogonal polynomials version in current study since the behavior of regression of gene expressions on phenotype can be easily interpreted.

The association study of gene expressions and phenotypes provides with a pilot research for gene network study. For example, we may first identify transcripts that are associated with the phenotypes of the disease. The common genes that have been discovered to be associated with multiple phenotypes may play key roles in disease development. Experiments may be carried out to verify their biological significance. We may treat the validated genes as seed genes and further search for other non-annotated genes that may be functionally connected with these genes. New genes may be identified by checking if their expression levels significantly correlate with that of the seed genes. Or, given the information on the genomic markers, we may map the seed genes as well as the genes with unknown functions jointly using so called eQTL mapping scheme, such as [21,19]. The genes that have been mapped to the same genomic loci with the seed genes are likely to be functionally related to the seed genes and contribute to the disease.

Acknowledgment

This research was supported by the National Institute of Health Grant R01-GM55321 and the National Science Foundation Grant DBI-0345205 to SX, and by the National Institute of Health SPECS Consortium Grant CA 114810-02.

References

1. Dudoit, S., Yang, Y.H., Callow, M.J., Speed, T.P.: Statistical methods for identifying differentially expressed genes in replicated cdna microarray experiments. Stat. Sinic. 12, 111–139 (2002)
2. Saban, M.R., Hellmich, H., Nguyen, N.B., Winston, J., Hammond, T.G., Saban, R.: Time course of lps- induced gene expression in a mouse model of genitourinary inflammation. Physiological Genomics 5, 147–160 (2001)
3. Blalock, E.M., Geddes, J.W., Chen, K.C., Porter, N.M., Markesbery, W.R., Landfield, P.W.: Incipient alzheimer's disease: Microarray correlation analyses reveal major transcriptional and tumor suppressor responses. Proceedings of the National Academy of Sciences of the United States of America 101, 2173–2178 (2004)
4. Tusher, V.G., Tibshirani, R., Chu, G.: Significance analysis of microarrays applied to the ionizing radiation response. Proceedings of the National Academy of Sciences of the United States of America 98, 5116–5121 (2001)
5. Efron, B., Tibshirani, R., Storey, J.D., Tusher, V.: Empirical bayes analysis of a microarray experiment. J. Am. Stat. Assoc. 96, 1151–1160 (2001)
6. Newton, M.A., Noueiry, A., Sarkar, D., Ahlquist, P.: Detecting differential gene expression with a semiparametric hierarchical mixture method. Biostatistics 5, 155–176 (2004)
7. Brown, M.P.S., Grundy, W.N., Lin, D., Cristianini, N., Sugnet, C.W., Furey, T.S., Ares, M., Haussler, D.: Knowledge-based analysis of microarray gene expression data by using support vector machines. Proceedings of the National Academy of Sciences of the United States of America 97, 262–267 (2000)
8. Herrero, J., Valencia, A., Dopazo, J.: A hierarchical unsupervised growing neural network for clustering gene expression patterns. Bioinformatics 17, 126–136 (2001)
9. Yeung, K.Y., Fraley, C., Murua, A., Raftery, A.E., Ruzzo, W.L.: Model-based clustering and data transformations for gene expression data. Bioinformatics 17, 977–987 (2001)
10. Lazzeroni, L., Owen, A.: Plaid models for gene expression data. Statistica Sinica 12, 61–86 (2002)
11. Huang, D.S., Pan, W.: Incorporating biological knowledge into distance-based clustering analysis of microarray gene expression data. Bioinformatics 22, 1259–1268 (2006)
12. Jia, Z., Xu, S.: Clustering expressed genes on the basis of their association with a quantitative phenotype. Genetical Research 86, 193–207 (2005)
13. Ghazalpour, A., Doss, S., Zhang, B., Wang, S., Plaisier, C., Castellanos, R., Brozell, A., Schadt, E.E., Drake, T.A., Lusis, A.J., Horvath, S.: Integrating genetic and network analysis to characterize genes related to mouse weight. Plos Genetics 2, 1182–1192 (2006)
14. Qu, Y., Xu, S.H.: Quantitative trait associated microarray gene expression data analysis. Molecular Biology and Evolution 23, 1558–1573 (2006)
15. Schwartz, G.: Estimating the dimensions of a model. Ann. Stat. 6, 461–464 (1978)

16. Hayes, J.G.: Numerical methods for curve and surface fitting. J. Inst. Math. Appl. 10, 144–152 (1974)
17. George, E.I., Mcculloch, R.E.: Variable selection via gibbs sampling. Journal of the American Statistical Association 88, 881–889 (1993)
18. Raftery, A.E., Lewis, S.M.: One long run with diagnostics: Implementation strategies for markov chain monte carlo. Statistical Science 7, 493–497 (1992)
19. Jia, Z., Xu, S.: Mapping quantitative trait loci for expression abundance. Genetics 176, 611–623 (2007)
20. Lan, H., Chen, M., Flowers, J.B., Yandell, B.S., Stapleton, D.S., Mata, C.M., Mui, E.T., Flowers, M.T., Schueler, K.L., Manly, K.F., Williams, R.W., Kendziorski, C., Attie, A.D.: Combined expression trait correlations and expression quantitative trait locus mapping. Plos Genetics 2, e6 (2006)
21. Schadt, E.E., Monks, S.A., Drake, T.A., Lusis, A.J., Che, N., Colinayo, V., Ruff, T.G., Milligan, S.B., Lamb, J.R., Cavet, G., Linsley, P.S., Mao, M., Stoughton, R.B., Friend, S.H.: Genetics of gene expression surveyed in maize, mouse and man. Nature 422, 297–302 (2003)