Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics and Data Analysis 53 (2009) 1661–1673
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Combining quantitative trait loci analyses and microarray data: An empirical likelihood approach

Dong Wang[a,*], Song Xi Chen[b]

[a] *Department of Statistics, University of Nebraska, Lincoln, NE 68583, United States*
[b] *Department of Statistics, Iowa State University, Ames, IA 50011, United States*

Available online 25 February 2008

Abstract

Selective transcriptional profiling is an attractive approach for alleviating the high cost of genetical genomics research as it requires only a subset of individuals in the QTL mapping study for microarray experiments. Current statistical methods for this approach are based on parametric models that might not be appropriate for all experiments. To provide a nonparametric method for analyzing data obtained in selective transcriptional profiling studies, an empirical-likelihood-based inference is derived for multi-sample comparisons when information is available on surrogate variables. The results show that when testing for the association between the transcriptional abundance of a given gene and a known QTL, using relatively inexpensive trait data on extra individuals significantly improves the power for the proposed test.

1. Introduction

Quantitative trait locus (QTL) mapping is a widely used technique in identifying genomic regions associated with quantitative traits in plants and animals. Traditional quantitative traits include crop yield, grain weight, animal fertility, body fat, disease resistance, and others. As microarray technology becomes widely available, there is great interest in using gene expression levels obtained through microarray experiments as quantitative traits to map expression QTLs (eQTLs, Schadt et al., 2003). Microarray technology enables researchers to obtain a snapshot of the transcriptional state of the whole genome by measuring the transcript abundance of thousands of genes at once. Combining it with QTL mapping provides a powerful approach in deciphering gene regulation networks (genetical genomics Jansen and Nap, 2001). Recent examples using this approach include DeCook et al. (2006), Kiekens et al. (2006), Zhang et al. (2006), Liang et al. (2007), Luo et al. (2007), West et al. (2007), and others.

However, microarray experiments do have the disadvantage of high cost, which can be a serious problem in QTL mapping studies because hundreds of plants or animals may have to be processed in order to attain reasonable power. On the other hand, traditional traits like crop yield or body fat are often relatively inexpensive to obtain, and researchers interested in certain biological pathways have routinely accumulated data regarding specific traits on a large number

* Corresponding author.
E-mail addresses: dwang3@unl.edu (D. Wang), songchen@iastate.edu (S.X. Chen).

0167-9473/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2008.02.023

of animals or plants. If a gene belongs to the pathway modulating a traditional trait, its expression levels are likely to be correlated with that trait of interest. Thus methods that can utilize trait data to enhance the power of eQTL analysis are especially attractive to researchers studying genetic regulation mechanisms concerning a traditional trait.

The selective transcriptional profiling approach seeks to improve the efficiency and affordability of genetical genomics. Suppose a major QTL for a traditional trait like crop yield or body fat has already been identified using a full panel of animals or plants. If, due to financial constraints, microarray experiments cannot be carried out on all individuals in the panel, a subset of individuals can be selected from each genotypic group on which the expression levels are measured using microarrays. By taking into account the QTL and trait information when analyzing the data, researchers can identify genes whose transcript abundance is associated with a quantitative trait of interest through linkage to a trait QTL with far fewer microarrays than the traditional genetical genomics approach. Conversely, one can achieve much greater power in the test for QTL association with the same number of microarrays.

Currently available methods such as the Wald test described in Wang and Nettleton (2006) are based on the assumption that the expression abundance of a given gene and the value of a traditional trait have a bivariate normal distribution. Also it is assumed that the covariance matrix is the same for individuals in different genotypic groups. Though these assumptions are reasonable over a wide range of experiments (perhaps after transformation), there are situations where these assumptions may be hard to justify. Thus it is desirable to develop a nonparametric method for analyzing data obtained by selective transcriptional profiling when strict parametric assumptions are questionable.

In this paper we propose an empirical-likelihood-based test for data obtained with the selective transcriptional profiling approach. Empirical likelihood is a nonparametric method of inference introduced by Owen (1988, 1990), which has properties analogous to that of parametric likelihood and has been successfully applied to a wide range of problems. Our proposed method is also related to the information recovery problem in studies with surrogate endpoints. In a surrogate variable study, data are composed of a validation sample and a nonvalidation sample. For the selective transcriptional profiling approach, the validation sample is made up of observations with measurement on the transcriptional abundance (true endpoint), as well as the surrogate or auxiliary variable representing a traditional trait and possibly also some covariate information. The nonvalidation sample is made up of observations with only information on the traditional trait and covariates. The goal is to make inference on some parameter β, which defines certain characteristics regarding the true end point (gene expression) and covariates. In the setting of clinical studies, Pepe (1992), Fleming et al. (1994) and others show that using the nonvalidation sample leads to more precise estimation of β, and the gain in efficiency increases with the size of the nonvalidation sample. Moreover, the method is more efficient when the information provided by the surrogate is highly "correlated" to the true end point.

The rest of this article is organized as follows. In Section 2, we introduce the concept of empirical likelihood and develop an empirical-likelihood-based inference for data obtained through selective transcriptional profiling. In Section 3, we report results from simulation studies, which demonstrate the advantage of the proposed method. Section 4 describes a real data example regarding barley disease resistance and gene expression. Section 5 is general discussion and theoretical derivations are provided in the Appendix.

2. Empirical-likelihood-based tests for selective transcriptional profiling

In this section we propose an empirical likelihood analysis for data obtained using the selective transcriptional profiling approach. Though the focus of this paper is on the nonparametric inference for expression QTL studies utilizing additional trait data, the general results can be applied to other multi-sample problems with auxiliary data.

2.1. Introduction to empirical likelihood

To introduce the concept of empirical likelihood, it is useful to consider the simple case of a mean parameter. This introduction is not technical. For a comprehensive overview on empirical likelihood, see Owen (2002).

Consider independent random variables $X_i, i = 1, \ldots, n$, from a common distribution with mean μ. The empirical likelihood is defined as

$$L(\mu) = \max \prod_{i=1}^{n} p_i,$$

where p_i is the probability weight placed on each observation X_i, and $\sum_i^n p_i = 1$. It is easy to show that without other constraints on p_i, $\prod_{i=1}^n p_i$ is maximized when all p_i's take the value $1/n$. If the interest is to test whether $\mu = \mu_0$, we shall maximize $\prod_{i=1}^n p_i$ under the additional constraint $\sum_{i=1}^n p_i(x_i - \mu_0) = 0$. If μ_0 is the real mean value of X_i, we can expect $L(\mu_0)$ to be close to $(1/n)^n$, i.e., the unconstrained maximum. But if μ_0 is significantly different from the true mean, the constraint $\sum_{i=1}^n p_i(x_i - \mu_0) = 0$ will force the probability weight to deviate from $1/n$, and thus cause the restricted maximum to be smaller. Moreover, Owen (1990) shows that $-2\log\{L(\mu_0)/n^{-n}\}$ has an asymptotic chi-square distribution when μ_0 is the true value of the mean. Thus empirical likelihood provides a nonparametric inference analogous to parametric likelihood, and its application has been extended to other parameter settings. The empirical-likelihood-based test for the mean and some other parameters has been implemented in the *emplik* package in the R software (http://www.r-project.org).

2.2. Empirical likelihood for selective transcriptional profiling

Now we consider the most common eQTL setting. For the sake of simplicity, we only present results for two QTL genotypes, which are directly applicable to back cross, doubled haploid or recombinant inbred lines. Results for cases with more than two genotypes can be similarly derived with necessary modifications. Let $X_i = (X_{i1}, X_{i2})$, $i = 1, \ldots, n_x + m_x$ be independent observations from individuals of genotype x and $Y_j = (Y_{j1}, Y_{j2})$, $j = 1, \ldots, n_y + m_y$ be observations from genotype y. In selective transcriptional profiling, X_{i1} and Y_{i1} are values of the traditional trait (auxiliary information) for the ith individual in genotypic group x and y respectively, while variables X_{i2} and Y_{i2} are the expression levels of a certain gene measured by microarrays for the corresponding plant or animal. Though we formulate our results for the expression of only one gene, the same analysis will be carried out for each of the thousands of genes on the microarray.

Here values for X_{i1}'s (trait) are available for all $n_x + m_x$ observations. Because microarray experiments are carried out on only a subset of individuals due to financial constraints, values for X_{i2}'s (transcriptional abundance) are only available for $i = 1, \ldots, n_x$, i.e., the validation sample. Accordingly $\{X_{(n_x+k)1}\}_{k=1}^{m_x}$ are the nonvalidation sample, for which $X_{(n_x+k)2}$'s are not available. We assume that $n_x/(n_x + m_x) \to c_x$ for a constant $c_x \in (0, 1)$ as $n_x \to \infty$. Similarly for genotype y, $\{Y_j\}_{j=1}^{n_y}$ are the validation sample, and $\{Y_{(n_y+l)1}\}_{l=1}^{m_y}$ are the nonvalidation sample.

Define β_x and β_y as the mean expression levels for individuals in the two genotypic groups respectively, and γ_x and γ_y are the corresponding means of the traditional trait known to be associated with the QTL. Then the information about these parameters is summarized in the following zero-mean estimating functions,

$$U_{xi}(\beta_x) = X_{i2} - \beta_x, \qquad g_{xi}(\gamma_x) = X_{i1} - \gamma_x,$$
$$U_{yj}(\beta_y) = Y_{j2} - \beta_y, \qquad g_{yj}(\gamma_y) = Y_{j1} - \gamma_y.$$

As we shall discuss later, estimating functions can take more general forms to accommodate a wide range of problems.

For the current setting, our interest is to test $H_0 : \beta_x = \beta_y$ vs. $H_a : \beta_x \neq \beta_y$ when auxiliary data (trait) are present. The advantage of using empirical likelihood is that auxiliary information can be easily incorporated to enhance the power of the test, which may be difficult for other nonparametric methods. To formulate empirical likelihood, denote

$$z \in \{x, y\} \quad \text{and} \quad Z \in \{X, Y\}.$$

Let $p_{z1}, \ldots, p_{zn_z}$ be the nonnegative weights placed on the validation samples $\{X_i\}_{i=1}^{n_x}$ or $\{Y_j\}_{j=1}^{n_y}$. Also let $q_{z1}, \ldots, q_{zm_z}$ be the nonnegative weights placed on the nonvalidation samples $\{X_{(n_x+k)1}\}_{k=1}^{m_x}$ or $\{Y_{(n_y+l)1}\}_{l=1}^{m_y}$. The empirical likelihood for parameter vector $(\beta_x, \gamma_x, \beta_y, \gamma_y)$ is

$$L(\beta_x, \gamma_x, \beta_y, \gamma_y) = \max\left(\prod_{i=1}^{n_x} p_{xi} \prod_{k=1}^{m_x} q_{xk} \prod_{j=1}^{n_y} p_{yj} \prod_{l=1}^{m_y} q_{yl}\right) \tag{1}$$

subject to

$$\sum_{i=1}^{n_z} p_{zi} = 1, \tag{2}$$

$$\sum_{k=1}^{m_z} q_{zk} = 1,$$

$$\sum_{i=1}^{n_z} p_{zi} \left(U(Z_i, \beta_z), g(Z_{i1}, \gamma_z)\right)^\tau = 0$$

and

$$\sum_{k=1}^{m_z} q_{zk} g(Z_{(n_z+k)1}, \gamma_z) = 0.$$

Here we use U^τ to denote the transpose of U.

When the null hypothesis is true, we also have the additional constraint

$$\beta_x = \beta_y := \beta. \tag{3}$$

Maximizing the empirical likelihood under constraints (2) and (3) gives the maximum empirical likelihood estimator under H_0, $(\tilde{\beta}, \tilde{\gamma}_x, \tilde{\gamma}_y)$. On the other hand, if we allow the estimators for β_x and β_y to take different values, we can obtain the maximum empirical likelihood estimator without constraint (3),

$$(\hat{\beta}_x, \hat{\gamma}_x, \hat{\beta}_y, \hat{\gamma}_y).$$

For given parameter values, the empirical likelihood in (1) can be easily evaluated using the *emplik* package in R. The unrestricted maximum empirical likelihood estimator can be obtained by searching over the parameter space for $\beta_x, \beta_y, \gamma_x, \gamma_y$ using quasi-Newton type methods. For the maximum empirical likelihood estimator under H_0, the search will be carried out over the restricted parameter space. This can be achieved by combining the R optimization function optim() and the el.test() function for empirical likelihood. Related R code is available from the first author.

2.3. Empirical-likelihood-based hypothesis testing with extra trait data

To test whether individuals with different genotypes have the same mean for gene expression, define

$$\ell(\beta_x, \gamma_x, \beta_y, \gamma_y) = -2\log\{L(\beta_x, \gamma_x, \beta_y, \gamma_y)\}.$$

Also define $\ell(\beta, \gamma_x, \gamma_y) = -2\log\{L(\beta, \gamma_x, \beta, \gamma_y)\}$ when the null hypothesis is true. The log empirical likelihood ratio statistic for testing $H_0 : \beta_x = \beta_y$ is

$$\mathcal{R}_0 = \ell(\tilde{\beta}, \tilde{\gamma}_x, \tilde{\gamma}_y) - \ell(\hat{\beta}_x, \hat{\gamma}_x, \hat{\beta}_y, \hat{\gamma}_y).$$

The following theorem is a nonparametric version of Wilks' theorem for the empirical likelihood ratio statistic.

Theorem 1. *Under the conditions given in the Appendix and* $\beta_x = \beta_y$, $\mathcal{R}_0 \xrightarrow{d} \chi^2_p$ *as* $\min(n_z, m_z) \to \infty$, $z \in \{x, y\}$.

This theorem can be used to provide calibration for the empirical likelihood ratio statistic when testing $H_0 : \beta_x = \beta_y$. Note that Wilks' theorem has been derived for the empirical likelihood ratio statistic in a wide range of problems. Our results suggest that this property of empirical likelihood also applies to data with surrogate variables.

To appreciate the fact that the power of the empirical-likelihood-based test is enhanced through utilizing trait data on additional individuals, define the covariance matrix of the trait value and the expression level for each individual as

$$\mathrm{Var}(Z_i) = \begin{bmatrix} \sigma_{z11} & \sigma_{z12} \\ \sigma_{z12} & \sigma_{z22} \end{bmatrix}.$$

Consider the local alternative (assuming $n_z = n_x = n_y$), $\beta_y - \beta_x = n_z^{-1/2} u$ with $0 < u < \infty$. The asymptotic power of the test using extra trait data is given by $P(\chi^2_1(\lambda) > \chi^2_{1,1-\alpha})$, where the noncentrality parameter for the noncentral chi-square distribution is

$$\lambda = \frac{u^2}{\sum_z \{\sigma_{z22} - (1 - c_z)\frac{\sigma^2_{z12}}{\sigma_{z11}}\}}.$$

When the nonvalidation data (extra trait data) are not used and the empirical likelihood is based only on individuals with expression data, the noncentrality parameter in the asymptotic power function becomes

$$\check{\lambda} = \frac{u^2}{\sum_z \sigma_{z22}}.$$

Clearly we achieve better power for the test of association between transcriptional abundance and the QTL of interest by including the extra trait data (nonvalidation data). The improvement in power is higher when c_z is smaller, that is, when the nonvalidation sample is large, and expression level is highly correlated with the trait. This is also the case in the parametric setting when bivariate normal distribution is assumed for the expression abundance and trait value.

2.4. Generalization to other settings

To this point, the discussion is limited to the most simple eQTL mapping problem. To accommodate more complex problems, we note that $U_{zi}(\beta_z)$ and $g_{zi}(\gamma_z)$ can be any zero-mean estimating functions satisfying the conditions given in the Appendix. Since many commonly used statistical models can be formulated using estimating functions, the result presented in Theorem 1 is also applicable to problems involving functional trait, categorical trait, or ordinal trait values (e.g., Hackett and Weller (1995), Lange and Whittaker (2001) and Ma et al. (2002)). Moreover, the auxiliary variable X_{i1} or Y_{i1} can also incorporate covariate information for each individual in addition to the trait value, thus accommodate different experimental designs.

For cases with more than two QTL genotypes, such as in F2 populations, results analogous to Theorem 1 can be obtained in a straightforward fashion. The general idea of using extra trait values to enhance the power of test still applies.

The performance of the empirical-likelihood-based method for finite sample size is studied with simulations presented in Section 3.

3. Simulation study

First we suppose that the expression level of a certain gene and the trait value known to be associated with a QTL have a bivariate normal distribution. There are 100 individuals with QTL genotype x and y respectively. The trait means are $\gamma_x = 0$ and $\gamma_y = 1$ for genotypes x and y. Suppose $\sigma_{z11} = \sigma_{z22} = 1$ and $\sigma_{z12} = .85$ for both genotypes. In this notation σ_{z11} is the variance for the trait value for individuals in genotypic group z, and σ_{z22} is the variance of expression abundance for genotype z individuals. Here we suppose that the two genotypic groups share the same covariance matrix. The trait value Z_{i1} is supposed to be known for all individuals, while the expression level as measured by microarray experiments are available on 30 randomly selected individuals in each genotype. By the simulation results not shown here, the value of γ_z does not appear to affect the performance of the test, while the test is more powerful if σ_{z11} or σ_{z22} is smaller.

Here we examine the performance of the empirical-likelihood-based test using both expression data on selected individuals and trait data on all individuals, and compare it with that of the empirical-likelihood-based test using only the expression data from 30 individuals in each genotypic group. For comparison with other nonparametric methods, we also carry out analysis using the Wilcoxon rank-sum test on selected individuals with expression data. In addition, we present results obtained using the Wald test described in Wang and Nettleton (2006), which uses the expression data on selected individuals as well as the trait data on all individuals. Note in this case all the parametric assumptions for the Wald test are satisfied. For a given difference in means $\beta_y - \beta_x$, ranging from 0 to 1 unit in increments of .25 units, 1000 independent samples were generated. The results are summarized in Table 1.

It can be seen from Table 1 that the Wald test is the most powerful of these four tests, which is expected as the parametric model assumptions for the Wald test are satisfied. But it is notable that the empirical-likelihood-based test with a surrogate variable (trait value) can achieve power that is only slightly lower than that of the Wald test under conditions most favorable to the parametric method. It is also obvious that using the trait value in the empirical likelihood inference leads to far superior performance than that of the empirical-likelihood-based test using only the expression data, which is in itself more powerful than the Wilcoxon rank-sum test. This confirms that the principle of using extra trait data to improve power also applies to our nonparametric method.

Table 1
The type I error rate and power of the Wald test, the empirical-likelihood(EL)-based test using both trait and expression data, the empirical-likelihood-based test and the Wilcoxon rank-sum test using only expression values from 30 randomly selected individuals out of 100 in each genotypic class

$\beta_y - \beta_x$	Wald	EL (trait & expression)	EL (expression)	Wilcoxon
0	.005	.008	.014	.012
	.048	.047	.063	.057
	.090	.099	.111	.102
.25	.121	.122	.064	.044
	.289	.280	.182	.179
	.392	.390	.258	.233
.50	.592	.590	.283	.249
	.820	.809	.549	.515
	.884	.878	.669	.618
.75	.944	.938	.643	.591
	.986	.983	.839	.800
	.996	.995	.913	.876
1.00	.997	.995	.881	.845
	1.000	1.000	.984	.969
	1.000	1.000	.984	.974

Within each genotypic group, expression level and trait are assumed to have a bivariate normal distribution. The type I error rate and power are reported for three test sizes: .01, .05 and .10, and results appear in that order.

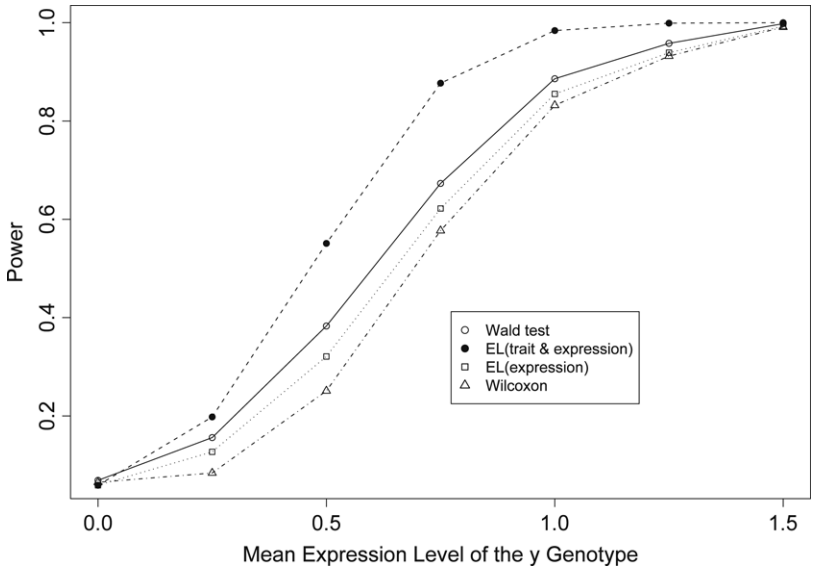

Fig. 1. The power of the Wald test, the empirical-likelihood-based test using both trait and expression data, the empirical-likelihood-based test and the Wilcoxon rank-sum test using only expression data for simulations using the same parameter setting of Table 3 but with a wider range of mean difference. The size of the test is.05. The expression level for individuals with genotype x is fixed at 0. The mean for individuals with genotype y varies from 0 to 1.5 in increments of .25.

Obviously, the interest in using the empirical likelihood method as opposed to using fully parametric models lies in that empirical likelihood does not require specific distributions. To explore conditions where the parametric assumptions for the Wald test no longer hold, we perform simulations for two such cases with results summarized in Tables 2 and 3, and Fig. 1.

For the results in Table 2, the trait value and expression level have a bivariate normal distribution for both genotypic groups, but with different covariance matrices. The distribution of trait and expression for genotype y individuals are

Table 2
The type I error rate and power of the Wald test, the empirical-likelihood(EL)-based test using both trait and expression data, the empirical-likelihood-based test and the Wilcoxon rank-sum test using only expression values from 30 randomly selected individuals out of 100 in each genotypic class

$\beta_y - \beta_x$	Wald	EL (trait & expression)	EL (expression)	Wilcoxon
0	.013	.012	.008	.011
	.053	.059	.047	.048
	.111	.115	.113	.105
.25	.045	.045	.039	.035
	.130	.139	.112	.099
	.219	.228	.205	.178
.50	.208	.228	.164	.138
	.411	.432	.381	.354
	.533	.544	.509	.478
.75	.518	.529	.430	.386
	.720	.738	.652	.610
	.808	.821	.765	.713
1.00	.789	.794	.720	.660
	.924	.933	.880	.848
	.958	.968	.936	.914

Within each genotypic group, expression level and trait are assumed to have a bivariate normal distribution, but the covariance matrices are different between genotypes. The type I error rate and power are reported for three test sizes: .01, .05 and .10, and results appear in that order.

Table 3
The type I error rate and power of the Wald test, the empirical-likelihood(EL)-based test using both trait and expression data, the empirical-likelihood-based test and the Wilcoxon rank-sum test using only expression values from 30 randomly selected individuals out of 100 in each genotypic class

$\beta_y - \beta_x$	Wald	EL (trait & expression)	EL (expression)	Wilcoxon
0	.018	.013	.015	.015
	.069	.059	.059	.065
	.121	.109	.124	.131
.25	.066	.075	.043	.026
	.156	.198	.127	.084
	.219	.276	.199	.143
.50	.218	.307	.157	.105
	.383	.551	.321	.251
	.491	.677	.450	.370
.75	.447	.703	.375	.291
	.673	.877	.622	.577
	.786	.931	.744	.701
1.00	.733	.925	.652	.597
	.886	.984	.855	.832
	.924	.993	.907	.889

For genotype y individuals, the trait and expression have a bivariate normal distribution, while for genotype x individuals, a skewed bivariate t-distribution is assumed. The type I error rate and power are reported for three test sizes: .01, .05 and .10, and results appear in that order.

the same as in the simulation for Table 1. For genotypic group x, however, the trait and expression have a bivariate normal distribution with mean $(0, 0)$, $\sigma_{x11} = .5$, $\sigma_{x22} = 2$, and $\sigma_{x12} = .01$. For the results in Table 3, the distribution for the trait and expression values for genotype y individuals is the same as before, but for individuals in genotypic group x, trait and expression values have a skewed bivariate t-distribution (Azzalini and Capitanio, 2003) with shape

parameters $\alpha = (4, 1)$, degrees of freedom of five, dispersion matrix

$$\bar{\Omega} = \begin{bmatrix} 1.00 & -1.30 \\ -1.30 & 2.25 \end{bmatrix},$$

and mean $(0, 0)$. In both of these settings, it can be seen that all four tests have satisfactory Type I errors, but the Wald test is less powerful than the empirical-likelihood-based test with auxiliary information. Thus the empirical-likelihood-based test using extra trait information can outperform the Wald test when the parametric assumptions for the Wald test are violated. Using extra trait data still improves the power relative to tests using expression values alone. These observations are also evident in Fig. 1, which report results for size .05 tests in a setting similar to Table 3 but with a wider range of mean difference.

4. Applications to barley spot blotch resistance and gene expression

Spot blotch caused by the fungus *Cochliobolus sativus* is one of the most common foliar diseases of barley in the midwest region of the United States. To identify genetic regions associated with barley resistance to spot blotch, Steffenson et al. (1996) carried out a QTL mapping study using 150 doubled haploid lines (DHLs) derived from the cross of two spring barley varieties. They considered the lesion size on the leaves of barley seedlings as the quantitative trait and identified a single major QTL associated with spot blotch resistance. The data regarding lesion size and genetic markers are available at the website http://www.genenetwork.org. The gene expression profiles of these 150 DHLs are studied by Luo et al. (2007) using Affymetrix Barley1 GeneChips.

In this section, we apply the proposed method in identifying genes whose expression is associated with the QTL controlling barley resistance to spot blotch. The genotype of marker CDO475 is used to represent the QTL genotype since it is very close to the position of the QTL identified by Steffenson et al. (1996). Among the 150 DHLs, 76 lines are of genotype A and 74 lines are of genotype B. To assess the performance of the proposed method, we randomly selected 25 DHLs of genotype A and 25 DHLs of genotype B respectively and assume that microarray data are available only on these 50 lines while the spot blotch lesion size is available on all 150 lines. Microarray data were normalized using the robust multichip average (RMA) method (Bolstad et al., 2003). We carried out tests for the association between gene expression and the QTL using the proposed empirical-likelihood-based test utilizing expression as well as lesion size (trait), the empirical-likelihood-based test and the Wilcoxon rank-sum test using only expression, and the Wald test on each of the 22 840 transcripts represented on the Barley1 GeneChip. The q-value method described in Storey and Tibshirani (2003) was used to control the false discovery rate (FDR) at specified levels. The number of genes whose expression was found to be associated with the CDO475 locus according to different methods is summarized in Table 4. Table 4 shows that the proposed empirical-likelihood-based test using both expression data on 50 selected lines and lesion size data on all 150 lines detected more genes than all the other methods. The fact that the proposed test detected more significant genes than the empirical-likelihood-based test using only expression data is consistent with the observation that superior power can be obtained by taking into account trait (auxiliary) data in the nonparametric setting. The reason that the proposed method also detected more significant genes than the Wald test might be due to the fact that the distribution of the spot blotch lesion size is more left skewed and flatter for plants of genotype A when compared with those of genotype B, which suggests that the joint distribution of lesion size and gene expression is likely to be different for the two genotypes. This is also consistent with results obtained in the simulation study, which shows that the proposed empirical-likelihood-based test can attain better power than the parametric Wald test under these conditions.

5. Conclusion and discussion

Selective transcriptional profiling can significantly reduce the cost associated with a large number of microarray chips in genetical genomics studies. However, there are occasions when the bivariate normal distribution and equal covariance assumption for the Wald test might be problematic. In this article, we develop empirical-likelihood-based inference for multi-sample comparisons with auxiliary data and apply it in the selective transcriptional profiling setting. It is shown that the idea of using extra trait data to improve the power in testing the association between a known QTL and transcriptional abundance also applies to nonparametric inference with empirical likelihood. When the parametric model assumptions are satisfied, the empirical-likelihood-based method using auxiliary data is only

Table 4
The number of genes whose transcriptional abundance is associated with the marker CDO475 according to the proposed empirical-likelihood-based test using both expression data on 50 selected DHLs and lesion size data on all 150 DHLs, the empirical-likelihood-based test and the Wilcoxon rank-sum test using only expression data on selected DHLs, or the Wald test using both expression and trait data is listed with the corresponding false discovery rate (FDR) level

FDR	Wald	EL (trait & expression)	EL (expression)	Wilcoxon
.01	259	273	228	180
.05	731	816	642	344
.10	1645	1751	1418	790
.15	2764	3003	2453	1733

slightly less powerful than the Wald test, and can be more powerful when the model assumptions are violated. Tests based on empirical likelihood are computation intensive (roughly fourteen hours for the Barley data on a personal computer), so one needs to balance improved power with additional computing time.

On the other hand, the Wald test often gives satisfactory Type I error rates even when the model assumptions do not hold exactly. Thus for ease of computation and explanation, the Wald test should be preferred if data are nearly normal, and the variance structures for the two genotypes are reasonably close. The empirical likelihood method utilizing extra trait data should be used if the data suggest substantial departures from the parametric assumptions of the Wald test.

To avoid making strong parametric assumptions on gene expression and trait values, the method presented here requires that the probability of selecting any individual is not strictly zero, that is, information on expression cannot be completely missing for a subset of individuals. This a common requirement for nonparametric methods in the missing data setting. Thus strategies of selecting only individuals with extreme trait values are not compatible with the proposed method since information on gene expression will be completely missing for individuals with intermediate trait values. We prefer randomly selecting individuals from each genotype for microarray experiments as in all examples discussed in this paper. However, it is possible to use different selection probabilities according to trait values, e.g., one can select individuals with extreme trait values with a higher probability. In that case the estimating functions need to be modified as $U_{zi}(\beta_z)/p(Z_{i1})$ and $g_{zi}(\gamma_z)/p(Z_{i1})$, where $p(Z_{i1})$ denotes the probability of selecting the ith individual for microarray experiment give the trait value Z_{i1}. Then results similar to Theorem 1 can be derived accordingly.

There is recent interest in applying empirical likelihood methods to genetic problems. Zou et al. (2002), Zou and Fine (2002), and Jin et al. (2007) developed a partial empirical likelihood method for QTL mapping. It will be of interest to see if their method can be combined with the selective transcriptional profiling approach when the location of the QTL is not certain.

Appendix A

A.1. Derivation of maximum empirical likelihood estimators

For the general case, suppose that the information for the distribution of $X_i = (X_{i1}, X_{i2})^\tau$ is summarized in an unknown p-dimensional parameter β_x via a p-dimensional estimating function $U(X_i, \beta_x)$, with $E\{U(X_i, \beta_x)\} = 0$. Furthermore, the auxiliary information involving only X_{i1} is summarized in a r-dimensional zero-mean estimating function $g(X_{i1}, \gamma_x)$ with a r-dimensional unknown parameter γ_x. Note in the most general formulation, X_{i1} may include both surrogate variables for X_{i2} and covariates that are always observable. We define $U(Y_j, \beta_y)$ and $g(Y_{j1}, \gamma_y)$ analogously with regard to $U(X_i, \beta_x)$ and $g(X_{i1}, \gamma_x)$. We denote $(\beta_{x0}, \gamma_{x0}, \beta_{y0}, \gamma_{y0})$ as the true values of the parameters. Recall that we use $z \in \{x, y\}$ and $Z \in \{X, Y\}$. The following conditions are needed for the derivation of the maximum empirical likelihood estimator and Theorem 1.

C1: Both $V_z(U) = E\{U(\beta_{z0})U^\tau(\beta_{z0})\}$ and $V_z(g) = E\{g(\gamma_{z0})g^\tau(\gamma_{z0})\}$ are positive definite, and the ranks of $E(\frac{\partial U_z(\beta_z)}{\partial \beta_z})$ and $E(\frac{\partial g_z(\gamma_z)}{\partial \gamma_z})$ are p and r respectively.

C2: $\frac{\partial^2 U_z(\beta_z)}{\partial \beta_z \partial \beta_z^\tau}$ is continuous in a neighborhood of β_{z0}, and both $\|\frac{\partial U_z(\beta_z)}{\partial \beta_z}\|$ and $\|U_z(\beta_z)\|^3$ are bounded in this neighborhood.

C3: $\frac{\partial^2 g_z(\gamma_z)}{\partial\gamma_z\partial\gamma_z^\tau}$ is continuous in a neighborhood of γ_{z0}, and both $\|\frac{\partial g_z(\gamma_z)}{\partial\gamma_z}\|$ and $\|g_z(\gamma_z)\|^3$ are bounded in this neighborhood.

C4: n_z and $m_z \to \infty$, and $n_z/(n_z+m_z) \to c_z \in (0,1)$ as $\min(n_z, m_z) \to \infty$. n_x and n_y are of the same order.

We denote $U_{zi}(\beta_z) = U(Z_i, \beta_z)$, $U'_{zi}(\beta_z) = \partial U_{zi}(\beta_z)/\partial\beta_z$, $g_{zi}(\gamma_z) = g(Z_{i1}, \gamma_z)$, and $g'_{zi}(\gamma_z) = \partial g_{zi}(\gamma_z)/\partial\gamma_z$. Using Lagrange multipliers as in Qin and Lawless (1994), the optimal weights for the parameter $(\beta_x, \gamma_x, \beta_y, \gamma_y)$ can be shown to be

$$p_{zi} = \frac{1}{n_z}\frac{1}{1 + t_{z1}^\tau U_{zi}(\beta_z) + t_{z2}^\tau g_{zi}(\gamma_z)} \quad \text{for } i = 1, \ldots, n_z, \tag{4}$$

$$q_{zk} = \frac{1}{m_z}\frac{1}{1 + t_{z3}^\tau g_{z(n_z+k)}(\gamma_z)} \quad \text{for } k = 1, \ldots, m_z, \tag{5}$$

where the Lagrange multipliers $t_{z\nu}$, $\nu = 1, 2, 3$, satisfy the following equations:

$$\sum_{i=1}^{n_z} \frac{U_{zi}(\beta_z)}{1 + t_{z1}^\tau U_{zi}(\beta_z) + t_{z2}^\tau g_{zi}(\gamma_z)} = 0, \tag{6}$$

$$\sum_{i=1}^{n_z} \frac{g_{zi}(\gamma_z)}{1 + t_{z1}^\tau U_{zi}(\beta_z) + t_{z2}^\tau g_{zi}(\gamma_z)} = 0 \tag{7}$$

and

$$\sum_{k=1}^{m_z} \frac{g_{z(n_z+k)}(\gamma_z)}{1 + t_{z3}^\tau g_{z(n_z+k)}(\gamma_z)} = 0. \tag{8}$$

Using (4) and (5), we can write the negative log empirical likelihood as

$$\begin{aligned}\ell(\beta_x, \gamma_x, \beta_y, \gamma_y) &= -2\log\{L(\beta_x, \gamma_x, \beta_y, \gamma_y)\} \\ &= 2\sum_z \sum_{i=1}^{n_z} \log\{1 + t_{z1}^\tau U_{zi}(\beta_z) + t_{z2}^\tau g_{zi}(\gamma_z)\} \\ &\quad + 2\sum_z \sum_{k=1}^{m_z} \log\{1 + t_{z3}^\tau g_{z(n_z+k)}(\gamma_z)\} + 2\sum_z \{n_z \log(n_z) + m_z \log(m_z)\}. \end{aligned} \tag{9}$$

When the null hypothesis is true, we also denote

$$\beta_x = \beta_y := \beta. \tag{10}$$

Then differentiate (9) with regard to β, γ_x and γ_y and use the results (6)–(8) to get

$$\sum_z \left\{ t_{z1}^\tau \sum_{i=1}^{n_z} \frac{U'_{zi}(\beta)}{1 + t_{z1}^\tau U_{zi}(\beta) + t_{z2}^\tau g_{zi}(\gamma_z)} \right\} = 0, \tag{11}$$

$$t_{z2}^\tau \sum_{i=1}^{n_z} \frac{g'_{zi}(\gamma_z)}{1 + t_{z1}^\tau U_{zi}(\beta) + t_{z2}^\tau g_{zi}(\gamma_z)} + t_{z3}^\tau \sum_{k=1}^{m_z} \frac{g'_{z(n_z+k)}(\gamma_z)}{1 + t_{z3}^\tau g_{z(n_z+k)}(\gamma_z)} = 0. \tag{12}$$

Let $(\tilde\beta, \tilde\gamma_x, \tilde\gamma_y, \tilde t_{x1}, \tilde t_{x2}, \tilde t_{x3}, \tilde t_{y1}, \tilde t_{y2}, \tilde t_{y3})$ be the solutions to (6)–(8) and (10)–(12). Accordingly, we have the maximum empirical likelihood estimator of the parameter under H_0 as $(\tilde\beta, \tilde\gamma_x, \tilde\gamma_y)$. On the other hand, if we allow the estimators for β_x and β_y to take different values and replace (11) with

$$t_{z1}^\tau \sum_{i=1}^{n_z} \frac{U'_{zi}(\beta_z)}{1 + t_{z1}^\tau U_{zi}(\beta_z) + t_{z2}^\tau g_{zi}(\gamma_z)} = 0, \tag{13}$$

we can obtain the maximum empirical likelihood estimator without constraint,

$$(\hat{\beta}_x, \hat{\gamma}_x, \hat{\beta}_y, \hat{\gamma}_y),$$

with the corresponding Lagrange multiplier $(\hat{t}_{x1}, \hat{t}_{x2}, \hat{t}_{x3}, \hat{t}_{y1}, \hat{t}_{y2}, \hat{t}_{y3})$, by solving (6)–(8), (12) and (13).

Proof of Theorem 1. The proof of Theorem 1 follows similar techniques used in Qin and Lawless (1994) and Chen et al. (2003). Only a brief outline is given here while complete details are available from the first author. We first derive the expansion of $\ell(\hat{\beta}_x, \hat{\gamma}_x, \hat{\beta}_y, \hat{\gamma}_y)$, in which the values of $\hat{\beta}_x$ and $\hat{\beta}_y$ can change freely with regard to each other. Define

$$T_{nz} = \left(n_z^{-1} \sum_{i=1}^{n_z} U_{zi}^{\tau}(\beta_{z0}), n_z^{-1} \sum_{i=1}^{n_z} g_{zi}^{\tau}(\gamma_{z0}), n_z^{-1} \sum_{k=1}^{m_z} g_{z(n_z+k)}^{\tau}(\gamma_{z0}) \right)^{\tau}.$$

Also define

$$\Sigma_{12} = \begin{bmatrix} \Sigma_{12x} & \\ & \Sigma_{12y} \end{bmatrix}, \qquad \Sigma_{22} = \begin{bmatrix} \Sigma_{22x} & \\ & \Sigma_{22y} \end{bmatrix}, \quad \text{where}$$

$$\Sigma_{12z} = \begin{bmatrix} EU_{zi}'(\beta_{z0}) & 0 & 0 \\ 0 & Eg_{zi}'(\gamma_{z0}) & \dfrac{1-c_z}{c_z} Eg_{z(n_z+k)}'(\gamma_{z0}) \end{bmatrix},$$

$$\Sigma_{22z} = \begin{bmatrix} E\{U_z(\beta_{z0})U_z^{\tau}(\beta_{z0})\} & E\{U_z(\beta_{z0})g_z^{\tau}(\gamma_{z0})\} & 0 \\ E\{g_z(\gamma_{z0})U_{zi}^{\tau}(\beta_{z0})\} & E\{g_z(\gamma_{z0})g_z^{\tau}(\gamma_{z0})\} & 0 \\ 0 & 0 & \dfrac{1-c_z}{c_z} E\{g_z(\gamma_{z0})g_z^{\tau}(\gamma_{z0})\} \end{bmatrix}$$

and

$$\Sigma_n = \begin{bmatrix} 0 & \Sigma_{12} \\ \Sigma_{12}^{\tau} & -\Sigma_{22} \end{bmatrix}.$$

By Taylor expansion around the true parameter value and some matrix manipulations, we can show

$$\ell(\hat{\beta}_x, \hat{\gamma}_x, \hat{\beta}_y, \hat{\gamma}_y) = (\sqrt{n}_x T_{nx}^{\tau}, \sqrt{n}_y T_{ny}^{\tau}) \{\Sigma_{22}^{-1} - \Sigma_{22}^{-1}\Sigma_{12}^{\tau}(\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^{\tau})^{-1}\Sigma_{12}\Sigma_{22}^{-1}\} \times (\sqrt{n}_x T_{nx}^{\tau}, \sqrt{n}_y T_{ny}^{\tau})^{\tau} + o_p(1).$$

For $\ell(\tilde{\beta}, \tilde{\gamma}_x, \tilde{\gamma}_y)$, the derivation is similar. Define

$$\Sigma_{12}^* = \begin{bmatrix} \Sigma_{12x}^* & \Sigma_{12y}^* \end{bmatrix}, \quad \text{where}$$

$$\Sigma_{12x}^* = \begin{bmatrix} EU_x'(\beta_0) & 0 & 0 \\ 0 & Eg_x'(\gamma_{x0}) & \dfrac{1-c_x}{c_x} Eg_x'(\gamma_{x0}) \\ 0 & 0 & 0 \end{bmatrix}$$

and

$$\Sigma_{12y}^* = \begin{bmatrix} EU_y'(\beta_0) & 0 & 0 \\ 0 & 0 & 0 \\ 0 & Eg_y'(\gamma_{y0}) & \dfrac{1-c_y}{c_y} Eg_y'(\gamma_{y0}) \end{bmatrix}.$$

We can then show

$$\ell(\tilde{\beta}, \tilde{\gamma}_x, \tilde{\gamma}_y) = (\sqrt{n}_x T_{nx}^{\tau}, \sqrt{n}_y T_{ny}^{\tau}) \{\Sigma_{22}^{-1} - \Sigma_{22}^{-1}\Sigma_{12}^{*\tau}(\Sigma_{12}^*\Sigma_{22}^{-1}\Sigma_{12}^{*\tau})^{-1}\Sigma_{12}^*\Sigma_{22}^{-1}\} \times (\sqrt{n}_x T_{nx}^{\tau}, \sqrt{n}_y T_{ny}^{\tau})^{\tau} + o_p(1).$$

Thus

$$\ell(\tilde{\beta}, \tilde{\gamma}_x, \tilde{\gamma}_y) - \ell(\hat{\beta}_x, \hat{\gamma}_x, \hat{\beta}_y, \hat{\gamma}_y) = (\sqrt{n}_x T_{nx}^{\tau}, \sqrt{n}_y T_{ny}^{\tau}) \Sigma_{22}^{-1/2} W \Sigma_{22}^{-1/2} (\sqrt{n}_x T_{nx}^{\tau}, \sqrt{n}_y T_{ny}^{\tau})^{\tau} + o_p(1),$$

where

$$W = \Sigma_{22}^{-1/2}\left\{\Sigma_{12}^{\tau}(\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^{\tau})^{-1}\Sigma_{12} - \Sigma_{12}^{*\tau}(\Sigma_{12}^{*}\Sigma_{22}^{-1}\Sigma_{12}^{*\tau})^{-1}\Sigma_{12}^{*}\right\}\Sigma_{22}^{-1/2}.$$

Note that $\Sigma_{22}^{-1/2}(\sqrt{n}_x T_{nx}^{\tau}, \sqrt{n}_y T_{ny}^{\tau})^{\tau} \xrightarrow{d} N(0, I_{2p+4r})$, $tr(W) = p$ and that W is symmetric and idempotent. This suggests that $\mathcal{R}_0$ has an asymptotic χ_p^2-distribution. □

A.2. Power properties in the general setting

To further simplify the notation, we may omit parameters β_z and γ_z in $U_z(\beta_z)$ and $g_z(\gamma_z)$ respectively. Also define $V(U_z) = E(U_z U_z^{\tau})$ and $V(g_z) = E(g_z g_z^{\tau})$. Consider the local alternative $\beta_y = \beta_x + n_z^{-1/2}u$, where $0 < \|u\| < \infty$. Here we suppose that n_x and n_y are of the same order. It can be shown that the asymptotic power for the test defined in Theorem 1 is $P(\chi_p^2(\lambda) > \chi_{p,1-\alpha}^2)$, where α is the size of the test and λ is the noncentrality parameter of the noncentral chi-square distribution with value

$$\lambda = u^{\tau}\left[\sum_z E^{-1}(U_z'^{\tau})\left\{V(U_z) - (1-c_z)E(U_z g_z^{\tau})V^{-1}(g_z)E(g_z U_z^{\tau})\right\}E^{-1}(U_z')\right]^{-1}u.$$

When the nonvalidation data are not utilized and the inference is based only on the validation data, the asymptotic power function becomes $P(\chi_p^2(\check{\lambda}) > \chi_{p,1-\alpha}^2)$, where the noncentrality parameter of the noncentral chi-square distribution is

$$\check{\lambda} = u^{\tau}\left[\sum_z E^{-1}(U_z'^{\tau})\{V(U_z)\}E^{-1}(U_z')\right]^{-1}u.$$

Since $\lambda > \check{\lambda}$, using nonvalidation data improves the power of the multi-sample test. The improvement in power is higher when c_z is smaller, that is, when the nonvalidation sample is large. Since the term $E(U_z g_z^{\tau})V^{-1}(g_z)E(g_z U_z^{\tau})$ can be seen as a measure of correlation between estimating functions U and g, it also suggests that the power is higher when the information contained in g is highly correlated to that contained in U.

References

Azzalini, A., Capitanio, A., 2003. Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. Journal of the Royal Statistical Society, Series B 65, 367–389.

Bolstad, B., Irizarry, R.A., Astrand, M., Speed, T., 2003. A comparison of normalization methods for high density oligonucleotide array data based on bias and variance. Bioinformatics 19, 185–193 (supplemental information).

Chen, S.X., Leung, H.Y., Qin, J., 2003. Information recovery in a study with surrogate endpoints. Journal of the American Statistical Association 98, 1052–1062.

DeCook, R., Lall, S., Nettleton, D., Howell, S.H., 2006. Genetic regulation of gene expression during shoot development in Arabidopsis. Genetics 172, 1155–1164.

Fleming, T., Prentice, R., Pepe, M., Glidden, D., 1994. Surrogate and auxiliary endpoints in clinical trials: With potential applications in cancer and AIDS research. Statistics in Medicine 13, 955–968.

Hackett, C.A., Weller, J.I., 1995. Genetic mapping of quantitative trait loci for traits with ordinal distributions. Biometrics 51, 1252–1263.

Jansen, R.C., Nap, J.N., 2001. Genetical genomics: The added value from segregation. Trends in Genetics 17, 388–391.

Jin, C., Fine, J., Yandell, B.S., 2007. A unified semiparametric framework for quantitative trait loci analyses, with application to spike phenotypes. Journal of the American Statistical Association 102, 56–57.

Kiekens, R., Vercauteren, A., Moerkerke, B., Goetghebeur, E., Van Den Daele, H., et al., 2006. Genome-wide screening for cis-regulatory variation using a classical diallel crossing scheme. Nucleic Acids Research 34, 3677–3686.

Lange, C., Whittaker, J.C., 2001. Mapping quantitative trait loci using generalized estimating equations. Genetics 159, 1325–1337.

Liang, Y., Jansen, M., Aronow, B., Geiger, H., Van Zant, G., 2007. The quantitative trait gene latexin influences the size of the hematopoietic stem cell population in mice. Nature Genetics 39, 178–188.

Luo, Z.W., Potokina, E., Druka, A., Wise, R., Waugh, R., et al., 2007. SFP genotyping from Affymetrix arrays is robust but largely detects cis-acting expression regulators. Genetics 176, 789–800.

Ma, C.-X., Casella, G., Wu, R., 2002. Functional mapping of quantitative trait loci underlying the character process: A theoretical framework. Genetics 161, 1751–1762.

Owen, A., 1988. Empirical likelihood ratio confidence intervals for a single functional. Biometrika 75, 237–249.

Owen, A., 1990. Empirical likelihood ratio confidence regions. The Annals of Statistics 18, 90–120.

Owen, A., 2002. Empirical Likelihood. Chapman and Hall, New York.
Pepe, M., 1992. Inference using surrogate outcome data and a validation sample. Biometrika 79, 355–365.
Qin, J., Lawless, J., 1994. Empirical likelihood and general estimating equations. The Annals of Statistics 22, 300–325.
Schadt, E.E., Monks, S.A., Drake, T.A., 2003. Genetics of gene expression surveyed in maize, mouse and man. Nature 422, 297–302.
Steffenson, B.J., Hayes, P.M., Kleinhofs, A., 1996. Genetics of seedling and adult plant resistance to net blotch (*Pyrenophora teres f. teres*) and spot blotch (*Cochliobolus sativus*) in barley. Theoretical and Applied Genetics 92, 552–558.
Storey, J.D., Tibshirani, R., 2003. Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences 100, 9440–9445.
Wang, D., Nettleton, D., 2006. Identifying genes associated with a quantitative trait or quantitative trait locus via selective transcriptional profiling. Biometrics 62, 504–514.
West, M.A.L., Kim, K., Kliebenstein, D.J., van Leeuwen, H., Michelmore, R.W., et al., 2007. Global eQTL mapping reveals the complex genetic architecture of transcript-level variation in Arabidopsis. Genetics 175, 1441–1450.
Zhang, Z.-Y., Ober, J.A., Kliebenstein, D.J., 2006. The gene controlling the quantitative trait locus epithiospecifier modifier 1 alters glucosinolate hydrolysis and insect resistance in Arabidopsis. Plant Cell 18, 1524–1536.
Zou, F., Fine, J.P., 2002. A note on a partial empirical likelihood. Biometrika 89, 958–961.
Zou, F., Fine, J.P., Yandell, B.S., 2002. On empirical likelihood for a semiparametric mixture model. Biometrika 89, 61–75.