# Shrinkage Estimation of Expression Fold Change As an Alternative to Testing Hypotheses of Equivalent Expression

Zahra Montazeri[*] Corey M. Yanofsky[†]
David R. Bickel[‡]

[*] Ottawa Institute of Systems Biology, Department of Biochemistry, Microbiology, and Immunology

[†] Ottawa Institute of Systems Biology, Department of Biochemistry, Microbiology, and Immunology

[‡] Ottawa Institute of Systems Biology, Department of Biochemistry, Microbiology, and Immunology, Department of Mathematics and Statistics, dbickel@uottawa.ca

# Shrinkage Estimation of Expression Fold Change As an Alternative to Testing Hypotheses of Equivalent Expression

Zahra Montazeri, Corey M. Yanofsky, and David R. Bickel

## Abstract

Research on analyzing microarray data has focused on the problem of identifying differentially expressed genes to the neglect of the problem of how to integrate evidence that a gene is differentially expressed with information on the extent of its differential expression. Consequently, researchers currently prioritize genes for further study either on the basis of volcano plots or, more commonly, according to simple estimates of the fold change after filtering the genes with an arbitrary statistical significance threshold. While the subjective and informal nature of the former practice precludes quantification of its reliability, the latter practice is equivalent to using a hard-threshold estimator of the expression ratio that is not known to perform well in terms of mean-squared error, the sum of estimator variance and squared estimator bias. On the basis of two distinct simulation studies and data from different microarray studies, we systematically compared the performance of several estimators representing both current practice and shrinkage. We find that the threshold-based estimators usually perform worse than the maximum-likelihood estimator (MLE) and they often perform far worse as quantified by estimated mean-squared risk. By contrast, the shrinkage estimators tend to perform as well as or better than the MLE and never much worse than the MLE, as expected from what is known about shrinkage. However, a Bayesian measure of performance based on the prior information that few genes are differentially expressed indicates that hard-threshold estimators perform about as well as the local false discovery rate (FDR), the best of the shrinkage estimators studied. Based on the ability of the latter to leverage information across genes, we conclude that the use of the local-FDR estimator of the fold change instead of informal or threshold-based combinations of statistical tests and non-shrinkage

estimators can be expected to substantially improve the reliability of gene prioritization at very little risk of doing so less reliably.

draft manuscript

# Shrinkage estimation of expression fold change as an alternative to testing hypotheses of equivalent expression

Zahra Montazeri[1], Corey M. Yanofsky[1], David R. Bickel[1,2,*]

## ABSTRACT

Research on analyzing microarray data has focused on the problem of identifying differentially expressed genes to the neglect of the problem of how to integrate evidence that a gene is differentially expressed with information on the extent of its differential expression. Consequently, researchers currently prioritize genes for further study either on the basis of volcano plots or, more commonly, according to simple estimates of the fold change after filtering the genes with an arbitrary statistical significance threshold. While the subjective and informal nature of the former practice precludes quantification of its reliability, the latter practice is equivalent to using a hard-threshold estimator of the expression ratio that is not known to perform well in terms of mean-squared error, the sum of estimator variance and squared estimator bias. On the basis of two distinct simulation studies and data from different microarray studies, we systematically compared the performance of several estimators representing both current practice and shrinkage. We find that the threshold-based estimators usually perform worse than the maximum-likelihood estimator (MLE) and they often perform far worse as quantified by estimated mean-squared risk. By contrast, the shrinkage estimators tend to perform as well as or better than the MLE and never much worse than the MLE, as expected from what is known about shrinkage. However, a Bayesian measure of performance based on the prior information that few genes are differentially expressed indicates that hard-threshold estimators perform about as well as the local false discovery rate (FDR), the best of the

[1] Ottawa Institute of Systems Biology; Department of Biochemistry, Microbiology, and Immunology; University of Ottawa; 451 Smyth Road; Ottawa, Ontario K1H 8M5

[2] Department of Mathematics and Statistics; University of Ottawa; 585 King Edward; Ottawa, Ontario K1N 6N5

[*] to whom correspondence should be addressed: dbickel@uottawa.ca

shrinkage estimators studied. Based on the ability of the latter to leverage information across genes, we conclude that the use of the local-FDR estimator of the fold change instead of informal or threshold-based combinations of statistical tests and non-shrinkage estimators can be expected to substantially improve the reliability of gene prioritization at very little risk of doing so less reliably.

*Subject headings:*

# 1. Introduction

## 1.1. Background

The accumulation of high-dimensional functional genomics data to solve biological problems as basic as reconstructing regulatory gene networks and as applied as understanding commercially important traits in crop species poses daunting obstacles to valid interpretation. Much research effort has been directed toward developing methods of reliably analyzing data for which there are hundreds or thousands of variables of interest. Transcriptional (gene expression) microarrays provide the main example, with measurements of thousands of genes, but the same statistical problems also plague metabolomic and proteomic studies. The large numbers of genes are especially problematic in the case of testing for each gene the null hypothesis that it is equivalently expressed across groups as opposed to differentially expressed: thousands of genes pose an extreme multiple comparisons problem for the most commonly used framework of hypothesis testing. While fully Bayesian approaches comply with the likelihood principle and thus clearly do not require adjustments for multiple comparisons, their inferences about each comparison might be improved by leveraging information across other comparisons (Scott and Berger 2006a); this prospect also motivates empirical Bayes methods of estimating local or global false discovery rates (Efron et al. 2001; Bickel 2004b, 2005).

To address the multiple comparisons problem in the Neyman-Pearson framework of hypothesis testing, substantial progress has been made in methodology for controlling Type I (false positive) error rates such as family-wise error rates and false discovery rates (Van der Laan et al. 2004). (The logic behind adjusting p-values for multiple comparisons also necessitates correcting estimated effect sizes in multiple comparisons problems. Bickel (2004a, 2008) developed such corrections to estimated levels of differential gene expression while others developed such corrections for genome-wide linkage scans (Sun and Bull 2005; Sun et al. 2006).)

Practitioners commonly use a p-value corrected for multiple testing in the same way as they would use an uncorrected p-value of a lone test; the assumption is that the p-value after correction may be treated as if it were a p-value that needed no correction. This is seen whenever a corrected p-value less than the conventional 5% level is interpreted as evidence against the null hypothesis; see, e.g., Craandijk and Schreuder (1979) and Dudbridge and Gusnanto (2008). If the (possibly corrected) p-value is less than the chosen threshold, a sufficiently high estimate of the average expression fold change across two conditions (commonly from a treatment condition to a control condition) flags a gene as a candidate for further investigation, with the fold change estimated using the sample mean. For example, the geometric mean expression ratio is estimated by the antilogarithm of the difference of log-transformed sample means across the two conditions. If, on the other hand, the p-value exceeds the threshold, the estimate is considered unreliable and the gene will not be considered further. Thus, genes not found to have statistically significant changes in expression are placed in the same category as those with small changes that are statistically significant. This practice of effectively treating genes with high p-values as if it were known that they are not differentially expressed, while not ideal, is nonetheless a reasonable use of limited resources that permit pursuing only a small subset of the thousands of genes represented on a microarray.

In 1925, no one could have anticipated the need in the face of thousands of hypotheses to treat statistically significant results of extremely small effect sizes in exactly the same way as non-significant results, and yet R. A. Fisher then laid the logical foundation of the way investigators in the post-genomic era test hypotheses of differential gene expression and act on the results of the tests. He convincingly argued that a treatment's effect size or parameter value is not even worth estimating if the accompanying p-value exceeds 0.05: "It is a useful preliminary before making a statistical estimate... to test if there is anything to justify estimation at all" (Fisher 2006, p. 300). His significance tests provided the scientific community with a simple way to determine whether or not there is sufficient data to infer that an observed effect would also appear in replicated experiments. Subsequent advances in statistics made possible by the computer revolution raise the prospect of moving beyond the ubiquitous practice of relying on a fold change estimate only if a sharp significance level has been achieved.

### 1.2. Shrinkage as an alternative to current practice

Research on analyzing microarray data has focused on the problem of identifying differentially expressed genes to the neglect of the equally important problem of estimating the level at which a gene is differentially expressed. As a result, even many of the most reliable methods for the analysis of gene expression microarray data have a common failure: strong evidence that an expression fold change or ratio differs from 1, as indicated by a high level of statistical significance, does not indicate that the expression ratio differs by a biologically significant amount from 1 (Bickel (2004a), Lewin et al. (2006)). A common approach to the problem is to select genes on the basis of a volcano plot like Figure 1, a graph of the statistical significance (p-value on $-\log_{10}$scale) versus the estimated expression ratio (e.g. Jin et al. (2001) , Chen et al. (2007)). However, this plot-based determination of which combinations of p-values and estimates merit gene selection is subjective, and it defies the automation that efficient screening requires.

Current solutions that do automate prioritization include those that either use the p-value to prioritize those genes that pass expression ratio estimate thresholds, a practice lacking theoretical and empirical justification, or use the expression ratio estimate to prioritize those genes that has a p-value threshold, a practice reminiscent of Fisher's method that enjoys some empirical support (e.g. Guo et al. (2006)). Such *hard-threshold* methods include those comparing a corrected p-value, estimated global or local false discovery rate, or approximate posterior probability to a some arbitrary or subjective value $\alpha$ that sharply separates low-priority genes from high-priority genes. Since a gene may be at the top or the bottom of the priority list depending on whether its p-value or expression ratio estimate is slightly above or slightly below the chosen threshold $\alpha$, hard thresholding has the same effect as does replacing all fold-change estimates of low-priority genes with 1, regardless of the measure $p$ of statistical significance (Figure 2). A less arbitrary approach "shrinks" the estimate of the fold change toward 1 to the extent that the variability is too high for precise estimation. This *shrinkage* approach to estimating levels of differential expression has been applied to microarray data using fully Bayesian (Theilhaber et al. 2001; Newton et al. 2004; Ishwaran and Rao 2003) and mixture-model (Gusnanto et al. 2005; Kauermann and Eilers 2004) estimators. More recently, the expression ratio was estimated via empirical Bayes (Bickel 2008; Hwang et al. 2009) and frequentist model averaging (Bickel and Yanofsky 2009) procedures. (Shrinkage properties of estimators of gene expression variance (Smyth 2004; McCarthy and Smyth 2009), gene-gene correlation (Schäfer and Strimmer 2005), and differential expression p-values (Ghosh 2006) have also been studied.) For a given gene, shrinkage toward equivalent expression results in an estimate of the expression ratio close to the maximum-likelihood estimates in the case of a small p-value, an estimate close to 1 in the case of a large p-value, and, unlike hard thresholding, in an intermediate estimate in the case of an intermediate p-value but without requiring subjective definitions of "small," "large," and "intermediate," as a *soft-threshold* approach would require. In this way, genes are automatically prioritized for further study on the basis of a single score, the shrinkage estimate of the fold change. For practical use in both basic and applied biological research,

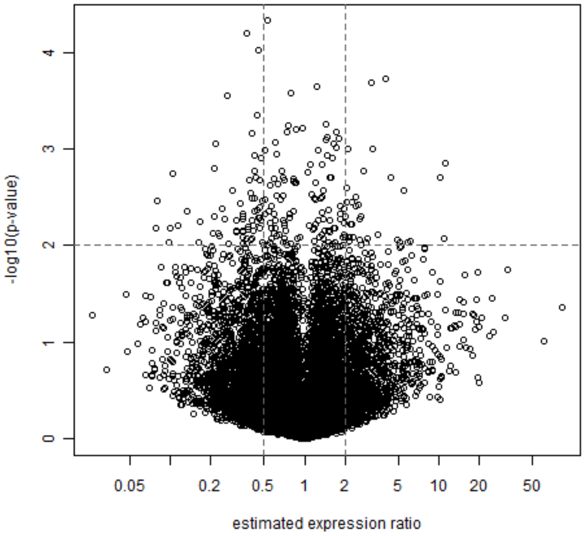


Fig. 1.— Volcano plot for an experiment on breast cancer cells, estrogen treatment versus control at 48 hours. The dashed lines correspond to some popular thresholds used to prioritize genes in the upper-left and upper-right rectangles for further study. See Section 3.2 for details regarding the data set.

a single score that contains the relevant information would often be preferred to the two scores needed for the volcano plot.

Section 2 presents a selection of representative hard-threshold and shrinkage methods of microarray data analysis for the identification of genes that are differentially expressed across experimental or observational conditions. The performance of these estimators is systematically studied by simulation in Section 3.1 and by empirical validation in Section 3.2. Finally, Section 4 provides an interpretation of the results.

## 2. Estimators of the level of differential gene expression

Let $x_{i,j}$ represent the natural logarithm of the measured expression intensity for the $i$th gene and the $j$th biological replicate of the control group; likewise, let $x'_{i,j}$ represent the logarithm of the measured expression intensity for the $i$th gene in the $j$th biological replicate of the treatment group. Let $n$ and $n'$ represent the number of biological replicates in the

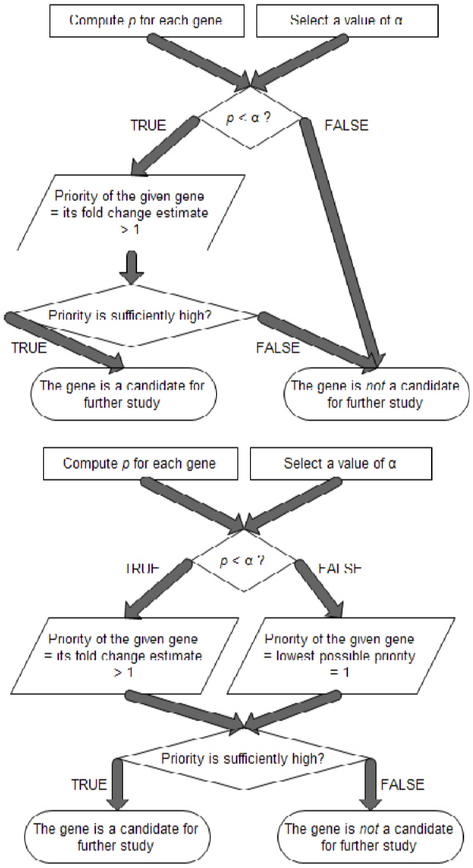

Compute $p$ for each gene
Select a value of α
$p < α$ ?
TRUE
FALSE
Priority of the given gene = its fold change estimate > 1
Priority is sufficiently high?
TRUE
FALSE
The gene is a candidate for further study
The gene is *not* a candidate for further study
Compute $p$ for each gene
Select a value of α
$p < α$ ?
TRUE
FALSE
Priority of the given gene = its fold change estimate > 1
Priority of the given gene = lowest possible priority = 1
Priority is sufficiently high?
TRUE
FALSE
The gene is a candidate for further study
The gene is *not* a candidate for further study

Fig. 2.— Two equivalent repesentations of hard thresholding, the current approach to prioritizing genes on the basis of their expression data. Here, $p$ is a (possibly adjusted) p-value or (approximate) posterior probability of equivalent expression, and $\alpha$ is a Type I error rate or other threshold. It can be seen that hard thresholding has exactly the same effect as setting fold-change estimates to 1 for all genes with non-significant expression data. The proposed approach instead obviates selection of $\alpha$ by simply setting the priority of each gene equal to its shrinkage estimate of the fold change.

control and treatment groups, respectively. Interest focuses on the expected difference of the difference logarithm of the expression intensity,

$$\mu_i = \mathrm{E}\left(X_i' - X_i\right),$$

where $X_i$ and $X_i'$ and the random variables of which $x_{i,j}$ and $x_{i,j}'$ are realizations and the expectation value is the sample space average. The true values of the expression ratio and fold change are $\exp\left(\mu_i\right)$ and $\exp\left(\left|\mu_i\right|\right)$, respectively; both quantities are geometric means over the sample space.

There are two experimental formats fitting the above description. For *paired data*, biological replicates are collected and subjected to the treatment and control conditions in a paired fashion to control variation due to unmeasured factors. By necessity $n' = n$ and the differences between paired data values,

$$y_{i,j} = x_{i,j}' - x_{i,j}$$

contains all of the information about $\mu_i$ in the data. Under the assumption that $X_i' - X_i$ is normally distributed, the maximum-likelihood estimate (MLE) of $\mu_i$ is the sample mean of the differences,

$$\hat{\mu}_{i,\mathrm{MLE}} = \bar{y}_i.$$

For *non-paired data*, it is assumed that $X_i'$ and $X_i$ are independent and normally distributed; the independence assumption is appropriate when the experimental protocol has excluded systematic errors, perhaps by randomization. The control and treatment groups may have different numbers of biological replicates, and the maximum-likelihood estimate of $\mu_i$ is the difference of the sample means,

$$\hat{\mu}_{i,\mathrm{MLE}} = \bar{x}_i' - \bar{x}_i.$$

Henceforth, we only explicitly write the gene index $i$ when it is needed for clarity.

Of course, the normality assumption never holds exactly, which is why any microarray data would fail a conventional test of normality given sufficient replication. In general, however, strong distributional assumptions are needed to reduce variance (at the expense of increased bias) in the context of the low levels of replication typical in microarray studies. Even for as many as six replicates ($n = 6$), statistical tests based on the particular assumption of normality have resulted in better inference than nonparametric tests (Bickel and Yanofsky 2009).

### 2.1. Hard-threshold estimators

Given a method of generating p-values to quantify the statistical significance of the evidence against equivalent expression, the general form for a hard-threshold estimator is,

$$\hat{\mu}_{\mathrm{HT}} = \begin{cases} \hat{\mu}_{\mathrm{MLE}} & \text{if p-value} < \alpha \\ 0 & \text{if p-value} \geq \alpha \end{cases}; \qquad (1)$$

that is, $\mu$ is estimated by the MLE if the (possibly adjusted) p-value is smaller than the significance threshold and by 0 otherwise. Different methods of generating p-values lead to different hard-threshold estimators. In the present analysis we set $\alpha = 0.05$ for all hard-threshold estimators.

#### *2.1.1. Statistical significance by error-rate control*

The standard measure of statistical significance is a p-value generated using a frequentist hypothesis test, a procedure which controls the probability of Type I error at level $\alpha$. The use of this p-value in equation (1) leads to the *raw p-value hard-threshold estimator.* The raw p-value controls the Type I error rate on a per-gene basis, but it will fail to control the probability that at least one Type I error occurs in the family of all $M$ hypotheses tested. To control this *family-wise error rate*, it is typical in differential gene expression experiments to adjust p-values for multiple comparisons, leading to the *adjusted p-value hard-threshold estimator.*

In the present analysis, we computed raw p-values for the above estimators using two-sided $t$-tests; for non-paired data, the variance was estimated separately for both groups and the Welch modification to the degrees of freedom was used. To adjust p-values, we used the step-down modification of the Šidàk single-step adjustment procedure as implemented in the *mt.rawp2adjp* function of the R package *multtest* (Dudoit et al. 2003).

#### *2.1.2. Empirical Bayes methods*

A popular alternative to Type I error rate control is the replacement of the p-value with an empirical Bayes estimate of the local false discovery rate (FDR) (Efron et al. 2001; Efron and Tibshirani 2002), a concept inspired by Benjamini and Hochberg's frequentist FDR control procedure (Benjamini and Hochberg 1995). The R package *locfdr* (Efron 2004) computes $\mathrm{LFDR}\,(p_i; \boldsymbol{p})$, the local false discovery rate associated with considering the $i^{\mathrm{th}}$

gene differentially expressed, where $p_i$ is the p-value of the $i^{\text{th}}$gene and $\boldsymbol{p}$ is the vector of the p-values of all genes; the p-values here are single-tailed. This function estimates the proportion of equivalently expressed genes to all genes within a small p-value neighborhood. Like a p-value, it falls in the range $[0, 1]$, and values closer to zero indicate stronger evidence of differential expression. When used in (1), it leads to the *local false discovery rate hard-threshold estimator.*

In the present analysis, we computed p-values for estimation of $\mathrm{LFDR}\,(p_i; \boldsymbol{p})$ using one-sided $t$-tests; as before, for non-paired data, the variance was estimated separately for both groups and the Welch modification to the degrees of freedom was used. See the Supplementary Information for a description of the estimation of $\mathrm{LFDR}\,(p_i; \boldsymbol{p})$.

## 2.2. Shrinkage estimators

### *2.2.1. Frequentist shrinkage estimators*

When a reasonably accurate initial estimate $\mu_0$ is available, it is possible to achieve lower mean squared error than the MLE by shrinking the MLE toward $\mu_0$ (Wiilink 2008). Starting with shrinkage estimators proposed in Thompson (968a) and Mehta and Srinivasan (1971), Wiilink (2008) proposed two shrinkage estimators, labeled $Q_1$ and $Q_2$, and showed that for univariate distributions they have smaller means squared error than the sample mean if the initial guess is appropriate. See the Supplementary Information for the definitions of these estimators.

### *2.2.2. Bayesian shrinkage model*

Scott and Berger (2006b) have described a fully Bayesian model for estimating the level of differential gene expression. The data are treated as a mixture of equivalently expressed genes whose true log-expression-ratios are zero and differentially expressed genes whose true log-expression-ratios are normally distributed around zero with variance $V$; the proportion of equivalently expressed genes to all genes is $\pi$. Observations are assumed to be normally distributed with variance $\sigma^2$. In effect, the model assumes that the observed data are drawn from a mixture of two normal distributions, one with variance $\sigma^2$ and one with variance $\sigma^2 + V$, both centered at 0. The model achieves shrinkage by two mechanisms: first, the data distribution contains a component explicitly modelling equivalently expressed genes as having a mean of zero; and second, even given that a gene is differentially expressed, the prior distribution of its expression is centered at a mean of zero, inducing further shrinkage.

See Supplementary Information for explicit descriptions of the prior distributions, likelihood function, and derivations of the posterior distribution of the quantities of interest.

The posterior distribution of the parameters $(V, \sigma^2, \pi)$ can easily be sampled from with a Markov chain Monte Carlo (MCMC) algorithm; we used Differential Evolution Monte Carlo (Ter Braak 2006), a Metropolis algorithm which automatically and efficiently generates proposals with the same covariance structure as the posterior distribution. Taking the average of $\text{E}_{\text{posterior}}\left(\mu_i | V, \sigma^2, \pi\right)$ over the MCMC samples of $(V, \sigma^2, \pi)$ yields a Monte Carlo approximation to $\text{E}_{\text{posterior}}\left(\mu_i\right)$, the expectation of the marginal posterior distribution of $\mu_i$.

### *2.2.3. Local-FDR-based shrinkage estimator*

Following Bickel (2008) and Bickel and Yanofsky (2009), we treat the local false discovery rate as an approximate posterior probability of equivalent expression for the purpose of estimating the degree of differential expression by the approximate posterior mean. This shrinkage estimator scales the MLE by the posterior probability that the gene is actually differentially expressed:

$$\mu_{\text{LFDR} \quad \text{shrinkage}} = \left[1 - \text{LFDR}\left(p_i | \boldsymbol{p}\right)\right] \hat{\mu}_{\text{MLE}}.$$

## 3. Quantification of estimator reliability

The simulations and parametric bootstrap sampling procedures described in Subsections 3.1 and 3.2 were used to quantify the performance of each of the estimators described in Section 2 as follows. The mean-squared error (MSE) for the $i$th gene was estimated by the empirical MSE, the sample mean of the squared error over the set of simulations or bootstrap samples, $\widehat{\text{MSE}}_i = \frac{1}{B}\sum_{b=1}^{B}\left(\widehat{\mu}_{b,i} - \mu_i\right)^2$, where $\mu_i$ is the true mean used to generate the simulated data (for bootstrap sampling, it is the MLE from the original data), $\widehat{\mu}_{b,i}$ is the estimate of $\mu_i$ for the $b^{\text{th}}$ simulated or bootstrap sample according to the estimator under consideration, and $B$ is the number of simulated or bootstrap samples. To quantify the performance of each estimator, we defined the *risk* associated with using an estimator as its total MSE over all genes and estimated it by $\widehat{\text{risk}} = \sum_{i=1}^{M} \widehat{\text{MSE}}_i$, where $M$ is the number of genes. See the Supporting Information for the quantification of uncertainty in $\widehat{\text{risk}}$ due to bootstrap or simulation sampling.

Bickel and Yanofsky (2009) used cross-validation to estimate the total MSE and the average of rescaled MSEs of *predicted observations*; herein, we summarize the MSEs of *parameter estimates* using the total MSE but not the average rescaled MSE because errors in estimates of genes with higher fold changes are of more consequence than those of genes with lower fold changes. Since the total MSE tends to give more weight to higher-priority genes than to those of lower priority, it has the advantage of reflecting errors in relative ranking between top-priority genes while effectively ignoring the irrelevant relative ranks of low-priority genes.

We conclude this section with a Bayesian method of quantifying estimator performance that leverages prior knowledge that only a few genes have notable differential expression.

## 3.1. Simulation studies

To compare estimators of differential expression levels, two classes of simulations were carried out: in one case, half of the simulated genes were equivalently expressed across conditions, and in the other case, no simulated genes had exactly equivalent expression across groups although most genes had very small levels of differential expression. In both cases, simulations were of paired data and $y_{i,j}$ values were simulated directly. For each simulation type, data sets with 2 and 8 biological replicates were simulated to investigate the effect of sample size on the estimators in the small-$n$ regime typical of laboratory experiments.

### *3.1.1. Case with half of the genes differentially expressed*

We first used the simulation design of M. Langaas and Ferkingstad (2005), drawing 200 independent simulations of 20,000 genes from a multivariate normal distribution, first with $n = 2$ and then with $n = 8$. Half of the genes were equivalently expressed, i.e., with true means set to 0. True means for the differentially expressed genes were drawn from a symmetric bitriangular density with boundaries $\log_2(1.2)$ and $\log_2(4)$. To construct the covariance matrix, genes were separated into groups of size 100, and pairwise correlations for all genes within a group were each set to 0.5; correlations were each 0 for pairs of genes not in the same group. Each gene's variance was set to 1.

### *3.1.2. Case of all genes with at least some differential expression*

For the case in which most genes have very small but nonzero levels of differential expression with no sharp distinction between differentially expressed and equivalently expressed genes, 200 independent simulations of 20,000 genes were drawn from univariate normal distributions, again with each sample of size $n = 2$ for one simulation and $n = 8$ for another simulation. Adapting the simulation design of Bickel (2008), the means of the distributions were

$$\mu_i = \begin{cases} -\left(\frac{10001-i}{10000}\right)^8 & i \in \{1, ..., 10^4\} \\ \left(\frac{i}{10000}\right)^8 & i \in \{10001, ..., 20^4\} \end{cases} .$$

The variance of the simulations were chosen to have a similar signal-to-noise ratio as in actual experiments, which in this case resulted simulation standard deviation of 0.23. (See the Supporting Information for details.) Reflecting biological reality, none of the genes is equivalently expressed, and therefore the null hypothesis that $\mu_i = 1$ is false for all $i$ and would be rejected by a test at any positive level $\alpha$ for sufficiently large $n$. For small $n$, however, many of the genes with low levels of differential expression will not be identified as differentially expressed.

### *3.1.3. Simulation results*

Figure 3 shows the results of the simulations in terms of risk relative to the risk of the MLE. These simulation results complement those of the two-sample simulation study in which Newton et al. (2004) found that their Bayesian shrinkage estimator of the fold change has lower MSE than that of the MLE.

## 3.2. Empirical validation

The estimators were evaluated by a frequentist method (parametric bootstrapping) and by a Bayesian method (posterior expected loss). The latter would be more appropriate given the prior knowledge that most genes are equivalently expressed.

### *3.2.1. Parametric bootstrap*

To evaluate the performance of the estimators in the context of real data, we applied the parametric bootstrap (*cf.* Van Der Laan and Bryan (2001)) to two experimental data sets. The first data set was from an experiment applying an estrogen treatment to cells of a human breast cancer cell line (Scholtens et al. 2004) (available from the Bioconductor project, http://www.bioconductor.org/). Two non-paired biological replicates were collected

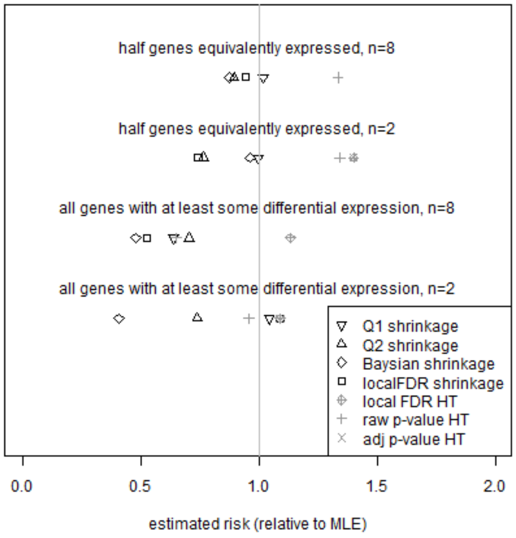


Fig. 3.— Estimated risk for all estimators relative to that of the MLE for the simulated data sets. The local FDR HT estimator for the first row and the adjusted p-value estimator for the first and third rows have relative risks greater than 2.0 and are not plotted.

for each of the four following states: 10 hours after treatment, 48 hours after treatment, control after 10 hours, and control after 48 hours. These data provide information about four differential genes expression levels: 10 hours after treatment versus control at 10 hours, 48 hours after treatment versus control at 48 hours, 48 hours after treatment versus 10 hours after treatment, and control at 48 hours versus control at 10 hours.

The second data set was a subset of data from the Microarray Quality Control (MAQC) study (Guo et al. 2006) in which rat liver cells were treated with comfrey. Six non-paired biological replicates were collected for both the treatment and control conditions. The data were analyzed on an Applied Biosystems platform, an Agilent platform, and two Affymetrix platforms in different laboratories, for a total of four sets of differential gene expression level estimates.

The parametric bootstrap is a strategy for estimating expectations over the sampling distribution when the parameters of that distribution are unknown. The parameters of the sampling distribution are estimated and then replicate data sets are sampled from the sampling distribution with the parameter values fixed to the estimates from the original data. The parametric bootstrap is preferred to the nonparametric bootstrap (Efron 1979) for small sample sizes. For the parametric bootstrap analysis, we modeled $X_i$ and $X_i'$ as normally distributed and estimated their means and variances using the usual unbiased estimators, and then generated 200 bootstrap samples for each of the eight possible sets of differential gene expression levels, four from each experiment. (To save time, 100 bootstrap samples were used for the hard-threshold estimators for the non-paired data.) The risk of each estimator was estimated from the resulting bootstrap samples. Figures 4 and 5 show the results of the bootstrap analyses in terms of risk relative to the risk of the MLE.

### *3.2.2. Assessment by posterior expected loss*

For a given posterior distribution for the parameters, the posterior expected squared error loss of an estimator can be calculated. The posterior expected squared error loss of an estimator is the squared difference between its estimate and the posterior mean, plus the posterior variance,

$$\mathrm{E}_{\,\mathrm{posterior}}\left[(\mu-\hat{\mu})^2\right]=\left[\mathrm{E}_{\,\mathrm{posterior}}\left(\mu\right)-\hat{\mu}\right]^2+\mathrm{var}_{\,\mathrm{posterior}}\left(\mu\right). \qquad (2)$$

See the Supplementary Material for the derivation of (2).

For the Bayesian shrinkage model, $\mathrm{var}_{\,\mathrm{posterior}}\left(\mu\right)$ is estimated as the average of $\mathrm{var}_{\,\mathrm{posterior}}\left(\mu|V,\sigma^2,\pi\right)$ over the MCMC samples of $(V,\sigma^2,\pi)$. For the local-FDR-based shrinkage estimator, there is

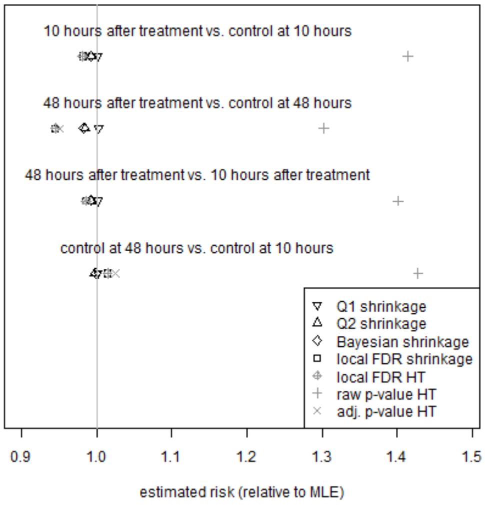


Fig. 4.— Estimated risk for all estimators relative to that of the MLE for the breast cancer data sets. The relative risk for the adjusted p-value hard-threshold estimator does not appear in the plot because it was greater than 1.5 in all cases.

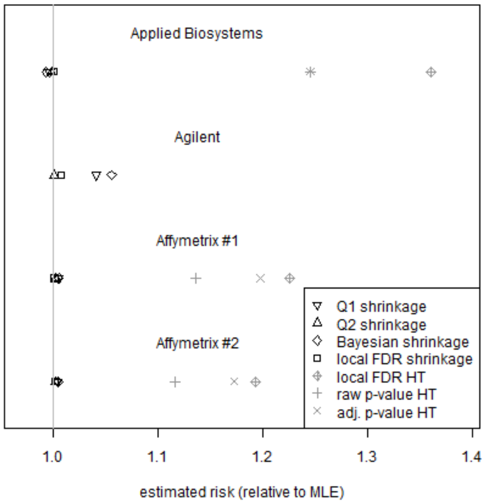


Fig. 5.— Estimated risk for all estimators relative to that of the MLE for the MAQC data sets. The relative risks for the hard-threshold estimators set were not plotted because they were greater than 1.9 in all cases.

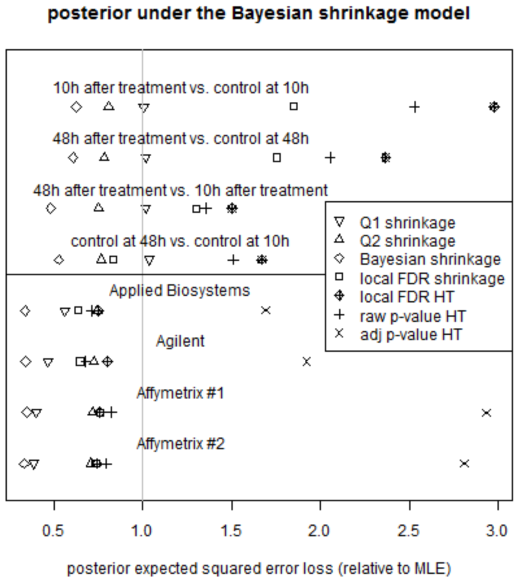


Fig. 6.— Posterior expected squared error loss under the Bayesian shrinkage model for all estimators relative to that of the MLE; upper panel shows results for the breast cancer data sets and lower panel shows results for the MAQC data sets.

no explicitly defined posterior distribution, but the local FDR can be interpreted as an approximation to an effective posterior distribution that has probability mass $\mathrm{LFDR}\,(p_i|\boldsymbol{p})$ at $\mu = 0$ and probability mass $[1 - \mathrm{LFDR}\,(p_i|\boldsymbol{p})]$ at $\mu = \hat{\mu}_{\mathrm{MLE}}$. For this posterior distribution, the variance of the mean is

$$\mathrm{var}_{\,\mathrm{posterior}}\,(\mu) = \mathrm{LFDR}\,(p_i|\boldsymbol{p})\,[1 - \mathrm{LFDR}\,(p_i|\boldsymbol{p})]\,(\hat{\mu}_{\mathrm{MLE}})^2\,.$$

Figures 6 and 7 show the total (over all genes) posterior expected squared error loss relative to that of the MLE for each estimator under the two posterior distributions for each of the eight real data sets.

## 4. Discussion and conclusions

We have contrasted hard-threshold approaches representing the current practice of estimation following testing (Figure 2) and shrinkage approaches to gene prioritization. The best hard-threshold estimator performed about as well as the best shrinkage estimator in one set of simulations, but in the other three sets of simulations, each of the shrinkage estimators

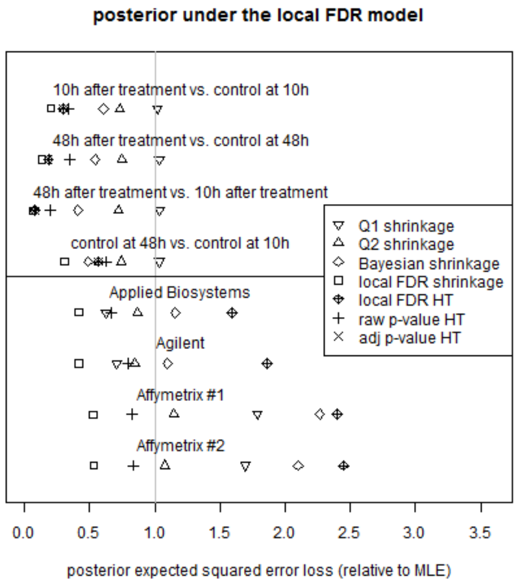


Fig. 7.— Posterior expected squared error loss under the local FDR model for all estimators relative to that of the MLE; upper panel shows results for the breast cancer data sets and lower panel shows results for the MAQC data sets. The relative posterior expected loss for the adjusted p-value hard-threshold estimator does not appear in the lower panel because it was greater than 5 in all cases.

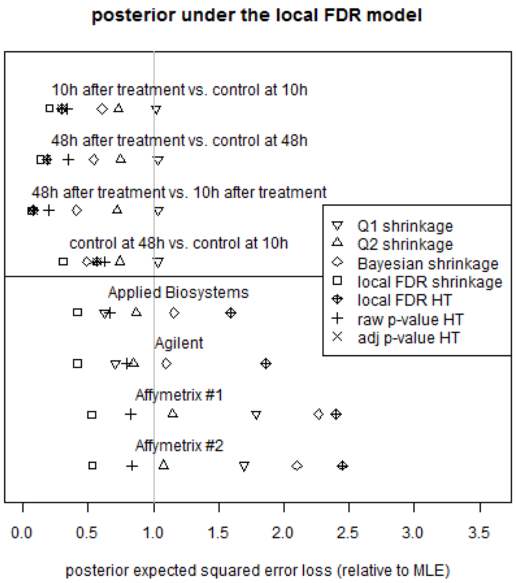


Fig. 8.— Posterior expected squared error loss under the local FDR model for all estimators relative to that of the MLE; upper panel shows results for the breast cancer data sets and lower panel shows results for the MAQC data sets. The relative posterior expected loss for the adjusted p-value hard-threshold estimator does not appear in the lower panel because it was greater than 5 in all cases.

proved to be substantially more reliable than each of the hard-threshold estimators (Figure 3). In the bootstrap analyses, we found that for the breast cancer data set (Figure 4), each estimator except for the raw p-value hard-threshold estimator performed essentially as good as the MLE. Local FDR shrinkage performed best in each case except for the comparison of control at 48 hours versus control at 10 hours, for which $Q_2$ shrinkage performed best. For the MAQC data set (Figure 5), we found that all shrinkage estimators performed essentially identically to the MLE, and all hard-threshold estimators performed notably worse than the MLE.

As equation (2) makes clear, the posterior expected squared error is its performance according to the Bayesian model used. Naturally, for a given model, its own posterior expectation has the minimum posterior expected squared error, as can be seen in Figures 6 and 7. From these figures, it is apparent that the Bayesian shrinkage model is most like the frequentist $Q_1$ and $Q_2$ estimators, while the local FDR shrinkage is more similar to the hard threshold estimators. The fully Bayesian model makes strong assumptions about the data: first, that different genes have identical sampling variances, and second, that the log-expressions of differentially expressed genes follow a normal distribution. In contrast, the local FDR model does not assume identical sampling variances and uses a flexible nonparametric estimate for the distribution of p-values of differentially expressed genes. Its key assumption is that most genes are equivalently expressed, an eminently reasonable assumption. Therefore, we regard Figure 7 as a more trustworthy indicator of actual estimator performance than Figure 6. Given prior knowledge in accord with the assumption of local FDR shrinkage that most genes are equivalently expressed, these results suggest that the hard threshold estimators and local FDR shrinkage are superior to the other estimators. The bootstrap and simulation results suggest that the hard-threshold estimators only perform well under this assumption, but that local FDR shrinkage performs well even in other cases.

The key observations of the risk estimates are that (i) no hard-threshold estimator ever had the lowest estimated risk, and they were frequently far worse than the MLE or any shrinkage estimator, and (ii) the shrinkage estimators often outperform the MLE and are never much worse. Given the known tendency of shrinkage estimators to perform well in terms of MSE and the lack of any indication that an estimator sensitive to the value of an arbitrary threshold would perform well, we conclude that the risk estimates accurately represent the relative reliabilities of the two classes of estimators except when only a small portion of genes have substantial differential expression, in which case the posterior expected loss according to the local FDR model better quantifies performance. Although hard threshold estimators were found to be competitive with the local FDR in terms of posterior expected loss, the latter has the advantage of borrowing strength across all genes rather than arbitrarily controlling Type I error rates. Thus, we recommend the replacement of threshold-based prioritization of genes with shrinkage-based prioritization. In particular, local FDR shrinkage had the best overall performance and is a good default choice, especially when only a few genes are expected to be differentially expressed at notable levels.

Opportunities for additional research abound. For example, since in $Q_1$ and $Q_2$ we used the tuning-parameter values preferred by Wiilink (2008), they have yet to be optimized for microarray data. They can be adjusted to give stronger shrinkage toward equivalent expression, which would decrease the posterior expected loss used here. Also, in some situations, researchers may wish to correct estimates of fold change for biases due to confounding (Kerr et al. 2001) or selection from multiple comparisons (Bickel 2004a, 2008); applying such bias corrections to the shrinkage estimates we considered invites further study.

## 5. Statement of contributions

CMY developed the algorithms for computing the posterior for the full Bayesian model and implemented the bootstrap sampling and evaluation by posterior expected square error. ZM extended $Q_1$ and $Q_2$ to the two-sample case and executed the simulations. DRB conceived the study on the basis of the equivalence between current practice and hard-threshold estimation. All authors participated in drafting the manuscript.

## 6. Acknowledgments

We thank Pei-Chun Hsieh for preparing the MAQC data for analysis. The *Biobase* package of Bioconductor (Gentleman et al. 2004) facilitated management of the expression data. This work was partially supported by the Canada Foundation for Innovation, the Ministry of Research and Innovation of Ontario, and the Faculty of Medicine of the University of Ottawa. *Conflicts of interest*: none declared.

## 7. Availability

R (R Development Core Team 2008) functions for hard-thresholding and shrinkage estimation of fold change are available from http://www.statomics.com.

## 8. Supplementary Information

Additional figures: http://www.davidbickel.com.

## REFERENCES

Benjamini, Y. and Hochberg, Y. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B*, 57:289–300.

Bickel, D. R. 2004a. Degrees of differential gene expression: Detecting biologically significant expression differences and estimating their magnitudes. *Bioinformatics (Oxford, England)*, 20:682–688.

Bickel, D. R. 2004b. Error-rate and decision-theoretic methods of multiple testing: Which genes have high objective probabilities of differential expression? *Statistical Applications in Genetics and Molecular Biology*, 3(1):8.

Bickel, D. R. 2005. Probabilities of spurious connections in gene networks: Application to expression time series. *Bioinformatics (Oxford, England)*, 21:1121–1128.

Bickel, D. R. 2008. Correcting the estimated level of differential expression for gene selection bias: Application to a microarray study. *Statistical Applications in Genetics and Molecular Biology*, 7(1):10.

Bickel, D. R. and Yanofsky, C. M. 2009. Validation of differential gene expression algorithms: Application comparing fold-change estimation to hypothesis testing. *Technical Report, Ottawa Institute of Systems Biology, COBRA Preprint Series, Article 50, available at tinyurl.com/bw9jod.*

Chen, J. J., Wang, S. J., Tsai, C. A., and Lin, C. J. 2007. Selection of differentially expressed genes in microarray data analysis. *Pharmacogenomics Journal*, 7(3):212–220.

Craandijk, A. and Schreuder, G. M. T. 1979. Association of juvenile disciform maculopathy with hla b15. *British Journal of Ophthalmology*, 63(10):678–679.

Dudbridge, F. and Gusnanto, A. 2008. Estimation of significance thresholds for genomewide association scans. *Genetic Epidemiology*, 32(3):227–234.

Dudoit, S., Shaffer, J. P., and Boldrick, J. C. 2003. Multiple hypothesis testing in microarray experiments. *Statistical Science*, 18(1):71–103.

Efron, B. 1979. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26.

Efron, B. 2004. Large-scale simultaneous hypothesis testing: The choice of a null hypothesis. *Journal of the American Statistical Association*, 99(465):96–104.

Efron, B. and Tibshirani, R. 2002. Empirical bayes methods and false discovery rates for microarrays. *Genetic epidemiology*, 23(1):70–86.

Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. 2001. Empirical bayes analysis of a microarray experiment. *J. Am. Stat. Assoc.*, 96(456):1151–1160.

Fisher, R. A. 2006. *Statistical Methods for Research Workers.* Cosmo Publications, New Delhi.

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A. J., Sawitzki, G., Smith, C., Smyth, G., Tierney, L., Yang, J. Y. H., and Zhang, J. 2004. Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology*, 5:R80.

Ghosh, D. 2006. Shrunken p-values for assessing differential expression with applications to genomic data analysis. *Biometrics*, 62(4):1099–1106.

Guo, L., Lobenhofer, E. K., Wang, C., Shippy, R., Harris, S. C., Zhang, L., Mei, N., Chen, T., Herman, D., Goodsaid, F. M., Hurban, P., Phillips, K. L., Xu, J., Deng, X., Sun, Y. A., Tong, W., Dragan, Y. P., and Shi, L. 2006. Rat toxicogenomic study reveals analytical consistency across microarray platforms. *Nat Biotech*, 24(9):1162–1169.

Gusnanto, A., Ploner, A., and Pawitan, Y. 2005. Fold-change estimation of differentially expressed genes using mixture mixed-model. *Statistical Applications in Genetics and Molecular Biology*, 4(1):i–22.

Hwang, J. T. G., Qiu, J., and Zhao, Z. 2009. Empirical bayes confidence intervals shrinking both means and variances. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 71(1):265–285.

Ishwaran, H. and Rao, J. S. 2003. Detecting differentially expressed genes in microarrays using bayesian model selection. *Journal of the American Statistical Association*, 98(462):438–455.

Jin, W., Riley, R. M., Wolfinger, R. D., White, K. P., Passador-Gurgell, G., and Gibson, G. 2001. The contributions of sex, genotype and age to transcriptional variance in drosophila melanogaster. *Nature Genetics*, 29(4):389–395.

Kauermann, G. and Eilers, P. 2004. Modeling microarray data using a threshold mixture model. *Biometrics*, 60(2):376–387.

Kerr, M. K., Martin, M., and Churchill, G. A. 2001. Analysis of variance for gene expression microarray data. *Journal of Computational Biology*, 7(6):819–837.

Lewin, A., Richardson, S., Marshall, C., Glazier, A., and Aitman, T. 2006. Bayesian modeling of differential gene expression. *Biometrics*, 62(1):1–9.

M. Langaas, B. H. L. and Ferkingstad, E. 2005. Estimating the proportion of true hypotheses, with application to dna microarray data. *Journal of the Royal Statistical Society. B*, 67.

McCarthy, D. J. and Smyth, G. K. 2009. Testing significance relative to a fold-change threshold is a treat. *Bioinformatics (Oxford, England)*, 25(6):765–771.

Mehta, J. and Srinivasan, R. 1971. Estimation of the mean by shrinkage to a point. *Journal of the American Statistical Association*, 66:86–90.

Newton, M. A., Noueiry, A., Sarkar, D., and Ahlquist, P. 2004. Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics*, 5(2):155–176.

R Development Core Team 2008. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Scholtens, D., Miron, A., Merchant, F. M., Miller, A., Miron, P. L., Iglehart, J. D., and Gentleman, R. 2004. Analyzing factorial designed microarray experiments. *J. Multivar. Anal.*, 90(1):19–43.

Schäfer, J. and Strimmer, K. 2005. A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, 4(1):1–30.

Scott, J. G. and Berger, J. O. 2006a. An exploration of aspects of bayesian multiple testing. *Journal of Statistical Planning and Inference*, 136(7 SPEC. ISS.):2144–2162.

Scott, J. G. and Berger, J. O. 2006b. An exploration of aspects of bayesian multiple testing. *Journal of Statistical Planning and Inference*, 136(7):2144–2162.

Smyth, G. K. 2004. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3(1).

Sun, L. and Bull, S. B. 2005. Reduction of selection bias in genomewide studies by resampling. *Genetic Epidemiology*, 28(4):352–367.

Sun, L., Craiu, R. V., Paterson, A. D., and Bull, S. B. 2006. Stratified false discovery control for large-scale hypothesis testing with application to genome-wide association studies. *Genetic Epidemiology*, 30(6):519–530.

Ter Braak, C. J. F. 2006. A markov chain monte carlo version of the genetic algorithm differential evolution: easy bayesian computing for real parameter spaces. *Statistics and Computing*, 16(3):239–249.

Theilhaber, J., Bushnell, S., Jackson, A., and Fuchs, R. 2001. Bayesian estimation of fold-changes in the analysis of gene expression: The pfold algorithm. *Journal of Computational Biology*, 8(6):585–614.

Thompson, J. 1968a. Some shrinkage techniques for estimating the mean. *Journal of the American Statistical Association*, 63:113–122.

Van Der Laan, M. J. and Bryan, J. 2001. Gene expression analysis with the parametric bootstrap. *Biostat*, 2(4):445–461.

Van der Laan, M. J., Dudoit, S., and Pollard, K. S. 2004. Augmentation procedures for control of the generalized family-wise error rate and tail probabilities for the proportion of false positives. *Stat. Appl. in Genet. and Mol. Biol.*, 3:15.

Wiilink, R. 2008. Shrinkage confidence intervals for the normal mean: using a guess for greater efficiency. *The Canadian Journal of Statistics*, 34(4):623–637.

This preprint was prepared with the AAS LaTeX macros v5.2.

# Shrinkage estimation of expression fold change as an alternative to testing hypotheses of equivalent expression
# Supplementary Information

July 3, 2009

Zahra Montazeri, Corey M. Yanofsky, David R. Bickel

## 1 Mathematical details

This section fills in the mathematical details of the estimation of the local false discovery rate, the frequentist shrinkage estimators, the Bayesian shrinkage model, the method for choosing the variances in the simulations, and the derivation of the expression for the posterior expected squared error.

### 1.1 Local false discovery rate estimation

The first step in the estimation of $\text{LFDR}\,(p_i;\boldsymbol{p})$ is the application of a nonparametric density estimator to $\boldsymbol{z}$, a vector calculated by applying the normal quantile function to the values in $\boldsymbol{p}$ (a vector of single-tailed p-values for all genes),

$$z_i = \Phi^{-1}\left(p_i\right),$$

where $\Phi^{-1}\left(\cdot\right)$ is the normal quantile function (that is, the inverse of the standard normal distribution function). The density function of the values in $\boldsymbol{z}$ is a two-component mixture for which one component is standard normal (i.e., the theoretical distribution of $z_i$ under the null hypothesis of equivalent expression) and the other is unknown:

$$f_{\mathbf{z}}\left(\cdot\right) = \lambda\varphi(\cdot) + (1-\lambda)\,g_{\mathbf{z}}\left(\cdot\right),$$

in which is $f_{\mathbf{z}}\left(\cdot\right)$ the density function of the $z_i$ values, $\lambda$ is the proportion of genes which are equivalently expressed, $\varphi(\cdot)$ is the density function of the standard normal distribution and $g_{\mathbf{z}}\left(\cdot\right)$ is the distribution of $\boldsymbol{z}$ values for differentially expressed genes. The density function $f_{\mathbf{z}}(\cdot)$ can be estimated using any consistent non-parametric density estimator. The assumption $g_{\mathbf{z}}(0) = 0$ allows $\lambda$ to be estimated,

$$\hat{\lambda} = \hat{f}_{\mathbf{z}}(0)\sqrt{2\pi},$$

and the local false discovery rate estimate is then the product of $\hat{\lambda}$ and ratio of the theoretical null density function to the estimated density function of the $z_i$ values,

$$\mathrm{LFDR}(p_i; \boldsymbol{p}) = \frac{\hat{\lambda}\varphi\left(\Phi^{-1}(p_i)\right)}{\hat{f}_{\mathbf{z}}\left(\Phi^{-1}(p_i)\right)}. \tag{1}$$

This expression is the ratio of the density of the null hypothesis component to the complete mixture density, and therefore estimates the proportion of equivalently expressed genes at any given p-value.

### 1.2 Frequentist shrinkage estimators

The two frequentist shrinkage estimators are,

$$Q_1(h, c) \equiv \mu_0 + \frac{\hat{\mu}_{\mathrm{MLE}} - \mu_0}{1 + hR} + cV, \tag{2}$$

and

$$Q_2(a, b, c) \equiv \hat{\mu}_{\mathrm{MLE}} - a(\hat{\mu}_{\mathrm{MLE}} - \mu_0)\exp\left(-\frac{b}{R}\right) + cV, \tag{3}$$

where $h$, $a$, $b$, and $c$ are tuning parameters chosen by the user and

$$R = \frac{S^2/n}{(\hat{\mu}_{\mathrm{MLE}} - \mu_0)^2}, \tag{4}$$

$$V \equiv \frac{S/\sqrt{n}}{1 + R} \times \mathrm{sign}(\hat{\mu}_{\mathrm{MLE}} - \mu_0). \tag{5}$$

It was shown that for univariate distributions, $Q_1$and $Q_2$ have smaller means squared error than the sample mean if the initial guess is appropriate, and that $Q_2(0.4, 0.01, 0.5)$ is suitable (in the sense of minimum MSE) for estimating the distribution mean when $|\mu - \mu_0| \leq 3.5\sigma/\sqrt{n}$.

These shrinkage estimators are suitable for use with paired data but must be modified for use with non-paired data. In that case, our goal is to shrink the estimate the difference of the means of two distributions toward 0. We replace the paired data MLE with the non-paired data MLE, set $S^2$ equal to the unbiased pooled variance estimator (which is appropriate under the assumption that both distributions have the same variance),

$$S^2 = \frac{(n' - 1)S_{x'}^2 + (n - 1)S_x^2}{n' + n - 2},$$

and replace $n$ in (4) and (5) with $n' + n$.

### 1.3 Bayesian shrinkage model

The model likelihood has the form,

$$f\left(\boldsymbol{w}|\sigma^2,\boldsymbol{\gamma},\boldsymbol{\mu}\right)=\prod_{i=1}^{M}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{\left(w_i-\mu_i\gamma_i\right)^2}{2\sigma^2}\right],$$

in which $w_i$ is the measured differential log-expression of the $i^{th}$ of $M$ genes (for paired data, $w_i=\bar{y}_i$; for non-paired data, $w_i=\bar{x}'_i-\bar{x}_i$) $\gamma_i$ is an indicator for differential expression (i.e., $\gamma_i=1$ indicates equivalent expression and $\gamma_i=0$ indicates indicates differential expression), $\mu_i$ is the log-expression level, and $\sigma^2$ is the sampling variance of the $w_i$ values (assumed equal for all $i$).
The prior distribution for the model is defined by the following distributions:

$$\begin{aligned}p\left(\sigma^2\right)&\propto\sigma^{-2},\\ p\left(\gamma_i=0|\pi\right)&=\pi,\\ p\left(\mu_i|\gamma_i=1,V,\sigma^2\right)&=\mathrm{N}\left(\mu_i|0,V\right),\end{aligned}$$

in which $\pi$ is the probability that a gene is equivalently expressed and $V$ is the variance of the normal prior distribution on $\mu_i$. Both of these parameters are estimated from the data, and therefore have prior distributions themselves,

$$\begin{aligned}p\left(\pi\right)&=1,\\ p\left(V|\sigma^2\right)&\propto\frac{1}{\sigma^2}\left(1+\frac{V}{\sigma^2}\right)^{-2}.\end{aligned}$$

The complete posterior distribution is,

$$p\left(\boldsymbol{\gamma},\boldsymbol{\mu},V,\sigma^2,\pi|\boldsymbol{w}\right)\propto\left(\sigma^2+V\right)^{-2}\cdot\left[\prod_{i=1}^{M}\pi^{1-\gamma_i}\left(1-\pi\right)^{\gamma_i}\right]\cdot f\left(\boldsymbol{w}|\sigma^2,\boldsymbol{\gamma},\boldsymbol{\mu}\right)\cdot\prod_{i|\gamma_i=1}\mathrm{N}\left(\mu_i|0,V\right).$$

The parameters $\boldsymbol{\gamma}$ can be summed over and the parameters $\boldsymbol{\mu}$ are analytically integrable, so they can be integrated out of the posterior distribution, leaving the marginal posterior distribution,

$$p\left(V,\sigma^2,\pi|\boldsymbol{w}\right)\propto\left(\sigma^2+V\right)^{-2}\cdot\left[\prod_{i=1}^{M}\pi^{1-\gamma_i}\left(1-\pi\right)^{\gamma_i}\right]\cdot f\left(\boldsymbol{w}|\sigma^2,\boldsymbol{\gamma},\boldsymbol{\mu}\right)\cdot\prod_{i|\gamma_i=1}\mathrm{N}\left(\mu_i|0,V\right).$$

This distribution can easily be sampled from with a Markov chain Monte Carlo (MCMC) algorithm; once posterior samples have been obtained, the key quantities of interest for estimating $\mu$ are,

$$p\left(\gamma_i = 0|V, \sigma^2, \pi, \boldsymbol{w}\right) = \left[1 + \frac{1-\pi}{\pi}\sqrt{\frac{\sigma^2}{\sigma^2+V}}\exp\left(\frac{w_i^2 V}{2\sigma^2\left(\sigma^2+V\right)}\right)\right]^{-1},$$

$$\mathrm{E}\left(\mu_i|\gamma_i = 1, V, \sigma^2, \pi, \boldsymbol{w}\right) = \frac{w_i V}{\sigma^2+V},$$

from which we can calculate the posterior distribution of $\mu_i$ unconditional on $\gamma_i$,

$$\begin{aligned}
&\mathrm{E}\left(\mu_i|V, \sigma^2, \pi, \boldsymbol{w}\right) = p\left(\gamma_i = 0|V, \sigma^2, \pi, \boldsymbol{w}\right)\mathrm{E}\left(\mu_i|\gamma_i = 0, V, \sigma^2, \pi, \boldsymbol{w}\right)\\
&+p\left(\gamma_i = 1|V, \sigma^2, \pi, \boldsymbol{w}\right)\mathrm{E}\left(\mu_i|\gamma_i = 1, V, \sigma^2, \pi, \boldsymbol{w}\right)\\
&= \left[1 - p\left(\gamma_i = 0|V, \sigma^2, \pi, \boldsymbol{w}\right)\right]\mathrm{E}\left(\mu_i|\gamma_i = 1, V, \sigma^2, \pi, \boldsymbol{w}\right) \qquad (6)\\
&= \left\{1 - \left[1 + \frac{1-\pi}{\pi}\sqrt{\frac{\sigma^2}{\sigma^2+V}}\exp\left(\frac{w_i^2 V}{2\sigma^2\left(\sigma^2+V\right)}\right)\right]^{-1}\right\}\cdot\frac{w_i V}{\sigma^2+V}.
\end{aligned}$$

Taking the average of $\mathrm{E}\left(\mu_i|V, \sigma^2, \pi, \boldsymbol{w}\right)$ over the MCMC samples of $\left(V, \sigma^2, \pi\right)$ yields a Monte Carlo estimate of $\mathrm{E}\left(\mu_i|\boldsymbol{w}\right)$, the expectation of the marginal posterior distribution of $\mu_i$.

## 1.4 Method for choosing simulation variances

For each real data set, we calculated a gene-wise estimate of the signal-to-noise ratio,

$$\mathrm{SNR}_i = \frac{|\hat{\mu}|_i}{\hat{\sigma}_i} = \frac{|\bar{x}'_i - \bar{x}_i|}{\sqrt{\frac{\sum_{j=1}^{n'}\left(x'_{i,j}-\bar{x}'_i\right)^2 + \sum_{j=1}^{n}\left(x_{i,j}-\bar{x}_i\right)^2}{n'+n-2}}},$$

that is, the ratio of the absolute value of the MLE and the square root of the usual unbiased pooled variance estimator. For a global signal-to-noise estimate, we desired a measure of the signal-to-noise ratio for the gene whose absolute value of the MLE was at the 90th quantile. Simply taking the $\mathrm{SNR}_i$ value of the gene with $|\hat{\mu}|_i = |\hat{\mu}|_{0.9}$ (numerical subscripts specifying quantiles) would not be robust, so we took the median of the estimated signal-to-noise ratios for the set of genes whose absolute value of the MLE was between the 85th and 95th quantile,

$$\mathrm{SNR}_{\mathrm{global}} = \mathrm{median}\left(\left\{\mathrm{SNR}_k \,|\, |\hat{\mu}|_{0.85} \le |\hat{\mu}|_k \le |\hat{\mu}|_{0.95}\right\}\right).$$

We calculated this global signal-to-ratio estimate for each real data set, and used the value that was smallest to set the variance of the simulations according to follows,

$$\frac{|\mu_{\mathrm{sim}}|_{0.9}}{\sigma_{\mathrm{sim}}} = \mathrm{SNR}_{\mathrm{global}},$$

which when rearranged gives,

$$\sigma^2_{\text{sim}} = \left( \frac{|\mu_{\text{sim}}|_{0.9}}{\text{SNR}_{\text{global}}} \right)^2 .$$

### 1.5 Derivation of the expression for the posterior expected squared error

The main text contains the equation,

$$\mathrm{E}_{\,\text{posterior}} \left[ (\mu - \hat{\mu})^2 \right] = \left[ \mathrm{E}_{\,\text{posterior}} \left( \mu \right) - \hat{\mu} \right]^2 + \text{var}_{\,\text{posterior}} \left( \mu \right) , \tag{7}$$

whose derivation is as follows:

$$\begin{aligned}
\mathrm{E}_{\,\text{posterior}} \left[ (\mu - \hat{\mu})^2 \right] &= \mathrm{E}_{\,\text{posterior}} \left( \hat{\mu}^2 - 2\mu\hat{\mu} + \mu^2 \right) \\
&= \hat{\mu}^2 - 2\hat{\mu}\, \mathrm{E}_{\,\text{posterior}} \left( \mu \right) + \mathrm{E}_{\,\text{posterior}} \left( \mu^2 \right) \\
&= \hat{\mu}^2 - 2\hat{\mu}\, \mathrm{E}_{\,\text{posterior}} \left( \mu \right) + \left[ \mathrm{E}_{\,\text{posterior}} \left( \mu \right) \right]^2 - \left[ \mathrm{E}_{\,\text{posterior}} \left( \mu \right) \right]^2 + \mathrm{E}_{\,\text{posterior}} \left( \mu^2 \right) \\
&= \left[ \mathrm{E}_{\,\text{posterior}} \left( \mu \right) - \hat{\mu} \right]^2 + \text{var}_{\,\text{posterior}} \left( \mu \right) .
\end{aligned}$$

## 2 Quantification of variance due to sampling

$$\begin{aligned}
\widehat{\text{risk}} &= \sum_{i=1}^{M} \widehat{\text{MSE}}_i \\
&= \frac{1}{B} \sum_{b=1}^{B} \left( \sum_{i=1}^{M} \left( \widehat{\mu}_{b,i} - \mu_i \right)^2 \right) \\
&= \frac{1}{B} \sum_{b=1}^{B} \text{sse}_b,
\end{aligned}$$

where

$$\text{sse}_b = \sum_{i=1}^{M} \left( \widehat{\mu}_{b,i} - \mu_i \right)^2 .$$

To quantify uncertainty in $\widehat{\text{risk}}$ over simulations, we considered its variance $\text{var}\,\widehat{\text{risk}}$, which we estimated by

$$\widehat{\text{var}\,\widehat{\text{risk}}} = \frac{1}{B} \left( \frac{1}{B-1} \sum_{b=1}^{B} \left( \text{sse}_b - \overline{\text{sse}_b} \right)^2 \right),$$

where

$$\overline{\text{sse}_b} = \frac{\sum_{b=1}^{B} \text{sse}_b}{B}.$$

We computed the ratio

$$\frac{\widehat{\operatorname{var}\widehat{\text{risk}}}}{\left(\widehat{\text{risk}}\right)^2}$$

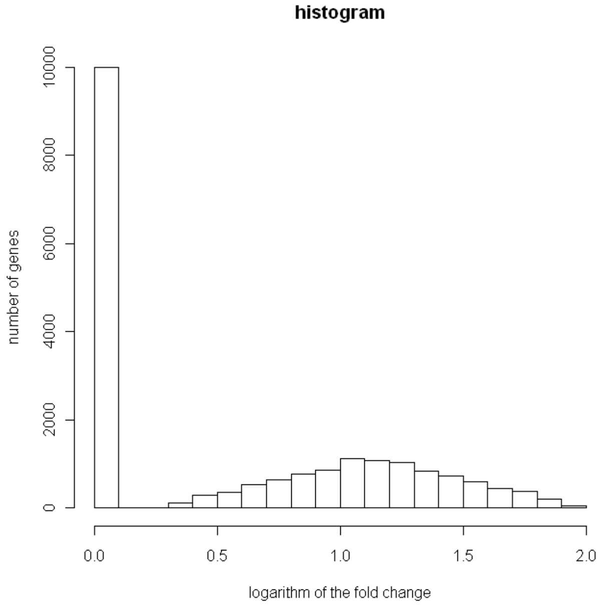
for each estimator and each set of simulations (Table 1). For estimators satisfying independence, uncertainty in the risk was negligible in the sense that $\widehat{\operatorname{var}\widehat{\text{risk}}} \ll \left(\widehat{\text{risk}}\right)^2$.

| Estimation | Case $I$, $n = 2$ | Case $I$, $= 8$ | Case $II$, $n = 2$ | Case $II$, $n = 8$ |
|---|---|---|---|---|
| $Q_1$ shrinkage | 1.6e-08 | 2.1e-08 | 5.9e-09 | 1.4e-06 |
| $Q_2$ shrinkage | 1.0e-08 | 4.4e-08 | 2.3e-08 | 1.3e-07 |
| MLE | 4.8e-09 | 5.0e-09 | 5.0e-09 | 5.0e-09 |
| Baysian shrinkage | 3.2e-07 | 9.5e-09 | 7.7e-07 | 8.7e-07 |
| Local FDR shrinkage | 5.3e-07 | 1.9e-07 | 3.9e-06 | 9.7e-07 |
| local FDR HT | 8.5e-07 | 3.7e-07 | 3.9e-06 | 2.2e-06 |
| raw p-value HT | 7.0e-07 | 2.7e-07 | 2.9e-06 | 6.2e-07 |
| adj.p-value HT | 8.5e-07 | 5.4e-07 | 3.8e-06 | 4.3e-06 |

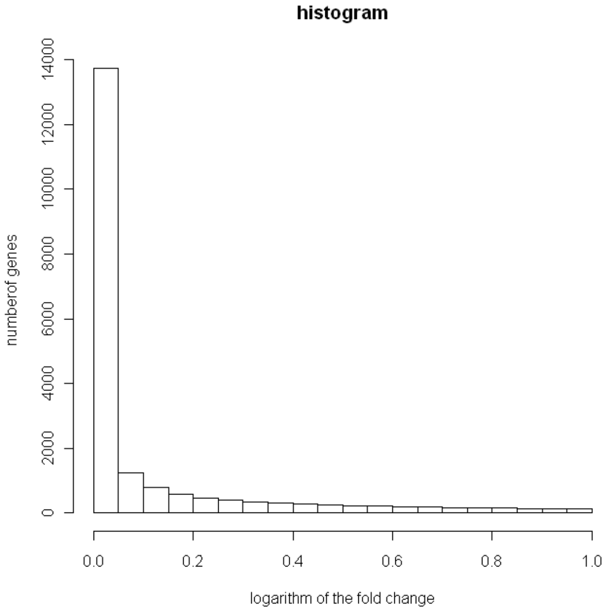
Table 1: Estimation of variance for risk estimate for each estimator and each set of simulation. Case I is the case that half of genes are differentially expressed and case II refers to the case of all genes with at least some differential expression.

## 3 Distributions of true means of simulated data

The two figures below show the logarithms of fold change values used to generate the simulated data.



Logarithm of the true fold change for the simulation of 20,000 genes where half of genes are equivalently expressed.

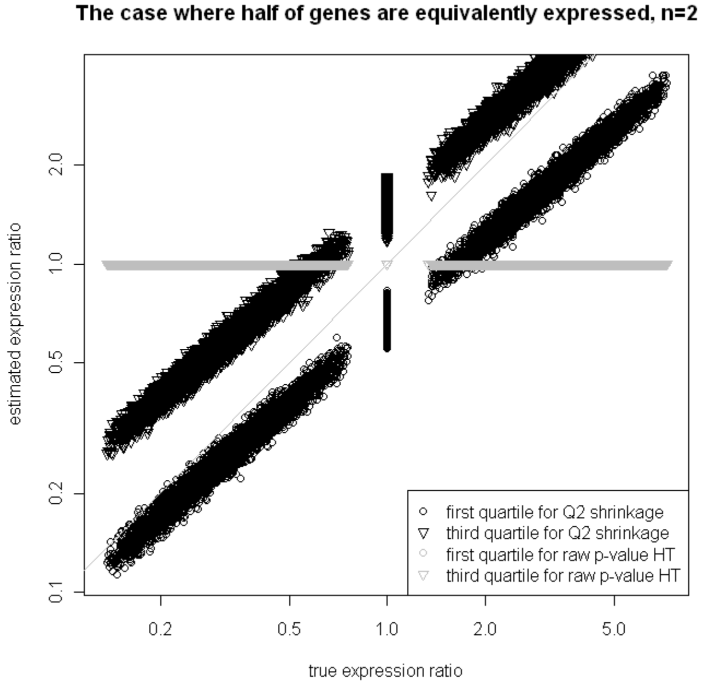


Logarithm of the true fold change for the simulation of 20,000 genes where all genes have at least some differential expression.

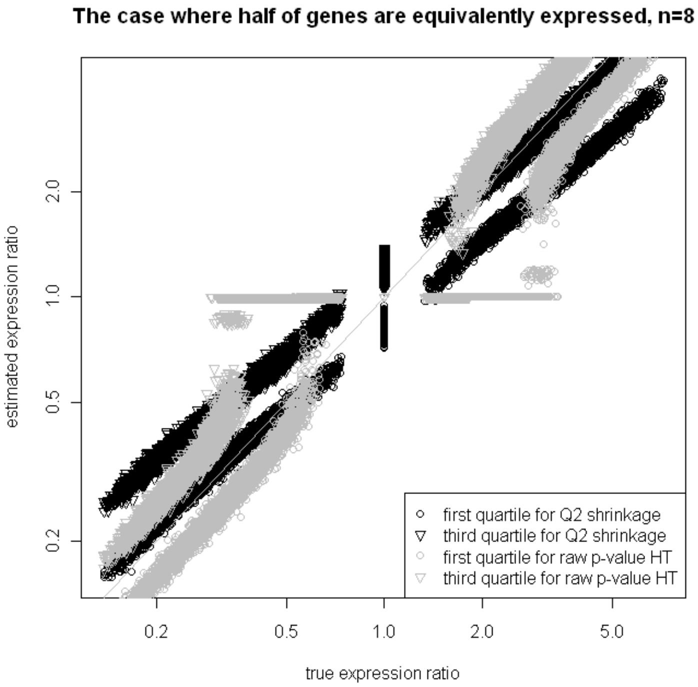
# 4 Comparison of shrinkage and hard-threshold estimation

All of the figures in this section compare the performance of the one of the best shrinkage estimators, $Q_2$, to the generally best-performing hard-threshold estimator, the raw p-value hard-threshold estimator. For the simulations, for each gene the estimators are summarized by their first and third quartile over the 200 simulations. These quartiles are plotted against the true expression ratio. For the real data sets, the two estimators are plotted on the $y$-axis and the MLE is plotted on the $x$-axis.

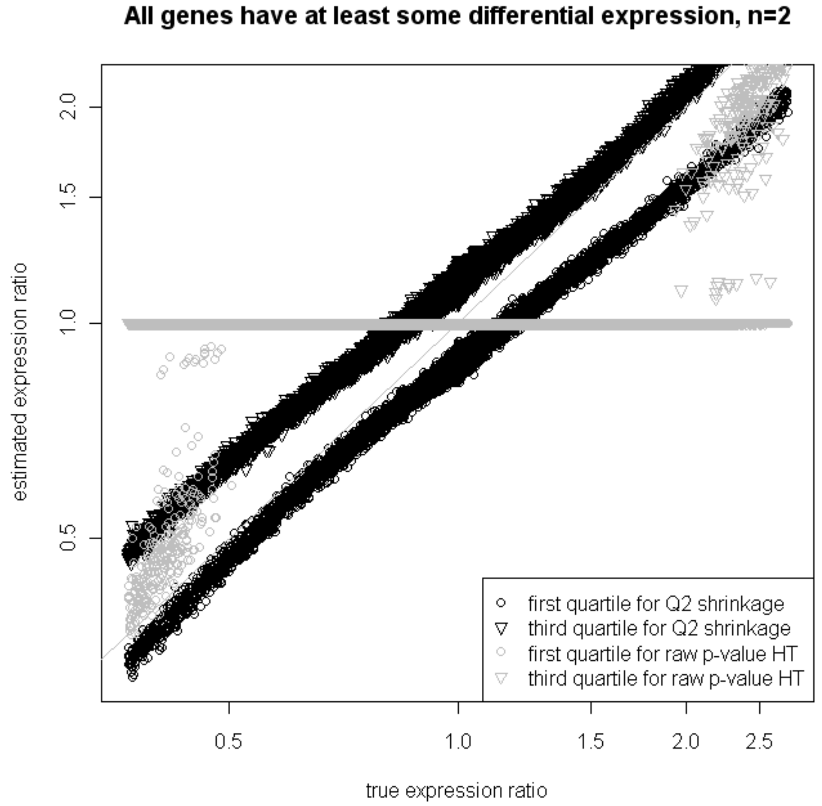
## 4.1 Simulations



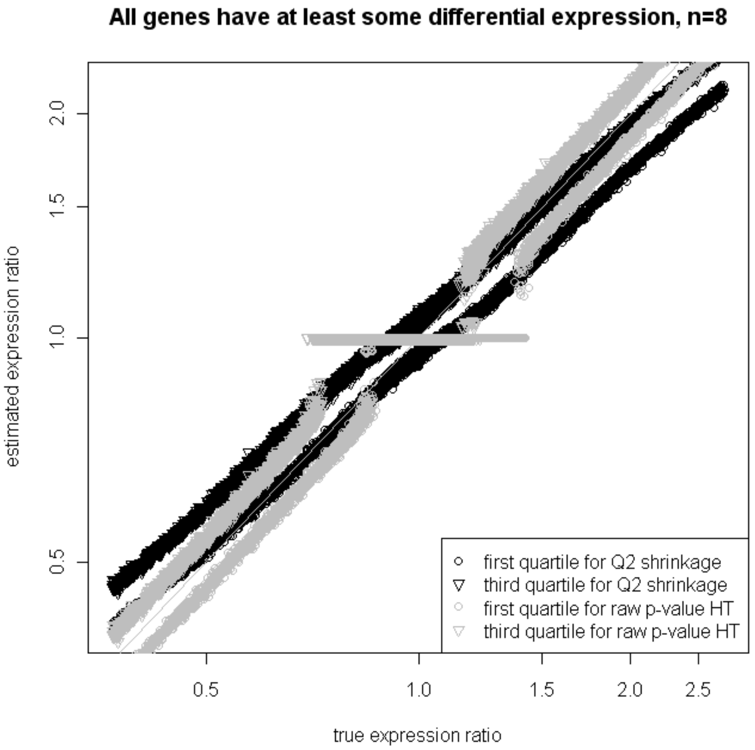
True expression ratio versus the first and third quartiles of both estimators over the 200 simulations for the case where half of genes are equivalently expressed and sample size is 2. When the sample size is small, the hard-threshold estimator cannot detect any differential expression.



True expression ratio versus the first and third quartiles of both estimators over the 200 simulations for the case where half of genes are equivalently expressed and sample size is 8. The hard-threshold estimator tends to collapse to zero at fold changes between about 1.5 and 3.0. For fold changes greater than 3.0 it tracks the true expression ratio. The shrinkage estimator has a bias that increases with the fold change.



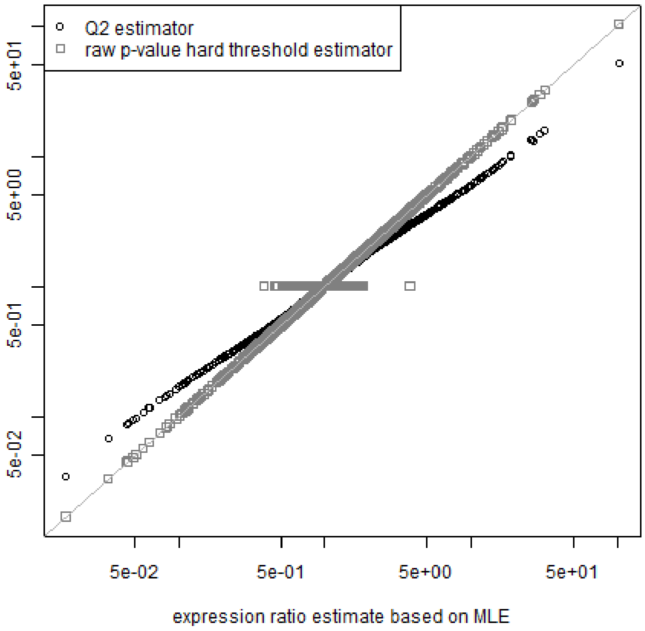
True expression ratio versus the first and third quartiles of both estimators over the 200 simulations for the case where all genes have at least some differential expression and sample size is 2. The three plots cover different ranges of fold changes. The hard-threshold estimator treats fold change below 10 as equivalently expressed and cannot consistently detect differential expression until the fold change is very large. The shrinkage estimator is biased away from zero when the true fold change is zero, but tracks the true fold change fairly closely once the fold change becomes large.



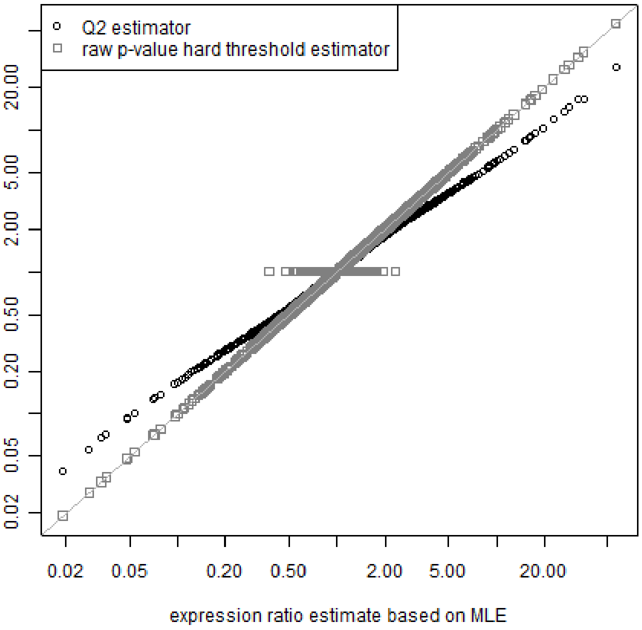
True expression ratio versus the first and third quartiles of both estimators over the 200 simulations for the case where all genes have at least some differential expression and sample size is 8. The three plots cover different ranges of fold changes. The hard-threshold estimator treats fold change below 1.5 as equivalently expressed and cannot consistently detect differential expression until the fold change is greater than around 3. The shrinkage estimator is biased away from zero when the true fold change is zero, but tracks the true fold change fairly closely once the fold change becomes large.

## 4.2 Real data sets

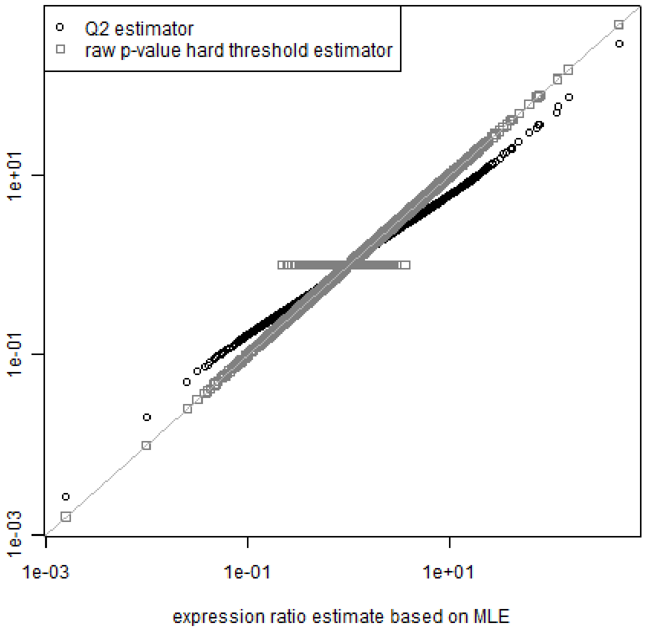
All figures of the real data sets display the same pattern: the shrinkage estimator is a biased version of the MLE, and there is some region within which the hard-threshold estimator may give an estimate of equivalent expression and beyond which the hard-threshold estimator is equal to the MLE.



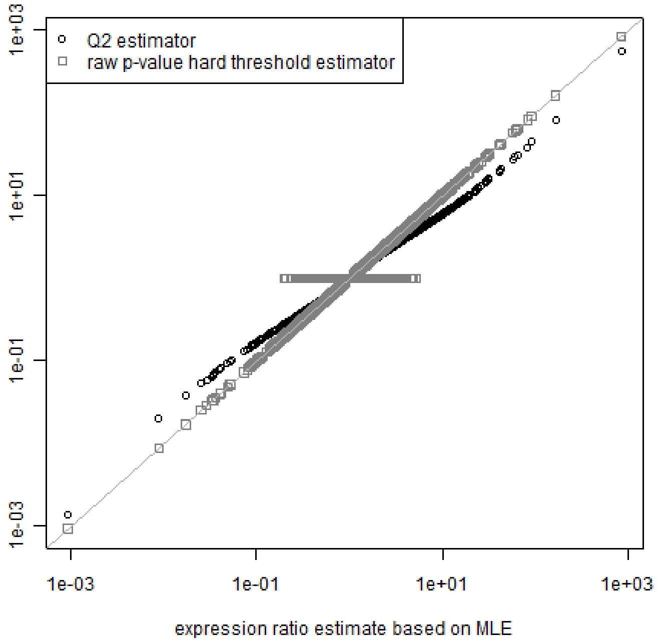
Maximum likelihood estimate versus both estimators for the first Affymetrix data set of the Microarray Quality Control experiment.



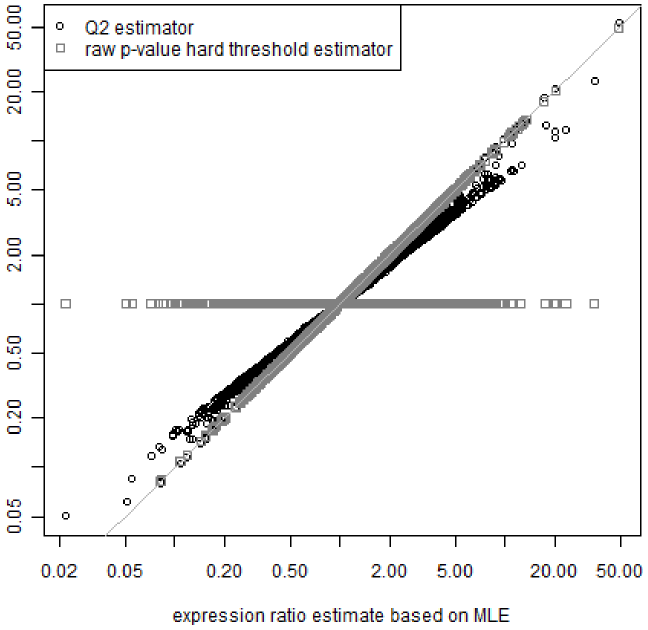
Maximum likelihood estimate versus both estimators for the second Affymetrix data set of the Microarray Quality Control experiment.



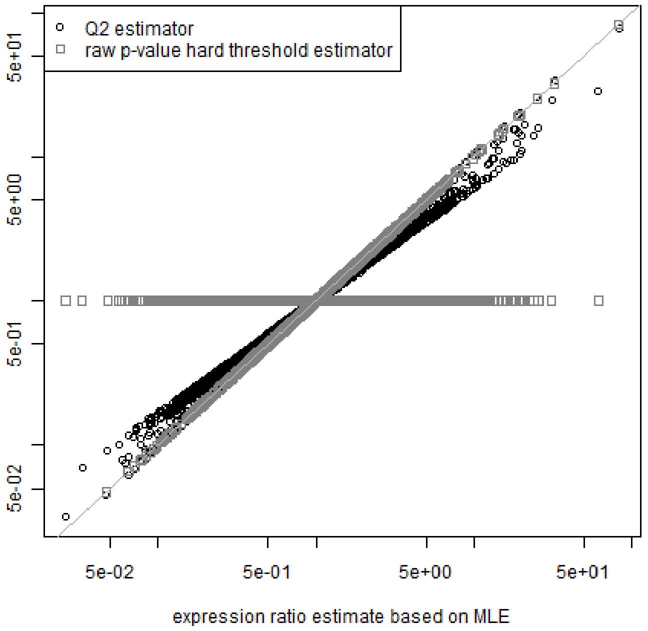
Maximum likelihood estimate versus both estimators for the Applied Biosystems data set of the Microarray Quality Control experiment.



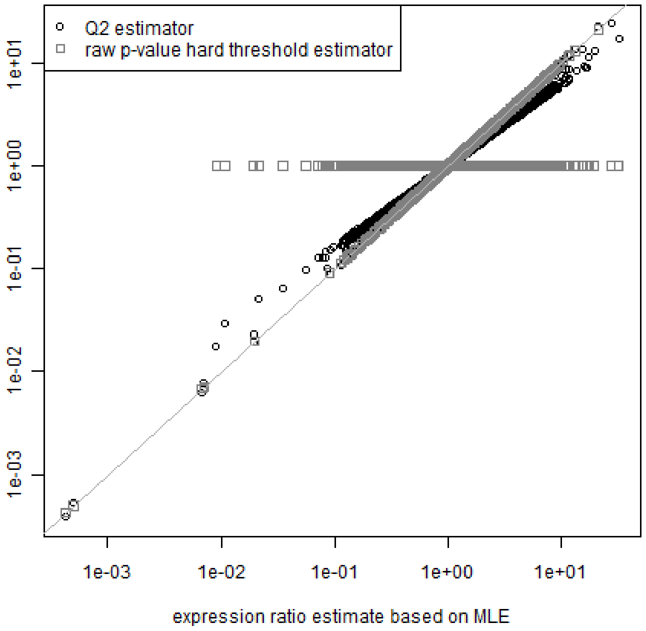
Maximum likelihood estimate versus both estimators for the Agilent data set of the Microarray Quality Control experiment.



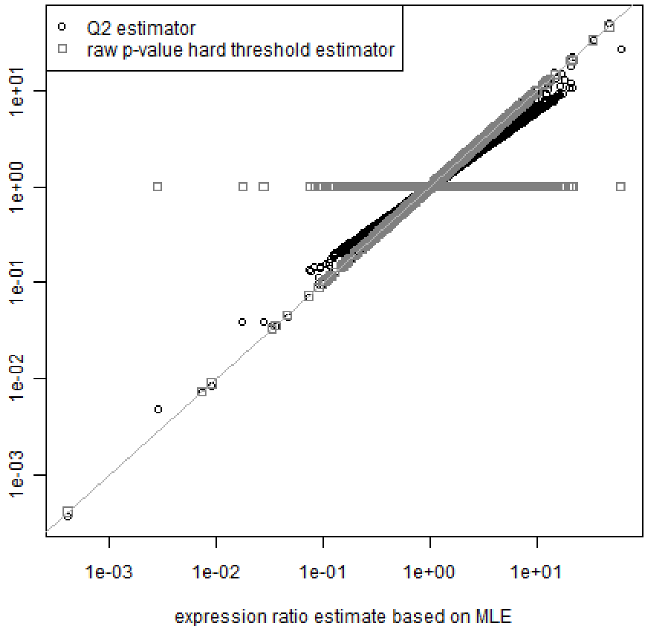
Maximum likelihood estimate versus both estimators for the 10 hours after treatment versus control at 10 hours data set of the breast cancer cell line experiment.



Maximum likelihood estimate versus both estimators for the 48 hours after treatment versus control at 48 hours data set of the breast cancer cell line experiment.



Maximum likelihood estimate versus both estimators for the 48 hours after treatment versus 10 hours after treatment data set of the breast cancer cell line experiment.



Maximum likelihood estimate versus both estimators for the control at 48 hours versus control at 10 hours data set of the breast cancer cell line experiment.