Biostatistics (2010), **11**, 2, *pp.* 281–289
doi:10.1093/biostatistics/kxp047
Advance Access publication on November 27, 2009

C

Bayesian ranking and selection methods using hierarchical mixture models in microarray studies

HISASHI NOMA*

Department of Biostatistics, Kyoto University School of Public Health, Yoshida Konoe-cho, Sakyo-ku, Kyoto 606-8501, Japan
nomahi@bstat.mbox.media.kyoto-u.ac.jp

SHIGEYUKI MATSUI

Department of Data Science, The Institute of Statistical Mathematics, 10-3 Midori-cho, Tachikawa, Tokyo 190-8562, Japan

TAKASHI OMORI, TOSIYA SATO

Department of Biostatistics, Kyoto University School of Public Health, Yoshida Konoe-cho, Sakyo-ku, Kyoto 606-8501, Japan

SUMMARY

The main purpose of microarray studies is screening to identify differentially expressed genes as candidates for further investigation. Because of limited resources in this stage, prioritizing or ranking genes is a relevant statistical task in microarray studies. In this article, we develop 3 empirical Bayes methods for gene ranking on the basis of differential expression, using hierarchical mixture models. These methods are based on (i) minimizing mean squared errors of estimation for parameters, (ii) minimizing mean squared errors of estimation for ranks of parameters, and (iii) maximizing sensitivity in selecting prespecified numbers of differential genes, with the largest effect. Our methods incorporate the mixture structures of differential and nondifferential components in empirical Bayes models to allow information borrowing across differential genes, with separation from nuisance, nondifferential genes. The accuracy of our ranking methods is compared with that of conventional methods through simulation studies. An application to a clinical study for breast cancer is provided.

Keywords: Empirical Bayes; Gene expression; Hierarchical mixture models; Microarrays; Ranking and selection.

1. INTRODUCTION

Recent developments in gene expression microarrays have enabled comprehensive screening of differentially expressed genes between different clinical subclasses. In genome-wide studies with microarrays, multiple testing is widely adopted, and statistically significant genes are reported as candidate genes

*To whom correspondence should be addressed.

for further investigation. Because of limited resources at this stage, prioritizing or ranking genes is needed.

There are at least 2 criteria on which genes are ranked; one is related to the probability of nondifferential expression and the other is related to the magnitude of differential expression or the strength of association with the clinical outcome. An example of the former criterion is the local false discovery rate, which represents the posterior probability of nondifferential expression of each gene (Efron *and others*, 2001; Efron, 2008). A similar or related statistic using hierarchical Bayes models was derived and discussed by Newton *and others* (2004). The ranking based on the local false discovery rate is optimal for selecting differentially expressed genes from the viewpoint of Bayesian decision theory (Berger, 1985; McLachlan *and others*, 2006).

For the latter criterion regarding magnitude of association, the fold change—which corresponds to the ratio or difference of mean expression levels between different clinical subtype classes—is commonly used (McLachlan *and others*, 2004; Guo *and others*, 2006; Choe *and others*, 2005). Some authors have reported that gene ranking based on the fold change is reproducible (MicroArray Quality Control [MAQC] Consortium, 2006) and accurate for large absolute changes in gene expression (Witten and Tibshirani, in preparation [Available on web: http://www-stat.stanford.edu/∼tibs/ftp/FCTComparison.pdf.]). However, the accuracy of gene ranking can be improved by "borrowing strength" across genes. Specifically, the ranking of empirical Bayes estimators can be more accurate than that of conventional statistics such as maximum likelihood estimators (Laird and Louis, 1989). Laird and Louis (1989) and Shen and Louis (1998) discussed optimal ranking in empirical Bayes inference. Lin *and others* (2006) discussed various loss functions in Bayesian optimal ranking and selection rules and derived optimal rules for selecting top-ranked features.

An important feature of microarray data is that a large proportion of the genes investigated are nondifferential. The incorporation of this feature of microarray data could allow for information sharing across differential genes separated from nuisance, nondifferential genes. In this article, we propose empirical Bayes methods for gene ranking on the basis of the strength of association under structural models that reflect this feature. Specifically, we assume a hierarchical mixture model with a 2-stage compound sampling model, in which the second stage of the model is a mixture distribution of differential and nondifferential components (Newton and Kendziorski, 2003; Lönnstedt and Speed, 2002; Gottardo *and others*, 2003).

We present the framework of the hierarchical mixture modeling and its empirical Bayes inference in Section 2, and develop 3 ranking and selection rules in Section 3. In Section 4, we compare the proposed rules with conventional or other methods through simulations. We describe the application to a breast cancer clinical study in Section 5. Discussion is provided in Section 6.

2. HIERARCHICAL MIXTURE MODELING

The gene expression data considered here comprise normalized log ratios from 2-color complementary DNA arrays or normalized log signals from oligonucleotide arrays (e.g. Affymetrix GeneChip). We consider a 2-class comparison problem: a binary response—for example, a poor prognosis and good prognosis, is compared on the basis of the expression levels of m candidate genes from n samples. For gene j, let θ_j be the parameter of interest, that is, the difference in the mean expression level between the 2 classes ($j = 1, \ldots, m$). As an estimator of θ_j, let Y_j be the fold change, which is the difference in the sample mean expression level obtained from n samples (Guo *and others*, 2006; Choe *and others*, 2005). We consider a 2-stage model with i.i.d. sampling from a 3-component mixture prior and from a normal gene-specific sampling model:

$$\begin{aligned} Y_j|\theta_j &\sim \mathrm{N}(\theta_j, \sigma_j^2), \\ \theta_j &\sim \pi_0\delta(\theta) + \pi_1 g_1(\theta|\xi_1) + \pi_2 g_2(\theta|\xi_2). \end{aligned} \tag{2.1}$$

Here, $\delta(\theta)$ is the Dirac delta function, representing nondifferential expression between 2 classes. The density functions $g_1(\theta|\boldsymbol{\xi}_1)$ and $g_2(\theta|\boldsymbol{\xi}_2)$ correspond to the "nonnull" components of underexpression and overexpression, respectively, for a particular class, for example, poor prognosis. The proportion π_i represents the mixing proportion ($i = 0, 1, 2$), with $\pi_0 + \pi_1 + \pi_2 = 1$. In the first-stage model in (2.1), we assume that the gene-specific variance σ_j^2 is known. We denote $Z_{ij} (i = 0, 1, 2;\ j = 1, 2, \ldots, m)$ as unobservable indicator random variables, such that $Z_{ij} = 1$ if gene j belongs to the ith component, and $Z_{ij} = 0$ otherwise. The π_i ($i = 0, 1, 2$) correspond to the probability of $Z_{ij} = 1$. An estimate of the hyperparameter $\boldsymbol{\eta} = (\pi_0,\ \pi_1,\ \boldsymbol{\xi}_1,\ \boldsymbol{\xi}_2)$ can be obtained by maximizing the marginal likelihood of Y_j (Carlin and Louis, 2009). We employ the expectation-maximization (EM) algorithm (Dempster *and others*, 1977) to cope with the unobservable indicator variable Z_{ij} in the mixture model. The hyperparameters of interest, that is, $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$ in the distribution of the effect sizes θ_j, can be estimated more stably by using fixed or, more generally, constrained estimates of the mixing proportions or by introducing prior distributions on the mixing proportions in the EM algorithm (e.g. Gottardo *and others*, 2003; Newton *and others*, 2001, 2004; Lo and Gottardo, 2007). In this article, we consider the former strategy and adopt a 2-stage procedure that estimates π_0 as in Storey (2002) and treat this estimate as a fixed value in the EM algorithm. We assume conjugate normal distributions $\mathrm{N}(\mu_1, \tau_1^2)$ and $\mathrm{N}(\mu_2, \tau_2^2)$ ($\mu_1 > 0$, $\mu_2 < 0$) for $g_1(\theta|\boldsymbol{\xi}_1)$ and $g_2(\theta|\boldsymbol{\xi}_2)$, respectively.

3. Posterior inference

For simplicity, our discussion here is restricted to identifying genes with the greatest positive θ_j, that is, overexpressed genes for a particular class. Its extension to the two-sided version—to obtain genes with the greatest absolute θ_j for both overexpressed and underexpressed genes—is straightforward.

3.1 *Ranking based on posterior mean*

From model (2.1), we obtain the posterior distribution:

$$p_j(\theta|y_j) = \Pr(Z_{0j} = 1|y_j)\, p_{0j}(\theta|y_j) + \Pr(Z_{1j} = 1|y_j)\, p_{1j}(\theta|y_j) + \Pr(Z_{2j} = 1|y_j)\, p_{2j}(\theta|y_j),$$

where p_{0j}, p_{1j}, and p_{2j} are the posterior densities of each component, which are obtained as $\delta(\theta)$ and

$$N\left(\frac{\tau_i^2 y_j + \sigma_j^2 \mu_i}{\tau_i^2 + \sigma_j^2}, \frac{\tau_i^2 \sigma_j^2}{\tau_i^2 + \sigma_j^2}\right), \quad (i = 1, 2),$$

respectively. The posterior probability that gene j belongs to the ith component ($i = 0, 1, 2$) is

$$\Pr(Z_{ij} = 1|y_j) = \frac{\pi_i h_{ij}(y_j|\boldsymbol{\xi}_i)}{\pi_0 h_{0j}(y_j) + \pi_1 h_{1j}(y_j|\boldsymbol{\xi}_1) + \pi_2 h_{2j}(y_j|\boldsymbol{\xi}_2)}. \tag{3.1}$$

Here, h_{0j}, h_{1j}, and h_{2j} are the marginal densities of the Y_js, in each component, namely $\mathrm{N}(0, \sigma_j^2)$, $\mathrm{N}(\mu_1, \sigma_j^2 + \tau_1^2)$, and $\mathrm{N}(\mu_2, \sigma_j^2 + \tau_2^2)$ for $i = 0, 1, 2$, respectively.

Minimizing the squared error loss, the Bayes estimator of θ_j is the posterior mean (PM) (Carlin and Louis, 2009). For the hierarchical mixture model, the PM is obtained as

$$\begin{aligned} E[\theta_j|Y_j] = {} & \Pr(Z_{0j} = 1|y_j) E[\theta_j|Z_{0j} = 1, Y_j] + \Pr(Z_{1j} = 1|y_j) E[\theta_j|Z_{1j} = 1, Y_j] \\ & + \Pr(Z_{2j} = 1|y_j) E[\theta_j|Z_{2j} = 1, Y_j]. \end{aligned} \tag{3.2}$$

This is the weighted average of the PMs of each component, where the weights are the posterior probabilities of component membership (3.1). The ranking is thus obtained via the magnitude of the PMs (the "PM" method).

3.2 *Ranking based on rank posterior means*

If the ranks of the parameters are the target feature, using the rank estimator is more appropriate than using the parameter estimator (Laird and Louis, 1989; Shen and Louis, 1998; Louis and Shen, 1999). We consider ranking within differential genes with positive effects, defined as

$$R_j = Z_{1j} \sum_{k=1}^{m} Z_{1k} I(\theta_j \geqslant \theta_k),$$

where R_j has a large value when gene j belongs to the nonnull component with positive effect and has a large θ_j value or is 0 when gene j belongs to the other components. For the squared error loss, the Bayes estimator of R_j is the PM:

$$E[R_j|y_j] = \Pr(Z_{1j} = 1 \mid y_j) \sum_{k=1}^{m} \Pr(Z_{1k} = 1 \mid y_j) \Pr(\theta_j \geqslant \theta_k | y_j, y_k). \tag{3.3}$$

Thus, gene ranking is obtained by the rank posterior means (the "RPM" method).

3.3 *Ranking based on tail-area posterior probability*

Because of the limited resources available in subsequent studies, the number of selected genes may be prespecified as a small number, K (Matsui *and others*, 2008). In this situation, we consider the K genes with the greatest positive effects as the target. Lin *and others* (2006) provided various loss functions and derived optimal ranking and selection rules via Bayesian decision theory; we generalize their rank-based misclassification loss function that equivalently penalizes to misclassifications between the true top K ranked genes and other differential genes:

$$L_{0/1}(K, R, R^{\text{est}}) = \frac{1}{m} \sum_{j=1}^{m} \{FP(K, R_j, R_j^{\text{est}}) + FN(K, R_j, R_j^{\text{est}})\},$$

where

$$FP(K, R_j, R_j^{\text{est}}) = I\{R_j \leqslant m_1 - K, R_j^{\text{est}} > m_1 - K\},$$

$$FN(K, R_j, R_j^{\text{est}}) = I\{R_j > m_1 - K, R_j^{\text{est}} \leqslant m_1 - K\}.$$

The R_j^{est} are estimators of the R_j, and $m_1 = \sum_j Z_{1j}$ is the number of genes generated from the nonnull component with positive effects in the m measured genes.

The derived optimal rule is to select K genes which have large values of

$$\tilde{P}_j(K) = \Pr(R_j > (m_1 - K)|\mathbf{y}),$$

as in Lin *and others* (2006). However, it is difficult to obtain the posterior distribution of R_j and to calculate $\tilde{P}_j(K)$ directly. Instead, we can use a simple computable approximation of $\tilde{P}_j(K)$. Define

$\gamma = K/(m_1 + 1)$ and let G_1 be the cumulative distribution function of g_1. Under the similar conditions of Theorem 5 described in Lin *and others* (2006), the approximation of $\tilde{P}_j(K)$ is obtained as

$$P*_j(K) = \Pr(\theta_j \geqslant \overline{G}_1^{-1}(\gamma)|Z_{1j} = 1, \mathbf{y})\Pr(Z_{1j} = 1|\mathbf{y}), \tag{3.4}$$

where

$$\overline{G}_1(t) = \frac{\sum_{j=1}^{m} \Pr(\theta_j \leqslant t|z_{1j} = 1, y_j)\Pr(Z_{1j} = 1|y_j)}{\sum_{j=1}^{m} \Pr(Z_{1j} = 1|y_j)}.$$

$P*_j(K)$ corresponds to the tail-area posterior probability (TPP) of θ_j. Proof of this approximation is provided in the supplementary material available at *Biostatistics* online. Here, we denote this rule as the "TPP" method. The quantity m_1 can be replaced by its estimator $\sum_j \Pr(Z_{1j} = 1|y_j)$ (McLachlan *and others*, 2004).

4. Simulation studies

We conducted a series of simulation studies to assess the performance of our proposed methods. Details of the simulations are presented in Section A of the supplementary material available at *Biostatistics* online.

In summary, the sensitivity and root mean squared error (RMSE) of all the proposed methods were better than those of the other methods. As was expected, the TPP method had the greatest sensitivity, the RPM method had the lowest RMSE values, and the posterior probability of differentially expressed (PPDE) (3.1) ranking had the lowest false-positive rate. The posterior mean under unimodal hierarchical model (PM_U) method, without invoking mixture models, had a lower sensitivity and a larger RMSE, compared with the PM method. The fold change had comparable sensitivity and RMSE values for large sample sizes, but not for small sample sizes. The fold change had very large false-positive rates even when the sample size was large because it does not guard against selecting null genes.

5. Application to a breast cancer study

We illustrate the proposed methods using the data set from a breast cancer clinical study (Wang *and others*, 2005). The data are available from the NCBI GEO database (GSE2034). This study was a large Affymetrix-based gene expression profiling study of 286 untreated patients with lymph node–negative primary breast cancer and analyzed estrogen receptor positive– and negative–patients separately. Here, we restrict our attention to the estrogen receptor positive patients. In this study, out of 22 283 genes, 60 genes were selected on the basis of statistical significance for predicting the risk of relapse. We considered comparison of patients who were relapse-free at 5 years (good prognostic group) and the other patients (poor prognostic group). During the follow-up period, of the 204 estrogen receptor positive patients, 138 patients were relapse-free at 5 years, while 66 developed distant metastasis.

The hyperparameters were estimated as $\hat{\pi}_0 = 0.769$, $\hat{\pi}_1 = 0.055$, $\hat{\pi}_2 = 0.176$, $\hat{\mu}_1 = 0.182$, $\hat{\tau}_1^2 = 0.021^2$, $\hat{\mu}_2 = -0.149$ and $\hat{\tau}_2^2 = 0.014^2$. The σ_j^2 were estimated on a gene-by-gene basis assuming a common variance between the 2 prognostic groups. Figure 1 represents comparison between the RPM statistic, which was shown to yield good gene ranking in the simulation studies in Section 4, and the other statistics for overexpressed genes in the poor prognostic group. Similar trends were observed in underexpressed genes for the poor prognostic group. Figure 1(a) indicates substantial discrepancy in gene ranking between the RPM statistic and the fold change. In Figure 1(b), top-ranked genes based on the RPM statistics had the smallest P-values, but low-ranked genes using the RPM statistic could also have very small P-values. Figures 1(c) and (d) indicate good agreement of gene ranking among the PM, RPM, and TPP statistics, especially, for the greatest values of these statistics (i.e. for top-ranked genes). For

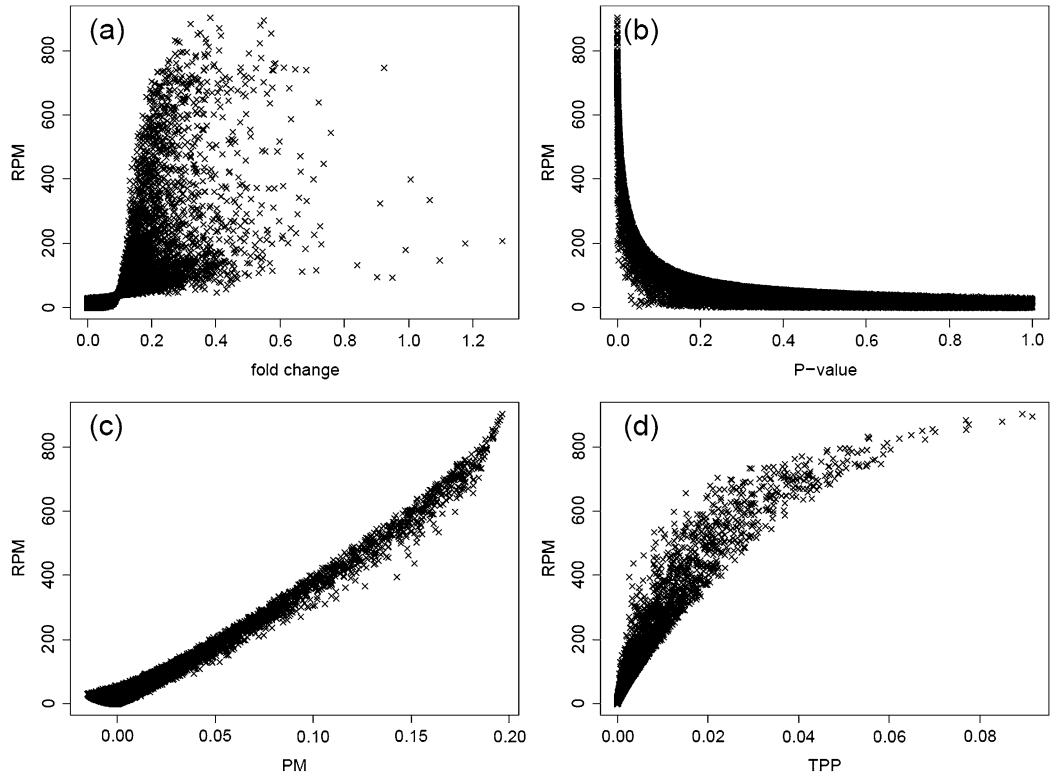

Fig. 1. Comparison of the RPM statistic with other statistics for the breast cancer data set. The 4 panels show scatter plots of the RPM statistic versus the fold change (a), the P-values from two-sample t-tests (b), the PM statistic (c), and the TPP statistic with $K = 30$ (d), for overexpressed genes in the poor prognostic group.

example, out of top 30 genes based on the RPM statistic, 27 and 21 genes were also selected in top 30 genes based on the PM and TPP statistics, respectively. The discrepancy in gene ranking between the proposed methods and fold change in Figure 1(a) would reflect very high false-positive rates for fold change as we found in our simulation study.

We also investigated overlap of top genes between the proposed methods and the 60 genes reported in the original paper (see Section B of the supplementary material available at *Biostatistics* online). There were 7 overlaps when the top 30 genes were selected by the RPM method for each of overexpression and underexpression in the poor prognostic group. 26 (43%) of the 60 genes reported in the original paper were among the top 1% of genes according to their RPM values. As indicated by Figure 1(b), there were fewer overlaps between the proposed methods and the t statistic. Because the sensitivity of the proposed methods was higher, gene ranking based on the proposed methods would be more reliable.

6. DISCUSSION

In microarray studies, prioritizing or ranking genes is an important statistical task. Because the number of simultaneous comparisons can go into the tens of thousands, gene ranking can suffer from a lack of accuracy. Sharing information across genes and incorporating the null/nonnull mixture structure are expected to be effective in improving accuracy. As seen in our simulations, the proposed PM, RPM, and TPP methods showed higher sensitivity and lower error of rank estimation for differential genes with

the greatest effects, compared to conventional methods. In addition, these methods had low false-positive rates. Although the results of our resampling studies indicated that such good performance can be compromised by violating the model assumptions, including independence among genes and the normality of gene expressions, our ranking methods, especially the RPM method, were still accurate compared to the other methods. Violation of normality can be handled by using other parametric distributions, as noted in Section 2. The greater accuracy of the PM method, compared to the PM_U method, shows that incorporating the null mixture component is effective in improving ranking accuracy. The PPDE ranking had the lowest false-positive rate, as was expected because of its theoretical optimality (Berger, 1985; McLachlan *and others*, 2006). These results are reasonable because a ranking method performs well for the criterion or loss function in gene ranking from which it is derived. Ranking methods should be selected according to the criterion of interest in gene ranking, that is, depending on the probability of nondifferential expression or the magnitude of differential expression.

The ranking accuracy of fold change would be asymptotically optimal because of the good performance in sensitivity and RMSE of gene ranking under large sample scenarios ($n/2 = 80$) in the simulations described in the supplementary material available at *Biostatistics* online. Further, the RMSE of fold change was lower than the proposed empirical Bayes methods under large sample scenarios in the resampling exercise in the supplementary material available at *Biostatistics* online. A similarly good performance using conventional methods, compared to those of empirical Bayes methods, has also been seen with large samples in other experiments (Greenland, 1993). In small sample scenarios, however, optimality was largely violated. Further, with respect to false-positive rates, the ranking of the fold change did not perform well, even with large samples. Although the MAQC project (MAQC Consortium, 2006) reported good reproducibility of ranking based on fold change, "accuracy" and "reproducibility" are different concepts as remarked by Witten and Tibshirani (in preparation). Hence, ranking via the fold change is not recommended for general use when the objective is to prevent false-positive detection.

For general practical use, the proposed 3 methods should be used according to the purpose of analysis. However, for the 3 proposed methods, the sensitivities in detecting differential genes with the greatest effects as well as the false-positive rates were comparable. Accordingly, we would recommend the RPM method.

As noted in Section 2, attempts to obtain stable estimates of the hyperparameters of interest, that is, $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$ in the distribution of the effect sizes θ_j, include the use of reasonable estimates of the mixing proportions such as treating π_0 as fixed quantities, invoking reasonable constraints or placing prior distributions on the mixing proportions in the EM algorithm. Comparison of these approaches is outside the scope of this paper, but it is an important subject for future research.

The R code for gene ranking is available in the supplementary material at *Biostatistics* online.

SUPPLEMENTARY MATERIAL

Supplementary material is available at http://biostatistics.oxfordjournals.org.

ACKNOWLEDGMENTS

The authors would like to thank Tomonori Oura for many valuable comments and advice on an earlier draft of this article, and the editor for helpful comments and suggestions. *Conflict of Interest:* None declared.

REFERENCES

BERGER, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*. New York: Springer.

CARLIN, B. P. AND LOUIS, T. A. (2009). *Bayesian Methods for Data Analysis*. New York: Chapman & Hall.

CHOE, S. E., BOUTROS, M., MICHELSON, A. M., CHURCH, G. M. AND HALFON, M. S. (2005). Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset. *Genome Biology* **6**, R16.

DEMPSTER, A. P., LAIRD, N. M. AND RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B* **39**, 1–38.

EFRON, B. (2008). Microarrays, empirical Bayes and the two-groups model (with discussion). *Statistical Science* **23**, 1–47.

EFRON, B., TIBSHIRANI, R., STOREY, J. AND TUSHER, V. (2001). Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association* **96**, 1151–1160.

GOTTARDO, R., PANNUCCI, J. A., KUSKE, C. R. AND BRETTIN, T. (2003). Statistical analysis of microarray data: a Bayesian approach. *Biostatistics* **4**, 597–620.

GREENLAND, S. (1993). Methods for epidemiologic analyses of multiple exposures: a review and comparative study of maximum-likelihood, preliminary testing, and empirical-Bayes regression. *Statistics in Medicine* **12**, 717–736.

GUO, L., LOBENHOFER, E. K., WANG, C., SHIPPY, R., HARRIS, S. C., ZHANG, L., MEI, N., CHEN, T., HERMAN, D., GOODSAID, F. M. *and others* (2006). Rat toxicogenomic study reveals analytical consistency across microarray platforms. *Nature Biotechnology* **24**, 1151–1161.

LAIRD, N. M. AND LOUIS, T. A. (1989). Empirical Bayes ranking methods. *Journal of Educational Statistics* **14**, 29–46.

LIN, R., LOUIS, T. A., PADDOCK, S. M. AND RIDGEWAY, G. (2006). Loss function based ranking in two-stage, hierarchical models. *Bayesian Analysis* **1**, 915–946.

LO, K. AND GOTTARDO, R. (2007). Flexible empirical Bayes models for differential gene expression. *Bioinformatics* **23**, 328–335.

LÖNNSTEDT, I. AND SPEED, T. P. (2002). Replicated microarray data. *Statistica Sinica* **12**, 31–46.

LOUIS, T. A. AND SHEN, W. (1999). Innovations in Bayes and empirical Bayes methods: estimating parameters, populations and ranks. *Statistics in Medicine* **18**, 2493–2505.

MAQC CONSORTIUM (2006). The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. *Nature Biotechnology* **24**, 1151–1161.

MATSUI, S., ZENG, S., YAMANAKA, T. AND SHAUGHNESSY, J. (2008). Sample size calculations based on ranking and selection in microarray experiments. *Biometrics* **64**, 217–226.

MCLACHLAN, G. J., BEAN, R. W. AND JONES, L. B.-T. (2006). A simple implementation of a normal mixture approach to differential gene expression in multiclass microarrays. *Bioinformatics* **22**, 1608–1615.

MCLACHLAN, G. J., DO, K.-A. AND AMBROISE, C. (2004). *Analyzing Microarray Gene Expression Data*. Hoboken, NJ: Wiley.

NEWTON, M. A. AND KENDSIORSKI, C. M. (2003). Parametric empirical Bayes methods for microarrays. In: Parmigiani, G., Garrett, E. S., Irizarry, R. A. and Zeger, S. (editors), *The Analysis of Gene Expression Data: Methods and Software*. New York: Springer, pp. 254–271.

NEWTON, M. A., KENDZIORSKI, C. M., RICHMOND, C. S., BLATTNER, F. R. AND TSUI, K. W. (2001). On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *Biostatistics* **8**, 37–52.

NEWTON, M. A., NOUEIRY, A., SARKAR, D. AND AHLQUIST, P. (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics* **5**, 155–176.

SHEN, W. AND LOUIS, T. A. (1998). Triple-goal estimates in two-stage hierarchical models. *Journal of the Royal Statistical Society, Series B* **60**, 455–471.

STOREY, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Series B* **64**, 479–498.

WANG, Y., KLIJN, J. G., ZHANG, Y., SIEUWERTS, A. M., LOOK, M. P., YANG, F., TALANTOV, D., TIMMERMANS, M., MEIJER-VAN GELDER, M. E., YU, J. *and others* (2005). Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet* **365**, 671–679.

[*Received June 27, 2009; revised September 18, 2009; accepted for publication October 20, 2009*]