Genet. Res., Camb. (2008), **90**, *pp.* 347–361. © 2008 Cambridge University Press
doi:10.1017/S0016672308009427 Printed in the United Kingdom

Extreme value theory in analysis of differential expression in microarrays where either only up- or down-regulated genes are relevant or expected

RENATA IVANEK[1]*, YRJÖ T. GRÖHN[1], MARTIN T. WELLS[2], SARITA RAENGPRADUB[3], MARK J. KAZMIERCZAK[4] AND MARTIN WIEDMANN[3]

[1] *Department of Population Medicine and Diagnostic Sciences, College of Veterinary Medicine, Cornell University, Ithaca, NY 14853, USA*
[2] *Department of Biological Statistics and Computational Biology, 301 Malott Hall, Cornell University, Ithaca, NY 14853, USA*
[3] *Department of Food Science, 412 Stocking Hall, Cornell University, Ithaca, NY 14853, USA*
[4] *Channing Laboratory, Harvard Medical School, 181 Longwood Avenue, Boston, MA 02115, USA*

(*Received 24 January 2008 and in revised form 30 May 2008*)

Summary

We propose an empirical Bayes method based on the extreme value theory (EVT) (BE) for the analysis of data from spotted microarrays where the interest of the investigator (e.g. to identify up-regulated gene markers of a disease) or the design of the experiment (e.g. in certain 'wild-type versus mutant' experiments) limits identification of differentially expressed genes to those regulated in a single direction (either up or down). In such experiments, unlike in genome-wide microarrays, analysis is restricted to the tail of the distribution (extremes) of all the genes in the genome. The EVT provides a platform to account for this extreme behaviour, and is therefore a natural candidate for inference about differential expression. We compared the performance of the developed BE method with two other empirical Bayes methods on two real 'wild-type versus mutant' datasets where a single direction of regulation was expected due to experimental design, and in a simulation study. The BE method appears to have a better fit to the real data. In the analysis of simulated data, the BE method showed better accuracy and precision while being robust to different characteristics of microarray experiments. The BE method, therefore, seems promising and useful for inference about differential expression in microarrays where either only up- or down-regulated genes are relevant or expected.

1. Introduction

A common task in microarray studies is to determine which genes are differentially expressed (DE) between two cell samples. Different rules for identifying DE genes have been adopted, many based on the gene-specific mean and the standard deviation (SD). Their main limitation originates from the low number of replicates. More recently, as an alternative to these methods, a number of empirical Bayes methods have been proposed (e.g. Efron *et al.*, 2001; Lonnstedt & Speed, 2002; Gottardo *et al.*, 2003; Newton *et al.*, 2004). Empirical Bayes methods seem natural in the context of microarray experiments, because they summarize information from the whole dataset into prior parameters to be combined with means and SDs at the gene level (Lonnstedt & Britton, 2005).

An empirical Bayes approach to differential expression is a normal mixture model proposed by Lonnstedt & Speed (2002). This approach was later extended by Smyth (2004) to general linear models and modified into an empirical Bayes normal model (and not mixture model) for variance regularization. Smyth's (2004) model represents a posterior odds statistic formulated in terms of a moderated *t*-statistic, in which posterior residual SDs are used in place of ordinary SDs; the model can also account for correlation among technical replicates (Smyth *et al.*, 2005). It is implemented in Limma (Smyth, 2005), a popular software package for gene expression analysis. Throughout this paper, we will refer to the empirical Bayes normal method offered in Limma

* Corresponding author. Department of Population Medicine and Diagnostic Sciences, College of Veterinary Medicine, Cornell University, Ithaca, NY 14853, USA. Tel: +1(607) 2533052. Fax. +1 (607) 2533083. e-mail: ri25@cornell.edu

https://doi.org/10.1017/S0016672308009427 Published online by Cambridge University Press

(Lonnstedt & Speed, 2002; Smyth, 2004; Smyth *et al.*, 2005) as the 'BN' method.

The underlying assumption of the BN method is normal distribution of mean ratios of expressions of DE genes, i.e. that the DE genes are symmetrically up- and down-regulated. This assumption is valid for genome-wide microarray experiments, where interest lies in detection of all DE genes between two samples, and typically, out of all the tested genes (the whole genome), only a fraction is expected to be DE, with similar numbers of up- and down-regulated genes. However, in a number of microarray experiments, only up- or down-regulated genes are of interest. Examples are certain experiments designed to identify novel markers for molecular diagnosis and therapy of diseases (Suzuki *et al.*, 2002; Kobayashi *et al.*, 2004). Furthermore, in some experiments, regulation is expected in one direction only (either up or down) due to the design of the experiments, such as in certain 'wild-type versus mutant' microarrays (Kazmierczak *et al.*, 2003; van Schaik *et al.*, 2007) and those designed to select DE genes among a group of potentially up- or down-regulated genes that have been pre-selected either as part of some genome-wide microarray or other (e.g. Hidden Markov Chain promoter search) screening methods (Kazmierczak *et al.*, 2003). In such experimental designs, where only up- or down-regulated genes are relevant or expected, the distributional assumption of normality of the BN method is obviously violated.

As a long-tailed alternative to the normal mixture model of Lonnstedt & Speed (2002), which assumes that the DE genes are symmetrically up- and down-regulated, Bhowmick *et al.* (2006) proposed a Laplace mixture model (hereinafter referred to as the 'BL' method). Bhowmick *et al.* (2006) showed that the asymmetric gene expression data fit better under the asymmetric BL method. However, while the performance of BL was similar to the Smyth (2004) and Lonnstedt & Speed (2002) methods, BL depends on a large number of replicates for acceptable parameter estimates. Hence, microarrays with a single relevant or expected direction of regulation lack an appropriate analytical tool.

Extreme value theory (EVT) has been traditionally used for risk and financial analysis, and studies of extreme events, such as intense rainfall and floods. An important feature of EVT is that it models the extreme behaviour (the tail of the distribution) rather than the average behaviour of the systems as classical statistics do. By focusing on the values located in the tail of the distribution that diverge extremely from the mean value of a dataset, EVT provides a natural framework for detection of DE genes in an experiment where regulation is relevant or expected in a single direction. Therefore, our objective was to develop a new statistical method, based on EVT, for analysis of differential expression in experiments interested in or expecting either up- or down-regulated genes. Application to two experimental datasets and simulation studies indicated comparatively better performance of the developed empirical Bayes extreme value distribution (EVD) mixture model (BE) as compared with the two existing empirical Bayes methods (BN and BL).

2. Methods

(i) *Model setup in the context of linear models*

Consider a microarray experiment comparing the expression (i.e. mRNA transcript) levels of wild-type and mutant cells, where mutant cells lack a gene encoding a positive regulator. We wish to identify genes that differ in their transcript levels between wild-type and mutant cells exposed to the same treatment. For the *j*th replicate of gene g on array *i*, we use the log ratio of the expressions:

$$Y_{gij} = \log_2 \frac{(\text{expression level in wild type})_{gij}}{(\text{expression level in mutant})_{gij}}.$$

Let us assume that we have *n* arrays, where each array has *G* number of genes spotted on it, and each gene is replicated *m* times. Therefore, the complete set of data from the experiment consists of Y_{gij}, $g=1, \ldots, G$, $i=1, \ldots, n$, and $j=1, \ldots, m$, and so $\mathbf{Y}_g$ can be viewed as a vector of *mn* log ratios observed for a gene *g*. We regard the Y_{gij} as random variables from a normal distribution with mean μ_g and variance σ_g^2, which, although not completely true, is convenient and found empirically to be roughly the case (Lonnstedt & Speed, 2002), i.e.

$$Y_{gij} \sim N(\mu_g, \sigma_g^2). \qquad (1)$$

A general microarray experiment can be represented by a linear model (Smyth, 2004), i.e. $E(\mathbf{Y}_g) = \mathbf{X}\boldsymbol{\beta}_g$, where $\mathbf{X}$ is an $nm \times k$ dimensional design matrix specifying experimental design and $\boldsymbol{\beta}_g$ is a vector of *k* regression coefficients. As we have *m* replicates of each gene on each array, our design matrix has *m* repeated rows corresponding to each set of *m* replicate spots. To simplify $E(\mathbf{Y}_g) = \mathbf{X}\boldsymbol{\beta}_g$, let $\bar{\mathbf{Y}}_g$ be the *n*-vector of array means $\bar{Y}_{gi}$ and let $\bar{\mathbf{X}}$ be the reduced $n \times k$ dimensional design matrix in which there is only one row instead of *m* rows for each array combination (Smyth *et al.*, 2005). Then, $E(\bar{\mathbf{Y}}_g) = \bar{\mathbf{X}}\boldsymbol{\beta}_g$. Now, let $\alpha_{gq} = \mathbf{c}^{\mathrm{T}}\boldsymbol{\beta}_g$, where $\mathbf{c}$ is a vector of *q* known constants defining contrasts of biological interest and α_{gq} is a particular contrast or linear combination of the regression coefficients and suppose that interest lies in testing H_0: $\alpha_{gq} = 0$ (Smyth *et al.*, 2005). For the rest of the paper, the subscript *q* will be suppressed for notational simplicity under the assumption that our

https://doi.org/10.1017/S0016672308009427 Published online by Cambridge University Press

hypothetical experiment has only one contrast of interest. We suppose that measurements from genes spotted on different arrays are independent. However, we acknowledge that within an array technical replicates of a gene are correlated. To account for correlation among replicates, we used a common correlation factor (Smyth *et al*., 2005). Let $\hat{\beta}_g$ be the estimator of $\boldsymbol{\beta}_g$ from fitting a linear model and let contrast estimator $\left(\hat{\alpha}_g\right)$ $\hat{\alpha}_g = \mathbf{c}^{\mathrm{T}}\hat{\beta}_g$. Here, $\hat{\alpha}_g$ is assumed to be a random variable from a normal distribution (as in (1)) with mean μ_g and variance $\nu_g^2\,\sigma_g^2$, where ν_g^2 is an unscaled variance of $\hat{\alpha}_g$ that also accounts for a common correlation factor among replicates (Smyth *et al*., 2005), i.e.

$$\hat{\alpha}_g|\mu_g, \sigma_g^2 \sim \mathrm{N}(\mu_g, \nu_g^2\sigma_g^2). \tag{2}$$

The residual variance from fitting a linear model (s_g^2) was assumed to follow a gamma distribution (G; equivalent to the scaled chi-square distribution assumed by Smyth, 2004):

$$s_g^2|\sigma_g^2 \sim \mathrm{G}\left(\frac{d_g}{2}, \frac{d_g}{2\sigma_g^2}\right), \tag{3}$$

where d_g is the residual degrees of freedom for the linear model for gene g.

(ii) *Empirical Bayes EVD mixture model* (*BE*)

The EVD has three parameters: a location parameter, a; a scale parameter, b; and a shape parameter, c. It encompasses three classes of distributions with types I, II and III widely known as the Gumbel, Fréchet and Weibull families, respectively (Coles, 2001). These are combined into a single-family model known as the generalized extreme value (GEV) family of distributions, where the shape parameter (c) determines the type; the Fréchet and Weibull families have $c > 0$ and $c < 0$, respectively, while the shape parameter of the Gumbel family has $c = 0$. Throughout the present paper, we will use EVD and GEV interchangeably, both indicating an unspecified family of EVD. In contrast, when we refer to a specific family we will use its name (i.e. Gumbel, Fréchet or Weibull distribution).

For an appropriately background corrected and normalized gene to be DE, its μ_g value should be statistically different from zero. Let I_g indicate whether the gene is DE ($\mu_g \neq 0$) or not ($\mu_g = 0$), such that

$$I_g = \begin{cases} 0, & \text{if } \mu_g = 0, \\ 1, & \text{if } \mu_g \neq 0. \end{cases}$$

We assume that

$$\mu_g|I_g = 0 \sim \delta(0), \tag{4}$$

where $\delta(0)$ denotes the distribution which places unit mass at $\mu_g = 0$. Because we expect DE genes to be either up- or down-regulated, we assumed an EVD prior on $\mu_g \neq 0$, i.e.

$$\mu_g|I_g = 1 \sim \mathrm{EVD}(a, b, c). \tag{5}$$

An inverse gamma (IG) prior distribution is assumed on σ_g^2, with σ_g^2 equivalent to a prior estimator s_0^2 with d_0 degrees of freedom (Smyth, 2004), i.e.

$$\sigma_g^2 \sim \mathrm{IG}\left(\frac{d_0}{2}, \frac{d_0 s_0^2}{2}\right). \tag{6}$$

We wish to know whether the gene is DE, i.e. what is $\Pr(I_g = 1|\hat{\alpha}_g, s_g^2)$ or what are the odds of differential expression (more specifically, the natural logarithm of the odds), BE_g:

$$\mathrm{BE}_g = \log\frac{\Pr\left[I_g = 1|(\hat{\alpha}_g, s_g^2)\right]}{\Pr\left[I_g = 0|(\hat{\alpha}_g, s_g^2)\right]}.$$

By the Bayes theorem,

$$\Pr[I_g = 1|(\hat{\alpha}_g, s_g^2)] = \frac{\Pr(I_g = 1, \hat{\alpha}_g, s_g^2)}{\Pr(\hat{\alpha}_g, s_g^2)}$$

$$= \frac{\Pr(I_g = 1)\Pr(\hat{\alpha}_g, s_g^2|I_g = 1)}{\Pr(I_g = 1)\Pr(\hat{\alpha}_g, s_g^2|I_g{=}1) + \Pr(I_g = 0)\Pr(\hat{\alpha}_g, s_g^2|I_g = 0)},$$

$$\Pr[I_g = 0|(\hat{\alpha}_g, s_g^2)] = \frac{\Pr(I_g = 0, \hat{\alpha}_g, s_g^2)}{\Pr(\hat{\alpha}_g, s_g^2)}$$

$$= \frac{\Pr(I_g = 0)\Pr(\hat{\alpha}_g, s_g^2|I_g = 0)}{\Pr(I_g{=}1)\Pr(\hat{\alpha}_g, s_g^2|I_g = 1) + \Pr(I_g = 0)\Pr(\hat{\alpha}_g, s_g^2|I_g = 0)}.$$

Then, the log Bayes' factor is

$$\mathrm{BE}_g = \log\left[\frac{\Pr(I_g = 1)}{[1 - \Pr(I_g = 1)]}\,\frac{\Pr[(\hat{\alpha}_g, s_g^2)|I_g = 1]}{\Pr[(\hat{\alpha}_g, s_g^2)|I_g = 0]}\right], \tag{7}$$

where $\Pr(I_g = 1)$ is the proportion of DE genes (pDE) in the experiment. The joint densities $\Pr(\hat{\alpha}_g, s_g^2|I_g = 1)$ and $\Pr(\hat{\alpha}_g, s_g{}^2|I_g = 0)$ are:

$$\Pr(\hat{\alpha}_g, s_g^2|I_g = 1) = f_{Ig=1}(\hat{\alpha}_g, s_g^2)$$
$$= \iint f(\hat{\alpha}_g|\mu_g, \sigma_g^2) f(\mu_g) f(s_g^2|\sigma_g^2) f(\sigma_g^2)\,\mathrm{d}\mu_g \mathrm{d}\sigma_g^2$$

$$\Pr(\hat{\alpha}_g, s_g^2|I_g = 0) = f_{Ig=0}(\hat{\alpha}_g, s_g^2)$$
$$= \int f(\hat{\alpha}_g|\mu_g = 0, \sigma_g^2) f(s_g^2|\sigma_g^2) f(\sigma_g^2)\mathrm{d}\sigma_g^2.$$

To evaluate the probabilities $\Pr(\hat{\alpha}_g, s_g^2|I_g = 1)$ and $\Pr(\hat{\alpha}_g, s_g^2|I_g = 0)$, and the BE statistic in (7), we used the distributions defined in equations (2)–(6), i.e. their respective likelihood or probability density functions:

$$f(\hat{\alpha}_g|\mu_g, \sigma_g^2) = \left(2\pi\nu_g^2\sigma_g^2\right)^{-1/2}\exp\left\{-\frac{1}{2\nu_g^2\sigma_g^2}\left(\hat{\alpha}_g - \mu_g\right)^2\right\}, \tag{8}$$

where $-\infty < \hat{\alpha}_g < \infty$, $-\infty < \mu_g < \infty$, $\nu_g \sigma_g > 0$;

$$f\left(s_g^2 | \sigma_g^2\right) = \frac{\left(d_g/2\sigma_g^2\right)^{d_g/2} \left(s_g^2\right)^{(d_g/2-1)}}{\Gamma\left(d_g/2\right)} \exp\left(-d_g s_g^2/2\sigma_g^2\right), \tag{9}$$

where $d_g/2 > 0$, $d_g/2\sigma_g^2 > 0$, $s_g^2 > 0$;

$$f_{Ig=0}(\mu_g) = \delta(0); \tag{10}$$

$$f_{Ig=1}(\mu_g | a, b, c) = \frac{1}{b}\left(1 + c\frac{\mu_g - a}{b}\right)^{-(1/c)-1} \times \exp\left\{-\left(1 + c\frac{\mu_g - a}{b}\right)^{-1/c}\right\} \tag{11}$$

defined on $\{\mu_g : 1 + (\mu_g - a)/b > 0\}$, where $-\infty < \mu_g < \infty$, $b > 0$ and $-\infty < \mathrm{c} < \infty$ (Coles, 2001; Panjer, 2006);

$$f\left(\sigma_g^2 | s_0^2, d_0\right) = \frac{\left(d_0 s_0^2/2\right)^{d_0/2} \left(\sigma_g^2\right)^{(-1-d_0/2)}}{\Gamma(d_0/2)} \exp\left(-d_0 s_0^2/2\sigma_g^2\right), \tag{12}$$

where $d_0/2 > 0$, $d_0 s_0^2/2 > 0$, $\sigma_g^2 > 0$.

It is not possible to integrate the integral $f_{Ig=1}(\hat{\alpha}_g, s_g^2)$. Therefore, numerical approximation through the Monte Carlo (MC) integration method (Tanner, 1996) was applied to obtain the posterior expectations of $f_{Ig=1}(\hat{\alpha}_g, s_g^2)$ and $f_{Ig=0}(\hat{\alpha}_g, s_g^2)$ (denoted as $\mathrm{E}(f_{Ig=1})$ and $\mathrm{E}(f_{Ig=0})$, respectively):

$$\mathrm{E}(f_{Ig=1}) = \frac{1}{r}\left(\sum_{i=1}^{r} f(\hat{\alpha}_g | \mu_{gi}, \sigma_{gi}^2) f(s_g^2 | \sigma_{gi}^2)\right), \tag{13}$$

$$\mathrm{E}(f_{Ig=0}) = \frac{1}{r}\left(\sum_{i=1}^{r} f(\hat{\alpha}_g | \mu_{gi} = 0, \sigma_{gi}^2) f(s_g^2 | \sigma_{gi}^2)\right), \tag{14}$$

where r denotes the number of iterations, while μ_{gi} and σ_{gi}^2 denote *iid* draws from prior distributions of gene-specific means and variances, respectively. Thereafter, the expected value of the BE statistic is:

$$\mathrm{BE}_g = \log\left[\frac{\mathrm{pDE}}{(1-\mathrm{pDE})} \frac{\mathrm{E}(f_{Ig=1})}{\mathrm{E}(f_{Ig=0})}\right]. \tag{15}$$

For an estimator to be useful, a measure of estimation error is required. The estimated Monte Carlo standard errors (MCSEs) of $\mathrm{E}(f_{Ig=1})$ and $\mathrm{E}(f_{Ig=0})$ (denoted as $\mathrm{SE}(f_{Ig=1})$ and $\mathrm{SE}(f_{Ig=0})$, respectively) were:

$$\mathrm{SE}(f_{Ig=1}) = \frac{1}{\sqrt{\mathrm{r}}} \times \sqrt{\frac{\sum_{i=1}^{r}[f(\hat{\alpha}_g | \mu_{gi}, \sigma_{gi}^2) f(s_g^2 | \sigma_{gi}^2) - \mathrm{E}(f_{Ig=1})]^2}{r-1}}, \tag{16}$$

$$\mathrm{SE}(f_{Ig=0}) = \frac{1}{\sqrt{\mathrm{r}}} \times \sqrt{\frac{\sum_{i=1}^{r}[f(\hat{\alpha}_g | \mu_{gi} = 0, \sigma_{gi}^2) f(s_g^2 | \sigma_{gi}^2) - \mathrm{E}(f_{Ig=0})]^2}{r-1}}. \tag{17}$$

Thereafter, by accounting for error propagation, the MCSE of the $\mathrm{BE_g}$ ($\mathrm{MCSE_g}$) was estimated by adding uncertainties in $\mathrm{E}(f_{Ig=1})$ and $\mathrm{E}(f_{Ig=0})$, i.e. $\mathrm{SE}(f_{Ig=1})$ and $\mathrm{SE}(f_{Ig=0})$, in quadrature (Taylor, 1982):

$$\mathrm{MCSE}_g = \sqrt{\left[\frac{\mathrm{SE}(f_{Ig=1})}{\mathrm{E}(f_{Ig=1})}\right]^2 + \left[\frac{\mathrm{SE}(f_{Ig=0})}{\mathrm{E}(f_{Ig=0})}\right]^2}. \tag{18}$$

BE statistics were evaluated on a PC running on Windows XP Pro SP2 with an AMD Athlon 64 X2 Dual Core Processor 4200+(2200 MHz) and 2048 Mb of RAM. For a dataset with 200 genes spotted in triplicate on four arrays (similar to datasets described in Section 3(i)), the computing time for the BE method involving 50 000 iterations was roughly 15 s, while the BN and BL methods took roughly 1 s each. For a considerably larger dataset, with 10 000 genes also spotted in triplicate on four arrays, the computing time for the BE method with 50 000 iterations was roughly 16 min, while it was roughly 1 min for each of the BN and BL methods. The R code implementing the proposed methodology is available on request from the first author.

(iii) *Estimation of hyperparameters*

The BE statistic defined in Section 2(ii) uses hyperparameters estimated from the data. The parameters s_0^2 and d_0 are estimated from s_g^2 following Smyth's (2004) procedure. The only exception to this rule was when d_0 could not be estimated as in Smyth (2004), indicating that there is no evidence that the underlying gene-specific variances σ_g^2 vary between genes, and so in Smyth (2004) d_0 was set to positive infinity. In evaluation of BE statistics, d_0 equal to positive infinity precludes generation of a prior on σ_g^2. Therefore, in such situations, instead of positive infinity we set d_0 to 10^{300}, which is close to the largest positive decimal number on a typical R platform. In terms of pDE, the BE method uses the same approach as the BN method; it fixes pDE and then estimates the remaining hyperparameters (a, b and c). Hyperparameters a, b and c are estimated simultaneously from $\hat{\alpha}_g$ values corresponding to a pDE fraction of genes with the highest moderated t-statistics. The method involves maximum likelihood estimation (MLE) fitting for the GEV distribution, with the Nelder–Mead optimization method (Nelder & Mead, 1965). The MLE procedure, subject to the limitations discussed in Section 5, provides the means and standard errors (SEs) for the parameters a, b and c. The means are

https://doi.org/10.1017/S0016672308009427 Published online by Cambridge University Press

used to estimate BE statistics, and the corresponding SEs provide an indication of the uncertainty around the estimated mean values. Genes with the top moderated *t*-statistics (estimated according to Smyth, 2004) were used instead of the top genes ranked by their ordinary *t*-statistics as they provided a more stable estimation of *a*, *b* and *c*.

3. Application to experimental data

(i) *Data description*

To analyse the performance of BE as compared with BN and BL, we used two datasets generated in a previous two-color cDNA microarray study conducted to identify genes regulated by the sigma factor σ^B in the bacterium *Listeria monocytogenes* (Kazmierczak *et al.*, 2003). In that study, an *L. monocytogenes sigB* null mutant (which lacks the σ^B protein) and a parent strain with intact *sigB* gene (wild-type) were exposed to two stress conditions, namely osmotic stress and stationary phase, to identify genes with transcript levels affected by the *sigB* deletion under these two conditions. For each stress condition, two independent RNA isolates (biological replicates) for both wild-type and *sigB* mutant cells were dye swapped for a total of four arrays per stress condition. Each array included 211 test genes, and a number of non-hybridizing and normalization controls (for details see Kazmierczak *et al.*, 2003) spotted in triplicate. Most (166) genes included on the array were identified by Hidden Markov Model promoter searches as being preceded by a putative σ^B-dependent promoter, while some genes (36) were included because of previous reports of their involvement in virulence or stress response. As σ^B is a positive regulator of gene expression with particular importance for regulating stress response and virulence genes, most genes in these two experiments are expected to show higher transcript levels in the wild-type strain as compared with the *sigB* deletion strain.

In their analysis, Kazmierczak *et al.* (2003) considered all individual spots as repetitions, generating 24 data points for each gene (3 spots per gene × 4 arrays × 2 channels per array), i.e. correlation among technical replicates was not considered. They reported findings for 208 of the 211 test genes as three genes were spotted twice. Prior to analysis, cross-slide mean normalization (without background correction) and flooring were performed. The analysis by the Significance Analysis of Microarrays (SAM) program (Tusher *et al.*, 2001) identified 51 (25%) and 41 (20%) genes with at least 1·5-fold different statistically significant expressions under osmotic stress and stationary-phase conditions, respectively.

Prior to our analysis of the two datasets of 211 genes, we performed the background correction and normalization. The median background fluorescence intensities are usually recommended for correction of the background noise because of their robustness to the outliers. We, however, used the mean background intensities because the distribution of median background intensities had a bimodal distribution with some spots having zero background while the others were in the higher range of intensities (above 2^8) (possibly due to the setup or limitations of the laser scanner used).

Two background correction procedures seemed appropriate for the data. The first, the normal-exponential convolution background correction model (NeBC) (performed with an offset of 100), involves fitting of the convolution of normal and exponential distributions to the foreground intensities using the background intensities as a covariate (also referred to as the normexp method in Smyth, 2005). The second procedure used was multiplicative background correction (MBC). This is a novel approach that involves logarithmic transformation of the intensity readings before the background correction and is found (via a series of examples) to be superior to the additive background correction and no background correction (Zhang *et al.*, 2006). Because MBC reportedly gives fewer false positives than conventional additive background correction (Zhang *et al.*, 2006) and because its performance has never been contrasted with NeBC, we used (and compared) both background correction models in our study.

The normalization appropriate for the data was the Lowess normalization (Cleveland & Devlin, 1988), with up-weighting of the background and normalization control spots, known to be non-DE (http://bioconductor.org/packages/1.8/bioc/vignettes/limma/inst/doc/usersguide.pdf). Application of the two background correction procedures (NeBC and MBC) to each of the two stress-condition datasets (osmotic stress and stationary phase) provided a total of four real model-datasets used in our analyses.

(ii) *Results*

In all four model-datasets, normalized and background corrected $\log_2$ ratios between genes' expression values in wild-type and mutant cells (Y_{gij}) were distributed asymmetrically around zero and heavily skewed to the right. This was expected because up-regulation was anticipated in most of the tested genes. It was therefore reasonable to assume that the distribution of mean expressions of DE genes follows EVD. Hence, the BE method could be applied for inference about differential expression.

A critical issue in the MC integration methodology, underlying the BE method, is to determine the number of iterations that can be safely used as a basis for

https://doi.org/10.1017/S0016672308009427 Published online by Cambridge University Press

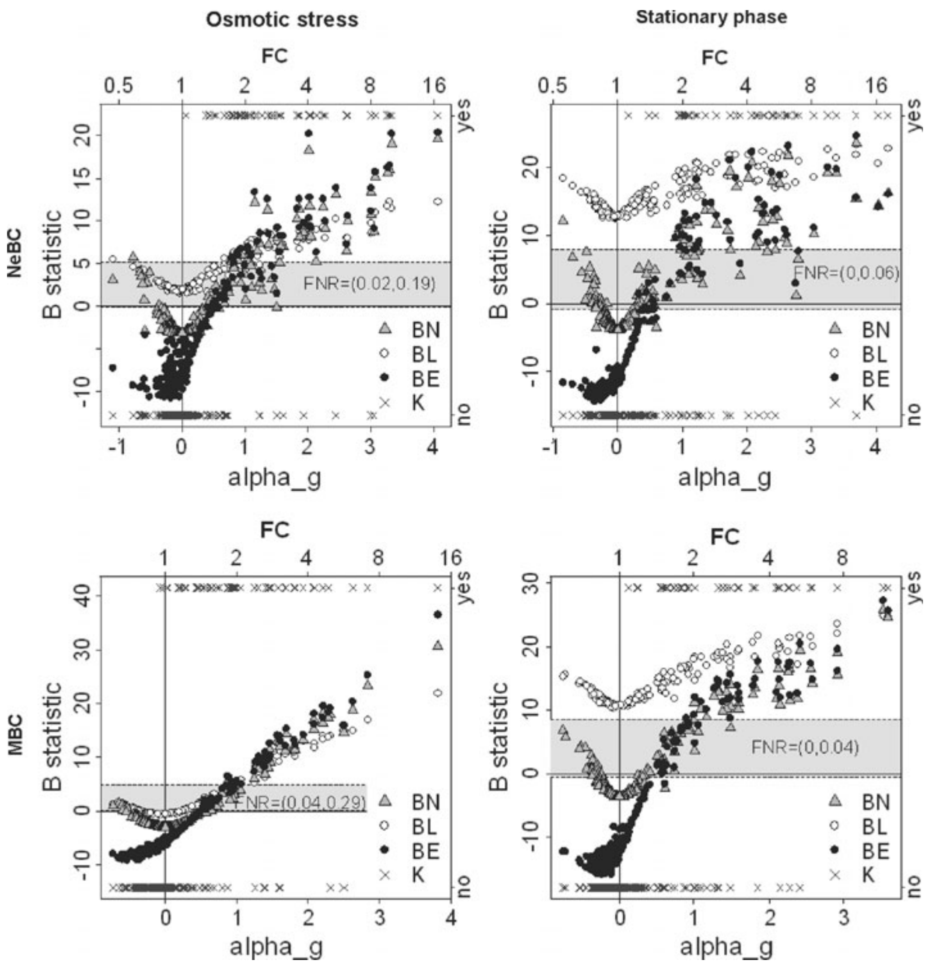

Fig. 1. The BN (Lonnstedt & Speed, 2002; Smyth, 2004), BL (Bhowmick *et al*., 2006) and empirical Bayes EVD mixture model (BE) statistics plotted against the contrast estimators from fitting linear models at the gene level, 'alpha_g' (denoted as $\hat{\alpha}_g$ in the text), also translated into fold changes (FC), and against results reported by Kazmierczak *et al*. (2003). 'K' and the associated right *y*-axis indicate whether Kazmierczak *et al*. (2003) reported a gene as DE ('yes') or not ('no'). 'NeBC' = normal-exponential convolution background correction method. 'MBC' = multiplicative background correction method. Two horizontal dashed lines (enclosing a shaded area) indicate the 5th and 95th percentiles of OT of the BE statistic estimated for the FDR fixed to 0. 'FNR = (,)' denotes false negative rate (5th and 95th percentiles) associated with the OT.

inference. We used 50 000 iterations as they provided reasonable accuracy of the approximated BE statistics. The achieved MCSEs varied for different genes and model-datasets. The medians, followed by the ranges in parentheses, of the achieved MCSEs were 0·05 (0·01–0·42) and 0·03 (0·01–0·82) for the osmotic stress datasets corrected with the NeBC and MBC methods, respectively, and 0·38 (0·02–0·92) and 0·18 (0·02–0·52) for the stationary-phase datasets corrected with the NeBC and MBC methods, respectively. In all four model-datasets, MCSEs were the lowest (<0·1) for the genes with the value of BE statistic around 0.

For each model-dataset, the gene-specific BN, BL and BE statistics were approximated. The biological meaning of the identified DE genes is important. Hence, for each of the four model-datasets in Fig. 1, we show the values of the BN, BL and BE statistics, plotted against contrast estimators from linear models ($\hat{\alpha}_g$) (also translated into fold changes, $2^{\hat{\alpha}_g}$, for more intuitive interpretation) and against previous results of Kazmierczak *et al*. (2003). In each model-dataset, genes that ranked very low with the BE statistic have a fold change below 1. At the same time, the BN statistic ranked high some of the genes with very low fold change, incorrectly suggesting down-regulation. The BL statistic gave ambiguous results with high values for most genes, particularly in the stationary-phase data. It should be noted that for approximation of the BN and BE statistics, we fixed the pDEs to those reported in Kazmierczak *et al*. (2003). Fixing pDEs to different values would change the BN and BE *vs*. fold change plots. Decreasing pDE would shift the plots to the right and down, whereas increasing pDE would shift the plots left and up on the *x*- and *y*-axes, respectively.

https://doi.org/10.1017/S0016672308009427 Published online by Cambridge University Press

Table 1. *Definitions of model parameters and hyperparameters in the empirical Bayes EVD mixture model (BE), and models of Lonnstedt & Speed (2002) modified by Smyth (2004) (BN) and Bhowmick* et al. *(2006) (BL)*

Notation	Parameter definition	Osmotic stress, NeBC[a]	Osmotic stress, MBC[b]	Stationary phase, NeBC	Stationary phase, MBC
G	Number of genes in the experiment	211	211	211	211
M	Number of technical replicates	3	3	3	3
N	Number of biological replicates (arrays)	4	4	4	4
	Correlation among technical replicates	0·65	0·43	0·65	0·42
pDE	Proportion of DE[c] genes	0·25	0·25	0·2	0·2
μ_g	Prior distribution of DE[c] gene means in BE	EVD[d] (1·27, 0·60, 0·15)	EVD (1·01, 0·47, 0·17)	EVD (1·61, 0·68, 0·06)	EVD (1·35, 0·53, 0·10)
σ^2_g	Prior distribution of gene variance in BE	IG[e] (1·61, 4·79)	IG (7·91, 0·47)	IG (1·41, 14·23)	IG (2·90, 3·83)
BNμ_g	Prior distribution of DE gene means in BN	N[f] (0, 11·95 × BNσ^2_g)	N (0, 7·06 × BNσ^2_g)	N (0, 30·31 × BNσ^2_g)	N (0, 15·65 × BNσ^2_g)
BNσ^2_g	Prior distribution of gene variance in BN	IG (1·61, 4·79)	IG (7·91, 0·47)	IG (1·41, 14·23)	IG (2·90, 3·83)
BLμ_g	Prior distribution of DE gene means in BL	L[g] (0·31, 2·43 × BLσ^2_g)	L (0·44, 2·54 × BLσ^2_g)	L (0·18, 5·98 × BLσ^2_g)	L (0·21, 4·65 × BLσ^2_g)
BLσ^2_g	Prior distribution of gene variance in BL	IG (1·9, 4·83)	IG (7·08, 0·75)	IG (1·12, 23·72)	IG (2·19, 7·94)
w[h]	Mixing probability	0·93	0·58	1	1

[a] NeBC = normal-exponential convolution background correction method; [b] MBC = multiplicative background correction method; [c] DE = differentially expressed; [d] EVD = extreme value distribution; [e] IG = inverse gamma distribution; [f] N = normal distribution; [g] L = Laplace distribution; [h] w = the probability that a gene is DE estimated as part of the BL method (note that the BN and BE statistics use a fixed, user-defined pDE).

Table 1 shows the characteristics of the data and the values of the hyperparameters estimated for each of the four model-datasets. The ambiguous results of the BL method are probably due, at least in part, to a very high estimated probability that a gene is DE ($w = 1$; Table 1). The prior variance distributions seem quite stable among the BN, BL and BE methods, except for the roughly double value of the scale parameter estimated for the BL method as compared with that estimated for the BN and BE methods. Contrary to that, the prior variances differ substantially between background correction methods, being narrower for the data corrected with MBC, which may explain the smoother plots of the BN, BL and BE statistics following MBC. Also, interestingly, correlation among technical replicates tends to be higher following the NeBC than MBC, demonstrating the difference between these two procedures.

In the BE statistic, a natural choice of the optimal threshold (OT) above which a gene could be considered DE is 0. However, the actual OT depends on the imposed criteria, such as the cost of a false positive and false negative. A typical approach in choosing a rule for interpretation of a statistical test is to control the type I error probability while maintaining a certain power. A sensible, powerful and easy to interpret (Verhoeven *et al.*, 2005) method to control type I error when multiple statistical tests are performed is the false discovery rate (FDR) (Benjamini & Hochberg, 1995). FDR is the expected proportion of errors among the genes selected to be DE. As a low FDR often comes at the cost of low sensitivity or power (i.e. a high false negative rate (FNR)), these should be controlled jointly (Pawitan *et al.*, 2005). Because Kazmierczak *et al.* (2003) considered genes that had been pre-selected for their expected differential expression, we chose an FDR = 0, i.e. no false positives were acceptable. The OT for BE (its 5th and 95th percentiles) was determined through simulation analysis for each of the four model-datasets (assuming that the pDEs reported in Kazmierczak *et al.* (2003) are true), and is shown in Fig. 1, together with the associated FNR. Genes whose BE statistic was above the 95th percentile of the OT could be considered DE with a high certainty. Genes with a BE statistic between the 5th and 95th percentiles of the OT are likely to be DE. BE did rank high (above the 95th percentile of the OT) some of the genes previously unidentified by Kazmierczak *et al.* (2003), whereas a few of the genes previously reported as DE by Kazmierczak *et al.* (2003) were ranked low (below the 5th percentile of the OT). However, the findings of the BE method have been validated by other independent studies for most of the genes, for which the result of the BE method differed from those reported by Kazmierczak *et al.* (2003) (elaborated in the Appendix).

https://doi.org/10.1017/S0016672308009427 Published online by Cambridge University Press

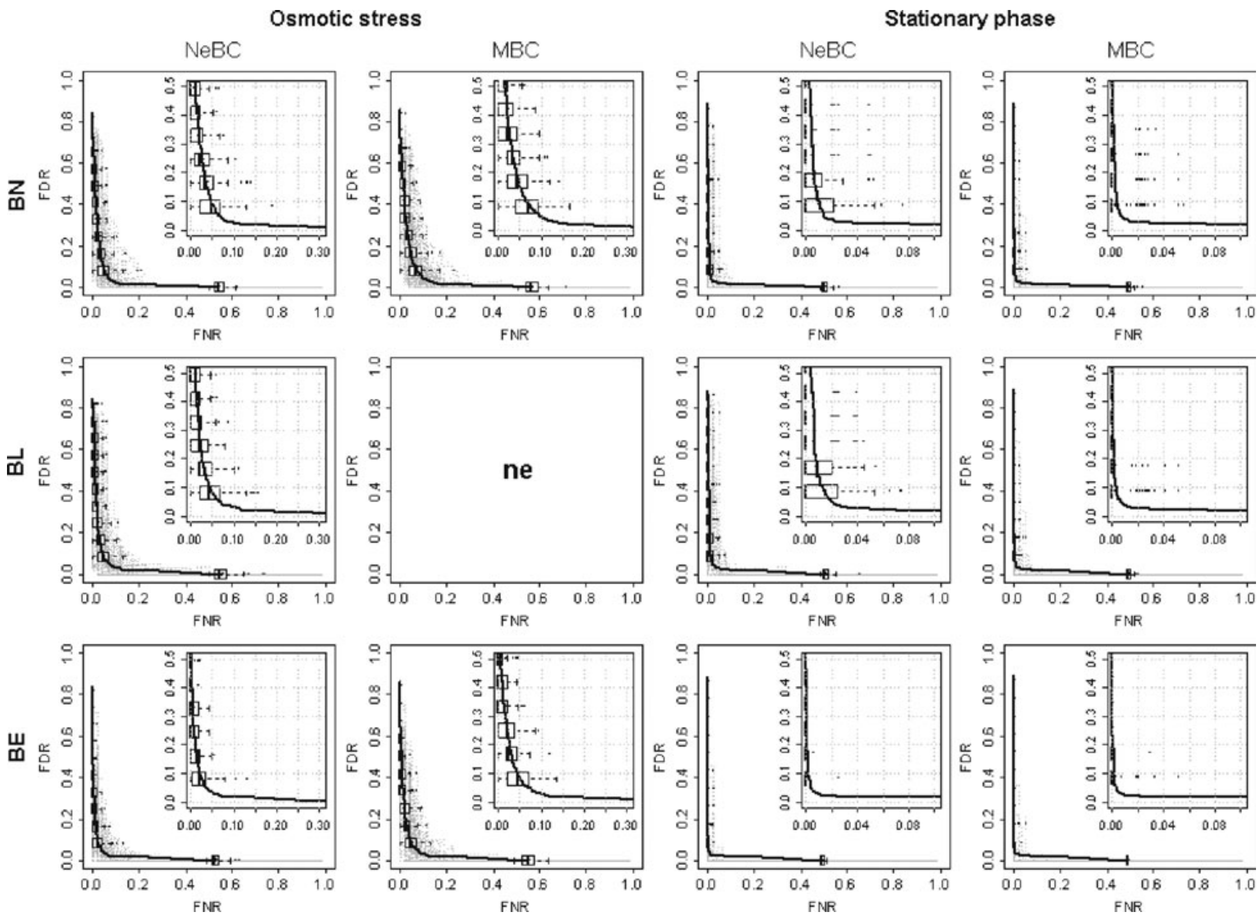

Fig. 2. FDR *vs*. FNR plots of 100 simulated datasets overlaid by the horizontal average curve and box plots showing horizontal spread of the performance of the BN (Lonnstedt & Speed, 2002; Smyth, 2004), BL (Bhowmick *et al*., 2006) and empirical Bayes EVD mixture model (BE) statistics in the four simulated model-datasets. 'NeBC' = normal-exponential convolution background correction method; 'MBC' = multiplicative background correction method; 'ne' = not estimated.

4. Simulation studies

Based on the hyperparameters estimated from the data and shown in Table 1, 100 datasets were simulated for each of the four model-datasets. To ease computational burden, we reduced the number of iterations in the MC integration model underlying the BE statistics from 50 000 to 5000. This reduction was warranted by our empirical observation that genes with BE around zero (a natural cutoff value for separation of DE and non-DE genes) have the lowest MCSEs. With 5000 iterations, the achieved MCSEs of these genes were below 0·2. So, the reduction did not jeopardize the ability of the BE to differentiate between DE and non-DE genes.

The performance of the BN, BL and BE statistics was tested under the Normal, Laplace and EVD models. The BE statistic did not do well under the asymmetric Laplace model and did even worse under the Normal model. That is expected because the Normal, and to a lesser extent the asymmetric Laplace, model permit both up- and down-regulated genes, while the EVD model allows one direction of expression (either up- or down-regulation) only. Because the Normal and Laplace models do not fit to data from experiments where only one direction of expression is expected, we only present results of the simulation under the EVD model.

Simulated datasets (100 of them) were generated under different scenarios. These were analysed with BN, BL and BE and the performance of the three statistics was evaluated based on the FDR *vs*. FNR plots, where the lower left corner indicates perfect performance, with zero false positives and false negatives. Figure 2 shows FDR *vs*. FNR plots of 100 simulated datasets overlaid by the horizontal average curve and box plots showing horizontal spread of the performance of the three statistics in the four simulated model-datasets. Averaging over simulated datasets horizontally mimics the situation when the experimenter chooses a tolerable FDR and tries to maximize power. Results for vertical averaging, corresponding to situations when an experimenter chooses an acceptable power while minimizing FDR, are not shown as they were very similar to the results for horizontal averaging. Figure 2 shows that the

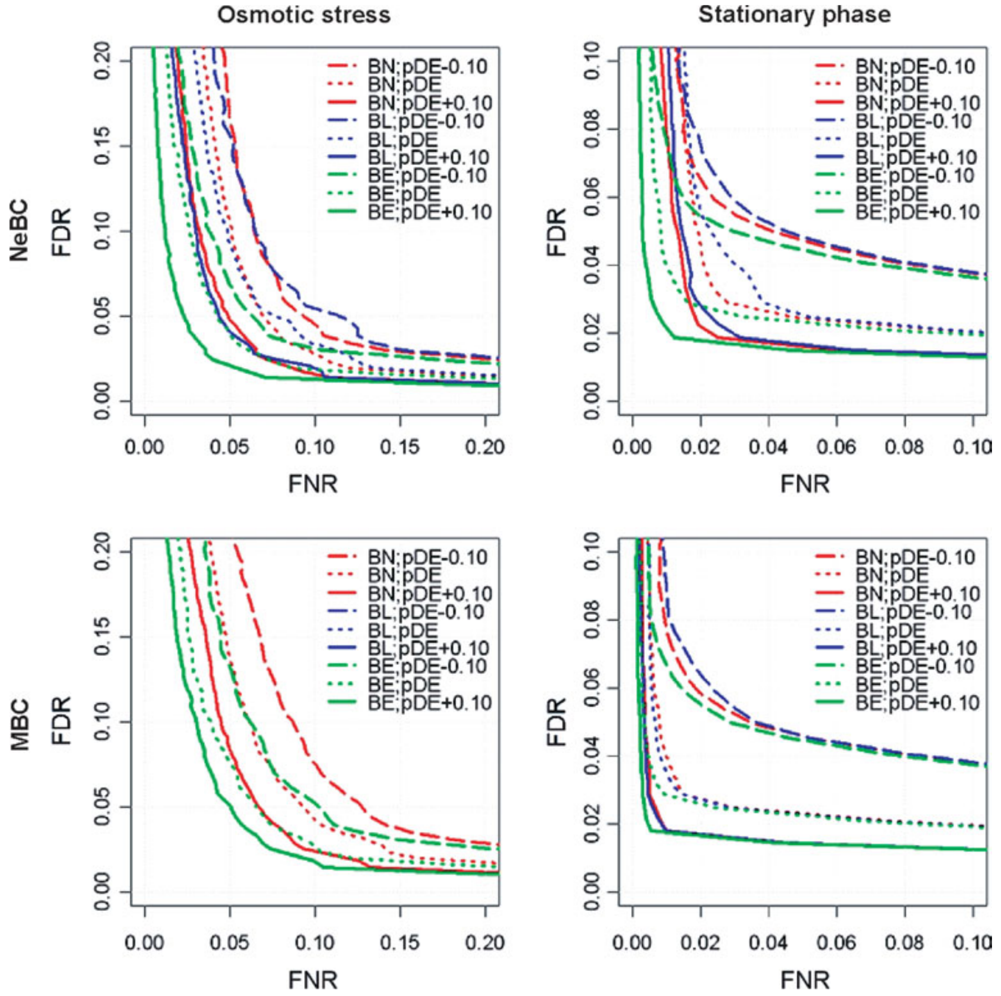

Fig. 3. FDR *vs.* FNR plots showing horizontal average of the BN (Lonnstedt & Speed, 2002; Smyth, 2004), BL (Bhowmick *et al*., 2006) and empirical Bayes EVD mixture model (BE) statistics from 100 simulated datasets in the four simulated model-datasets with the proportion of DE genes (pDE) reported in Kazmierczak *et al*. (2003) (25 and 20%, in osmotic stress and stationary-phase data, respectively) and simulating deviations from pDE of −10% and +10%. 'NeBC' = normal-exponential convolution background correction method; 'MBC' = multiplicative background correction method.

three statistics do very well in all four model-datasets. The only exception was BL: analysis failed for several simulated datasets reproduced from the osmotic model-dataset background corrected with MBC. To assure fair comparison, the performance of the BL method in analysis of this model-dataset is not plotted in Fig. 2. A curve showing average FNR for different values of FDR (horizontal average curve) in Fig. 2 indicates that BE, on average, has better accuracy (correct classification of genes as DE and non-DE) in the simulated data than BN and BL. The horizontal box plots of the achieved FNR for various FDR show narrower spread of BE compared with the other two statistics, indicating better precision (repeatability of classification success) in the simulated data.

Robustness of the BE statistic, as compared with the BN and BL statistics, was tested by varying several key characteristics of a microarray experiment (pDE, n, m and G). The pDE was increased and decreased by +10% and −10%, respectively. Figure 3 shows, not surprisingly, that as the pDE decreases all three B statistics perform worse, while they do better as the pDE increases. However, overall, BE consistently performed the best in these simulated datasets. In the osmotic stress data background corrected with MBC, BL failed under all three simulation scenarios (pDE−0·10, pDE and pDE+0·10) in several of 100 simulated datasets and thus is not shown.

To test the robustness of the model to the number of arrays (n), usually representing the number of biological replicates, we simulated the four model-datasets with $n=2$, 4 and 8. Figure 4 shows that BE consistently performed better than BN and BL, but the margin of BE superiority became smaller as the number of arrays increased. Interestingly, the improvement in the performance gained with the larger number of arrays was more apparent in the model-datasets corrected with MBC. The wiggly plot of

https://doi.org/10.1017/S0016672308009427 Published online by Cambridge University Press

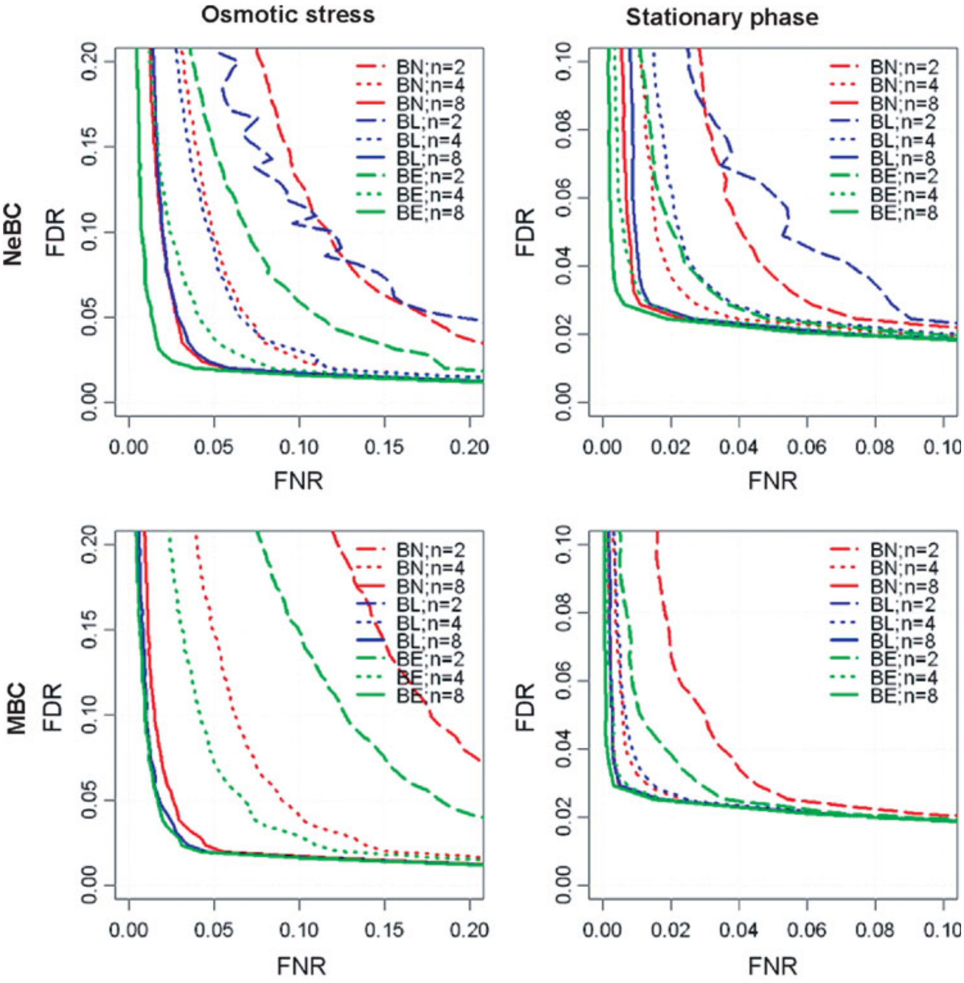

Fig. 4. FDR *vs*. FNR plots showing horizontal average of the BN (Lonnstedt & Speed, 2002; Smyth, 2004), BL (Bhowmick *et al*., 2006) and empirical Bayes EVD mixture model (BE) statistics from 100 simulated datasets in the four simulated model-datasets simulating the number of biological replicates (n) equal to 2, 4 and 8. 'NeBC' = normal-exponential convolution background correction method; 'MBC' = multiplicative background correction method.

BL for $n=2$ in osmotic stress and stationary-phase data corrected with NeBC is a consequence of outliers; in a few simulated datasets, there was a high FNR estimated for different values of FDR. When osmotic stress and stationary-phase data corrected with MBC were used, BL failed in the analysis of several simulated datasets with $n=2$ and 4 and $n=2$, respectively.

We also tested the sensitivity of the model to the number of technical replicates (m); we simulated $m=1$, 3 and 6 (Fig. 5). BE did better in all four simulated model-datasets, and under all simulated numbers of technical replicates. BL failed in the analysis of several simulations of the MBC osmotic stress dataset when simulated with all three tested values of m (1, 3 and 6). Interestingly, in the stationary-phase data (and to a lesser extent in the osmotic stress data) corrected with NeBC, the average performance of all three statistics was better under the simulation with $m=3$ than $m=6$. This may be a consequence of overfitting to the empirical data in estimation of hyperparameters. The performance of all three statistics, particularly BN and BL, was remarkably good in the simulated stationary-phase data corrected with MBC.

Finally, we also tested the performance of the BE method in experiments characterized with a high number of tested genes but with only a few DE genes (Fig. 6). We ran the models with combinations ($G=200$, pDE$=0{\cdot}05$) and ($G=1000$, pDE$=0{\cdot}01$) (note equal number (10) of DE genes in both the settings). BE did better than the other two statistics although the margin was not great, particularly in the stationary-phase data. That is expected because as the number of genes increases and only a few genes are DE, the distribution of gene-specific means becomes closer to the normal distribution. BL failed in the analysis of several realizations of the osmotic stress data corrected with MBC under the scenario with $G=200$ and pDE$=0{\cdot}05$.

https://doi.org/10.1017/S0016672308009427 Published online by Cambridge University Press

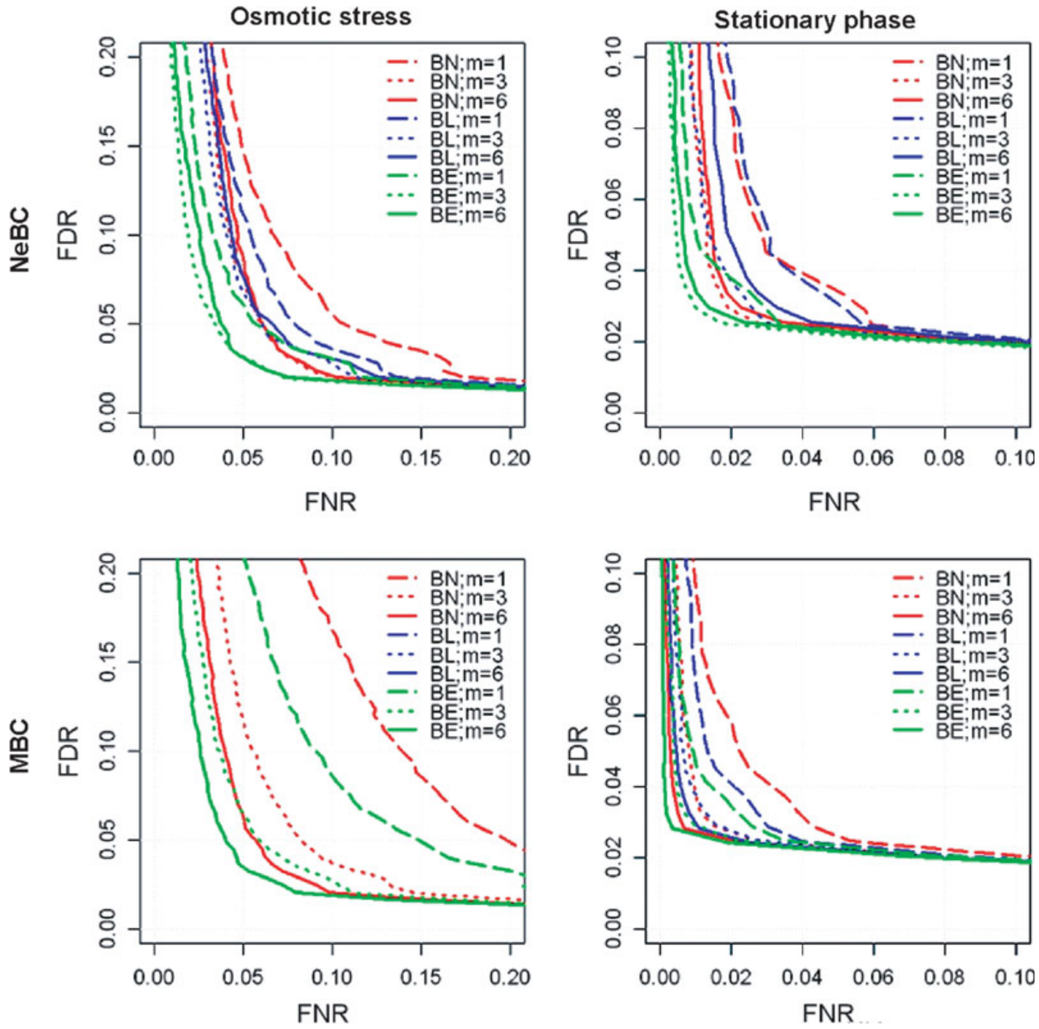

Fig. 5. FDR *vs*. FNR plots showing horizontal average of the BN (Lonnstedt & Speed, 2002; Smyth, 2004), BL (Bhowmick *et al*., 2006) and empirical Bayes EVD mixture model (BE) statistics from 100 simulated datasets in the four simulated model-datasets simulating the number of technical replicates (*m*) equal to 1, 3 and 6. 'NeBC' = normal-exponential convolution background correction method; 'MBC' = multiplicative background correction method.

As illustrated above, in all simulated settings, the performance of the BE model was at least as good as the performance of the other two models. BL failed in several simulated datasets under several simulation settings, mostly due to errors in integration and/or hyperparameter estimation. We would like to add here that we performed simulation analysis on the datasets generated under the modified Normal model that allows only mean ratios of expressions larger than 1 (i.e. higher gene expression in the wild-type as compared with the mutant). The results were similar to those obtained from simulation under the EVD model, so they are not shown here.

5. Discussion

We proposed a novel method (BE) for analysis of microarray data from experiments where only up- or down-regulated genes were expected or relevant. Its merit originates from an empirical Bayes framework and foundation in the EVT, as well as its good accuracy, precision and stability demonstrated in the simulated datasets. The main limitations of the BE method pertain to its sensitivity to the assumed pDE and the higher computational cost of its underlying numerical approximation through the MC integration method.

The main asset of empirical Bayes methods in analysis of differential expression is their use of information from hundreds or thousands of simultaneously tested genes to support testing at the gene level. However, empirical Bayes methods do make distributional assumptions, particularly when the choice of priors for the parameters is limited to conjugate priors (Lonnstedt, 2001). We stepped out of the conjugate prior class, and in experiments where only up- or only down-regulated genes are expected or relevant, assumed that the mean ratios of expressions

https://doi.org/10.1017/S0016672308009427 Published online by Cambridge University Press

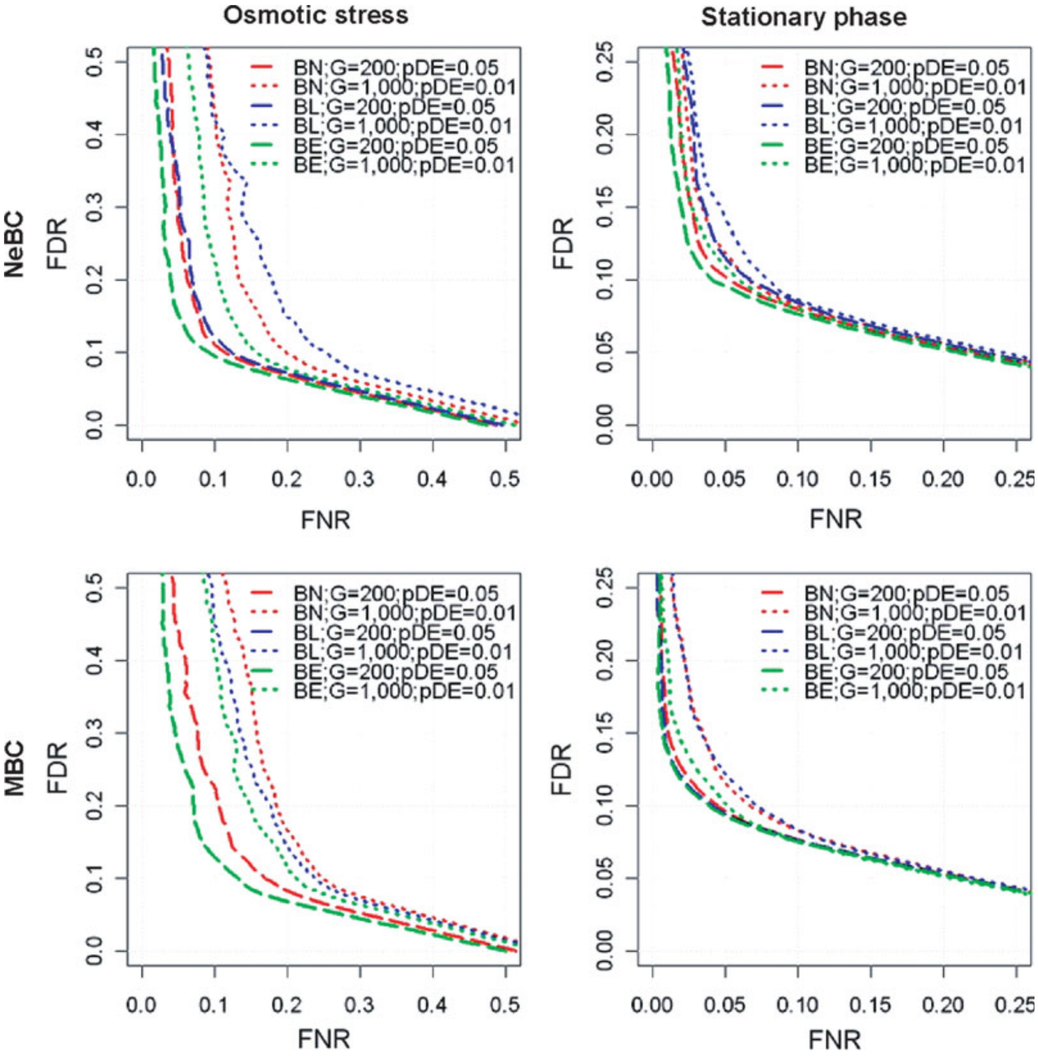

Fig. 6. FDR *vs*. FNR plots showing horizontal average of the BN (Lonnstedt & Speed, 2002; Smyth, 2004), BL (Bhowmick *et al*., 2006) and empirical Bayes EVD mixture model (BE) statistics from 100 simulated datasets in the four simulated model-datasets simulating an experiment with a low number of DE genes (10). 'NeBC' = normal-exponential convolution background correction method; 'MBC' = multiplicative background correction method.

of DE genes follow the EVD. In microarrays expecting a single direction of regulation, the validity of this assumption is based on the fact that the distribution of the true genes' mean ratios of expressions (the log of it) cannot be symmetric around zero. In the experiments where only a certain direction of regulation is relevant, it is advantageous to use the BE over the BN method even if the gene expression data are more or less symmetric around zero. That is because, being restricted to one direction of expression, due to the design of the experiment or the interest of an investigator, the genes characterized by such expressions are actually 'selected' beyond a threshold expression, and so, they show extreme behaviour.

The three families in EVD (Gumbel, Fréchet and Weibull) have distinct forms of tail behaviour, with finite limit distributions in its upper end-point for the Weibull distribution, and infinite ones for the Gumbel and Fréchet distributions (Coles, 2001). Furthermore, the density in the upper end-point decays exponentially for the Gumbel distribution and polynomially for the Fréchet distribution (Coles, 2001). While, consequently, the three families give quite different representations of the extreme value behaviour, it may not be obvious which best represents the data and the problem at hand. It is thus convenient to use a single family GEV rather than 'guessing' which of the three original EVD families is proper, because the data themselves (through inference about the shape parameter) determine the most appropriate type of tail behaviour (Coles, 2001).

To estimate the hyperparameters of the EVD, we applied the MLE procedure. While many techniques have been proposed for estimation of hyperparameters in extreme value models (including graphical techniques and moment-based and likelihood-based

https://doi.org/10.1017/S0016672308009427 Published online by Cambridge University Press

methods), the MLE approach is particularly appealing due to its all-around utility and adaptability (Coles, 2001). Indeed, our analyses showed that BE works well even when hyperparameters a, b and c are estimated from a small sample of genes ($\sim$10) related to smaller pDE and/or a smaller number of tested genes G. However, caution is needed when MLE is used to estimate means and SEs of EVD hyperparameters. Because the end-points of the GEV distributions are functions of the parameter values, regularity conditions required for the usual asymptotic properties associated with the maximum likelihood (ML) estimator are not satisfied by the GEV model (Coles, 2001). Particularly, in the experiments involving detection of down-regulated genes, if the value of the shape parameter (c) lies between -1 and $-0{\cdot}5$, ML estimators are generally obtainable, but do not have the standard asymptotic properties, while when $c<-1$, ML estimators are unlikely to be obtainable (Smith, 1985). Conversely, when $c>-0{\cdot}5$, ML estimators are regular (have the usual asymptotic properties) (Smith, 1985). Therefore, in experiments designed to detect up-regulated genes, where $c>0$, the usual asymptotic properties immediately apply. For detection of down-regulated genes, it is reasonable to negate the contrast estimators from gene-specific linear models before estimation of the MLE hyperparameters and subsequent approximation of BE statistics, i.e. to switch from modelling minima to modelling maxima.

While only one direction of differential expression is often expected in the wild-type *vs.* mutant experiments, it is possible to observe a few genes with their expression in an opposite direction. For example, while the sigma factor σ^B is a positive regulator of gene expression, some genes show higher transcript levels in a *sigB* null mutant, likely representing indirect effects, which may, however, be biologically relevant. For example, in the bacterium *Staphylococcus aureus*, genes encoding a number of exoenzymes and toxins show higher transcript levels in a *sigB* null mutant strain as compared with the parent strain with an intact *sigB* gene, suggesting indirect negative regulation of these genes by σ^B (Bischoff *et al.*, 2004). To test differential expression of a suspected biologically relevant gene regulated in the direction opposite to the direction of regulation of the majority of the genes in the experiment, its contrast estimator should be negated prior to applying the BE method. Otherwise, the BE method would rank this gene low, indicating non-differential expression.

Because there is no conjugate structure between the normal likelihood of a gene being DE and the EVD prior, it is impossible to estimate BE analytically. Therefore, the MC integration approximation technique was applied. While this allowed us to base the choice of the prior on the nature of the problem, rather than on the available conjugate priors, it came at the price of a higher, but still reasonable (e.g. 16 min for 50 000 iterations in a dataset with 10 000 genes), computational cost related to a large number of tested genes and (usually) a large number of iterations required to achieve a reasonable accuracy of the BE statistics (because the accuracy increases only as the square root of the number of iterations). The MCSEs of the estimated BE statistics differed for various genes, being the lowest for genes with a BE statistic around zero. That is as expected because genes with a high mean expression ratio, corresponding to the upper tail of the EVD prior, will likely have a high BE statistic. Similarly, genes with a high variance, corresponding to the upper tail of the variance prior, will likely have a low BE statistic. At each iteration, the probability of drawing a sample from a tail is lower than from a body of the prior distribution, so these genes require more iterations to achieve sufficiently low MCSEs. If the BE statistic is used to identify DE genes, as in the present study, an even lower number of iterations may be sufficient. However, when BE is used with the purpose of precise ranking of genes by their strength of differential expression, a larger number of iterations may be necessary.

As discussed by Smyth (2004), it is possible to estimate pDE from the data, for example, from posterior odds through MLE or based on the P-values from t-statistics (moderated or ordinary) (such as in Langaas *et al.*, 2005 and Wu *et al.*, 2006). Nevertheless, estimation of pDE is unstable (related to collapse in estimation of posterior odds caused by boundary values of pDE$=$1 and pDE$=$0 having positive probability), and may be sensitive to the particular prior distribution assumed for the contrast estimators from linear models and to dependence between the genes (Smyth, 2004). To bypass these problems, Smyth (2004) and Lonnstedt & Speed, (2002) fixed pDE and then estimated the remaining hyperparameter. This same approach has been adopted in the BE method. A wrong choice of pDE would not change the shape of the BE statistic *vs.* fold change plot (i.e. ranking of the genes would stay intact), but it would move the BE up and down on the y-axis (as in Lonnstedt & Speed, 2002). Also, through influence of the pDE on the hyperparameters of EVD, it would vary fold change corresponding to BE$=$0. Accepting that BE$=$0 is a natural cutoff separating DE from non-DE, the fold change corresponding to this point represents the 'meaningful fold change cutoff' above which a gene could be considered meaningfully DE. This 'meaningful fold change cutoff' could give some indication of whether the chosen pDE is wrong. For example, in experiments expecting only up-regulated genes, the

https://doi.org/10.1017/S0016672308009427 Published online by Cambridge University Press

'meaningful fold change cutoff' ⩽1 would indicate that the selected pDE is too high.

The accuracy and precision of BN, BL and BE were assessed based on the achieved FDR and FNR over 100 simulated datasets generated under the EVD model. In an experiment, a researcher could choose a tolerable FDR and try to minimize FNR (i.e. maximize power) or choose a tolerable FNR and try to minimize FDR. In the simulated datasets, we mimicked both situations, and in both, the accuracy of the BE method was at least as good as or better than the other two tested methods. In addition to being more accurate, BE was also more precise than BN and BL in the simulated datasets. We also tested the sensitivity of BE to several key characteristics of a microarray experiment, namely, the pDE, the number of arrays (usually representing the number of biological replicates) and the number of technical replicates, as well as its performance in experiments with a high number of tested genes but with only a few DE genes. Overall, BE performed the best in all four simulated model-datasets and under all the tested simulated scenarios. When applied to real datasets, BE performed well, with most of the findings being validated by other independent studies (elaborated in the Appendix). This 'reality check' is a valuable addition to simulation studies because datasets used in simulation analyses have been replicated from the same model as the BE is built upon, thus limiting ability of the simulation studies to give a completely unbiased evaluation of BE performance.

6. Conclusions

In microarray experiments where only one direction of expression is expected or relevant (only up- or down-regulation), such as in certain experiments involving wild-type *vs.* knockout design (Kazmierczak *et al.*, 2003; van Schaik *et al.*, 2007) and some restricted coverage arrays, including those performed with the purpose of identifying novel gene markers and drug targets (Suzuki *et al.*, 2002; Kobayashi *et al.*, 2004), the nature of genes' true mean ratios of expressions is extreme. EVT was designed specifically to study extreme behaviour, and is therefore an excellent candidate for use in analysis of differential expression in such experiments. In this paper, we proposed a new empirical Bayes method, BE, for analysis of differential expression that is based on the EVT, and we compared its performance with two other empirical Bayes methods reportedly used in analysis of such data. The first of the two methods, BN, is a very popular method that is, however, based on a distributional assumption invalid for experiments where only one direction of expression is anticipated or of interest. The distributional assumption of the other method, BL, is valid, but the method is unstable and its performance is still comparable with BN. Based on the series of simulation analyses and analyses of two real datasets, we believe that the BE method has greater accuracy, precision and stability than the BN and BL methods. It thus seems promising and useful for analysis of differential expression in experiments where only up- or down-regulated genes are relevant or expected. Because custom, restricted-coverage microarray experiments, including those described here, are likely to become much more common in the future due to their possible use in therapeutic and diagnostic applications (Liu-Stratton *et al.*, 2004), development and use of appropriate bioinformatics tools will become vital, and the advantages of the BE method may become even more evident.

This work was funded by the National Institutes of Health (RO1-AI052151-01A1 to Kathryn J. Boor and M.W.) and the United States Department of Agriculture (USDA) National Needs Fellowship grant (2002-38420-11738S to S.R.).

Appendix

The appendix for this paper is available online at the journal's website (http://journals.cambridge.org/grh).

References

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

Bhowmick, D., Davison, A. C., Goldstein, D. R. & Ruffieux, Y. (2006). A Laplace mixture model for identification of differential expression in experiments. *Biostatistics* **7**, 630–641.

Bischoff, M., Dunman, P., Kormanec, J., Macapagal, D., Murphy, E., Mounts, W., Berger-Bachi, B. & Projan, S. (2004). Microarray-based analysis of the *Staphylococcus aureus* σ^B regulon. *Journal of Bacteriology* **186**, 4085–4099.

Cleveland, W. S. & Devlin, S. J. (1988). Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American Statistical Association* **83**, 596–610.

Coles, S. (2001). *An Introduction to Statistical Modeling of Extreme Values*. Springer Series in Statistics. Berlin: Springer.

Efron, B., Tibshirani, R., Storey, J. D. & 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. & Brettin, T. (2003). Statistical analysis of microarray data: a Bayesian approach. *Biostatistics* **4**, 597–620.

Kazmierczak, M. J., Mithoe, S. C., Boor, K. J. & Wiedmann, M. (2003). *Listeria monocytogenes* σ^B regulates stress response and virulence functions. *Journal of Bacteriology* **185**, 5722–5734.

Kobayashi, K., Nishioka, M., Kohno, T., Nakamoto, M., Maeshima, A., Aoyagi, K., Sasaki, H., Takenoshita, S.,

https://doi.org/10.1017/S0016672308009427 Published online by Cambridge University Press

Sugimura, H. & Yokota, J. (2004). Identification of genes whose expression is upregulated in lung adenocarcinoma cells in comparison with type II alveolar cells and bronchiolar epithelial cells *in vivo*. Oncogene **23**, 3089–3096.

Langaas, M., Lindqvist, B. H. & Ferkingstad, E. (2005). Estimating the proportion of true null hypotheses, with application to DNA microarray data. *Journal of the Royal Statistical Society, Series B* **67**, 555–572.

Liu-Stratton, Y., Roy, S. & Sen, C. K. (2004). DNA microarray technology in nutraceutical and food safety. *Toxicology Letters* **150**, 29–42.

Lonnstedt, I. (2001). Replicated microarray data. Licentiate Thesis. Department of Mathematics, Uppsala University.

Lonnstedt, I. & Britton, T. (2005). Hierarchical Bayes models for cDNA microarray gene expression. *Biostatistics* **6**, 279–291.

Lonnstedt, I. & Speed, T. P. (2002). Replicated microarray data. *Statistica Sinica* **12**, 31–46.

Nelder, J. A. & Mead, R. (1965). A simplex algorithm for function minimization. *Computer Journal* **7**, 308–313.

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

Panjer, H. H. (2006). Fitting extreme value models. In *Operational Risk: Modeling Analytics* (ed. H. H. Panjar), pp. 383–393. New York: John Wiley and Sons.

Pawitan, Y., Michiels, S., Koscielny, S., Gusnanto, A. & Ploner, A. (2005). False discovery rate, sensitivity and sample size for microarray studies. *Bioinformatics* **21**, 3017–3024.

Smith, R. L. (1985). Maximum likelihood estimation in a class of non-regular cases. *Biometrika* **72**, 67–90.

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**, Article 3.

Smyth, G. K. (2005). Limma: linear models for microarray data. In *Bioinformatics and Computational Biology Solutions Using R and Bioconductor* (ed. R. Gentleman, V. Carey, S. Dudoit, R. Irizarry & W. Huber), pp. 397–420. New York: Springer.

Smyth, G. K., Michaud, J. & Scott, H. (2005). The use of within-array replicate spots for assessing differential expression in microarray experiments. *Bioinformatics* **21**, 2067–2075.

Suzuki, H., Gabrielson, E., Chen, W., Anbazhagan, R., van Engeland, M., Weijenberg, M. P., Herman, J. G. & Baylin, S. B. (2002). A genomic screen for genes upregulated by demethylation and histone deacetylase inhibition in human colorectal cancer. *Nature Genetics* **31**, 141–149.

Tanner, M. A. (1996). *Tools for Statistical Inference. Methods for the Exploration of Posterior Distributions and Likelihood Functions*. New York: Springer-Verlag.

Taylor, J. R. (1982). *An Introduction to Error Analysis. The Study of Uncertainties in Physical Measurements*. Herndon, VA: University Science Books.

Tusher, V. G., Tibshirani, R. & Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the USA* **98**, 5116–5121.

van Schaik, W., van der Voort, M., Molenaar, D., Moezelaar, R., de Vos, W. M. & Abee, T. (2007). Identification of the σ^{B} regulon of *Bacillus cereus* and conservation of σ^{B}-regulated genes in low-GC-content Gram-positive bacteria. *Journal of Bacteriology* **18**, 4384–4390.

Verhoeven, K. J. F., Simonsen, K. L. & McIntyre, L. M. (2005). Implementing false discovery rate control: increasing your power. *Oikos* **108**, 643–647.

Wu, B., Guan, Z. & Zhao, H. (2006). Parametric and nonparametric FDR estimation revisited. *Biometrics* **62**, 735–744.

Zhang, D., Zhang, M. & Wells, M. T. (2006). Multiplicative background correction for spotted microarrays to improve reproducibility. *Genetical Research* **87**, 195–206.

https://doi.org/10.1017/S0016672308009427 Published online by Cambridge University Press