

Use of Within-Array Replicate Spots for Assessing Differential Expression in Microarray Experiments*

Gordon K. Smyth, Joëlle Michaud and Hamish S. Scott
Walter and Eliza Hall Institute of Medical Research
Melbourne, Vic 3050, Australia

26 December 2004
with corrections 22 June 2006 and 29 January 2013

Abstract

Motivation. Spotted arrays are often printed with probes in duplicate or triplicate, but current methods for assessing differential expression are not able to make full use of the resulting information. Usual practice is to average the duplicate or triplicate results for each probe before assessing differential expression. This loses valuable information about gene-wise variability.

Results. A method is proposed for extracting more information from within-array replicate spots in microarray experiments by estimating the strength of the correlation between them. The method involves fitting separate linear models to the expression data for each gene but with a common value for the between-replicate correlation. The method greatly improves the precision with which the genewise variances are estimated and thereby improves inference methods designed to identify differentially expressed genes. The method may be combined with empirical Bayes methods for moderating the genewise variances between genes. The method is validated using data from a microarray experiment involving calibration and ratio control spots in conjunction with spiked-in RNA. Comparing results for calibration and ratio control spots shows that the common correlation method results in substantially better discrimination of differentially expressed genes from those which are not. The spike-in experiment also confirms that the results may be further improved by empirical Bayes smoothing of the variances when the sample size is small.

Availability. The methodology is implemented in the *limma* software package for R, available from the CRAN repository <http://www.r-project.org>.

1 Introduction

Microarrays measure the mRNA expression of tens of thousands of genes in a single hybridization experiment. Designed experiments involving two or more microarrays hybridized with RNA from different sources generate expression profiles which can help classify the genes according to functional groups or molecular pathways. Although much attention has been given to the statistical analysis of microarray data many problems are still unresolved (Nguyen et al., 2002; Smyth et al., 2003; Parmigiani et al., 2003; Speed, 2003; Causton et al., 2003; Firestein and Pissetsky, 2002; Tilstone, 2003). Particular challenges and opportunities arise from the multiplicity of genes and the possibilities for parallel inference.

A standard analysis method is to fit the same statistical model separately to the expression measurements for each gene (Wolfinger et al., 2001; Yang and Speed, 2003). A number of authors have noted that inference for each individual gene can be made more reliable by making use of information generated from the whole ensemble of genes (Newton et al., 2001; Tusher et al., 2001; Efron et al., 2001; Efron and Tibshirani, 2002; Lönnstedt and Speed, 2002; Kendzioriski et al., 2003; Newton et al., 2004; Smyth, 2004). Such methods have not as yet been applied to experimental designs in which there are technical or biological replicates leading to multiple strata of random variation for each gene. This article develops a between-gene moderation method appropriate for a particular type of technical replication, that of within-array replicate spots. The method proposed is particularly simple in that a suitably chosen parameter is constrained to be common between the genes. The treatment proposed here for within-array replicates may be combined with moderation methods designed for a single error strata.

Spotted microarrays are produced by printing cDNA or oligonucleotide sequences on glass slides using a robotic printer. The spots are laid down using a printhead made up of capillary print tips or pins or inkjets. The DNA is prepared in 96-well or 384-well plates ready for printing, nor-

* *Bioinformatics* **21** no. 9, 2005, pages 2067–2075.

mally one well for each distinct probe. The robot acquires DNA by dipping the tips of the print head into the wells of the plate before depositing the DNA on the glass slide. In most cases only a small proportion of the DNA in each well is actually printed onto the arrays and any excess is discarded. Provided that there is space on the array, there is no cost, apart from printing time, in printing two or more spots on the arrays from each well. This is accomplished by programming the robotic printer to dip the print head more than once into the same set of wells. This results in a printed array in which each gene appears two or more times a fixed distance apart. Usually multiple printing produces two spots of each gene but an arbitrarily large number of replicate spots may be printed if there is sufficient space to accommodate them on the arrays.

Normally the replicate spots are printed side-by-side on the same row or side-by-side on the same column or on the top and bottom halves of the array. Any intensity or log-ratio measurements made from the replicate spots will be positively correlated through being observed on the same array. Replicate spots which are side-by-side are likely to be very highly correlated since not only are they printed with the same gene but are spatially close together and therefore likely to share many common causes including local effects on the array surfaces as well as hybridization and labelling effects. Indeed the value of having multiple prints of each clone on an array has often been questioned given the low within-array variability compared to between array variability (Tran et al., 2002). Replicate spots in the top and bottom halves of the array are also likely to be positively correlated but less so than side-by-side replicates.

Replicate spots are often used as a quality assessment tool since disagreement between replicates is strong evidence that at least one of the spots is affected by a local artifact. Repeatability of the log-ratios across replicate spots within arrays can be used as a basis for removing outlier spots (Tseng et al., 2001; Hoffmann et al., 2002; Yang et al., 2002; Jenssen et al., 2002; Lyne et al., 2003; König et al., 2004), to construct spot quality measures (Beissbarth et al., 2000) or to evaluate the effectiveness of a spot quality scheme (Wang et al., 2001). It is almost universal practice to average the log-intensities or log-ratios obtained from within-array replicate spots before conducting formal statistical analysis of differential expression (Andrews et al., 2000; Tseng et al., 2001; Berwanger et al., 2002; Hoffmann et al., 2002; Yang et al., 2002; Kaynak et al., 2003; Lyne et al., 2003), although averaging can cause complications when some of the log-ratios are missing or when there are spot-specific quality weights. Many public microarray database programs, such as the Stanford Microarray Database, automatically average log-ratios from duplicate spots. A relatively small number of studies have used within-array replicate level information to improve the assessment of differential expression (Baggerly et al., 2001;

Boer et al., 2001; Fan et al., 2004).

The method developed in this paper extracts more information from the within-array replicate spots by estimating the correlation between them. A simple model is explored in which the between-replicate correlation is taken to be constant across genes. The method uses a consensus estimator of the correlation across genes in such a way that the correlation can be taken to be known at the individual gene level. Compared with simply averaging replicate spots, this method greatly improves the precision with which the genewise variances are estimated and thereby improves inference methods designed to identify differentially expressed genes.

The method is validated using data from a microarray experiment involving calibration and ratio control spots in conjunction with spiked-in RNA. Comparing results for calibration and ratio control spots shows that the within-array correlation method results in substantially better discrimination of differentially expressed genes from those which are not compared with simply averaging the replicate spots. On this data the proposed method increases power to detect differential expression when it is present without incurring a greater rate of Type I errors when it is not.

2 cDNA Microarray Preparation Methods

2.1 Spike-in Control Spots

This paper uses data from a set of 26 cDNA microarrays which were printed and hybridized as part of a study on human transcription factors. The paper presents data only from the spike-in control spots.

The arrays were printed at the Australian Genome Research Facility with the Hs8k cDNA clone library from Research Genetics and a selection of control spots. Each array was printed with twelve sets of the Lucidea Universal Scorecard system (Amersham). Spots were printed in duplicate, side-by-side by rows, including the twelve sets of ScoreCard spots.

The RNA samples hybridized to the arrays included ScoreCard spike mixes according to the Lucidea ScoreCard User's Guide. The ScoreCard system includes calibration and ratio control spots designed to generate pre-determined fold changes. Each set of ScoreCard spots includes ten calibration spots labeled here Calib1 to Calib10 which have theoretical fold change one and are expressed at successively decreasing intensities. The ratio controls have fold changes as follows: three fold up and down at low intensity (3UL and 3DL), three fold up and down at high intensity (3UH and 3DH), ten fold up and down at low intensity (10UL and 10DL) and ten fold up and down at high intensity (10UH

and 10DH). The same spike mix was applied to all the arrays, so the arrays can be treated as a set of replicate arrays for the purposes of the ScoreCard spots.

2.2 Hybridization

50 μ g of total RNA extracted from HeLa cells and 1 μ l of either reference or test spike mRNA was reverse transcribed using an anchored oligo(dT) primer and 200 units of Superscript II reverse transcriptase (Invitrogen) in the presence of 25mM dATP, 25mM dCTP, 25mM dGTP, 15mM aminoallyl-dUTP (SIGMA #A0410) and 10mM dTTP. The single strand cDNA was purified using QIAquick PCR purification kit (Qiagen) and labelled with CyDye post-labelling dye (Amersham) for an hour. After a second purification as above, both Cy-5 and Cy-3 labelled cDNAs were pooled and mixed to 25 μ g of human Cot1 DNA, 38 μ g of polyA DNA and 50 μ g of salmon sperm DNA. The mixture was concentrated using a vacuum dryer and resuspended in 50% formamide, 5x SSC and 0.1% SDS.

The arrays were incubated in 50% formamide, 5x SSC, 0.1% SDS and 10mg/ml BSA for 45 minutes at 42°C, rinsed with distilled water and dried using an air gun. The labelled cDNA mixture was denatured at 95°C for 5 minutes, incubated at 45°C for 30 minutes and cooled to room temperature before being pipetted onto the array. The slides were incubated overnight at 42°C in hybridisation chamber (Corning) placed in a water bath. After incubation the slides were washed in 1x SSC/0.2% SDS solution for 5 minutes, in 0.1xSSC/0.2%SDS solution for 5 minutes, and twice in 0.1X SSC for 2 minutes. The slides were then spun dry using a centrifuge.

2.3 Image Analysis and Normalization

The arrays were scanned using a Genepix 4000B scanner with adjusted settings in order to obtain a similar green and red overall intensity. The images were analysed using the SPOT software (Buckley, 2000). Foreground intensities were background corrected using the ‘morph’ background measure and the scorecard spot log-ratios were normalized using global loess normalization with the default smoothing span of 0.3 (Yang et al., 2001; Smyth and Speed, 2003).

3 The Balanced Single Sample Problem

3.1 Individual correlations

For simplicity, consider first a series of n replicate two-color microarray experiments, each array hybridized with RNA from the same two sources. Suppose that each gene is replicated m times on each array at a fixed distance apart. Im-

age analysis and normalization of the microarray data will yield a log-ratio of expression y_{gij} for each spot. Here y_{gij} is the log-ratio for gene $g = 1, \dots, G$, array $i = 1, \dots, n$ and replicate $j = 1, \dots, m$. Usually y_{gij} is a normalized version of $\log_2(R_{gij}/G_{gij})$ where R_{gij} is the measured red intensity while G_{gij} is the measured green intensity for that spot. Assume that

$$E(y_{gij}) = \mu_g$$

where μ_g is the true log-ratio of the expression levels for gene g . Interest lies in estimating μ_g and especially in testing $H_0 : \mu_g = 0$.

It is reasonable to assume that observations made on different arrays for a given gene are independent or nearly so. On the other hand, replicate observations made on the same array are likely to be correlated, perhaps highly so. For the remainder of this article, the term ‘replicate spots’ will be taken to refer to spots on the same array. Let ρ_g , be the correlation between replicate spots for gene g . We will assume that

$$\text{var}(y_{gij}) = \sigma_g^2$$

and

$$\text{cor}(y_{gij}, y_{gij'}) = \rho_g$$

for $j \neq j'$. Observations with different i are assumed independent. Observations on different genes on the same array are also likely to be correlated. The correlations between genes however are highly problematic to estimate, because of the very large number of genes compared to the number of arrays, and so these correlations are left unspecified in this article. If the replicate spots are close together we expect ρ_g to be large, perhaps close to unity. If the replicate spots are far apart, the correlation will be much smaller. Note that ρ_g is constrained according to $-1/(m-1) \leq \rho_g \leq 1$ by the requirement that the covariance matrix of the y_{gij} be non-negative definite.

It will be further assumed in this article that the y_{gij} are normally distributed. Although this assumption will be used in deriving specific results in this article, most of the results of this article do not depend on normality for their usefulness. See further comments on this issue in the discussion section.

For each gene g write \bar{y}_{gi} for the sample mean of the replicate observations on array i and \bar{y}_g for the overall sample mean across all arrays. For each gene let s_g^B be the between-arrays standard deviation,

$$(s_g^B)^2 = \frac{m}{n-1} \sum_{i=1}^n (\bar{y}_{gi} - \bar{y}_g)^2$$

and s_g^W be the within-arrays standard deviation,

$$(s_g^W)^2 = \frac{1}{n(m-1)} \sum_{i=1}^n \sum_{j=1}^m (y_{gij} - \bar{y}_{gi})^2.$$

Then \bar{y}_g , s_g^B and s_g^W are mutually independent and sufficient for μ_g , σ_g and ρ_g with

$$\begin{aligned}\bar{y}_g &\sim N\left(\mu_g, \frac{\sigma_g^2}{nm}\{1 + (m-1)\rho_g\}\right) \\ (s_g^B)^2 &\sim \sigma_g^2\{1 + (m-1)\rho_g\}\frac{\chi_{n-1}^2}{n-1} \\ (s_g^W)^2 &\sim \sigma_g^2\{1 - \rho_g\}\frac{\chi_{n(m-1)}^2}{n(m-1)}\end{aligned}$$

Under this model, inference about μ_g can be conducted entirely using \bar{y}_g and s_g^B . The within standard deviation s_g^W does not contribute any further information. The maximum likelihood estimator of μ_g is

$$\hat{\mu}_g = \bar{y}_g$$

and the most powerful test statistic for testing $H_0 : \mu_g = 0$ is

$$t_g = \frac{\bar{y}_g}{s_g^B/\sqrt{nm}}. \quad (1)$$

If $\mu_g = 0$ then $t \sim t_{n-1}$. This explains why it is usual practice to average the replicate spots before undertaking inference for microarrays with within-array replication.

It is useful for later reference to consider the estimation of ρ_g even though it does not contribute here to inference about μ_g . Let

$$\theta_g = \frac{1}{2} \log\left(\frac{1 + (m-1)\rho_g}{1 - \rho_g}\right).$$

Note that θ_g is a monotonic increasing transformation of ρ_g which takes values on the whole real line. The transformation is reversed by

$$\rho_g = \frac{\exp(2\theta_g) - 1}{\exp(2\theta_g) + m - 1}$$

which reduces to $\rho_g = \tanh \theta_g$ when $m = 2$. The residual maximum likelihood (REML) (Searle et al., 1992) estimator of θ_g is

$$\hat{\theta}_g = \log(s_g^B/s_g^W)$$

which is distributed as $\theta_g + \frac{1}{2} \log F_{n-1, n(m-1)}$. This shows that

$$E(\hat{\theta}_g) = \theta_g + b(n-1, n\{m-1\})$$

where the bias is determined by the function

$$b(f_1, f_2) = \frac{1}{2} \left\{ \psi\left(\frac{f_1}{2}\right) - \log\left(\frac{f_1}{2}\right) - \psi\left(\frac{f_2}{2}\right) + \log\left(\frac{f_2}{2}\right) \right\}$$

where ψ is the digamma function. The variance is

$$\text{var}(\hat{\theta}_g) = v(n-1, n\{m-1\})$$

with

$$v(f_1, f_2) = \frac{1}{4} \left\{ \psi\left(\frac{f_1}{2}\right) + \psi\left(\frac{f_2}{2}\right) \right\}$$

where ψ is the trigamma function. The distribution of $\hat{\theta}_g$ is somewhat skewed to the left because of the differing degrees of freedom for s_g^B and s_g^W . In the worst case with $n = m = 2$ the bias of $\hat{\theta}_g$ is -0.35 .

3.2 Common correlation

Now we make the simplifying assumption that the between-replicate correlation is common across genes, $\rho_g = \rho$ for all g . This assumption is motivated by the belief that the correlation springs mainly from the physical proximity of the replicate spots on the same array. The robotic printing ensures that the spacing between the replicate spots is the same for all genes and all arrays. In practice it will not be necessary that the assumption be precisely true but rather that the correlations be sufficiently stable to make the common correlation model a useful one. This is likely to be true when the between and within standard deviations s_g^B and s_g^W are positively associated across genes, meaning that the correlations are much more stable than the variances. This has been true in all microarray experiments seen by the authors so far.

Let

$$\theta = \frac{1}{2} \log\left(\frac{1 + (m-1)\rho}{1 - \rho}\right).$$

If observations on different genes were independent then the REML estimator of θ would be

$$\hat{\theta} = \frac{1}{2} \log\left\{\frac{\sum_{g=1}^G (s_g^B)^2}{\sum_{g=1}^G (s_g^W)^2}\right\} \quad (2)$$

which would be distributed as $\theta + \frac{1}{2} \log F_{G(n-1), Gn(m-1)}$. This estimator remains consistent as $n \rightarrow \infty$ even if the genes are not independent because it requires only that the mean of the $(s_g^B)^2$ and the mean of the $(s_g^W)^2$ converge to quantities in the ratio of $1 + (m-1)\rho$ to $1 - \rho$. For the same reasons it requires only weak assumptions on the dependence between genes to be consistent as $G \rightarrow \infty$. In practice this estimator is likely to be very accurate if the number of genes G is large. Under the assumption of independence between genes, the bias is

$$b(G\{n-1\}, Gn\{m-1\})$$

and the variance is

$$\text{var}(\hat{\theta}_g) = v(G\{n-1\}, Gn\{m-1\})$$

both of which are very small when G is large. For example if $G = 1000$ and $n = m = 2$ the above bias is minimal at -0.00025 while the standard deviation is 0.016 .

The fact that the correlation is common between genes does not change the estimator $\hat{\mu}_g$ for each gene but does change substantially inference about σ_g^2 . Because the common correlation can be estimated very accurately from the ensemble of genes, ρ may be treated as known to a very good approximation when undertaking inference about each individual gene. This means that s_g^W can contribute to the estimator of σ_g^2 , improving the precision with which we judge whether μ_g is nonzero. The REML estimator of σ_g^2 is approximately

$$s_g^2 = \frac{1}{nm-1} \left\{ \frac{(n-1)(s_g^B)^2}{1+(m-1)\hat{\rho}} + \frac{n(m-1)(s_g^W)^2}{1-\hat{\rho}} \right\} \quad (3)$$

which is approximately distributed as $\sigma^2 \chi_{nm-1}^2 / (nm-1)$. The test statistic for testing $H_0 : \mu_g = 0$ now becomes

$$t_g = \frac{\bar{y}_g}{s_g \{ [1 + (m-1)\hat{\rho}] / (nm) \}^{1/2}} \quad (4)$$

and this follows a t_{nm-1} distribution under the null hypothesis. The number of degrees of freedom associated with the test statistic is more than doubled compared with Section 3.1 even in the most conservative case when there are $m = 2$ within-array replicates.

4 Results for Balanced linear models

Section 3 considered only replicate arrays comparing two RNA sources. The results of Section 3 generalize easily to arbitrarily complicated microarray experiments comparing two or more RNA sources. Let \mathbf{y}_g be the vector containing the nm log-ratios or log-intensities observed for gene g . A general microarray experiment can be represented by a linear model

$$E(\mathbf{y}_g) = X\boldsymbol{\beta}_g$$

where X is a known $nm \times k$ dimensional design matrix specifying the experimental design and $\boldsymbol{\beta}_g$ is a vector of k regression coefficients (Yang and Speed, 2003; Smyth, 2004). In the order for the linear model to be identifiable we assume that $k < n$ and that the matrix X is of full column rank. When there are m replicates of each gene on each array, there will be m repeated rows of the design matrix X corresponding to each set of m replicate spots. The covariance matrix is

$$\text{var}(\mathbf{y}_g) = \sigma_g^2 R_g$$

where R_g is the block diagonal matrix with n blocks equal to the $m \times m$ correlation matrix

$$\begin{pmatrix} 1 & \rho_g & \cdots & \rho_g \\ \rho_g & 1 & \cdots & \rho_g \\ \vdots & \vdots & \ddots & \vdots \\ \rho_g & \cdots & \rho_g & 1 \end{pmatrix}$$

Let $\alpha_g = \mathbf{c}^T \boldsymbol{\beta}_g$, where \mathbf{c} is a vector of known constants, be a particular contrast or linear combination of the regression coefficients and suppose that interest lies in testing $H_0 : \alpha_g = 0$. This formulation is sufficiently general to accommodate a wide variety of microarray experiments including dye-swaps, time course experiments and factorial experiments. It is also applicable to single-channel microarray experiments for which y_{gij} is a normalized version of $\log_2 I_{gij}$ where I_{gij} is the measured intensity for that spot.

Generalizing from replicate arrays to the linear model causes little extra complication for the theory of Section 2. Let $\bar{\mathbf{y}}_g$ be the n -vector of array means \bar{y}_{gi} and let \bar{X} be the reduced $n \times k$ dimensional design matrix in which there is only one row instead of m rows for each gene by array combination. Then

$$E(\bar{\mathbf{y}}_g) = \bar{X}\boldsymbol{\beta}_g \quad (5)$$

To generalize the results of Section 2.2, we simply generalize the between-arrays standard deviation s_g^B to be m times the residual standard error which arises from fitting the linear model (5). This mean square is now on $n - k$ instead of $n - 1$ degrees of freedom. Let $\hat{\boldsymbol{\beta}}_g$ be the estimator of $\boldsymbol{\beta}_g$ from fitting this model and write $\hat{\alpha}_g = \mathbf{c}^T \hat{\boldsymbol{\beta}}_g$. The estimate $\hat{\boldsymbol{\beta}}_g$ from the reduced linear model (5) is the same as that from the full linear model for \mathbf{y}_g . Note that

$$\text{var} \hat{\alpha}_g = \sigma_g^2 u_g^2$$

with

$$u_g^2 = \mathbf{c}^T (X^T R_g^{-1} X)^{-1} \mathbf{c} = \frac{1 + (m-1)\rho_g}{m} \mathbf{c}^T (\bar{X}^T \bar{X})^{-1} \mathbf{c}$$

The t -statistic (1) arising from the individual correlation model is now

$$t_g = \frac{\hat{\alpha}_g}{s_g^B u_g}$$

which is on $n - k$ degrees of freedom.

Assuming now that $\rho_g = \rho$, the common correlation estimator (2) would now be distributed as $\theta + \frac{1}{2} \log F_{G(n-k), Gmn}$ if the genes were independent. The pooled variance estimator (3) now becomes

$$s_g^2 = \frac{1}{nm-k} \left\{ \frac{(n-k)(s_g^B)^2}{1+(m-1)\hat{\rho}} + \frac{n(m-1)(s_g^W)^2}{1-\hat{\rho}} \right\}$$

which is on $mn - k$ degrees of freedom. The t -statistic (4) now becomes

$$t_g = \frac{\hat{\alpha}_g}{s_g u}$$

on $mn - k$ degrees of freedom and is used to test $H_0 : \alpha_g = 0$.

It can be seen that the relative difference in degrees of freedom between s_g^B and s_g^W can be large if k is larger than one and especially if k is not much smaller than n . This means

that the gain in degrees of freedom of s_g over s_g^B which results from assuming common correlations is especially important for larger values of k , i.e., for designed experiments involving a larger number of distinct RNA sources to be compared.

5 Results for Unbalanced models

Suppose that now there are spot-specific weights w_{gij} associated with the observations so that

$$\text{var } y_{gij} = \sigma_g^2 / w_{gij}$$

The weights may arise from quality assessment or quality filtering of the spots (Smyth and Speed, 2003). In general the weights are non-negative but may be permitted to take value zero corresponding to log-ratios or log-intensities which are missing. The linear model is as before

$$E(\mathbf{y}_g) = X\boldsymbol{\beta}_g$$

but the covariance matrix is now

$$\text{var}(\mathbf{y}) = \sigma_g^2 \text{diag}(w_{gij}^{-1/2}) R_g \text{diag}(w_{gij}^{-1/2}).$$

Unlike in Section 3 the estimator $\hat{\boldsymbol{\beta}}_g$ of $\boldsymbol{\beta}_g$ is now somewhat dependent on the estimated value for ρ_g . This produces an unbalanced statistical model in which there are no non-iterative formulae for the REML estimators of σ_g or ρ_g . On the other hand, iterative computational procedures are readily available to compute the numerical REML estimates $\hat{\sigma}_g$ and $\hat{\rho}_g$ for any given data set (Pinheiro and Bates, 2000).

Assume now that $\rho_g = \rho$. Even assuming independence between the genes, exact REML estimation of the common correlation would require iterative computation using the entire data set. This is at best very unattractive computationally and would in most cases involve prohibitive memory storage requirements. An alternative and much easier strategy is to estimate the common correlation ρ by combining the individual correlation estimators $\hat{\rho}_g$. The fact that this estimation method is not fully efficient is not important when the number of genes is large. Let

$$\hat{\theta}_g = \frac{1}{2} \log \left\{ \frac{1 + (m-1)\hat{\rho}_g}{1 - \hat{\rho}_g} \right\}$$

where $\hat{\rho}_g$ is the REML estimator of ρ_g from the data for gene g . By analogy with the balanced case we can conclude that

$$E(\hat{\theta}_g) \approx \theta + b(d_g^B, d_g^W)$$

where d_g^B is the between array degrees of freedom and d_g^W is the within array degrees of freedom for gene g . A combined estimator of θ is

$$\bar{\theta} = \frac{1}{G} \sum_{g=1}^G \left\{ \hat{\theta}_g - b(d_g^B, d_g^W) \right\}$$

This estimator is consistent for θ as $n \rightarrow \infty$ regardless of the dependence structure between the genes and is consistent as $G \rightarrow \infty$ given weak assumptions on the dependence structure. The estimator of ρ is recovered by

$$\hat{\rho} = \frac{\exp(2\bar{\theta}) - 1}{\exp(2\bar{\theta}) + m - 1}.$$

Having estimated the common correlation, the regression coefficients $\hat{\boldsymbol{\beta}}_g$ can be estimated by weighted least squares of \mathbf{y}_g on X with weight matrix

$$W_g = \text{diag} \left(w_{gij}^{1/2} \right) \hat{R}^{-1} \text{diag} \left(w_{gij}^{1/2} \right)$$

where \hat{R} is equal to R_g with $\hat{\rho}$ substituted for ρ_g . The weighted least squares estimator is

$$\hat{\boldsymbol{\beta}}_g = (X^T W_g X)^{-1} X^T W_g \mathbf{y}_g$$

and the approximate REML estimator of σ_g^2 is the residual variance

$$s_g^2 = \frac{1}{nm - k} (\mathbf{y}_g - X\hat{\boldsymbol{\beta}}_g)^T W_g (\mathbf{y}_g - X\hat{\boldsymbol{\beta}}_g).$$

The test statistic for testing $H_0 : \alpha = 0$ is

$$t_g = \frac{\hat{\alpha}_g}{s_g u_g}$$

where

$$u_g^2 = \mathbf{c}^T (X^T W_g X)^{-1} \mathbf{c}$$

is the unscaled variance of $\hat{\alpha}$. The t -statistic is on $nm - k$ degrees of freedom. If $w_{gij} = 1$ then s_g^2 and t_g reduce to the same forms as in the balanced case in Section 3 apart from differences in the estimation of ρ , specifically the replacement of $\hat{\theta}$ with $\bar{\theta}$.

Note that the t -statistic t_g is not sensitive to small changes in the correlation correlation $\hat{\rho}$, since the estimated residual variance s_g^2 will tend to compensate. This reassures us that the common correlation model will not lead to misleading results if it fails to be exactly correct for some genes.

The efficiency of $\bar{\theta}$ relative to the REML estimator can be computed for the balanced case under the assumption of independence between genes. When $G = 1000$ and $n = m = 2$, the standard deviation is of $\bar{\theta}$ is 0.029 showing that its relative efficiency compared to the REML estimator is about 30%. This is the worst case; efficiency increases with the number of arrays. For example the efficiency is 70% if there are $n = 6$ arrays. For the purposes of the methodology of this paper, these are acceptable efficiencies.

6 Combining with Empirical Bayes Moderation

A number of authors have shown that one can improve on the use of t -statistics for assessing differential expression in microarray experiments by using appropriate shrinkage methods to moderate the genewise sample variances (Tusher et al., 2001; Baldi and Long, 2001; Efron et al., 2001; Lönnstedt and Speed, 2002; Broberg, 2003; Smyth, 2004). We show here that the empirical Bayes method of Smyth (2004) combines in a natural way with the methods of this paper.

In the separate correlation model of Section 2.1, an inverse Gamma prior would be applied to the between-array variances $\sigma_g^2\{1 + (m - 1)\rho_g\}$ yielding posterior variances

$$(\tilde{s}_g^B)^2 = \frac{(n - 1)(s_g^B)^2 + d_0(s_0^B)^2}{n - 1 + d_0}$$

where $(s_0^B)^2$ is the prior value and d_0 the prior degrees of freedom. Replacing the sample variance in (1) with the posterior variance produces the moderated t -statistic

$$\tilde{t}_g = \frac{\bar{y}_g}{\tilde{s}_g^B / \sqrt{nm}}$$

which follows a t -distribution on $n - 1 + d_0$ degrees of freedom if $\mu_g = 0$ (Smyth, 2004). In the common correlation model of Section 2.2 an inverse Gamma prior would be applied to σ_g^2 yielding posterior variances

$$\tilde{s}_g^2 = \frac{(nm - 1)s_g^2 + d_0s_0^2}{nm - 1 + d_0}.$$

Replacing the sample variance in (4) with the posterior variance produces the moderated t -statistic

$$\tilde{t}_g = \frac{\bar{y}_g}{\tilde{s}_g[\{1 + (m - 1)\hat{\rho}\}/(nm)]^{1/2}}$$

which is t -distributed on $nm - 1 + d_0$ degrees of freedom if $\mu_g = 0$. The same technique could be applied to the individual and common correlation models of Sections 3 and 4. In Section 4 an inverse Gamma prior for σ_g^2 would lead to posterior variances

$$\tilde{s}_g^2 = \frac{(nm - k)s_g^2 + d_0s_0^2}{nm - k + d_0}$$

and to moderated t -statistics

$$\tilde{t}_g = \frac{\hat{\alpha}_g}{\tilde{s}_g u_g}$$

on $nm - k + d_0$ degrees of freedom. The use of empirical Bayes results in effect in a further d_0 degrees of freedom for the estimation of the genewise sample variances, where d_0 is

estimated from the data. The common correlation methodology proposed in this paper and the use of empirical Bayes to smooth the variances are complementary techniques in the sense that using both techniques together results in the greatest possible increase in the effective degrees of freedom for estimating the variances.

Note that empirical Bayes smoothing could in principle be applied to the correlations as well as the variances. In fact, smoothing the between and within variances $(s_g^B)^2$ and $(s_g^W)^2$ independently leads immediately to smoothed correlation estimators from

$$\tilde{\theta}_g = \log(\tilde{s}_g^B / \tilde{s}_g^W).$$

This however turns out to be equivalent to averaging the log-ratios over replicate spots and then applying smoothing to the variances, i.e., empirical Bayes smoothing of the correlations does not add more information over that of smoothing the variances alone. So it appears that to get an extra benefit it is necessary to smooth the correlations to a *greater* degree than the variances, e.g., by setting them equal as in this paper.

7 Results with Spike-in Data

The methodology is demonstrated on a set of 26 microarrays for which the differential expression status of a set of control spots is known. Figure 1 shows boxplots of t -statistics for the scorecard series of control spots. There are twelve t -statistics in each box. The grey filled boxplots, on the left of each pair of boxplots, show statistics computed using common correlations while the white boxplots on the right of each pair show statistics computed by averaging the duplicate spots.

The t -statistics produced by averaging the duplicate spots are on fewer degrees of freedom than those produced by the common correlation method, meaning that they are not directly comparable on the basis of magnitudes alone. One way to compare the t -statistics would be to compute p -values. The vertical axis in the plot actually shows z -score equivalents of the t -statistics, i.e., the standard normal deviate which has the same p -value as has the t -statistic. The z -scores puts t -statistics with different degrees of freedom on the same scale. Comparing z -scores is equivalent to comparing p -values but the z -scores are better suited to graphical presentation.

An ideal test statistic will show z -score values which are randomly distributed about zero with as little variability as possible for the calibration spots and z -scores as far from zero as possible for the ratio controls. The better the separation between the calibration values and the ratio values, the better the performance of the statistic. The plot shows that the t -statistics computed assuming common correlations give much larger absolute z -scores for the differentially expressed

genes while maintaining a similar null distribution for the non-differentially expressed spots. This shows that the common correlation t -statistics have greater power for detecting differential expression while producing no more false positives on average.

The right panel of Figure 1 shows the results with empirical Bayes smoothing of the sample variations while the left panel shows results with ordinary t -statistics. The relatively large number of arrays here means that the sample variances are fairly reliable so that use of empirical Bayes changes the picture only slightly.

The minimum number of arrays for which t -statistics can be computed is two, this being the minimum number to return a degree of freedom for error when the duplicate spot values are averaged. In order to examine this extreme situation, we separated the 26 arrays into 13 pairs of arrays and computed t -statistics for each pair. Figure 2 shows the results. Each boxplot in Figure 2 represents 156 values, i.e., twelve values for each of the thirteen pairs. The fact that t -statistics are computed from only two arrays instead of all 26 means that they are less able to distinguish which spots are differentially expressed, however the t -statistics using common correlations do markedly better. As before, the common correlation t -statistics have greater power to detect differential expression while having a similar null distribution (Figure 2, left panel). The relative gain of the common correlation method compared with averaging the duplicate spots is perhaps even greater here with $n = 2$ than with the larger sample size.

With only two arrays for each t -statistic, the sample variances are rather unreliably estimated. When the duplicate spots are averaged, the sample variances have in fact only one degree of freedom. In this situation, empirical Bayes smoothing can be expected to make a large impact on the reliability of the statistics. The right panel of Figure 2 shows the same results as the left but with empirical Bayes smoothing of the variances. The empirical Bayes method greatly improves the performance of both statistics, with and without common correlations, and the separation of calibration and ratio values is improved relative to the left panel. The comparison between common and individual correlations is no longer clear cut because the z -scores for ratio controls without common correlations are so variable, sometimes larger and sometimes smaller than the statistics with common correlations. The important observation here is that the common correlation (white) boxes are noticeably more compact than the grey boxes for all intensities of calibration spots, i.e., the rate of false discoveries is reduced. Furthermore, assuming common correlations also gives larger median z -scores for all types of ratio controls except for 10DL, meaning that the common correlation method gives greater power in most cases as well as better control of type I errors.

8 Discussion

This paper shows that setting the between-replicate correlation constant across genes is a useful strategy. Results using spike-in probes show that the statistics assuming common correlations give clearly improved assessment of differential expression. Any bias which is introduced by assuming correlations to be equal seems to be more than offset by an increase in the precision with which the genewise variances are estimated. When the number of arrays is small, the spike-in results were further by empirical Bayes smoothing of the sample variances as in Smyth (2004).

The authors have applied the methodology to a variety of microarray experiments with arrays printed in several different laboratories with several different clone libraries. Our experience has been that correlations between side-by-side duplicates are estimated typically in the range 0.7–0.9, suggesting that side-by-side duplicates share about half of their variability as measured by squared correlation. Correlations between replicates in top and bottom halves of array are typically estimated in the range 0.5–0.6, suggesting that duplicates at the maximum distance apart still share about a quarter of their variability. These observations are consistent with the idea that spots which are further apart should be less highly correlated.

In most experiments the genewise correlation estimates $\hat{\rho}_g$ are found to be too variable across genes to be compatible with a common true correlation and the theoretical scaled $F_{n-k,n}$ sampling variability for $(1 + \hat{\rho}_g)/(1 - \hat{\rho}_g)$ (data not shown). In other words the assumption of constant correlation across genes does not appear to be strictly tenable. On the other hand, the between and within sample variances s_g^B and s_g^W have been found to be positively associated, meaning that the correlation estimates $\hat{\rho}_g$ are less variable, relative to the theoretical F -distribution, than are the sample variances s_g^2 relative to their theoretical chisquare distribution. So the assumption of constant correlation appears to be valid in practice in the relative sense that the correlations are more nearly constant than are the variances themselves.

The effectiveness of the common correlation model seems to be due to three main characteristics. Firstly, the estimated common correlation is very stable being a consensus estimator derived from a large number of genes. This stability results in a favorable variance-bias trade-off, especially for small data sets. Secondly, the correlation is a nuisance parameter rather than a quantity of primary interest. It has been noted that the genewise t -statistics are not sensitive to small changes in the correlation estimate, so it is not necessary to track small differences in the genewise correlations provided that the common correlation estimate is broadly correct. Thirdly, the common correlation model causes genes with poor quality data to be down-weighted. Good quality data seems to give rise to consistently high correlations be-

tween replicate spots. Even for arrays with a lot of poor quality spots, the common correlation is generally large and positive. Those genes which do give rise to low or even negative correlations seem to do so most often because of poor quality data, for example artifacts on the surface of the arrays which affect only one of a set of replicate spots. Holding the correlation fixed forces the estimated residual variance for these genes to be relatively large to reflect the disagreement between the replicate spots. This means that statistical significance for these genes is downweighted, a phenomenon which seems conservative and desirable. Allowing each gene to estimate its own correlation would cause disagreements between replicate spots to be disregarded.

The formal calculations in this paper have assumed normality of the expression log-ratios as well as independence and constant variances across arrays for each gene. There are several reasons to expect the methodology to remain useful even for data which deviates from normality. Firstly, apart from the bias correction $b(f_1, f_2)$ which is relatively small in magnitude, the estimators derived here remain consistent given only the first and second moments of the response distribution. Secondly, the estimation procedures can be modified to make them more resistant to outliers. A simple method which has proved effective is to estimate θ not from $\bar{\theta}$ in Section 5 but from a robust mean of the $\hat{\theta}_g - b(d_g^B, d_g^W)$. This has the effect of ignoring a small proportion of aberrant correlation estimates. The default estimator used in the limma software package is the trimmed mean removing 15% of the values from each tail.

Thirdly, and perhaps most importantly, the most basic purpose of differential expression analyses for microarray data is to rank the genes in terms of evidence for differential expression (Smyth et al., 2003). An effective ranking, which reliably ranks the truly differentially expressed genes near the top, is even more important than the ability to decide which genes are significantly differentially expressed. It is more important than that the p -values for different genes are correctly ordered than that the p -values have the correct uniform null distribution. On this measure, the common correlation method has clear advantages over alternatives even when the underlying model is not exactly correct. It is more effective than averaging the replicate spots because it takes into account deviations between replicates when estimating the precision of the data for each gene. Compared with rank-based or permutation tests, the parametric method described here has the advantage of greater resolution, i.e., lack of granularity in the estimators and p -values, allowing genes to be more finely graduated for small or moderate sized data sets.

For the reasons explained above, the application of the methods described here is not limited to high quality data sets for which normality might be reasonable nor to very large data sets for which rigorous checking of the distributional assumptions might be feasible. In fact the benefits,

relative to alternative methods such as averaging the replicate spots or simply ranking genes on fold change, may be most pronounced in experiments with very few arrays or with poor quality data. This expectation is born out by the spike-in experimental data with $n = 2$.

As with any statistical modelling technique, it is assumed that appropriate quality assessment has been done of the data before application of the method proposed here. It has been described in Section 5 how the method is capable of incorporating spot and array quality weights which might arise from such quality assessment.

The method used in this paper differs from previous work on empirical Bayes or shrinkage estimators in that a suitably chosen parameter is simply set equal across genes. The idea works here because the correlation parameter is of secondary interest from an inferential point of view and because it is relatively stable across genes. The technique is applicable to other situations involving mixed model analyses of microarray data such as those with technical as well as biological replication or the separate channel analyses described by Jin et al. (2001) and Wolfinger et al. (2001). These situations have within-block or within-spot correlations for which consensus estimators might be used across genes.

The methods described in this paper, are implemented in the software package limma for the R computing environment (Smyth et al., 2004). Limma is part of the Bioconductor project at <http://www.bioconductor.org> (Gentleman et al., 2004).

Acknowledgements

The authors thank Lisa Martin, Melanie O’Keefe and Cathy Jensen of the Australian Genome Research Facility for printing the microarrays used in the study described in the paper. Thanks are due to Terry Speed and Matthew Ritchie for valuable discussions. This research was supported by NHMRC Grants 257501 and 257529.

References

- Andrews, J., G. G. Bouffard, C. Cheadle, J. Lu, K. G. Becker, and B. Oliver (2000). Gene discovery using computational and microarray analysis of transcription in the drosophila melanogaster testis. *Genome Res* 10(12), 2030–43.
- Baggerly, K. A., K. R. Coombes, K. R. Hess, D. N. Stivers, L. V. Abruzzo, and W. Zhang (2001). Identifying differentially expressed genes in cDNA microarray experiments. *J Comput Biol* 8(6), 639–59.
- Baldi, P. and A. D. Long (2001). A Bayesian framework for the analysis of microarray expression data: regularized

- t*-test and statistical inferences of gene changes. *Bioinformatics* 17, 509–519.
- Beissbarth, T., K. Fellenberg, B. Brors, R. Arribas-Prat, J. Boer, N. C. Hauser, M. Scheideler, J. D. Hoheisel, G. Schutz, A. Poustka, and M. Vingron (2000). Processing and quality control of DNA array hybridization data. *Bioinformatics* 16(11), 1014–22.
- Berwanger, B., O. Hartmann, E. Bergmann, S. Bernard, D. Nielsen, M. Krause, A. Kartal, D. Flynn, R. Wiedemeyer, M. Schwab, H. Schafer, H. Christiansen, and M. Eilers (2002). Loss of a FYN-regulated differentiation and growth arrest pathway in advanced stage neuroblastoma. *Cancer Cell* 2(5), 377–86.
- Boer, J. M., W. K. Huber, H. Sültmann, F. Wilmer, A. von Heydebreck, S. Haas, B. Korn, B. Gunawan, A. Vente, L. Fuzesi, M. Vingron, and A. Poustka (2001). Identification and classification of differentially expressed genes in renal cell carcinoma by expression profiling on a global human 31,500-element cDNA array. *Genome Res* 11(11), 1861–70.
- Broberg, P. (2003). Statistical methods for ranking differentially expressed genes. *Genome Biology* 4, R41.
- Buckley, M. J. (2000). *Spot User’s Guide*. Sydney, Australia: CSIRO Mathematical and Information Sciences. Software manual available from <http://www.cmis.csiro.au/iap/Spot/spotmanual.htm>.
- Causton, H. C., J. Quackenbush, and A. Brazma (2003). *Microarray gene expression data analysis : a beginner’s guide*. Malden, MA: Blackwell Pub.
- Efron, B. and R. Tibshirani (2002). Empirical Bayes methods and false discovery rates for microarrays. *Genet Epidemiol* 23(1), 70–86.
- Efron, B., R. Tibshirani, J. D. Storey, and V. Tusher (2001). Empirical Bayes analysis of a microarray experiment. *J. Amer. Statist. Assoc.* 96(456), 1151–1160.
- Fan, J., P. Tam, G. V. Woude, and Y. Ren (2004). Normalization and analysis of cDNA microarrays using within-array replications applied to neuroblastoma cell response to a cytokine. *Proc Natl Acad Sci U S A* 101(5), 1135–40.
- Firestein, G. S. and D. S. Pisetsky (2002). DNA microarrays: boundless technology or bound by technology? guidelines for studies using microarray technology. *Arthritis Rheum* 46(4), 859–61.
- Gentleman, R. C., V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. Yang, and J. Zhang (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology* 5(10), R80.
- Hoffmann, K. F., D. A. Johnston, and D. W. Dunne (2002). Identification of schistosoma mansoni gender-associated gene transcripts by cDNA microarray profiling. *Genome Biol* 3(8), Research0041.
- Jenssen, T. K., M. Langaas, W. P. Kuo, B. Smith-Sorensen, O. Myklebost, and E. Hovig (2002). Analysis of repeatability in spotted cDNA microarrays. *Nucleic Acids Res* 30(14), 3235–44.
- Jin, W., R. M. Riley, R. D. Wolfinger, K. P. White, G. Passador-Gurgel, and G. Gibson (2001). The contributions of sex, genotype and age to transcriptional variance in drosophila melanogaster. *Nat Genet* 29(4), 389–95.
- Kaynak, B., A. von Heydebreck, S. Mebus, D. Seelow, S. Hennig, J. Vogel, H. P. Sperling, R. Pregla, V. Alexi-Meskishvili, R. Hetzer, P. E. Lange, M. Vingron, H. Lehrach, and S. Sperling (2003). Genome-wide array analysis of normal and malformed human hearts. *Circulation* 107(19), 2467–74.
- Kendzioriski, C. M., M. A. Newton, H. Lan, and M. N. Gould (2003). On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles. *Stat Med* 22(24), 3899–914.
- König, R., D. Baldessari, N. Pollet, C. Niehrs, and R. Eils (2004). Reliability of gene expression ratios for cDNA microarrays in multiconditional experiments with a reference design. *Nucleic Acids Res* 32(3), e29.
- Lönnstedt, I. and T. Speed (2002). Replicated microarray data. *Statist. Sinica* 12(1), 31–46.
- Lyne, R., G. Burns, J. Mata, C. J. Penkett, G. Rustici, D. Chen, C. Langford, D. Vetrie, and J. Bahler (2003). Whole-genome microarrays of fission yeast: characteristics, accuracy, reproducibility, and processing of array data. *BMC Genomics* 4(1), 27.
- Newton, M. A., C. M. Kendzioriski, C. S. Richmond, F. R. Blattner, and K. W. Tsui (2001). On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *J Comput Biol* 8(1), 37–52.
- Newton, M. A., A. Noueir, D. Sarkar, and P. Ahlquist (2004). Detecting differential gene expression with a semiparametric hierarchical mixture method. *Biostatistics* 5(2), 155–176.

- Nguyen, D. V., A. B. Arpat, N. Wang, and R. J. Carroll (2002). DNA microarray experiments: biological and technological aspects. *Biometrics* 58(4), 701–717.
- Parmigiani, G., E. S. Garrett, R. A. Irizarry, and S. L. Zeger (Eds.) (2003). *The analysis of gene expression data*. Statistics for Biology and Health. New York: Springer-Verlag.
- Pinheiro, J. C. and D. M. Bates (2000). *Mixed-effects models in S and S-PLUS*. Statistics and computing. New York: Springer.
- Searle, S. R., G. Casella, and C. E. McCulloch (1992). *Variance components*. Wiley series in probability and mathematical statistics. Applied probability and statistics. New York: Wiley.
- 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), Article 1.
- Smyth, G. K. and T. Speed (2003). Normalization of cDNA microarray data. *Methods* 31(4), 265–73.
- Smyth, G. K., N. Thorne, and J. Wettenhall (2004). *Limma: Linear Models for Microarray, User's Guide*. Software manual available from <http://bioinf.wehi.edu.au/limma>.
- Smyth, G. K., Y. H. Yang, and T. Speed (2003). Statistical issues in cDNA microarray data analysis. *Methods Mol Biol* 224, 111–36.
- Speed, T. P. (Ed.) (2003). *Statistical analysis of gene expression microarray data*. Interdisciplinary statistics. Boca Raton, FL: Chapman & Hall/CRC.
- Tilstone, C. (2003). DNA microarrays: vital statistics. *Nature* 424(6949), 610–2.
- Tran, P. H., D. A. Peiffer, Y. Shin, L. M. Meek, J. P. Brody, and K. W. Cho (2002). Microarray optimizations: increasing spot accuracy and automated identification of true microarray signals. *Nucleic Acids Res* 30(12), e54.
- Tseng, G. C., M. K. Oh, L. Rohlin, J. C. Liao, and W. H. Wong (2001). Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Res* 29(12), 2549–57.
- Tusher, V. G., R. Tibshirani, and G. Chu (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A* 98(9), 5116–21.
- Wang, X., S. Ghosh, and S. W. Guo (2001). Quantitative quality control in microarray image processing and data acquisition. *Nucleic Acids Res* 29(15), E75–5.
- Wolfinger, R. D., G. Gibson, E. D. Wolfinger, L. Bennett, H. Hamadeh, P. Bushel, C. Afshari, and R. S. Paules (2001). Assessing gene significance from cDNA microarray expression data via mixed models. *J Comput Biol* 8(6), 625–37.
- Yang, I. V., E. Chen, J. P. Haseleman, W. Liang, B. C. Frank, S. Wang, V. Sharov, A. I. Saeed, J. White, J. Li, N. H. Lee, T. J. Yeatman, and J. Quackenbush (2002). Within the fold: assessing differential expression measures and reproducibility in microarray assays. *Genome Biol* 3(11), research0062.
- Yang, Y. H., S. Dudoit, P. Luu, and T. P. Speed (2001). Normalization for cDNA microarray data. In M. L. Bittner, Y. Chen, A. N. Dorsel, and E. R. Dougherty (Eds.), *Microarrays: Optical Technologies and Informatics*, Volume 4266 of *Proceedings of SPIE*, pp. 141–152.
- Yang, Y. H. and T. P. Speed (2003). Design and analysis of comparative microarray experiments. In T. P. Speed (Ed.), *Statistical Analysis of Gene Expression Microarray Data*, pp. 35–91. Chapman & Hall/CRC Press.

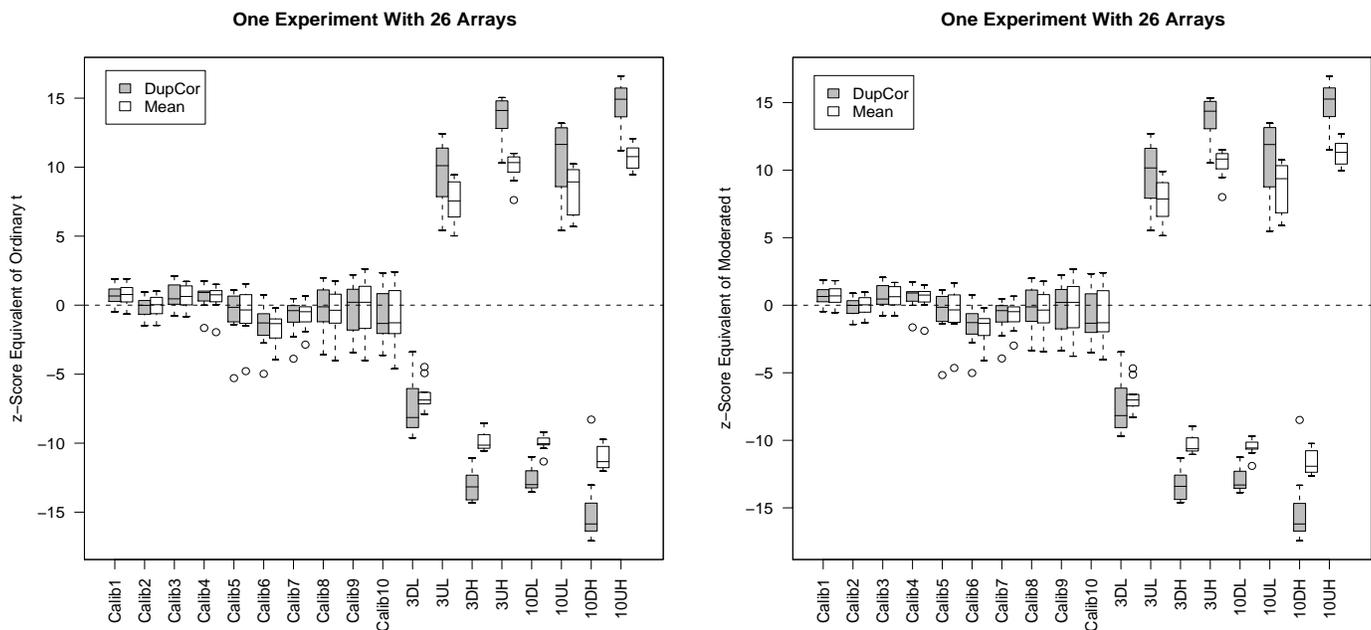


Figure 1: Boxplots of Z -score equivalents of ordinary t -statistics (on the left) and of moderated t -statistics (on the right) for different types of spike-in spot-pairs. The grey filled boxes are for statistics based on estimated between-replicate correlations while the unfilled (white) boxes are for statistics based on averaging the replicate observations. Statistics are calculated from the whole series of 26 arrays. Control spots labeled Calib1–10 are non-differentially expressed calibration spots at increasing dilutions and therefore decreasing intensities. Control spots labeled 3DL and 3UL are ratio controls designed to be 3-fold down-regulated and 3-fold up-regulated respectively. Control spots labeled 3DH and 3UH are similar but at high rather than low intensity. Control spots labeled 10DL, 10UL, 10DH and 10UH are similar but are 10-fold up or down-regulated.

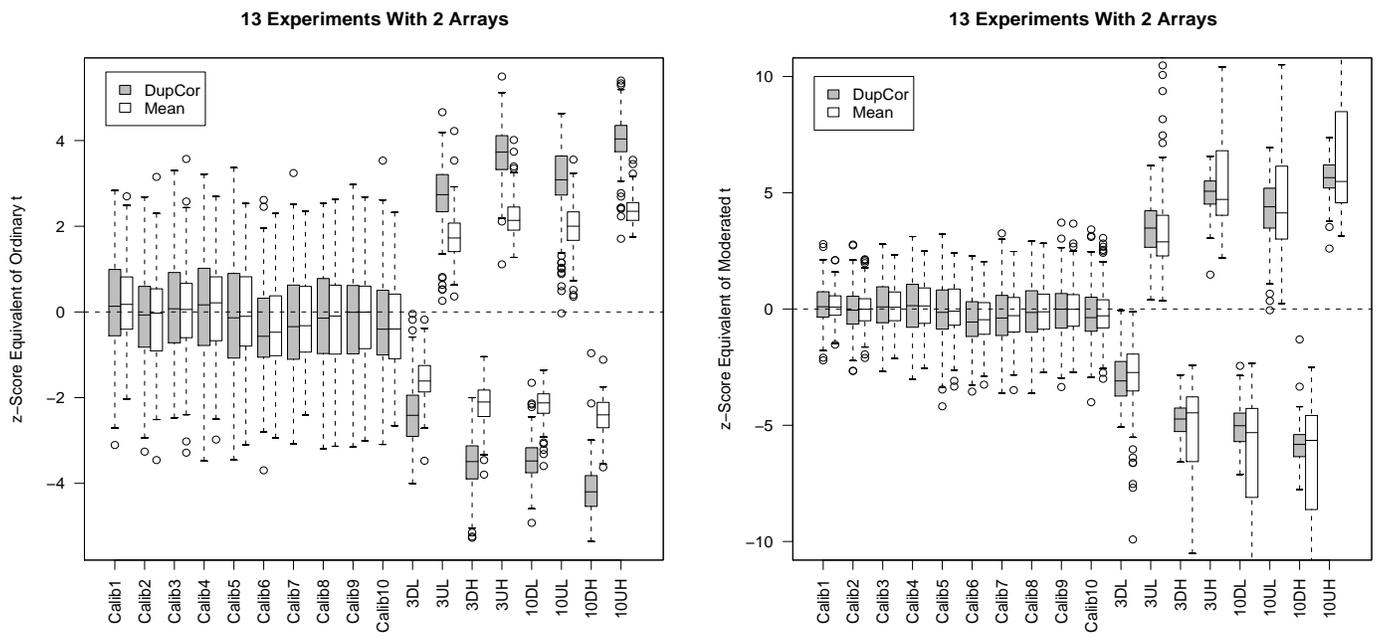


Figure 2: Boxplots of Z -score equivalents of ordinary t -statistics (on the left) and of moderated t -statistics (on the right) for different types of spike-in spot-pairs. The grey filled boxes are for statistics based on estimated between-replicate correlations while the unfilled (white) boxes are for statistics based on averaging the replicate observations. Statistics are calculated from two arrays. The boxes include statistics from 13 such sets of two arrays.