BIOINFORMATICS ORIGINAL PAPER Vol. 22 no. 7 2006, pages 802–807
doi:10.1093/bioinformatics/btk047

Gene expression

Combining signals from spotted cDNA microarrays obtained at different scanning intensities

H. P. Piepho[1,*], B. Keller[1], N. Hoecker[2] and F. Hochholdinger[2]

[1]Bioinformatics Unit, Institute for Crop Production and Grassland Research, University of Hohenheim, Fruwirthstrasse 23, 70599 Stuttgart, Germany and [2]Center for Plant Molecular Biology, Department of General Genetics, University of Tuebingen, Auf der Morgenstelle 28, 72076 Tuebingen, Germany

Received on December 9, 2005; revised on December 22, 2005; accepted on January 6, 2006
Advance Access publication January 17, 2006
Associate Editor: Steen Knudsen

ABSTRACT

Motivation: The analysis of spotted cDNA microarrays involves scanning of color signals from fluorescent dyes. A common problem is that a given scanning intensity is not usually optimal for all spotted cDNAs. Specifically, some spots may be at the saturation limit, resulting in poor separation of signals from different tissues or conditions. The problem may be addressed by multiple scans with varying scanning intensities. Multiple scanning intensities raise the question of how to combine different signals from the same spot, particularly when measurement error is not negligible.

Results: This paper suggests a non-linear latent regression model for this purpose. It corrects for biases caused by the saturation limit and efficiently combines data from multiple scans. Combining multiple scans also allows reduction of technical error particularly for cDNA spots with low signal. The procedure is exemplified using cDNA expression data from maize.

Availability: All methods were implemented using standard procedures available in the SAS/STAT module of the SAS System. Programming statements are available from the first author upon request.

Contact: piepho@uni-hohenheim.de

Supplementary information: The supplementary data are available at *Bioinformatics* online.

INTRODUCTION

When scanning fluorescent spot signals (Cy3/Cy5) from cDNA microarrays, the choice of an optimal scanning intensity is critical. A particular choice of scanning intensity may be optimal or near-optimal only for a subset of the spotted cDNAs, while other cDNAs are at the detection limit or saturation limit (Lyng *et al*., 2004). For this reason, it has been suggested to scan the same microarray repeatedly using different scanning intensities. If there are G genes and S scanning intensities, the procedure results in a $G \times S$ array of signals. One possible route for analysis is to select an optimal scanning intensity for each cDNA spot and then base the analysis for a spot only on the signal detected at this one intensity. This approach has several drawbacks. Choice of the optimal scanning intensity is difficult, and some kind of automated analysis is needed. Also, looking at several scanning intensities simultaneously exacerbates the multplicity problem (Korn *et al*., 2004). Moreover, using only data from one intensity is a waste of information, particularly when error variance is large. A more promising approach is to integrate the results of multiple scans into a single expression value per gene and dye (Romualdi *et al*., 2003; Garcia de la Nava *et al*., 2004). An obvious procedure is to compute, for each cDNA spot, some kind of average across scanning intensities. The simple average is not necessarily the best method for doing this, since signals may be more informative for some scanning intensities than for others. Specifically, for spots with signals near the saturation threshold, the information content is relatively low, while signals with intermediate intensity are more informative.

The central idea of this paper is to devise an adjustment for observations near the saturation limit based on a statistical model. The main assumption is that the observed signal at a given scanning intensity and the amount of expression product are connected by some non-linear relationship. The amount of the gene product enters the model as a latent variable, thus giving rise to a non-linear latent regression model. The approach is exemplified using cDNA gene expression data from maize.

MATERIALS AND METHODS

The data

The data are from a large project aimed at unraveling the molecular causes of heterosis in maize. Details of the experimental design are described in Keller *et al*. (2005). Kernels of the maize inbred lines UH002, UH005, UH250, UH301, and their eight reciprocal hybrids were germinated on filter paper (20×70 cm Grade 603 N, Sartorius, Göttingen, Germany) in a phytochamber at 26°C, with a 16 h light, 8 h dark cycle and 60% humidity. RNA was extracted from primary roots 3.5 days after germination using Trizol reagent (Invitrogen, Karlsruhe, Germany) and the Oligotex mRNA Kit (Qiagen, Hilden, Germany), according to the manufacturer's instructions. mRNA concentrations were quantified with RiboGreen (Molecular Probes, Leiden, Netherlands) using a fluorescence spectrophotometer F-2000 (Hitachi, Japan). mRNA was transcribed into cDNA using random hexamer primer and reverse transcriptase 'SuperscriptII' (both Invitrogen). cDNA was indirectly labelled with Cy3 or Cy5 fluorescent dye (Amersham Biosciences, Little Chalfont, UK) and hybridized with a 12.160 element microarray chip (Generation II VersionB) generated at the Iowa State Microarray Facility (Ames, IA; http://www.plantgenomics.iastate.edu/maizechip/), as described in Hedge *et al*. (2000). We analyzed a total of 49 arrays. The four parents were replicated more often than the hybrids because this optimized accuracy for estimating heterosis (Keller *et al*., 2005).

*To whom correspondence should be addressed.

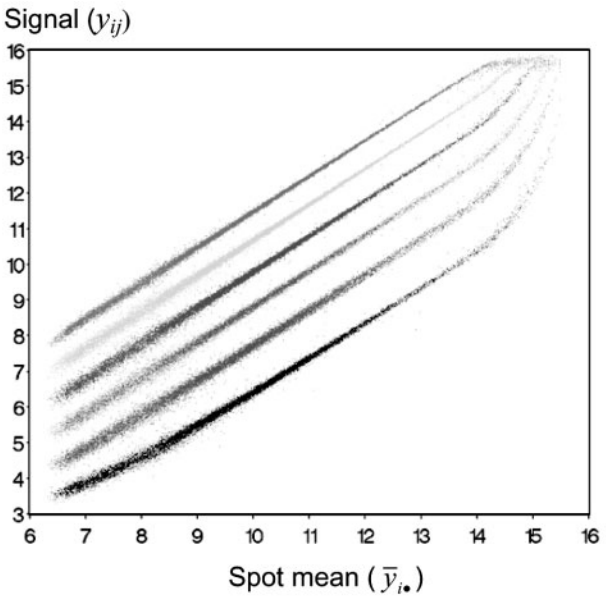

Fig. 1. Plot of signals at six different scanning intensities (y_{ij}) versus spot mean ($\bar{y}_{i\bullet}$) for 11 664 cDNA spots on chip 288. Genotype: UH005x301. Intensities: 20, 30, 40, 50, 60, 70. Channel: Cy5.

Scanning of the cDNA arrays was performed with a GMS 418 Array Scanner (Genetic MicroSystems, Woburn, MA). Each scan was obtained with a laser power of 90%. A series of six scans, in ascending order of PMT gain, was performed at 10 μm resolution. Initial settings for Cy3 and Cy5 were ~10–20% and were increased by 5 or 10% increments. Fluorescent signal intensities were determined and quantified using ImaGene 5.6 (Biodiscovery, Inc., Mariana Del Rey, CA). In total 11 446 spots of this maize microarray chip were subjected to data analysis.

Scanning intensities were $\log_2$-transformed. For spots with signals at all six scanning intensities, we computed the mean signal on the log-scale and then plotted each log-signal versus its corresponding spot mean (Fig. 1). The plot suggests that there is a marked saturation level between 15 and 16 units, which is reached for a number of spots at the highest scanning intensities. Also, in the region of linear response, the variance tends to increase towards lower signal intensities. The relationship between any two intensities is linear except for the area around the detection limit. Below the detection limit, the scatter of spots runs parallel for all intensities, and the slope of each scatter is very close to unity.

The model

The strong linearity below the saturation limit suggests that up to a region close to the saturation limit it is reasonable to assume a linear relationship between expected signal (μ) and a latent variable (η) related to the amount of gene expression product and the scanning intensity. The parallelism of scatters for the six intensities in Figure 1 further suggests that an additive model of the following form may be used:

$$\eta_{ij} = \alpha_j + g_i, \tag{1}$$

where η_{ij} is the latent value for the i-th spot at the j-th scanning intensity, α_j is the main effect of the j-th intensity ($j = 1, \ldots, J = 6$), subject to the identifiability constraint $\alpha_6 = 0$, and g_i is the main effect of the i-th spot ($i = 1, \ldots, I = 11\,446$). The rationale for this model choice is that a plot of η_{ij} versus the spot mean $(\eta_{i1} + \eta_{i2} + \eta_{i3} + \eta_{i4} + \eta_{i5} + \eta_{i6})/6$ will be linear with slope equal to unity, which corresponds to the pattern in Figure 1.

It appears to be difficult to include the deviant behavior near the saturation point in a single functional model, so one may resort to a segmented regression model (Seber and Wild, 1989; Schabenberger and Pierce, 2002). The simplest such model, also known as the 'broken stick' model, has two linear regression lines with different slopes connected at the change point. This has the undesirable property of an abrupt change in slope at the change point. A model with a smooth non-linear transition between linear part and an asymptotic plateau may be obtained by fitting an exponential function near the saturation limit. The model can be written as

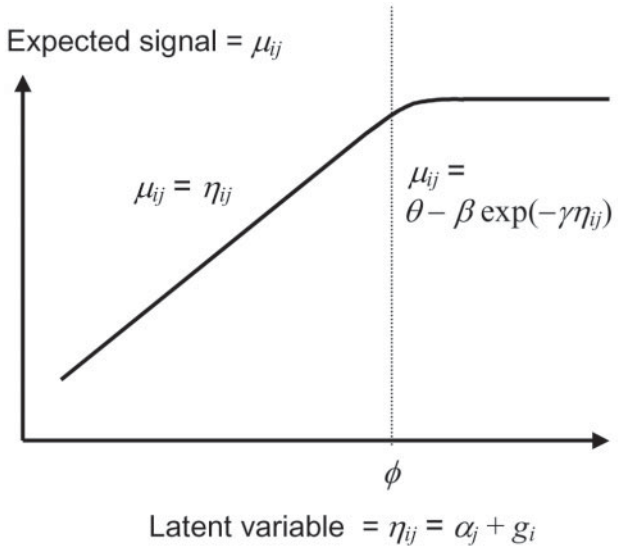

Fig. 2. Sketch of segmented model in Equation (2).

$$\mu_{ij} = \eta_{ij} \quad \text{when } \eta_{ij} < \phi \text{ and} \tag{2a}$$

$$\mu_{ij} = \theta - \beta \exp(-\gamma\eta_{ij}) \quad \text{when } \eta_{ij} \geq \phi, \tag{2b}$$

where μ_{ij} is the expected signal for the i-th spot at the j-th scanning intensity, θ is the saturation limit, ϕ is the break point, and β and γ are regression parameters. The model is depicted in Figure 2. The observed signal is modelled as

$$y_{ij} = \mu_{ij} + e_{ij}, \tag{3}$$

where e_{ij} is a random measurement error term distributed with zero mean.

We require a smooth transition at the break point $\eta_{ij} = \phi$ in the sense that function value and first derivative of the connected segments must coincide (Seber and Wild, 1989). This requirement gives rise to two constraints on the parameters. Specifically, in order to have a smooth transition in the first derivative, we require

$$\beta\gamma \exp(-\gamma\phi) = 1 \Leftrightarrow \beta = [\gamma \exp(-\gamma\phi)]^{-1}. \tag{4}$$

In order to have no jumps at the change point, we further require

$$\phi = \theta - \beta \exp(-\gamma\phi) \Leftrightarrow \phi = \theta - [\gamma \exp(-\gamma\phi)]^{-1} \exp(-\gamma\phi) = \theta - \gamma^{-1}. \tag{5}$$

Thus, apart from intensity effects α_j and spot effects g_i, only the effects θ and γ need to be estimated.

ALGORITHM

One might consider fitting model (2) directly in a single step by non-linear regression. This would be computationally very expensive, however, because the dimension of the problem nearly equals the number of spots (~12 000 in the case of the maize cDNA microarrays used in this paper). The space required to hold and

manipulate the design matrix and the data usually exceeds the memory available on standard PCs. Even if memory permits fitting of the model by standard non-linear model procedures, computing time will typically be unacceptably long for routine use. We therefore sought a simple alternative.

The main idea is to fit the model by alternating least squares (Kroonenberg and De Leeuw, 1980), keeping one set of parameters fixed, while estimating the other, and vice versa. The first set of parameters, denoted as intensity set, consists of the intensity effects α_j and the two non-linear parameters θ and γ, while the other set, denoted as spot set, contains the spot effects g_i. When the spot set is kept fixed, optimization is with respect to only $J + 1$ parameters, which is computationally inexpensive. When the intensity set is kept fixed, there are I parameters to be estimated. The key idea is that each of these parameters can be fitted individually, because conditionally on the intensity set, all information on g_i is confined to the data of the i-th spot. The algorithm suggested in this paper alternates between the two sets, fitting parameters of the one set, while keeping parameters in the other set fixed. Each of the alternating optimization steps is itself trivial computationally. The total computational cost is governed by the total number of alternating steps. Our experience is that several dozens of iterations are needed until convergence.

The signal versus spot mean plot (Fig. 1) shows some heterogeneity of variance, the variance increasing with decreasing expected signal. Inspection of Figure 1 reveals that heterogeneity is mainly a problem for spots with weak mean signal. One can model the variance as a function of the spot effect and re-estimate all parameters by weighted least squares (WLS), possibly iterating between estimation of the expected signal and the variance (Carroll and Ruppert, 1988). We suggest to use only one or two iterations, following the suggestion by Carroll and Ruppert (1988). The scatterplot of variance (log-scale) versus spot effect may be somewhat irregular (Fig. 3), so we use locally weighted smoothing splines (LOESS; Cleveland, 1979) for fitting the variance function.

A further aspect of multiple scan data is that occasionally a signal may be missing. Thus, any proposed method should be able to handle unbalanced data in an efficient way. Standard procedures for linear and non-linear regression (Searle, 1987; Seber and Wild, 1989) are well suited for this task. We initialize spot and intensity effects based on a fit of the additive model in Equation (1). To speed up computing time, intensity effects are estimated by absorbing spot effects into the least squares equations (Engel, 1990; Harville, 1997; SAS Institute, 1999; see Appendix A in the online Supplementary Data). The parameter γ, which appears in the exponent of the non-linear response function, is the most sensitive parameter in iterations. Therefore, we determine convergence based on the relative change in this parameter.

The following algorithm is proposed to fit model (2):

(1) Initialize intensity and spot effects by fitting an additive two-way model [Equation (1)] to the data. When fitting the model, absorb spot effects to save computing time. Initialize θ by $\max(y_{ij})$. For γ, try several starting values between 0 and 5. Initialize weights $w_i = 1$. Set $Z = 0$. Set $Z_{\max}$ equal to the desired number of WLS steps.

(2) Keep the spot set (g_i) fixed. Estimate the intensity set (θ, γ, α_j) by weighted non-linear least squares using weights w_i. Use the whole data set in this optimization.

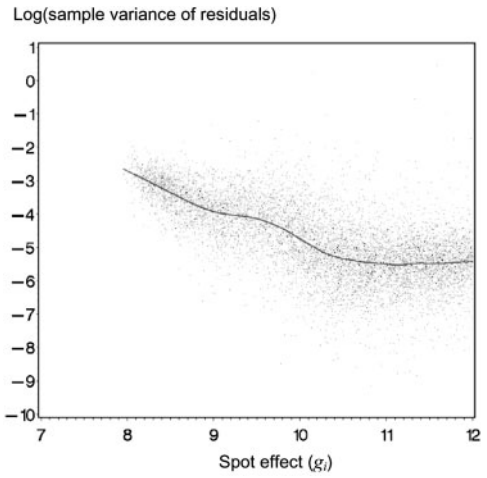

Fig. 3. Plot of logged sample variance of residuals from least squares fit of model (2) versus spot effect estimate with LOESS fit for chip 288. Genotype: UH005x301. Intensities: 20, 30, 40, 50, 60, 70. Channel: Cy5.

(3) Keep the intensity set (θ, γ, α_j) fixed and estimate the spot set (g_i) by nonlinear least squares. Perform the optimization separately for each spot.

(4) If the relative change in the estimate of γ is smaller than some small value ϵ, proceed to step (5). Otherwise return to step (2).

(5) If $Z = Z_{\max}$, terminate the algorithm. Otherwise, do the following. Based on last fit, compute least squares residuals. From the residuals, compute the sample variance per spot. Fit a LOESS curve to predict log-variance from spot effects. Based on the predicted variance, compute weights w_i per spot as the inverse of the variance. Set $Z := Z + 1$ and go back to step (2).

In order to evaluate the algorithm, we compared our method with a standard analysis based on a single scan. For each array, we selected the scan intensity that yielded the best separation among genotypes based on visual inspection, thus mimicking a procedure likely to be applied in practice when multiple scans are available. The data were analyzed using a mixed model approach described in detail by Keller *et al*. (2005). We tested pairwise contrasts using a Wald-type t-test based on the mixed model fit. We compared P-values of these tests obtained for single-scan data and multiple scans. The P-value, which assesses the significance of a test, is a measure that integrates accuracy of a method: If compared with a standard method P-values of a new method are decreased, this indicates increased accuracy of the new method.

IMPLEMENTATION

We used procedures of the SAS system for all computations (SAS Institute, 1999). Non-linear regression was performed with the NLIN procedure. The LOESS fit for the logarithmically transformed sample variance was computed using the LOESS procedure. Programming statements are available from the first author upon request.

RESULTS

The plot of logarithmically transformed variance of residuals versus spot effect shows marked heterogeneity of variance. The plot for

Table 1. Parameter estimates of intensity effects α_j and nonlinear parameters θ and γ for chip 288

Parameter	Estimate[a]	Standard error
α_1	−5.035	0.000745
α_2	−3.744	0.000745
α_3	−2.655	0.000745
α_4	−1.682	0.000746
α_5	−0.797	0.000748
θ	15.598	0.00900
γ	1.999	0.1024

Genotype: UH005x301. Intensities: 20, 30, 40, 50, 60, 70. Channel: Cy5.
[a]By WLS, using LOESS fit of variance function.

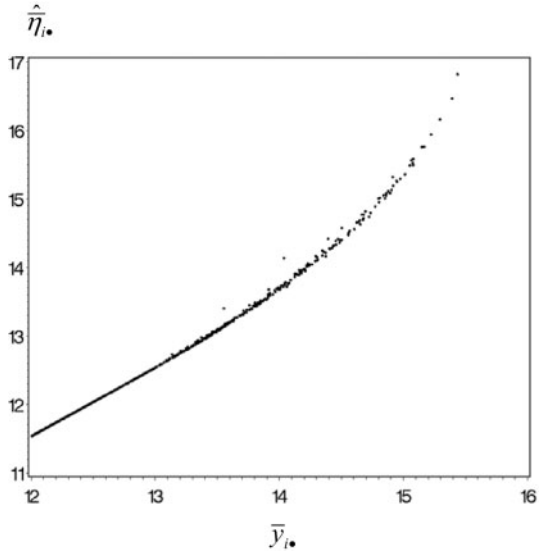

Fig. 4. Plot of model-based spot mean $\hat{\bar{\eta}}_{i\bullet}$ versus simple spot mean $\bar{y}_{i\bullet}$ for spots near the saturation limit (chip 288). Genotype: UH005x301. Intensities: 20, 30, 40, 50, 60, 70. Channel: Cy5.

example shown in Figure 3 also indicates that the variance stabilizes for high signal values. The weighted estimates of the intensity effects for one array are reported in Table 1. The small standard errors support our claim that intensity effect estimates can approximately be regarded as known. The threshold is estimated as $\hat{\theta} = 15.698$, while $\hat{\gamma} = 1.999$. The resulting estimate of the breakpoint is $\hat{\phi} = 15.1977$. For spots near the saturation limit, we computed estimates $\hat{\bar{\eta}}_{i\bullet}$ based on the fitted model and compared these with the simple spot means $\bar{y}_{i\bullet}$ (Fig. 4). It is seen that the model-based adjustment results in an upward shift compared with the naïve spot means. We also plotted estimates of μ_{ij} versus $\bar{\mu}_{i\bullet}$ for all six intensities (Fig. 5). The resulting curves closely resemble the plot of y_{ij} versus $\bar{y}_{i\bullet}$ (Fig. 1), suggesting that the model fit is very good.

For 12 contrasts we evaluated the *P*-values of *t*-tests based on single and multiple scans. In most cases, multiple scans led to more significances (Table 2). For two contrasts, we plotted *P*-values of all spots for both analyses. When both tests yield identical result, all spots should fall on a straight line with slope equal to unity. The examples in Figures 6 and 7 show two cases, where *P*-values of

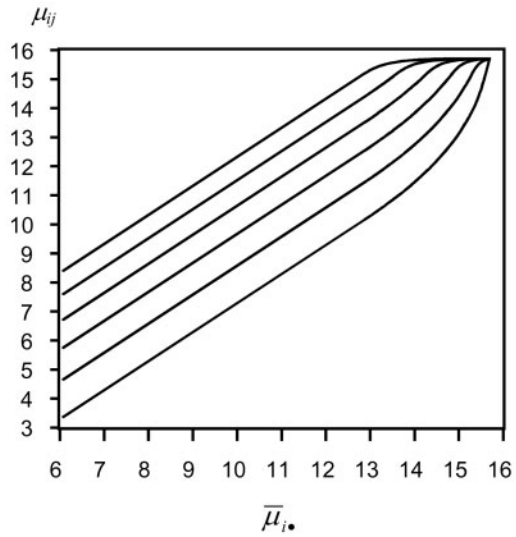

Fig. 5. Plot of estimates of μ_{ij} versus $\bar{\mu}_{i\bullet}$ for all six intensities for chip 288. Genotype: UH005x301. Intensities: 20, 30, 40, 50, 60, 70. Channel: Cy5.

Table 2. Percentage of spots significant at $\alpha = 5\%$ (not adjusted for multiplicity) for twelve pairwise contrasts among parents and hybrids.

Contrast among genotypes		Percentage significant	
		Single scan	multiple scans
(1)	UH005x005 versus UH005x301	4.44	5.32
(2)	UH005x005 versus UH301x005	1.09	1.13
(3)	UH005x301 versus UH301x301	0.74	0.50
(4)	UH301x005 versus UH301x301	0.86	0.91
(5)	UH002x002 versus UH002x250	4.48	5.61
(6)	UH002x002 versus UH250x002	9.28	10.03
(7)	UH002x250 versus UH250x250	7.95	13.38
(8)	UH250x002 versus UH250x250	7.14	9.92
(9)	UH002x002 versus UH002x301	12.75	11.22
(10)	UH002x002 versus UH301x002	11.19	9.49
(11)	UH002x301 versus UH301x301	5.54	6.69
(12)	UH301x002 versus UH301x301	5.49	7.06

multiple scans were reduced compared with single scans, thus yielding more significant results.

DISCUSSION

The motivation for developing the method proposed in this paper was to address the problem of scanning spots near the saturation limit in the presence of measurement error. Some alternative suggestions have been made on how to use multiple scans to reduce variability of microarray expression data. Our approach is novel in that non-linearity near the saturation limit is explicitly modeled, thus allowing an unbiased adjustment for saturated spots. Also, heterogeneity of variance, which was dramatic in the example used in this paper, is accounted for by a WLS approach. Essentially, our non-linear model linearly extrapolates the observed signal beyond the saturation limit and estimates what would have been observed in the absence of a saturation limit (Garcia de la Nava

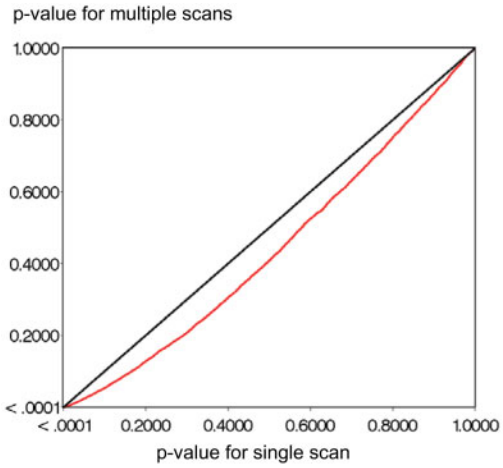

Fig. 6. *Q–Q* plot of *P*-values (red line) for multiple scans versus single scans for contrast UH002x250 versus UH250x250 (contrast 7 in Table 2). Black line: reference line with slope equal to unity.

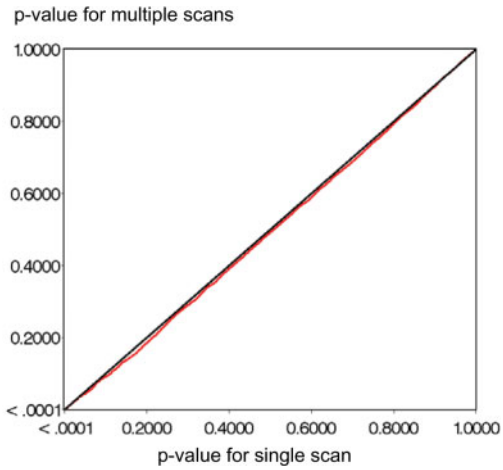

Fig. 7. *Q–Q* plot of *P*-values for multiple scans versus single scans for contrast UH005x005 versus UH301x005 (contrast 2 in Table 2). Black line: reference line with slope equal to unity.

et al., 2004). The procedure requires multiple scans at different scanning intensities.

A main issue addressed by our method is non-linearity near the saturation limit and the difficulty of determining whether or not a spot signal is at the saturation limit. The difficulty arises from the presence of measurement error. If one could unequivocally identify observations in the saturation region ($\eta_{ij} > \phi$), one could delete these and fit the additive model (2a) to the remaining data. Provided for each spot at least one observation obeys the inequality $\eta_{ij} > \phi$, one could thus obtain unbiased estimates of spot effects g_i. In practice, this procedure may be difficult to implement when measurement error is not negligible at the saturation limit, as was the case for the maize data (Fig. 1). Correct classification of an observation as being at the saturation limit is difficult, since both η_{ij} and ϕ are not directly observable, and one is likely to miss some signals for which $\eta_{ij} > \phi$, while some signals are kept for which $\eta_{ij} < \phi$. False classification of observations owing to measurement error may yield a downward bias for spots close to the saturation limit.

Romualdi *et al*. (2003) propose to use either the maximum or the mean pixel-intensity of several scans for further analysis, arguing that the detection of differentially expressed genes can be considerably enhanced. However, no correction for saturated pixels is provided. Lyng *et al*. (2004) estimate intensity effects as the mean ratio of untransformed signals at two scanning intensities, assuming a constant ratio across spots not at the saturation limit. This is essentially equivalent to our assumption of additive intensity effects on a logarithmic scale. The approach by Lyng *et al*. (2004) allows only two scanning intensities, which may not be enough to accommodate optimal scanning intensities for all cDNA spots. Moreover, their mean ratio estimator does not allow to handle heterogeneity of variance (see Appendix B in the online Supplementary Data), which appears to be similarly pronounced in their data as in our maize data. Taking logarithms and working on an additive scale does not fully resolve the problem. We have demonstrated, however, how a simple WLS approach can be implemented to account for heterogeneity of variance.

A correction method for saturation at pixel-level, where a single scanning of an array is sufficient, is provided by Dodd *et al*. (2004). Similar to Lyng *et al*. (2004), a linear relationship is estimated, though not between the spots of two scans, but between the pixels of the Cy3 and the Cy5 channel of a certain spot. Thus the pixel of the channel, where no saturation is present, is used for calculating the corrected value for the saturated pixel. The threshold for pixels regarded as saturated is set at a fixed value, which might introduce misclassification, as discussed above. Also, when the number of saturated pixels is high or saturation occurs in both channels, the performance of the procedure might not be optimal.

Garcia de la Nava *et al*. (2004) provide an approach that adjusts signals not only for saturation but for quantization, i.e. the process of assigning each possible value for the effective amount of hybridization product a discrete value of signal intensity. Each array is scanned at two different intensities. Signal intensities of both scans are related to one another by a linear or a gamma curve using robust regression. With the parameter estimates a maximum likelihood approach is used to gain corrected signal values. Variance heterogeneity does not seem to be an important issue in their data and so is neglected.

This paper has focussed on non-linearities near the saturation limit. In principle, non-linearities could also occur near the detection limit, though this was not a serious issue with our data. Extension of our algorithm to cater for non-linearities near the detection limit is straightforward. One would simply need to add a non-linear segment at low intensities, yielding a model with three segments: the middle segment is linear, while the two outer segments are non-linear. This model would have two additional non-linear parameters compared with model (2).

With multiple laser scans, photobleaching is a potential problem, though we did not observe this effect in our experiments. Photobleaching occurs mainly when scanning intensities are very high, reducing the emitted signal remaining for subsequent scans. As a precaution, we therefore always started with the lowest scanning intensity and finished with the highest. While we did not find evidence of photobleaching with our scans, it should be stressed that our proposed model will automatically account for such effects. Assuming that photobleaching affects all spots in a similar way,

the intensity effects α_j will capture any signal reductions across spots.

Analysis of the maize data revealed that variance increases dramatically as signal intensity decreases (Figs 1 and 3). Thus, repeated scans are also useful for spots with signals well removed from the saturation limit in that they provide measurement replication and thus reduce the standard error of mean signals. Multiple scans are therefore of benefit not only for spots with very high signal, for which saturation is a problem, but also for spots with very low signal, for which measurement error is the primary issue.

Conflict of Interest: none declared.

REFERENCES

Carroll,R.J. and Ruppert,D. (1988) *Transformation and Weighting in Regression*. Chapman and Hall, London.

Cleveland,W.S. (1979) Robust locally weighted regression and smoothing scatterplots. *J. Am. Stat. Assoc.*, **74**, 829–836.

Dodd,L.E. *et al.* (2004) Correcting log ratios for signal saturation in cDNA microarrays. *Bioinformatics*, **20**, 2685–2693.

Engel,B. (1990) The analysis of unbalanced linear models with variance components. *Stat. Neerl.*, **44**, 195–219.

Garcia de la Nava,J. *et al.* (2004) Saturation and quantization reduction in microarray experiments using two scans at different sensitivities. *Stat. Appl. Genet. Mol. Biol.*, **3**, 11.

Harville,D.A. (1997) *Matrix Algebra from a Statistician's Perspective*. Springer, Berlin.

Hedge,P. *et al.* (2000) A concise guide to cDNA microarray analysis. *Biotechniques*, **29**, 548–562.

Keller,B. *et al.* (2005) Designing a microarray experiment to estimate heterosis. *Theoret. Appl. Genet.* **111**, 57–64.

Korn,E.L. *et al.* (2004) Controlling the number of false discoveries: application to high-dimensional genomic data. *J. Stat. Plan. Inference*, **124**, 379–398.

Kroonenberg,P.M. and De Leeuw,J. (1980) Principal component analysis of three-mode data by means of alternating least squares algorithms. *Psychometrika*, **45**, 69–97.

Lyng,H. *et al.* (2004) Profound influence of microarray scanner characteristics on gene expression ratios: analysis and procedure for correction. *BMC Genomics*, **5**, 10.

Mood,A., Graybill,F.A. and Boes,D.C. (1974) *Introduction to the Theory of Statistics*. 3rd edn. McGraw-Hill, New York.

Romualdi,C. *et al.* (2003) Improved detection of differentially expressed genes in microarray experiments through multiple scanning and image integration. *Nucleic Acids Res.*, **31**, e149.

SAS Institute, Inc. (1999), *SAS/STAT User's Guide*. Version 8. SAS Institute, Cary, NC.

Schabenberger,O. and Pierce,F.J. (2002) *Contemporary Statistical Models for the Plant and Soil Sciences*. CRC Press, Boca Raton.

Searle,S.R. (1987) *Linear Models for Unbalanced Data.* Wiley, New York.

Seber,G.A.F. and Wild,C.J. (1989) *Nonlinear Regression*. Wiley, New York.