BIOINFORMATICS

Vol. 20 no. 13 2004, pages 2026–2034
doi:10.1093/bioinformatics/bth193

DNA microarray normalization methods can remove bias from differential protein expression analysis of 2D difference gel electrophoresis results

David P. Kreil[1,*], Natasha A. Karp[2] and Kathryn S. Lilley[2]

[1]Department of Genetics/Inference Group (Cavendish Laboratory) and [2]Department of Biochemistry, University of Cambridge, Downing Street, Cambridge CB23EH, UK

Received on September 23, 2003; revised and accepted on February 26, 2004
Advance Access publication March 25, 2004

ABSTRACT

Motivation: Two-dimensional Difference Gel Electrophoresis (DIGE) measures expression differences for thousands of proteins in parallel. In contrast to DNA microarray analysis, however, there have been few systematic studies on the validity of differential protein expression analysis, and the effects of normalization methods have not yet been investigated. To address this need, we assessed a series of same–same comparisons, evaluating how random experimental variance influenced differential expression analysis.

Results: The strong fluctuations observed were reflected in large discrepancies between the distributions of the spot intensities for different gels. Correct normalization for pooling of multiple gels for analysis is, therefore, essential. We show that both dye-specific background levels and the differences in scale of the spot intensity distributions must be accounted for. A variance stabilizing transform that had been developed for DNA microarray analysis combined with a robust Z-score allowed the determination of gel-independent signal thresholds based on the empirical distributions from same–same comparisons. In contrast, similar thresholds holding up to cross-validation could not be proposed for data normalized using methods established in the field of proteomics.

Availability: Software is available on request from the authors.

Contact: D.Kreil@gen.cam.ac.uk

Supplementary information: There is supplementary material available online at http://www.flychip.org.uk/kreil/pub/2dgels/

INTRODUCTION

Two-dimensional (2D) polyacrylamide gel electrophoresis is a high-throughput method used for the measurement of changes in the expression levels of thousands of individual proteins in parallel, affording a global view of the state of a proteome (Lilley *et al*., 2002). Significant advances have been realized by coupling 2D gel analysis with mass spectrometry. Protein spots can rapidly be identified through in-gel digestion, subsequent mass spectrometry and database searching (Fey and Larsen, 2001).

There are methods for the quantitative study of protein expression that do not use 2D gels, e.g. exploiting isotopic labelling of samples. Approaches include labelling proteins after extraction with a chemical tag which can be supplied in several stable isotopic forms—e.g. ICAT (Li *et al*., 2003)—and differential incorporation of a stable isotope during the growth of an organism (Krijgsveld *et al*., 2003). These methods can largely be viewed as complementary techniques to modern 2D gel experiments, having the potential to give information on sets of proteins which are poorly represented on 2D gels, such as integral membrane proteins. Both techniques are still relatively new and there is little information in the literature about the experimental variance associated with their use.

In early comparative 2D gel experiments, each of the samples to be analysed was run in a separate gel. For a long time, the low dynamic range and high variability of the traditional silver stain limited quantitative work. Only recently has fluorescent labelling been employed in the field of proteomics. The post-electrophoretic fluorescent stain SYPRO Ruby gives a dynamic range of four orders of magnitude (Lopez *et al*., 2000; Malone *et al*., 2001). High gel-to-gel variation, however, makes the detection of corresponding spots unreliable, and the quantification of true differences using this method is very difficult (Asirvatham *et al*., 2002; Alban *et al*., 2003).

The task of detecting true changes in protein expression has been greatly simplified by the introduction of Difference Gel Electrophoresis (DIGE) by Ünlü *et al*. (1997). The method has then been commercialized by Amersham Biosciences, and has only last year become widely available to researchers. In this approach, the samples to be compared are labelled with spectrally resolvable fluorescent dyes. The labelled samples are mixed before running them in a single 2D gel. Using the

*To whom correspondence should be addressed.

cyanine dyes Cy2, Cy3, and Cy5, up to three samples can be examined in parallel, each dye giving an *independent channel* of measurement. The extent of sample loss during protein separation in the first gel-dimension varies strongly from gel to gel. Variation of spot intensities due to gel-specific experimental factors, however, will be the same for all samples run on a particular DIGE gel. Consequently, the relative amounts of a particular protein in the samples will be unaffected. This is exploited by differential in-gel analysis (DIA). For each spot, the intensities in the respective dye channels can be compared directly, giving ratios of fluorescence intensities as the primary indicators of differential protein expression (Gharbi *et al*., 2002; Yan *et al*., 2002).

For the combination of multiple gels into an experiment, aliquots of all the samples to be compared can be pooled into an *internal standard*, which is run in one channel on all the gels for standardization purposes. This naturally extends intra-gel comparisons of DIA to also allow inter-gel analysis (Alban *et al*., 2003). Yet, converting the multiple gel images from either type of analysis into expression difference calls for individual proteins remains a complex challenge and is an area of active research. Due to its ability to co-detect spots in multiple images, the DeCyder software system has become a widely used analysis platform, and is also used in this laboratory for image analysis and quantification (Amersham Biosciences, 2002, http://www.amershambiosciences.com/). A comparison with other gel analysis tools is not attempted in this study. The issues discussed here will similarly affect other gel analysis software.

In this study, we focus on the influence of experimental error in a typical DIA set-up. Our findings, however, equally apply to multi-gel experiments standardized by a common reference sample, although further issues that are not discussed here will be of concern.

In an assessment of experimental error, two main classes of errors must be distinguished: (a) random fluctuations (*noise*) and (b) systematic trends (*bias*). Replicate experiments are used to reduce uncertainty arising from noise. To combine data from multiple gels, it is well appreciated that data must be normalized, e.g. to compensate for differences in overall system gains. After normalization, the signal distributions from multiple gels must be similar for meaningful statistical analysis. In contrast to the field of DNA microarray analysis, however, the effectiveness of different normalization methods has received little explicit attention in published analyses of protein expression. Remaining system bias, such as any dye-specific effects, must be removed before any statistical analysis can proceed. Ideally, bias would be controlled at the stage of the experiment in the laboratory. If this is not possible, it must be compensated by normalization. It should be emphasized that no amount of replication can make up for a lack of control for bias.

This study, for the first time, quantifies both types of error in protein expression measurement by DIGE, and examines how they influence differential expression analysis. Improved methods dealing with the complications encountered are introduced and validated. We also give principled advice on the interpretation of DIGE analysis results.

SYSTEM AND METHODS

Experimental error was studied by a series of six same–same comparisons. Aliquots of the same protein sample were individually labelled with one of the three fluorescent cyanine dyes developed for DIGE, giving a total of 18 samples. For each of the six gels, samples labelled with different dyes were then pooled and separated by 2D gel electrophoresis. Details of experimental methods are available in the Online Supplement.

Image quantification for DIA

Image analysis was performed using DeCyder V4.0 (Amersham Biosciences, Sweden), a 2D gel analysis software package designed specifically to be used with DIGE. The estimated number of spots for each co-detection procedure was set to 2000. As recommended, a conservative exclusion filter rejecting spots with a slope greater than one was applied to remove artefacts caused by dust particles. Spot intensities are spot volumes, i.e. a particular spot intensity is obtained by integrating the pixel intensities over the spot area.

The analysis has also been repeated with a preview version of DeCyder V5.0 with no major differences observed.

DeCyder DIA data transformation

The DeCyder DIA module processes fluorescent spot intensities in two steps. First, local estimates of background fluorescence are subtracted from the measured spot intensities, then signals are adjusted to compensate for dye-specific system gain to give normalized spot intensities. This procedure is common to analysis tools in the field of proteomics, with differences only in the implementation of each step.

Background estimates in DeCyder are formed for each spot by taking the 10th percentile of the pixel values observed on the spot boundary. For overlapping spots, this may lead to an overestimate of background fluorescence (cf. Fig. 1a). To compensate for dye-specific system gain, one channel is used as reference, and the other channels are then rescaled. This implicitly assumes that there are no dye-specific effects other than that one dye is brighter by a constant factor. In the case of two channels with intensities (volumes) V_1 and V_2, and the second channel being used as reference, the first channel will be rescaled to

$$V_1' = aV_1 \tag{1}$$

with the constant factor a yet to be determined. The relative intensities V_2/V_1' unfortunately are asymmetric with respect to over- or under-expression of a particular protein. Analysing the data transformed to log-space is preferable:

$$R = \log V_2 - \log V_1' = \log(V_2/V_1'). \tag{2}$$

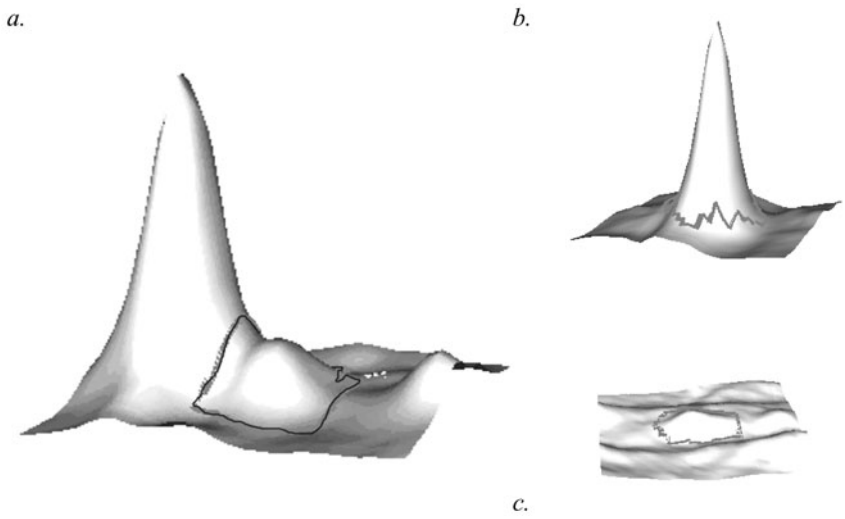

Fig. 1. DeCyder 3D display of spot image data. Spots are typically scored for smooth cone profiles and lack of irregularities. Spot boundaries are shown by a dark-grey line at the base. (**a**) Overlapping spots are very common. If several spots overlap, estimation of background from the pixel intensities on the spot boundary becomes problematic. (**b**) This protein spot has a typical smooth cone profile (**c**) An example of a low-intensity spot is shown. These are particularly difficult to validate.

As a difference of log-transformed intensities, R is symmetric with respect to over- or under-expression. For non-differentially expressed proteins, R is zero.

To determine the rescaling factor a, a normal distribution is fitted to the mode of the empirical distribution of R, excluding data that by an empirical criterion appears not to belong to the main peak, and using a standard least squares gradient descent (Amersham Biosciences, 2002). The factor a is then chosen to shift the mean of the fitted normal distribution to zero, reflecting the assumption that most proteins are not expected to be differentially expressed.

The software then transforms the log-ratio R back to linear space,

$$E = \begin{cases} V_2/V_1' & \text{for } R > 0, \text{ i.e., } V_2/V_1' > 1 \\ -V_1'/V_2 & \text{for } R < 0, \text{ i.e., } V_1'/V_2 > 1 \end{cases} \quad (3)$$

As this creates a disjoint distribution of values, however, we prefer to work with the data in log-space.

Offset/scale normalization method

Following an observation from cDNA microarray analysis that estimates of local background can be unreliable and that, *in lieu* of effective local estimates, the simple model of a globally constant background can be successful (Brown *et al*., 2001), we wanted to transform each channel

$$V' = aV + b, \quad (4)$$

where the scaling factor a adjusts for dye-specific system gain and the additive offset b corrects for different background fluorescence intensities. Unfortunately, at present there is no method to export spot intensities from DeCyder that had not yet had a local background estimate subtracted. Consequently, the random error that the local background estimation introduced could not be removed.

The transform (4) will, however, compensate for any constant additive bias present after the subtraction of local background estimates. The transform parameters a and b in (4) are determined by an iterative trimmed least squares maximum likelihood estimate assuming that most proteins are not expected to be differentially expressed. Our implementation uses the code published by Huber *et al*. (2002, cf. http://www.bioconductor.org/repository/devel/package/html/vsn.html) for the normalization of DNA microarray data. Adaptations by this laboratory include automatic convergence detection, which is available now in the current release of the published code. Both additive and multiplicative noise are explicitly allowed for in the model employed,

$$y = \alpha + \beta\mu e^{\eta} + \varepsilon \quad \eta \sim \mathcal{N}(0, \sigma_\eta^2) \;\; \varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2), \quad (5)$$

where y is the observed signal, μ the true expression level, α a constant background term, β the channel gain, and η and ε are normally distributed noise terms.

It should be noted that, if additive noise ε is also permitted, it is not a log-transform that decouples the variance from the signal intensity (a property desired for statistical analysis), but an asinh-transform (Munson, 2001; Durbin *et al*., 2002; Huber *et al*., 2002; Durbin and Rocke, 2003). The difference between the two transforms is pronounced for signal values close to zero, but disappears quickly for larger signal levels. For the gels encountered in this study, normalized signal levels were high, so that the asinh-transformed data were identical to the log-transformed data within measurement error. In other

experimental situations, however, this may not be the case, and use of the log transform will lead to inflated variation at low signal levels of the log-transformed normalized data. We suggest using an asinh-transform instead of the log in such cases, rather than try and deal with the signal-dependent variance explicitly (the latter approach has been adopted by Tonge *et al*., 2001).

Standardization of difference signal distributions from multiple gels

To standardize for the varying degrees of scale of the distributions from multiple gels, we compute

$$Z = (R - M_R)/S_R \quad M_R = \text{median}(R) \quad S_R = \text{MAD}_{\text{adj}}(R) \tag{6}$$

with the log-ratio R, a robust estimate of location M_R and of scale S_R. MAD_{adj} is the median absolute deviation adjusted for asymptotically normal consistency.

Validation of statistically determined thresholds

In a 'leave-one-out approach' to cross-validation, thresholds for experimental variance were empirically determined from data pooled from five of the six gels. The thresholds were then tested on the remaining, sixth gel. This was repeated so that each gel was once excluded from the pool and used as a test set. This is a common approach to validation when the number of sample sets is too small to allow full bootstrap sampling.

RESULTS AND DISCUSSION

Differently labelled aliquots of the same sample have been compared by DIGE in multiple independent experiments. After normalization for dye-channel specific differences, the fluorescent dye intensities are expected to show equal amounts of protein in each channel for any particular resolved spot. Whichever pair of dyes is being considered, any deviations from the expected log-ratio of zero, therefore, reflect experimental error. Two main classes of errors must be distinguished: random fluctuations (*noise*), and systematic trends (*bias*). Both types of error can clearly be seen in the non-normalized data when the spot intensities of one channel are plotted against the respective intensities of another channel (Fig. 2a and b). The increase of scatter at high intensities observed in Figure 2a is indicative of multiplicative noise, and justifies a log-transform, which also allows more appropriate visualization of data spanning several orders of magnitude. In the logarithmic plot, two deviations from channel identity are not noise: (a) the Cy5 channel has a higher system gain than the Cy3 channel, and (b) low-intensity spots are significantly brighter in Cy3 than expected by their Cy5 signal. Normalization needs to remove both effects in order for a statistical analysis of differences between channels to be meaningful.

Effectiveness of bias removal by alternative normalization methods

The scale normalization of DeCyder DIA successfully equalizes the dye-specific differences in system gain. A bias for low-intensity spots, however, is present after scale normalization and can best be viewed as a plot of channel difference as a function of channel average (Fig. 2c).

It could be argued that the spots affected by bias might be artefacts of the spot-finding stage of image analysis. The hypothesis was tested by manually reviewing all spots of a particular gel. This is a laborious process, which requires a skilled operator to judge whether a spot looks 'real' by examination of the raw image data (Fig. 1). Decisions are particularly difficult for low-intensity spots (Fig. 1c). Consequentially, as it is low-intensity spots that show the bias, many affected spots are removed in conservative manual review. A large number of spots that manual review confirms as real, however, still show the same type of bias.

Similar forms of bias have been observed in studies examining normalization transforms for DNA microarrays (Cui *et al*., 2002, http://www.jax.org/staff/churchill/labsite/research/expression/Cui-Transform.pdf), suggesting that a different normalization transform might resolve the issue. Indeed, using a normalization transform by Huber *et al*. (2002) that explicitly allows for an offset correction to adjust for dye-specific background fluorescence we could completely remove the bias (Fig. 2d). Table 1 highlights the impact that the choice of normalization methods makes on the perceived relative protein expression—under-expression in the Cy5 channel judging from scale normalized data, but non-differential expression according to offset/scale normalization. Here, we have picked an arbitrary spot of moderately low intensity from the set of spots that had been confirmed in manual review.

Dye-specific background fluorescence and system response can explain the systematic differences observed between the channels. The gel itself, the glass plate support, scratches and various non-protein matter will show fluorescence in the wavelength region of concern. To assess dye-specific background fluorescence, three empty gels have been scanned (see Methods section of the Online Supplement). For all the gels examined, background fluorescence consistently was strongest in Cy2 (5159 ± 545), slightly lower for Cy3 (4588 ± 836) and very low for Cy5 (212 ± 121). In contrast to DNA microarrays, where the fluorescence outside of spot areas may differ from non-specific spot signals (cf. Brown *et al*., 2001), the background fluorescence of gel and support as measured here will contribute additively to the fluorescence of protein spots. This explains why a shift between channels is observed in raw data and a normalization of channels is required to remove that difference.

Spatial variation in background was clearly visible and hence justifies attempts at local background subtraction. It was largest for Cy3 (205 ± 41), moderate for Cy2 (165 ± 37)

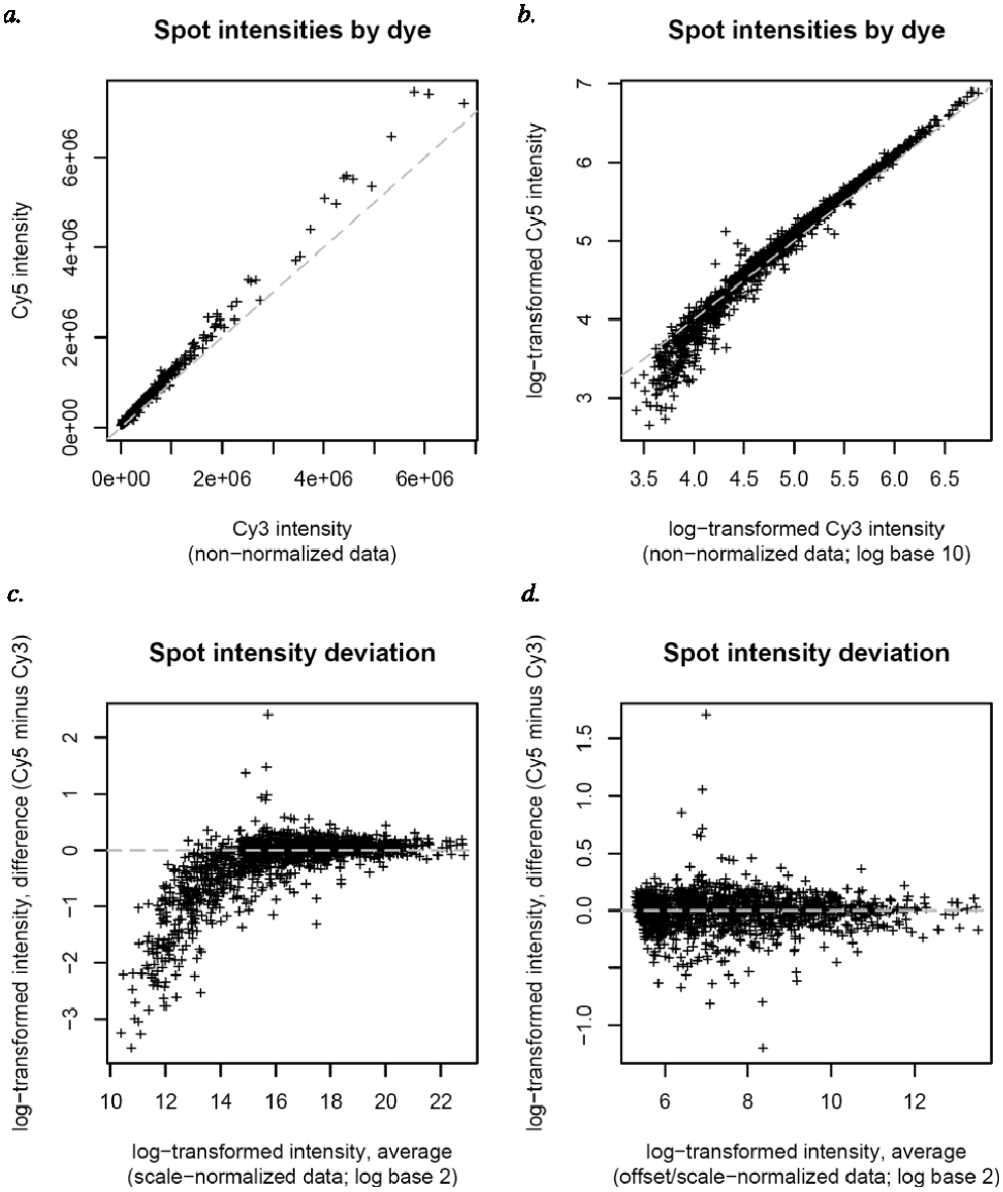

Fig. 2. One of the six gels examined was arbitrarily chosen for these and all subsequent figures (Gel number 4, see Supplement). (**a**, **b**) plot the non-normalized spot intensities of one channel (Cy3) versus the respective intensities of another channel (Cy5). The data are as exported from DeCyder, i.e. after subtraction of estimates of local background fluorescence. (**c**, **d**) show channel differences versus channel averages for normalized data. Traditionally, the channel difference (also 'log-ratio'), $\log_2(\text{Cy5}) - \log_2(\text{Cy3}) = \log_2(\text{Cy5/Cy3})$, is plotted for easy interpretation as fold-change. On that scale, $+1$ means 2-fold over-expression of protein in the Cy5-channel compared to the Cy3-channel; -1 means 2-fold under-expression. (a) On a linear scale, the expected relationship between the two channels is seen. Dye-specific differences in system gain are reflected in that the trend seen in the data deviates from the grey dashed line, which indicates channel identity. The scatter is the result of experimental noise. (b) On a logarithmic scale, a bias for low intensity spots also becomes apparent. See text for discussion. (c) Channel difference shown as a function of channel average for scale normalized Cy3/Cy5 spot intensities. (d) Channel difference shown as a function of channel average for offset/scale normalized Cy3/Cy5 spot intensities.

and lowest for Cy5 (42 ± 35). As DeCyder already subtracts estimates of local background fluorescence, if these estimates were unbiased, no dye-specific trends would be expected in the normalized data. Spatial variation was lower than the differences observed between gels, suggesting that appropriate normalization between gels was more critical than correct estimation of local backgrounds. Offset/scale normalization will compensate for such differences between gels and may hence also be a good choice for laboratories that work with multiple gels of one dye only. (The method presented here

Table 1. Effect of failure to compensate for dye-specific background fluorescence, illustrated using an arbitrary spot of moderately low intensity from the set of spots that had been confirmed in manual review

	Channel 1 Cy3	Channel 2 Cy5
Spot intensities after subtraction of DeCyder estimates for local background fluorescence	4652	2377
Spot intensities after scale normalization with $a = 1.18$	5492	2377
DeCyder ratio ($\log_2$-ratio)	$E = -1.66$ ($R = -0.73$)	
Spot intensities after offset/scale normalization with $a_{1/2} = 1.86/1.54 \times 10^{-3}$, $b_{1/2} = 32.8/39.5$	41.45	43.16
Offset/scale normalized ratio ($\log_2$-ratio)	$E = 1.04$ ($R = 0.06$)	

requires, however, that spots from the different gels can fairly reliably be matched to one another.)

Validation of significance estimates of expression difference calls

Given normalized data, a major question of interest is the determination of differentially expressed proteins. Due to the inherent noise in the system, there is a chance of false difference calls (type I error) as well as a likelihood of missing true differences (type II error). Traditionally, a method is desired that indicates whether a certain protein is differentially expressed, given a certain tolerance for type I and type II errors per test. It should be emphasized that the error rates specified are for the test of a specific protein spot. By ranking differential signals, however, many hundreds of spots are being evaluated, which must be accounted for in assessing overall significance of a result. A conservative view will make use of Bonferroni correction, while less stringent correction procedures have recently been discussed for the analysis of microarray data (Tusher *et al*., 2001; Dudoit *et al*., 2002).

An established approach to assessing the significance of a differential expression signal for a specific protein, which is supported by DeCyder, fits a normal distribution to the difference signals of all the spots from a same–same experiment. Under the assumption that non-differential signals of future experiments will follow a very similar distribution, the parameters of the fitted normal distribution can then be used in the traditional way to assess significance of extreme values. Thresholds of ± 2 SD as suggested by Amersham Biosciences (2002) then correspond to a type I error rate of 4.6% per test. Under the assumption that a particular protein is not differentially expressed, there is a chance of 4.6% that a log-ratio beyond these thresholds is observed, wrongly classifying the protein as 'differentially expressed'. The two main requirements for this approach to be valid are that (a) the distributions of difference signals of multiple gels are sufficiently similar after normalization and (b) a normal distribution is a good approximation to this distribution.

Comparison of the frequency histogram of difference signals with the fitted normal distribution (Fig. 3a) shows strong deviations from normality in both asymmetry (skew) and the presence of heavy tails (positive kurtosis). This is also seen very clearly in a quantile–quantile (QQ) plot, where quantiles of the observed distribution are plotted against the quantiles of the fitted normal distribution (Fig. 3b). For good agreement, the QQ plot will form a straight line of slope one.

The corresponding plots for offset/scale normalized data (Fig. 3c and d) present a markedly improved picture. Heavy tails, however, are still present. The improvement is caused by the compensation for the dye-specific bias for low-intensity spots. This can be shown by progressive exclusion of low-volume spots from the scale normalized data. Normality improves comparably (Online supplement), at the cost, however, of losing sensitivity. Approximately 50% of all data had to be excluded to remove the bias this way.

In all cases, thresholds based on parameters of a fitted normal distribution are inappropriate due to the heavy tails of the log-ratio signal distribution. If distributions from multiple gels were sufficiently similar, however, the 5th and the 95th percentile could, e.g. be used as empirical thresholds for an empirical per-family error rate of 10% (cf. Tusher *et al*., 2001). The considerably varying spreads of the distributions from multiple gels, however, have to be standardized before such an approach can be fruitful.

Thresholds that held up to cross-validation could not be determined at all for data with bias, as the bias varied strongly between different gels. Using a robust Z-score, offset/scale normalized data distributions, however, are matched well enough across multiple gels such that empirical thresholds that hold up to cross-validation can be specified for empirical per-family error rates of 10% or higher (Table 2). As can be seen from the QQ plot, the empirical distributions can be employed somewhat beyond the region where they are well approximated by a normal distribution (Fig. 3d). For error rates lower than 10%, cross-validation showed reduced reliability due to the high gel-to-gel variation in the tips of the distribution tails (Table 2 and Fig. 3d). Instead of using a robust Z-score, by quantile normalization after robust regression, the distributions could be equalized over their entire range. Such an approach would then have to be assessed with spike-in experiments. The same is true for the popular methods using loss smoothers. In general, future transfers of concepts developed in the context of DNA microarray analysis to proteomics seem promising.

To conclude, from the study of variation both within and across gels of the series of same–same experiments, we find:

(1) Correct normalization for differences in fluorescent background is essential. An ideal solution would

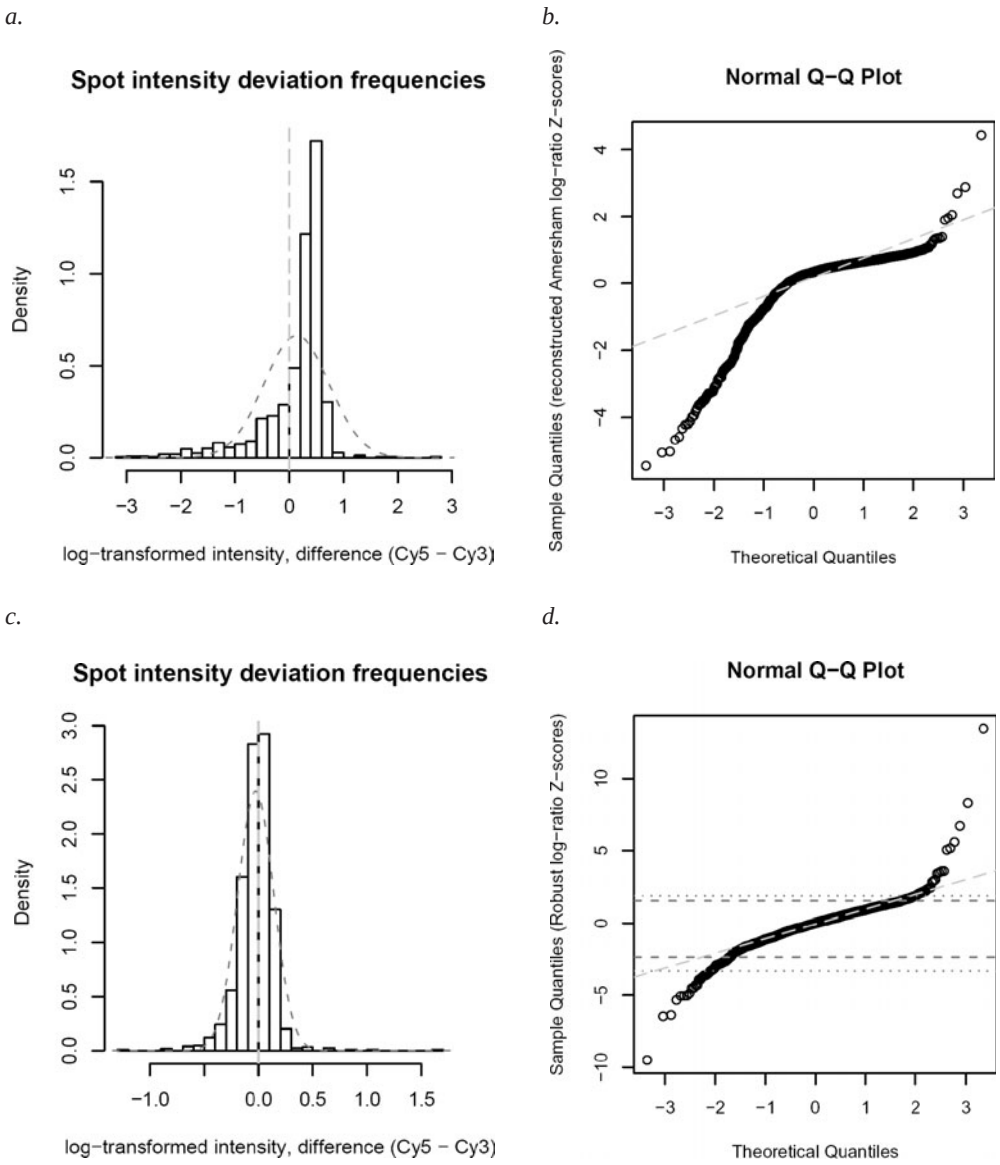

Fig. 3. Distribution of difference signals ($\log_2$-ratios) in normalized data. (**a, b**) show the results of scale normalization, (**c, d**) are for offset/scale normalization. In either case, strong deviations from normality can be observed. Consequently, empirical score thresholds must be determined. (a, c) Histogram showing relative frequencies of difference signal values, compared with a best-fit normal distribution (grey dashed curve). (b, d) Normal quantile–quantile (QQ) plot: Quantiles of the observed distribution are plotted against the quantiles of the fitted normal distribution. For good agreement, the QQ plot will form a straight line of slope one: The light-grey long-dashed line corresponds to identity. Axes values denote multiples of the standard deviation. The grey dashed horizontal lines correspond to score thresholds for 10% empirical per-family error rates, the grey dotted lines represent the cutoffs for 5% error rates. For scale normalized data, no such cutoffs stood up to cross-validation. See Table 2, and text for discussion.

also account for spatial variation in background fluorescence. Data processed using the present DeCyder implementation of local background correction, however, showed clear dye-specific bias, which is most apparent for spots of low intensity. To resolve this problem, we recommend subjecting the data to the offset/scale normalization presented here. This allows the use of all spots, increasing the sensitivity of the analysis. An alternative, less favourable solution is removal of all spots of low intensity at the expense of sensitivity, after which the data need to be renormalized. In this study, about half the spots would have had to be sacrificed to remove the bias this way. This can be acceptable for certain experiments which focus on high-intensity spots, only (e.g. with the intent of spot identification by mass spectrometry).

Table 2. Cross-validation results for Cy3/Cy5 (see Supplement for Cy2 results), shown for two different empirical per-family error rates (p*)

Gel excluded	$p^* = 5\%$ Thresholds from pool		False positives (%)	$p^* = 10\%$ Thresholds from pool		False positives (%)
	Lower	Upper		Lower	Upper	
1	−3.30	1.80	6.0	−2.39	1.49	10.7
2	−3.26	1.89	4.4	−2.28	1.53	10.1
3	−3.32	1.90	2.9	−2.35	1.53	8.6
4	−1.92	1.86	4.0	−2.37	1.52	8.8
5	−2.86	1.93	***6.9***	−2.20	1.54	11.6
6	−3.34	1.80	5.8	−2.38	1.49	10.9
Mean ± SD	**−3.00 ± 0.56**	**1.86 ± 0.05**	**5.0 ± 1.5**	**−2.33 ± 0.07**	**1.52 ± 0.02**	**10.1 ± 1.2**

(2) For the assessment of the significance of expression difference calls for specific protein spots, parametric thresholds based on fits of normal distributions to the data are not appropriate, as the observed distributions deviate widely from a normal distribution. Thresholds based on empirical distributions may be used. The feasibility of that approach has been validated in this study, and it was shown that after suitable normalization or other removal of bias, and the introduction of a robust Z-score, the thresholds -2.33 ± 0.07 and 1.52 ± 0.02 could successfully be used at an empirical per-family error rate of 10%. It should be noted that these scores may vary from laboratory to laboratory, also depending on the reagents and protocols employed. The method presented here, however, can be applied by individual groups to determine their own validated threshold values for the dye-pairs and protocols they use.

ACKNOWLEDGEMENTS

We thank S. Coulthurst for the provision of sample, and are grateful to W. Huber and I. Currie for helpful discussions. We also greatly appreciated the thorough and constructive criticism of our manuscript by the three anonymous reviewers appointed by the journal. D.P.K. acknowledges support by a Medical Research Council research fellowship (grant G81/555), N.A.K. acknowledges funding by the BBSRC (grant 8/G420). K.S.L. is a recipient of a Wellcome VIP award. N.A.K. and K.S.L. acknowledge support by the BBSRC Investigating Gene Function initiative.

REFERENCES

Alban,A., David,S.O., Bjorkesten,L., Andersson,C., Sloge,E., Lewis,S. and Currie,I. (2003) A novel experimental design for comparative two-dimensional gel analysis: two-dimensional difference gel electrophoresis incorporating a pooled internal standard. *Proteomics*, **3**, 36–44.

Amersham Biosciences (2002) Ettan DIGE User manual for DeCyder v4.

Asirvatham,V.S., Watson,B.S. and Sumner,L.W. (2002) Analytical and biological variances associated with proteomic studies of *Medicago truncatula* by two-dimensional polyacrylamide gel electrophoresis. *Proteomics*, **2**, 960–968.

Brown,C.S., Goodwin,P.C. and Sorger,P.K. (2001) Image metrics in the statistical analysis of DNA microarray data. *Proc. Natl Acad. Sci., USA*, **98**, 8944–8949.

Cui,X., Kerr,M.K. and Churchill,G.A. (2002) Data Transformation for cDNA Microarray Data, The Jackson Laboratory, Maine, USA.

Dudoit,S., Yang,Y.H., Callow,M.J. and Speed,T.P. (2002) Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Stat. Sin.*, **12**, 111–139.

Durbin,B., Hardin,J., Hawkins,D.M. and Rocke,D.M. (2002) A variance-stabilizing transformation for gene-expression microarray data. *Bioinformatics*, **18**, 105S–110S.

Durbin,B. and Rocke,D. (2003) Estimation of transformation parameters for microarray data. *Bioinformatics*, **19**, 1360–1367.

Fey,S.J. and Larsen,P.M. (2001) 2D or not 2D. Two-dimensional gel electrophoresis. *Curr. Opin. Chem. Biol.*, **5**, 26–33.

Gharbi,S., Gaffney,P., Yang,A., Zvelebil,M.J., Cramer,R., Waterfield,M.D. and Timms,J.F. (2002) Evaluation of two-dimensional differential gel electrophoresis for proteomic expression analysis of a model breast cancer cell system. *Mol. Cell. Proteom.*, **1**, 91–98.

Huber,W., von Heydebreck,A., Sültmann,H., Poustka,A. and Vingron,M. (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression, *Bioinformatics*, **18**, 96S–104S.

Krijgsveld,J., Ketting,R.F., Mahmoudi,T., Johansen,J., Artal-Sanz,M., Verrijzer,C.P., Plasterk,R.H.A. and Heck,A.J.R. (2003) Metabolic labeling of *C.elegans* and *D.melanogaster* for quantitative proteomics, *Nat. Biotechnol.*, **21**, 927–931.

Li,J., Steen,H. and Gygi,S.P. (2003) Protein profiling with cleavable Isotope-Coded Affinity Tag (cICAT) reagents: the yeast salinity stress response. *Mol. Cell Proteom.*, **2**, 1198–1204.

Lilley,K.S., Razzaq,A. and Dupree,P. (2002) Two-dimensional gel electrophoresis: recent advances in sample preparation, detection and quantitation, *Curr. Opin. Chem. Biol.*, **6**, 46–50.

Lopez,M.F., Berggren,K., Chernokalskaya,E., Lazarev,A., Robinson,M. and Patton,W.F. (2000) A comparison of silver stain and SYPRO Ruby protein gel stain with respect to protein detection in two-dimensional gels and identification by peptide mass profiling. *Electrophoresis*, **21**, 3673–3683.

Malone,J., Radabaugh,M., Leimgruber,R. and Gerstenecker,G. (2001) Practical aspects of fluorescent staining for proteomics applications. *Electrophoresis*, **22**, 919–932.

Munson,P. (2001) A 'consistency' test for determining the significance of gene expression changes on replicate samples and two convenient variance-stabilizing transformations. In Gene-Logic Workshop of Low Level Analysis of Affymetrix GeneChip Data.

Tonge,R., Shaw,J., Middleton,B., Rowlinson,R., Rayner,S., Young,J., Pognan,F., Hawkins,E., Currie,I. and Davison,M. (2001) Validation and development of fluorescence two-dimensional differential gel electrophoresis proteomics technology. *Proteomics*, **1**, 377–396.

Tusher,V.G., Tibshirani,R. and Chu,G. (2001) Significance analysis of microarrays applied to the ionizing radiation response, *Proc. Natl Acad. Sci., USA*, **98**, 5116–5121.

Ünlü,M., Morgan,M.E. and Minden,J.S. (1997) Difference gel electrophoresis: a single gel method for detecting changes in protein extracts. *Electrophoresis*, **18**, 2071–2077.

Yan,J.X., Devenish,A.T., Wait,R., Stone,T., Lewis,S. and Fowler,S. (2002) Fluorescence two-dimensional difference gel electrophoresis and mass spectrometry based proteomic analysis of *Escherichia coli*. *Proteomics*, **2**, 1682–1698.