

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

David P Kreil¹, Natasha A Karp², and Kathryn S Lilley²

1: Department of Genetics / Inference Group (Cavendish Laboratory),
University of Cambridge

2: Department of Biochemistry, University of Cambridge

Correspondence: D P Kreil,
University of Cambridge, Department of Genetics,
Downing Street, Cambridge CB2 3EH, U.K.
Tel: +44 1223 764107, Fax: +44 1223 333992 FAO Kreil

E-mail: D.Kreil@gen.cam.ac.uk
nak23@mole.bio.cam.ac.uk
K.S.Lilley@bioc.cam.ac.uk

Running title: Bias in differential protein expression analysis

Abbreviations: DIGE: Difference Gel Electrophoresis
DIA: Differential In-gel Analysis

Keywords: Expression proteomics, Differential in-gel analysis,
Normalization, System bias, Differential protein expression

ONLINE SUPPLEMENT – Revised 17th of February, 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: This document is the Online Supplement. There is additional data available at <http://www.flychips.org.uk/kreil/pub/2dgels/>.

Online Supplement

Methods

Sample choice and preparation

The gram-negative enteric bacterium *Erwinia carotovora* was chosen as a simple system to provide ample amounts of protein. Although, in general, the complexity of gel images of proteins from higher organism is greater due to the larger variety of isoforms reflecting post-translational modifications, the behaviour of the proteins travelling through the gel is the same, irrespective of their source. Our results therefore equally apply to studies of other organisms. Difference gel electrophoresis is indeed successfully applied to higher organisms (*e.g.*, Somiari et al., 2003). There are certain sample types, however, where a few protein species dominate, for example serum or the secreted fraction of proteins from some systems, and pre-fractionation methods are often used to enrich for lower abundance species. It is important then to also assess the reproducibility of any enrichment methods applied. We wish to point out that the robust normalization method by Huber *et al.* (2002), which was employed for our study, explicitly allows for a reasonably large proportion of ‘outlier’ spots which are excluded from affecting the normalization process. As long as the *number of spots* from the strongly varying fraction of proteins is clearly less than 50% of all spots, the method described in our manuscript can directly be applied.

Wildtype cultures of *E. carotovora* were grown and proteins harvested following standard protocols: Bacterial samples were grown in liquid broth media (10 g/l Bacto Tryptone, 5 g/l Bacto Yeast Extract, 5 g/l sodium chloride) at 30°C and 300 rpm overnight and harvested by centrifugation for 10 min at 4°C at 5000 rpm. Cells were resuspended in lysis buffer (8M urea, 4% w/w CHAPS, 5 mM magnesium acetate, 10 mM Tris pH 8.0 and protease inhibitor cocktail set I at 1x concentration (Calbiochem, Germany) and lysed by sonication (3x10 s pulses on ice). Cell debris was removed by discarding the pellet formed after centrifugation for 10 min at 4°C at 4500 rpm. To harvest the soluble protein fraction the sample was centrifuged at 13 000 rpm for 10 minutes at 4°C and the pellet discarded. The protein concentration was determined using the Bio-Rad DC protein assay as described by the manufacturers (Bio-Rad, UK).

CyDye labelling

Samples were labelled using fluorescent cyanine dyes developed for 2-D DIGE (Amersham Biosciences, UK) following the manufacturer’s recommended protocols. 50 µg of protein were labelled with 400 pmol of amine reactive cyanine dyes, freshly dissolved in anhydrous dimethyl formamide. The labelling reaction was incubated at room temperature in the dark for 30 minutes and the reaction was terminated by addition of 10nmol lysine. Equal volume of 2x sample buffer (7 M urea, 2 M thiourea, 2% amidosulfobetaine-14, 20 mg/ml DTT and 2% Pharmalytes 3-10) were added to each of the labelled protein samples and the two samples were mixed. Rehydration buffer (7 M urea, 2 M thiourea, 2% amidosulfobetaine-14, 2 mg/ml DTT and 1% Pharmalytes 3-10) was added to make up the volume to 250 µl.

2-D Protein separation by 2-D gel electrophoresis

13 cm Immobilised linear pH gradient (IPG) strips, pH 3-10 (Amersham Biosciences, UK) were rehydrated with Cy-labelled samples for 10 hours at 20°C at 20 volts using the IPGphor II apparatus

following manufacturer's instructions (Amersham Biosciences, Sweden). Isoelectric focusing was performed for a total of 41,700 Vh at 20°C at 10 mA. Prior to SDS-PAGE, the strips were each equilibrated for 15 min in 100 mM Tris pH 6.8, 30% glycerol, 8 M urea, 1% SDS, and 0.2 mg/ml bromophenol blue on a rocking table. The strips were loaded onto a 12% 13 cm (1 mm thick) acrylamide gel. The strips were overlaid with 1% agarose in SDS running buffer containing a 5 milligrams of bromophenol blue. The gels were run at 20 mV for 15 minutes and then at 40 mV at 20°C until the bromophenol blue dye front had run off the bottom of the gels. A running buffer of 25 mM Tris pH 8.3, 192 mM glycine, and 0.1% SDS was used.

Gel imaging

Labelled proteins were visualised using a Typhoon 9410 imager (Amersham Biosciences, UK). The Cy2/Cy3/Cy5 images were, respectively, scanned using 488/532/633 nm lasers for excitation and 520/580/670 nm emission filters with a 40/30/30 nm band-pass. All gels were scanned at 100 µm resolution. The gain of the photo-multiplier tube was adjusted to give bright images without saturation. Prior to analysis, images were cropped to remove areas outside the gel using ImageQuant V5.2 (Amersham Biosciences, UK).

For the measurement of background fluorescence, it was ensured that scanner performance was typical by comparison of scans of the same gels performed by another laboratory. Gels were scanned both before and after the gel had been run as an empty gel. Background fluorescence and its variance were considerably lower after the gels had been run, hence only the values measured in these scans were reported. For each empty gel, we obtained the mean and the standard deviation of the image pixel intensities, giving a measure of the average background fluorescence and its spatial variance for each gel. We report the mean and standard deviation of both the average background fluorescence and the spatial variance, calculated from the three independently measured gels to give an indication of typical experimental variation.

Abbreviations

CHAPS	3-[(3-cholamidopropyl)-dimethylamino]-1-propanesulfonate
TRIS	tris-(hydroxymethyl)-aminomethane
DTT	dithiothreitol
SDS	sodium dodecylsulfate
PAGE	polyacrylamide-gel electrophoresis

Results

Deletion of low volume spots removes their dye-specific bias (Fig. S1).

Spots of low molecular weight are not an explanation for the observed bias, they have a similar signal distribution to other spots (Fig. S2).

Figure S3 shows application of the offset/scale normalization to data from a same–same comparison of human samples, demonstrating its efficiency for proteins from higher organisms.

Results involving Cy2

Figure S4 shows application of the offset/scale normalization to other dye combinations (Cy2/Cy5 and Cy3/Cy2). Table S1 shows the result of a cross-validation of empirical score thresholds for empirical per-family error rates of 5% and 10% for Cy2/Cy5.

Additional figure panels

QQ-plots comparing the signal distributions of all six gels are also available at <http://www.flychip.org.uk/kreil/pub/2dgels/>.

Raw data supporting the manuscript

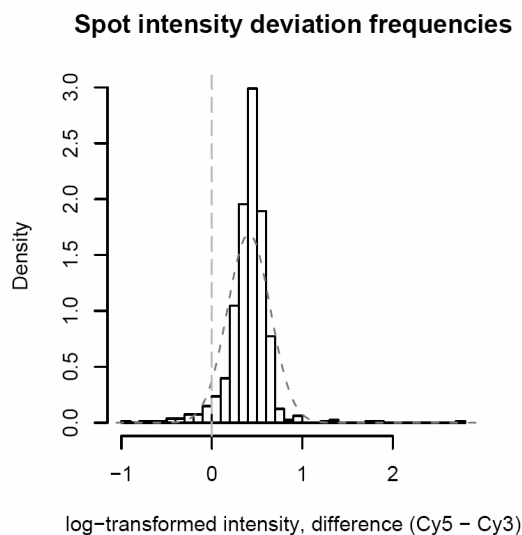
2D gel images and DeCyder output indicating manually flagged spots are available from <http://www.flychip.org.uk/kreil/pub/2dgels/>.

References

Somiari,R.I., Sullivan,A., Russell,S., Somiari,S., Hu,H., Jordan,R., George,A., Katenhusen,R., Buchowiecka,A., Arciero,C., Brzeski,H., Hooke,J. and Shriver,C. (2003) High-throughput proteomic analysis of human infiltrating ductal carcinoma of the breast, *Proteomics*, **3**, 1863–1873.

Figures and Tables

a.



b.

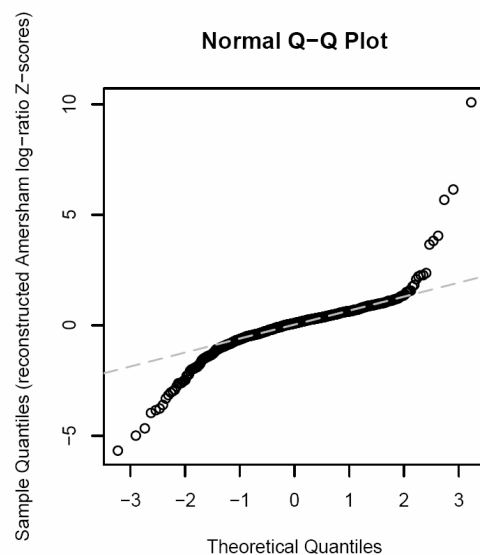


Figure S1: Distribution of difference signals in scale normalized data with spot intensities of 40,000 or higher. *a.* Histogram showing relative frequencies of difference signal values, compared to a best-fit normal distribution (grey dashed curve). The shift of the distribution centre from zero would be reset by re-normalization in DeCyder. *b.* Quantile–quantile plot. The grey dashed line corresponds to identity.

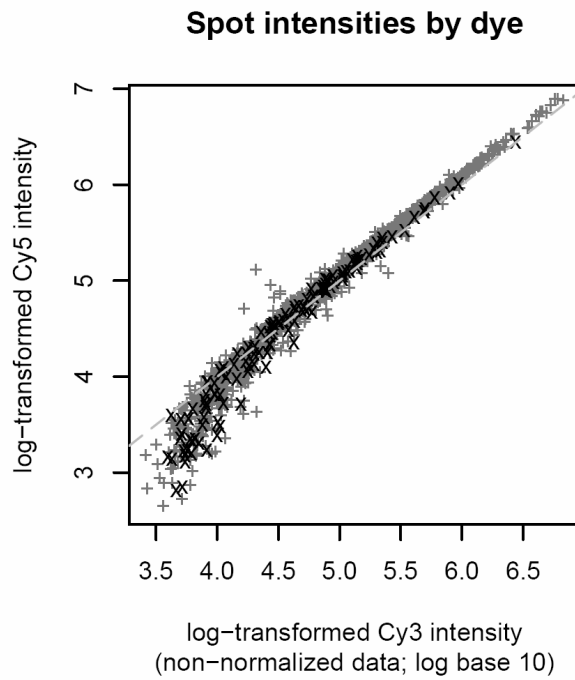


Figure S2: Demonstration that expression levels of proteins of low and high molecular weights show similar distributions, equally affected by dye-specific bias. The 10% of proteins with the lowest molecular weights are represented by black 'x' symbols, the others by grey '+' symbols.

a.

b.

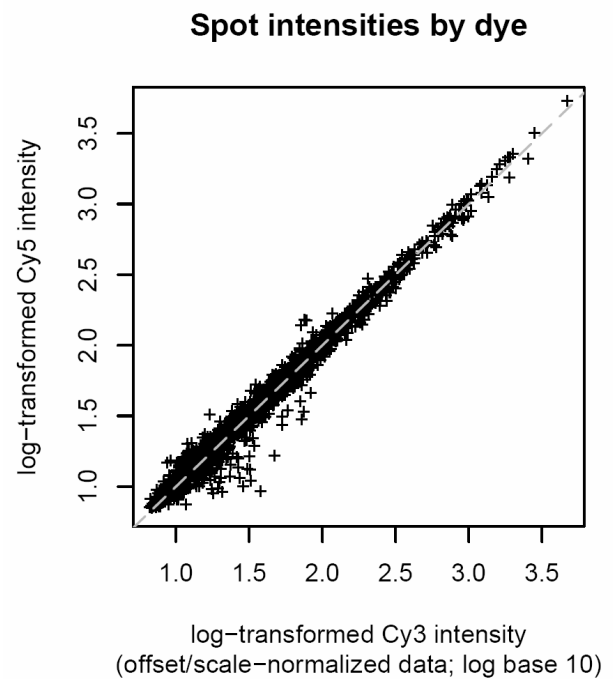
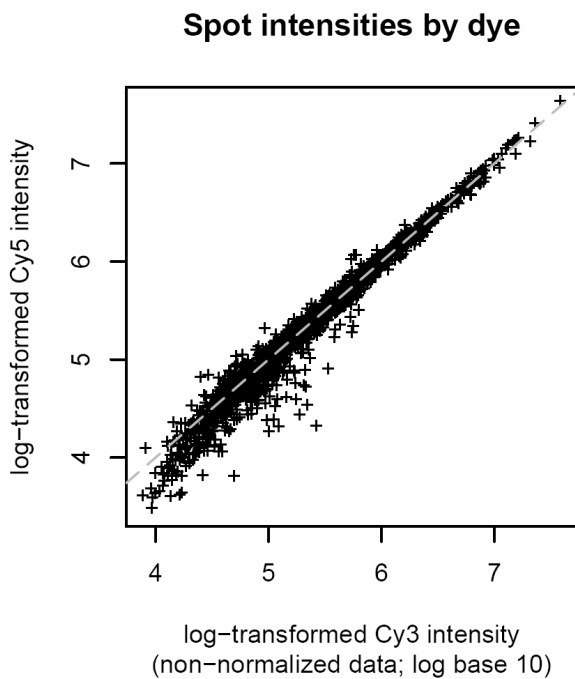


Figure S3: Application to gels of human proteins. *a.* Non-normalized data. *b.* Data after offset/scale normalization. See also legend to Fig. 2.

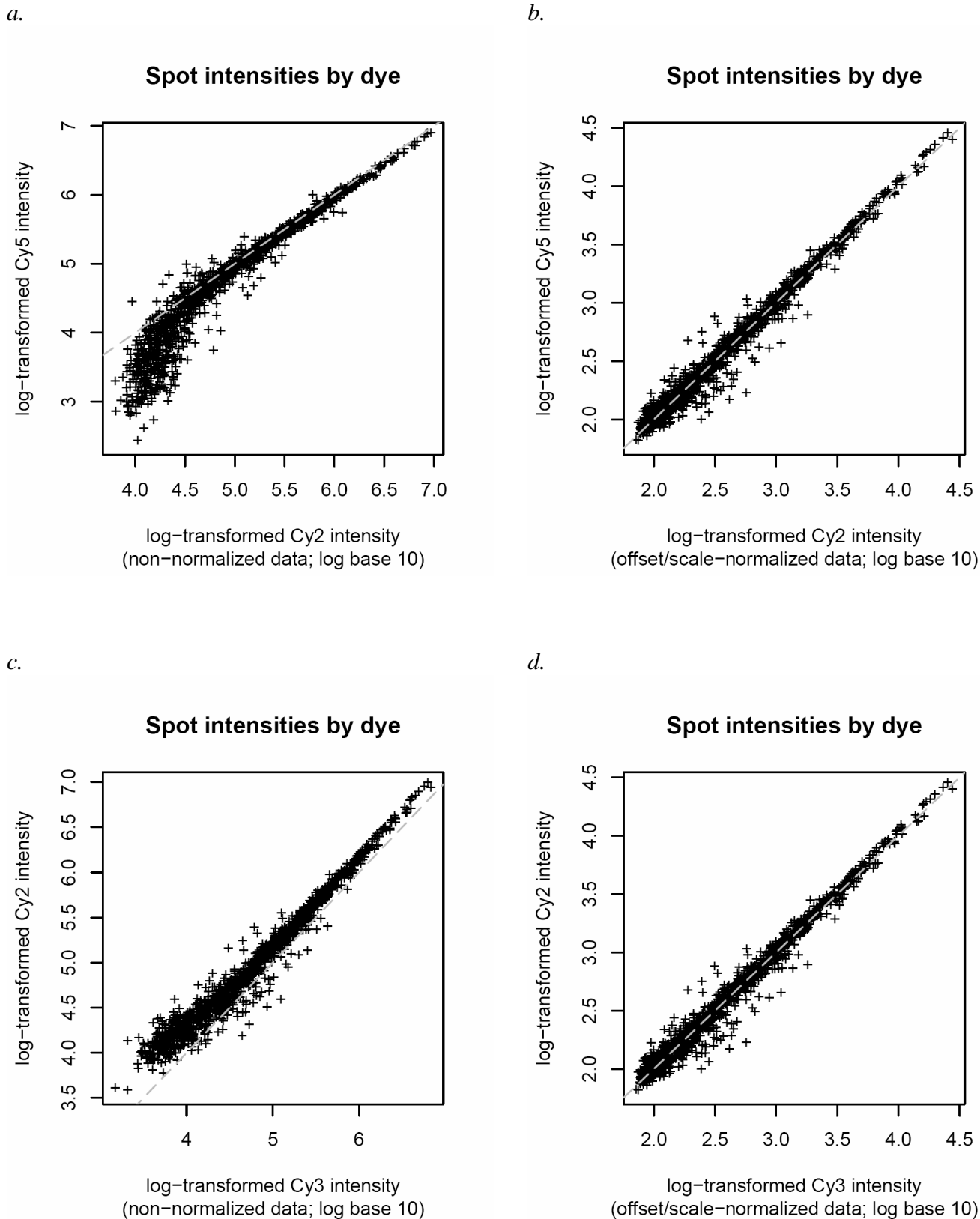


Figure S4: Application to other dye combinations. *a/b.* Cy2 vs Cy5. *a/c.* Cy3 vs Cy2. *a/c.* Non-normalized data. *b/d.* Data after offset/scale normalization. See also legend to Fig. 2.

Table S1: Cross-validation results for Cy2 / Cy5, shown for two different empirical per-family error rates.

Gel excluded	$p^* = 5\%$			$p^* = 10\%$		
	Thresholds from pool		False positives [%]	Thresholds from pool		False positives [%]
	Lower	Upper		Lower	Upper	
1	-2.75	2.34	4.4	-1.96	1.85	8.3
2	-2.60	2.30	6.2	-1.94	1.78	11.8
3	-2.78	2.34	3.3	-2.01	1.82	7.6
4	-2.77	2.29	4.9	-1.97	1.77	11.2
5	-2.32	2.51	7.5	-1.72	1.92	12.6
6	-2.80	2.29	4.0	-1.99	1.79	9.4
mean \pm SD	-2.67\pm0.19	2.35\pm0.08	5.1\pm1.6	-1.93\pm0.11	1.82\pm0.06	10.1\pm2.0