Clinically Validated Benchmarking of Normalisation Techniques for Two-Colour Oligonucleotide Spotted Microarray Slides

Jennifer Listgarten,[1,3] Kathryn Graham,[1,2] Sambasivarao Damaraju,[1,2] Carol Cass,[1,2] John Mackey,[1,2] Brent Zanke[4]

[1]PolyomX Program, Cross Cancer Institute, University of Alberta, Alberta Cancer Board, Edmonton, Alberta, Canada; [2]Department of Oncology, University of Alberta, Alberta Cancer Board, Edmonton, Alberta, Canada; [3]Department of Computer Science, Toronto, Ontario, Canada; [4]Cancer Care Ontario, Toronto, Ontario, Canada

Abstract: Acquisition of microarray data is prone to systematic errors. A correction, called normalisation, must be applied to the data before further analysis is performed. With many normalisation techniques published and in use, the best way of executing this correction remains an open question. In this study, a variety of single-slide normalisation techniques, and different parameter settings for these techniques, were compared over many replicated microarray experiments. Different normalisation techniques were assessed through the distribution of the standard deviation of replicates from one biological sample across different slides. It is shown that local normalisation outperformed global normalisation and that intensity-based 'lowess' outperformed trimmed mean and median normalisation techniques. Overall, the top performing normalisation technique was a print-tip-based lowess with zero robust iterations. Lastly, we validated this evaluation methodology by examining the ability to predict oestrogen receptor-positive and -negative breast cancer samples with data that had been normalised using different techniques.

Keywords: microarray, normalisation, oligonucleotide, oestrogen receptor, validation

Introduction

Oligonucleotide and cDNA microarray technologies allow simultaneous measurement of thousands of gene transcripts from a single sample. This high-throughput technology allows for large-scale screening of individual genes implicated in various biological conditions (Botwell and Sambrook 2003). After analysing medically relevant samples, statistical models of gene expression can be built, which can then be used for subsequent medical diagnosis or prognosis (van't Veer et al 2002). A short probe of cDNA or oligonucleotide for each gene is spotted onto a glass slide. Typically, the RNA sample of interest is converted to cDNA and labelled with one fluorescent dye, while a reference sample is labelled with a different fluorescent dye, and the two are mixed and hybridised to the microarray slide. The measurement of interest is then the ratio of the fluorescence level of the sample relative to that of the reference. Use of a reference channel helps to control for sources of experimental variation such as spot size and probe quality. The reference cDNA may make biological sense, as is seen where two closely related cell lines are hybridised to a slide, or individual patient cancer samples are hybridised with a pool of all patient cancer samples used as the reference. Alternatively, the reference RNA may be an ad hoc mixture of RNAs chosen to maximise spot hybridisation. In this study, we have used the latter technique.

Microarray data are prone to systematic experimental errors that introduce a signal imbalance between the two data channels. This error results from variation in RNA quality, efficiency of dye labelling, photobleaching and physical properties of the scanning in each of the two channels, which in turn is complicated by varying scanning parameters. Two-channel microarray normalisation, a critical step to further analysis (Zien et al 2001; Bilban et al 2002; Tahi et al 2002; Yang et al 2002), corrects these errors. Many normalisation techniques for microarray data now exist (Tseng et al 2001; Zien et al 2001; Bilban et al 2002; Kroll and Wölfl 2002; Yang et al 2002; Bolstad et al 2003), although there is no known optimal choice. There are no standard methods of evaluating normalisation strategies; this impedes objective proof of validity across published techniques. In this paper we systematically compared a variety of single-slide, global and local, intensity-based and non-intensity-based normalisation techniques for two-channel oligonucleotide microarray data. While the utility of a particular normalisation technique is dependent on the type of slides used and the nature of experiment conducted, a quantitative benchmarking exercise of this nature has not been reported previously in the microarray literature.

Evaluation of microarray normalisation techniques has typically only been useful for pairwise comparisons of microarray slides (eg using correlation between slides hybridised with the same sample (Schuchhardt et al 2000; Tahi et al 2002)). Furthermore, correlation measures, such as Pearson correlation, do not allow assessment of absolute agreement between datasets; they only allow measurement of the degree to which two datasets are related in a linear fashion. Most techniques assume that the data should appear in a certain way (eg with a horizontal best-fit line in a Ratio-Intensity plot) and then judge the merits of a particular normalisation technique by plotting the data after normalisation to determine if the data satisfy the original assumptions. Kroll and Wölfl (2002) introduced Rank Intensity Plots (RIPs) to evaluate different normalisation techniques. Similar to

histogram matching, RIPs are useful as a visual, exploratory tool rather than a quantitative evaluation technique. The main concern in using it for a rigorous evaluation of normalisation is that it matches distributions of intensity ratios. It compares the *distribution of ratios* on one slide with that of another, or several others, and does not compare *individual repeated gene measurements* with each other. Thus it assumes that different RNA samples should have identical overall expression profiles and that matching a given gene to itself is not critical. Such evaluation of normalisation techniques, while useful, should not be used as an objective, final method of evaluation. Very recently, Park et al (2003) performed a similar comparison to the one reported here. As in the present study, they used variation across replicates to assess different normalisation techniques (global, local, intensity-dependent: linear and non-linear, scaling), visualising the results with dot plots. Variation was assessed in two ways: (1) a pooled variance estimator and (2) analysis of variance. They also conducted simulation studies in an effort to examine the bias and mean squared error in addition to variance. We discuss their results in the Discussion and Conclusion.

However, their study differs from the present study in that it did not examine the effects of different parameter settings for each given technique and it did not examine the effects of normalisation on downstream analysis. This latter step is a crucial validation test of evaluation of normalisation techniques on the basis of variation over replicates.

Evaluation of microarray normalisation should have the following properties:

1. It should be bias free and make no underlying assumptions about the biology or experimental setup. The evaluation method must not make the same assumptions as the normalisation technique. Many normalisation techniques assume that overall expression distributions of different samples are identical or, at least, highly similar. Most force all ratio distributions to have a centre (eg mean) of zero in logarithmic scale. If the evaluation of such normalisation techniques is a measure of how close all distribution centres are to zero, then the evaluation is simply determining how well the algorithm forced the data to adhere to its possibly incorrect assumptions.
2. It should not require any external measurements, such as Northern blot analysis or quantitative RT-PCR, since not all genes can feasibly be measured in this way.

In this work we satisfied these requirements for evaluation by using the standard deviation of repeated measurements over one gene in one biological sample hybridised to three different slides. This standard deviation was calculated for each unique spot on a slide, defined by block, column and row, across repeated measurements using aliquots from the same sample. We repeated this experiment in multiple samples, each with multiple aliquots. The resulting distribution of standard deviations of replicate measurements was examined. The smaller the overall standard deviations of replicates, the better the normalisation was taken to be.

Materials and methods

Microarray hybridisation

Microarray slides of the Operon (Alamadea, CA, USA) Human 70-mer oligonucleotide set, version 1.1, representing 13 971 genes, were printed by the Gene Array Facility of Genome

British Columbia. The oligos were printed in duplicate on each ArrayIt SuperAmine slide (TeleChem, Sunnyvale, CA, USA) using a Microgrid TAS 2 (Biorobotics, Woburn, MA, USA) array printer.

Tumour samples were collected from patients undergoing surgery for primary breast cancer. Patients prospectively provided written informed consent for tissue banking and analysis plans that were approved by the local Institutional Review Board. To ensure RNA integrity, the time from devitalisation to storage of the sample in liquid nitrogen did not exceed 20 min. Total RNA was isolated from 25 samples using Trizol followed by purification on an RNeasy column (Qiagen, Mississauga, Ontario, Canada), according to the manufacturer's recommendations. Human total RNA prepared from ten human tumour cell lines (Stratagene, La Jolla, CA, USA) was used as a reference sample.

Microarray slides were probed in triplicate with labelled cDNA prepared from 30 μg each of tumour and reference total RNA. Superscript II was used to prepare cDNA, which was then labelled with Cy3 or Cy5 using an indirect amino allyl technique (Botwell and Sambrook 2003). After hybridisation, the microarray slides were scanned with an Axon 4000B using GenePix 3.0 software.

Processing and evaluation methods

The benchmarking dataset was derived from 75 oligonucleotide microarray slides containing 28 704 spots; genes were in duplicate, side-by-side. GenePix gpr (tab-delimited text files) files for this dataset are available at http://www.cs.toronto.edu/~jenn/normalizationStudy/normalization.htm. Each slide had 48 blocks with 23 rows and 26 columns. All spots in a given block were printed using the same print-tip. RNA from 25 breast tumour samples, each arrayed in triplicate, was used in this study. Thus, for each patient/spot combination, a maximum of three replicate measurements were available. For the purposes of this experiment, duplicate gene measurements on a single slide were treated separately since the focus of the study was to assess inter-slide variability as a result of normalisation. Thus spots on a slide were uniquely identified by block, column and row. Replicates were considered to be spots from the same sample (with the same block, column and row) hybridised on *different* slides.

After hybridisation and scanning, image processing was performed in GenePixPro 3.0. Background estimation used local measurements surrounding each spot (GenePixPro 3.0 User's Manual). Median foreground and background intensities were exported for each spot for each channel and their difference treated as the channel intensity. Saturation levels were also exported. Spots on each slide were filtered on the basis of the following criteria: (1) GenePixPro flag was greater than zero (ie the spot-finding algorithm in GenePix found a spot with diameter greater than 50 μm and less than 300 μm and composite intensity greater than zero); (2) saturation in either channel was less than five percent; (3) Cy5 (red) background subtracted intensity was less than 70; and (4) Cy3 (green) background subtracted intensity was less than 60. This left an average of 5 416 spots per slide (this is typical for the slides being used). Next, the ratio measurements, $y_{sni} = r_{sni}/g_{sni}$ (where s = 1,…, 25 denotes the biological sample, n = 1,…,3 denotes the slide replicate number of the sample, and i = 1,…,28 704 denotes a unique three-tuple of block, column and row) on each slide were normalised, one slide at a time (one block at a time for local methods). All normalisation was performed inside a custom database set up for and by the PolyomX

project (www.polyomx.org), using Perl, MySQL and R Statistical Language. Post-processing was done in Matlab™.

Evaluation of different normalisation techniques used the standard deviation of replicates from one biological sample across either two or three different slides (if a spot was only present on one slide, then the spot was ignored). Thus, for one gene, *i*, and for one biological sample, *s*, the standard deviation is:

$$\sigma_{si} = \sqrt{\frac{\sum_{n=1}^{3} (v_{sni} - \bar{v})^2}{(n-1)}} \qquad (1)$$

where $v_{sni} = \log_2(y_{sni})$, and the sum is only over spots that pass the filters. A given normalisation technique was assessed by examining the distribution of the log-transformed σ_{si} over all samples, *s*, and spots, *i*. The log transformation was performed to improve visualisation of the results. We call this distribution the 'distribution of standard deviation errors', or the DSDE. The DSDE is parameterised by the normalisation technique, *T*. Thus, DSDE = DSDE(T_j), where *j* denotes the different normalisation techniques. For the purposes of comparison, we display the DSDE using a histogram of the σ_{si} where the counts are converted to percentages (for example, Figure 1). If one DSDE(T_1) lies systematically to the left of another DSDE(T_2) (ie it is shifted by a negative amount), we say that normalisation technique, T_1 is better than technique T_2. Since this strict requirement is often not completely satisfied, we are sometimes more lenient in the ranking of different normalisation techniques.

While use of only three observations per calculation may cause concern about instability, the validity of these calculations is supported by the second part of this study where prediction accuracy is examined. Agreement between these completely different assessments of normalisation leads us to believe that the standard deviations computed are sufficient and representative of the true underlying differences.

Note that it is common in the microarray community to use the coefficient of variation (the standard deviation divided by the mean, denoted by CV) as a measure of reproducibility. However, the CV can only be applied to single channel data: two-colour ratios should be logged, and the log of a ratio equal to one is zero. The CV for these cases would be undefined. Thus, we here use the standard deviation. The assumption behind use of the CV is that the standard deviation scales linearly with the mean. Whether or not this is the case for microarray data and, in particular, the data in this study, is irrelevant since we are looking at the distribution of standard deviations across *all spots* on a slide, regardless of their mean. Since each normalisation technique is using exactly the same dataset, this should not bias the comparison.

A second assessment was done on a subset of the normalisation techniques. The microarray data, which were derived from breast cancer tumour samples, were used to build a predictive model for oestrogen receptor (OR) status. Each normalisation technique applied to slides in this study led to a different dataset. Normalisation techniques were ranked according to how high their respective dataset's predictive accuracy was using the Nearest Shrunken Centroid (NSC) model (Tibshirani et al 2002).

Normalisation techniques studied

Because the reference RNA used in this study was pooled from ten different cell lines, we did not expect the true average ratio to lie near one. We did not seek the absolute expression ratios on individual slides, but instead sought the relative expression ratios on different slides.

We chose normalisation techniques that could be applied to any one slide, that do not require any specific internal or external controls, set of genes, matched dye-reversals or pre-selected set of hybridised slides.

Each normalisation technique was applied in a global manner, ie over the whole slide, as well as in a local manner, ie block by block (Yang et al 2002). We implemented a trimmed mean normalisation that set the trimmed mean log intensity ratio to zero for trim = 0, 0.05, 0.10, 0.30. A mean with a trim of 0.05 is the mean of all values after having discarded the smallest and largest five percent. We also used a median normalisation where the median log intensity ratio was set to zero. The commonly used intensity-based, 'lowess' (locally weighted regression) normalisation (Yang et al 2002) using a variety of parameter settings was also implemented. Lowess normalises different intensity ranges separately, but in a similar way to mean normalisation, and in such a way that the transition is smooth as one varies from one intensity range to another. Two parameters for the lowess were varied:

1. The smoothing fraction, *S*, which determines how local the linear regression is. For example, if $S = 0.4$, a typical setting for microarray normalisation, then each point on the regression curve is calculated using only the 40 percent of points that lie closest to it. We used $S = 0.1, 0.4, 0.7$ and 0.9.
2. The number of robust iterations, *R*, in which outliers are discarded from the regression set, as, for example, one can set as a parameter in the Statistical Language R function 'lowess'. We set $R = 0$ and 20.

Finally, we tried to improve upon all of the aforementioned techniques by scaling the data distributions. This was accomplished by dividing every log intensity ratio on a given slide (block for local normalisation) by the median absolute deviation (MAD – a robust standard deviation) of logged intensity ratios on that slide (block for local normalisation), inspired by Yang et al (2002). Scaling was performed as a post-processing step after the main normalisation technique had been applied. For slides hybridised with the same biological sample, it intuitively makes sense that their distributions should be scaled to the same width since they should have identical values. However, it is difficult to know the appropriate scaling just as it is difficult to choose the value to shift the distributions to (eg shifting the mean to one). Since the standard deviations on all slides were close in value to one before normalization, we chose, in an admittedly ad hoc manner, to drive the standard deviation of each slide to one.

Other normalisation techniques are being rapidly published. We feel that the techniques listed above are representative of the core ideas in the area of normalisation: (1) global versus local (accounting for spatial biases); (2) intensity-dependent normalisation, or not; and (3) scaling of data. Newer techniques are typically modifications or extensions of these ideas. Park et al (2003) provide a list of some of the recent extensions.

Results and discussion

Comparison of mean-like normalisation techniques

Figure 1 shows a comparison of global normalisation techniques with different trim values, global median, global lowess based normalisation and no normalisation. Performing no normalisation is clearly inferior as we would expect. Lowess is superior to the mean and median techniques. The mean, with different trim, and the median perform extremely similarly, and it is difficult to choose a clear ordering of these. Upon closer examination, one may venture to say that the median outperforms all trimmed means, and that the larger the trim, the better the normalisation. An analogous comparison of the local techniques reveals the same pattern (data not shown).

Comparison of intensity-based lowess normalisation techniques

Figures 2 and 3, respectively, show a comparison of global and local normalisation techniques. Each figure shows lowess normalisation with different values of the smoothing parameter, *S*, different values of the number of robust iterations, *R*, no normalisation and median normalization. Of the global techniques, we observe that median normalisation performs only better than no normalisation. The different lowess normalisation techniques all perform extremely similarly, and it is difficult to choose a clear ordering of these. Upon close examination, use of twenty robust iterations appears to be better than zero, and a larger smoothing parameter appears better than a smaller one. Figure 3 shows that performing no normalisation is the worst performer. The other techniques are very difficult to differentiate from each other, though median normalisation appears to perform more poorly than the lowess. The different lowess normalisations are intertwined, but the two with $S = 0.1$ are worse than the others, and between these two, use of zero robust iterations appears possibly to be worse than use of twenty iterations.

Global versus local normalisation

Figure 4 shows a comparison of global versus local for median normalisation and two lowess normalisations. Both lowess normalisations outperform all other techniques, with the local lowess performing better than the global lowess. Similarly, the local median performs better than the global median. However, these results should be interpreted with caution: as a given normalisation technique, say median normalisation, progresses from using all spots on the slide (called global in this paper), to using fewer and fewer spots at a time, (called local in this paper), the DSDE analysis may become less trustworthy. Taken to the extreme limit of using only one spot at a time for normalisation (as opposed to one slide, or one print-tip block), local normalisation would, in this case, force every gene ratio to be one. Thus, the standard deviation would be zero, and the normalisation perfect according to this assessment. In the present study, an average of 86 spots per block, out of a possible 598, were used for normalisation. Thus it seems unlikely that such a pathological situation exists. Furthermore, these results closely match the results in the next section, where prediction accuracy is examined. Were a normalisation scheme to overfit the data (ie force the data to adhere to some fixed, incorrect pattern), no discriminative power would be present in the resulting dataset, and such a normalisation scheme would appear worse; this does not occur as the reader will shortly see. In particular, the normalisation schemes most

prone to overfitting (ie local and intensity-dependent techniques) end up having superior classification accuracies.

The overall results presented here closely match those results reported by Park et al (2003), who found that intensity-dependent normalisation often performed better than global normalisation, and that linear and non-linear methods performed similarly (which is somewhat akin to our changing the smoothing parameter for the lowess method). They did not examine the sensitivity of the measured reproducibility, to changes in parameters for each of their techniques. Next we examine the effects of scaling, and the effects of normalisation on downstream analysis.

Effect of scaling

Figure 5 shows a comparison of global mean and a global lowess normalisation performed with and without scaling as a post-processing step. In all cases, scaling the data clearly had an adverse effect on the reproducibility of ratio measurements. In fact, scaling was worse than not performing any normalisation. An analogous comparison for local methods reveals the same pattern (data not shown).

Effect of normalisation on downstream analysis

Without a gold standard by which to assess each and every gene, it is very difficult to properly assess the value of any normalisation technique. Measures of variation ultimately confer only precision, not accuracy. Since the goal of many microarray studies is to make sense of gene patterns in different datasets, we took our normalisation study one step further by analysing our dataset, which comprises breast cancer samples, for prediction of OR status, a known, dominant signal in microarray data (Gruvberger et al 2001; West et al 2001; van't Veer et al 2002). By using the same dataset, but varying the normalisation technique applied to all slides in the dataset, we hoped to shed further light on which normalisation technique resulted in the most accurate representation of the dataset. Thus, for a subset of the normalisation techniques studied, we used the nearest shrunken centroid (Tibshirani et al 2002) method to build a predictive model of OR status. The geometric mean of replicates within a single slide was used as the expression ratio for that gene. The log of this mean ratio was used as a feature in the predictive model. Each slide was considered to be a unique sample (though there were three repeat samples per patient, each with the same OR status). Any genes that were not present in at least 40% of tumour samples were removed from analysis; this left 2295 distinct genes.

Nearest shrunken centroid, a supervised analytical technique, is particularly suitable for microarray data because it naturally handles missing and noisy data. A known class category (such as oestrogen receptor positive/negative), along with labelled instances of the data (eg microarray data for individual patients along with corresponding OR status), are used to build a predictive model of the data for the specified class. The model is then validated, in the present case, using leave-one-out cross-validation (where one sample at a time is left out for testing and the remainder of the data used to train, until every sample has been left out). The nearest centroid method is a simple, classical technique where the multivariate mean of each class is calculated, after the variables have been standardised by their pooled within-class standard deviations. A new sample is classified according to which centroid its standardised variables lies closest to. The nearest shrunken centroid

method is a modification of this: individual components of centroids that lie close (within some threshold) to the overall centroid for that component are shrunken to match the mean, and hence play no discriminatory role. Features further away (above some threshold) are all shrunken toward the overall mean by the same amount. This has the effect of reducing the role of noisy genes and embedding feature selection into the algorithm in a very natural way. By performing varying amounts of shrinkage, all datasets, except the one that was not normalised, were able to achieve 100% cross-validation accuracy (data not shown). Thus, for the purposes of assessing normalisation, we omitted the shrinkage component because we did not want the algorithm to be so robust as to overcome the deficits of any given normalisation technique; thus, we reverted to the simple nearest centroid method.

Table 1 shows the results of using the NSC technique for prediction of ER status. With no normalisation, the predictive accuracy was only 68%. With use of any normalisation technique the accuracy immediately jumped to 88%. The ranking of normalisation techniques here according to predictive accuracy closely mirrors the coarse findings in the previous section. In particular, the local techniques marginally outperform the global techniques, and the lowess outperforms both the mean and median techniques. For the global and local lowess, performing zero robust iterations marginally improved predictive performance over 20 robust iterations. A smoothing parameter, $S = 0.7$ marginally outperforms S = 0.4. Contrary to our other observations, scaling seems to have little effect on the predictive accuracy.

A summary of how the normalisation and cross-validation were performed together, as well as a brief discussion of how cross-validation would change in the case of multi-slide normalisation, appear in the Appendix. Because each slide was normalised independently, it is not problematic having two replicates for a sample used for training, while one is being tested. However, out of curiosity, we also tried threefold cross-validation, where each fold consisted of three replicates from one biological sample (data not shown). Accuracies were a few percent lower on the whole. The relative ordering induced on the normalisation techniques was almost unchanged, with only one difference: 'Global, Lowess, R = 20, S = 0.7, Scale = 0' was one percent higher than 'Global, Lowess, R = 0, S = 0.7, Scale = 0' and 'Global, Lowess, R = 0, S = 0.4, Scale = 0', which were tied with each other. In the earlier results, this former was one percent lower than these two latter, which were also previously tied with each other.

We have systematically compared a variety of global and local, two-channel microarray normalisation techniques. Adjusting particular algorithm parameters, such as the trim value for mean normalisation, or the smoothing parameter for lowess, had little effect on the results. Overall, the top performing normalisation technique was a local (print-tip) based lowess. We validated these results based on variation over replicate experiments, through examination of the changes in predictive accuracy for estrogen receptor status in breast cancer samples, for different normalisation techniques.

Acknowledgements

The authors thank Jenny Bryan and Jochen Brumm for useful discussions, comments and suggestions. They also thank Lillian Cook, Jennifer Dufour, Sherry Purdue and Cheryl Santos for excellent technical support, and Adrian Driga for his database expertise. This study was supported by the Alberta Cancer Board, the Alberta Cancer Foundation and Alberta Health and Wellness.

References

Bilban M, Buehler L, Head S, Desoye G, Quaranta V. 2002. Normalizing DNA microarray data. *Curr Issues Mol Biol*, 4:57–64.

Bolstad BM, Irizarry RA, Astrand M, Speed TP. 2003. A comparison of normalization methods for high density oligonucleotide array data based on bias and variance. *Bioinformatics*, 19:85–193.

Botwell D, Sambrook J, eds. 2003. DNA microarrays: a cloning manual. Cold Spring Harbour, NY: Cold Spring Harbour Laboratory Pr.

Gruvberger S, Ringner M, Chen Y, Panavally S, Saal LH, Borg A, Ferno M, Peterson C, Meltzer PS. 2001. Estrogen receptor status in breast cancer is associated with remarkably distinct gene expression patterns. *Cancer Res*, 15:5979–84.

Kroll T, Wölfl S. 2002. Ranking: a closer look on globalization methods for normalization of gene expression arrays. *Nucleic Acids Res*, 30(11):E50.

Park T, Yi SG, Kang SH, Lee S, Lee YS, Simon R. 2003. Evaluation of normalization methods for microarray data. *BMC Bioinformatics*, 4:33.

Schuchhardt J, Beule D, Malik A, Wolski E, Eickhoff H, Lehrach H, Herzel H. 2000. Normalization strategies for cDNA microarrays. *Nucleic Acids Res*, 28(10):E47.

Tahi F, Achddou B, Decraene C, Alibert O, Guiot H, Auffray C, PiétuG. 2002. Automatic quantitation of hybridization signals on cDNA arrays. *BioTechniques*, 32:1386–97.

Tibshirani R, Hastie T, Narasimhan B, Chu G. 2002. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc Natl Acad Sci USA*, 99:6567–72.

Tseng G, Oh M, Rohlin L, Liao J, Wong W. 2001. Issues in cDNA microrarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Res*, 29:2549–57.

van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT et al. 2002. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415:530–6.

West M, Blanchette C, Dressman H, Huang E, Ishida S, Spang R, Zuzan H, Olson JA Jr, Marks JR, Nevins J. 2001. Predicting the clinical status of human breast cancer by using gene expression profiles. *Proc Natl Acad Sci USA*, 98:11462–7.

Yang Y, Dudoit S, Luu P, Lin D, Peng V, Ngai J, Speed T. 2002. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res*, 30(4):E15.

Zien A, Aigner T, Zimmer R, Lengauer T. 2001. Centralization: a new method for the normalization of gene expression data. *Bioinformatics*, 17 Suppl 1:S323–31.

Table 1 Comparison of Oestrogen Receptor Predictive Accuracy

Normalization Technique	Predictive Accuracy
Local, Lowess, R=0, S=0.7, Scale=0	100%
Local, Lowess, R=20, S=0.7, Scale=0	99%
Global, Lowess, R=0, S=0.4, Scale=1	99%
Global, Lowess, R=0, S=0.4, Scale=0	97%
Global, Lowess, R=0, S=0.7, Scale=0	97%
Global, Lowess, R=20, S=0.7, Scale=0	96%
Local, Mean, Trim=0.00, Scale=0	89%
Local, Median, Scale=0	89%
Global, Mean, Trim=0.00, Scale=0	88%
Global, Median, Scale=0	88%
No Normalization	68%

Figures 1–5 Comparison of different normalization techniques using the DSDE. Better techniques are those shifted more to the left.

Figure 1 Comparison of Global, Mean-Like Normalization Techniques

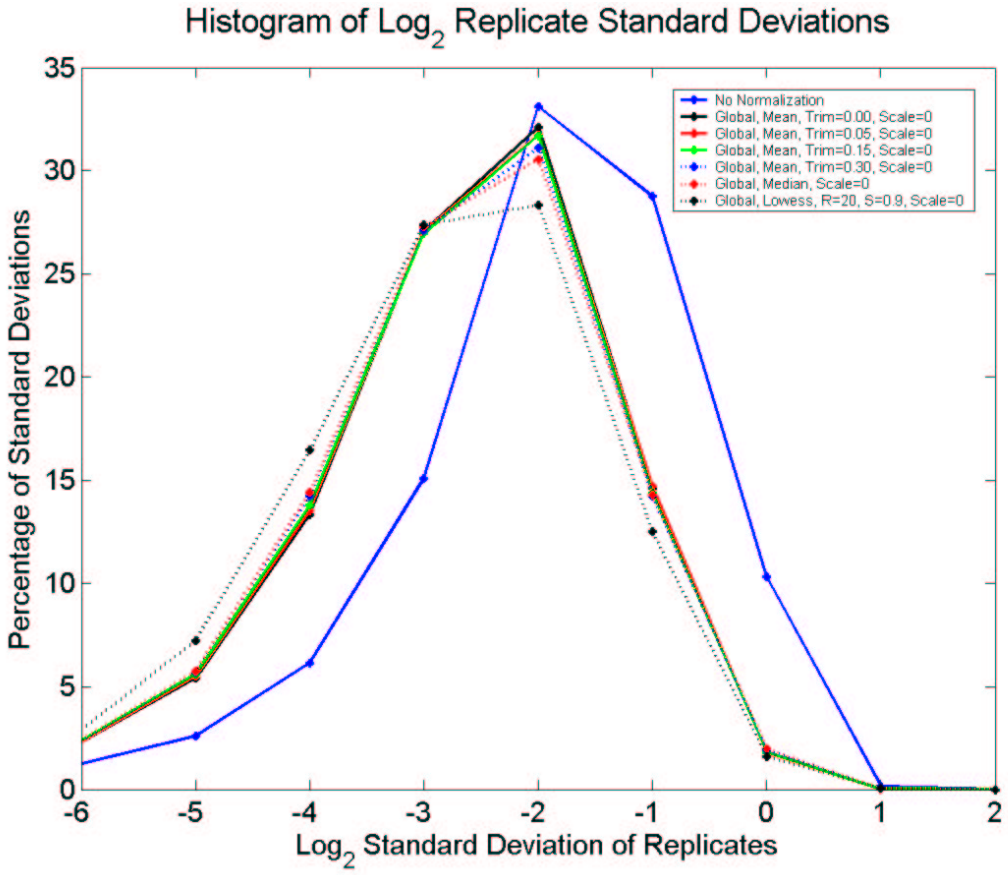


Figure 2 Comparison of Global, Lowess Normalization Techniques

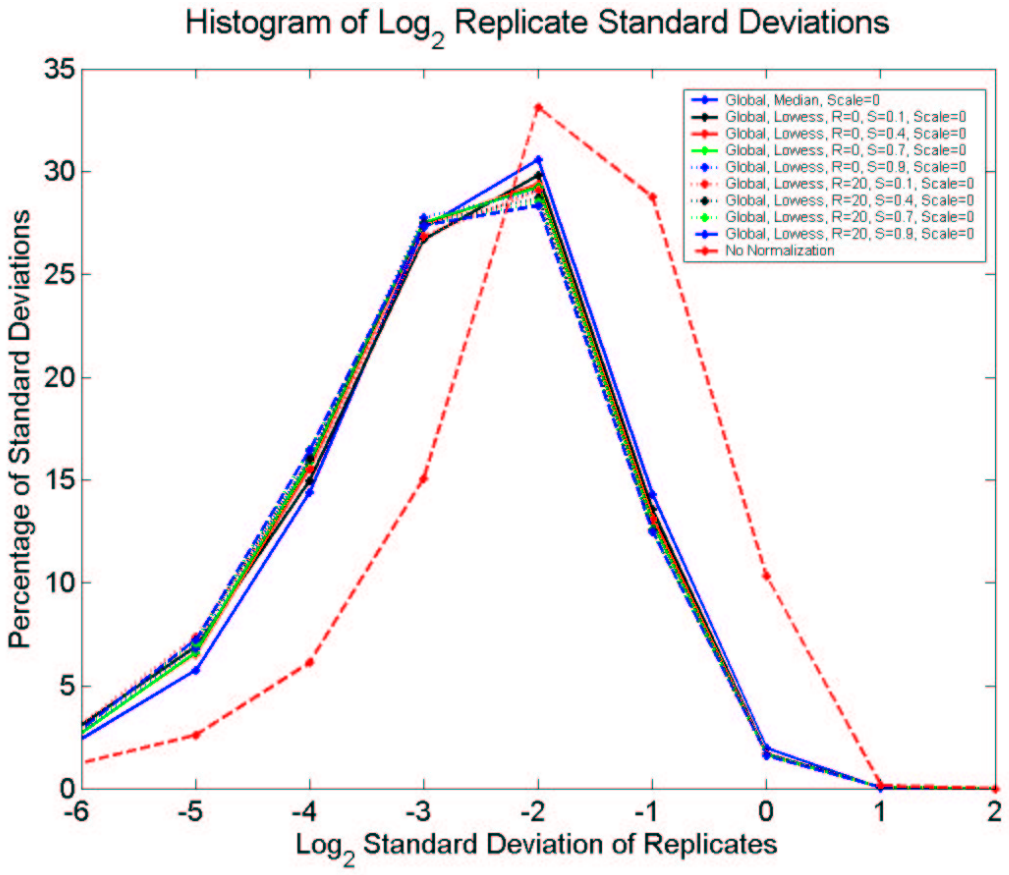


Figure 3 Comparison of Local, Lowess Normalization Techniques

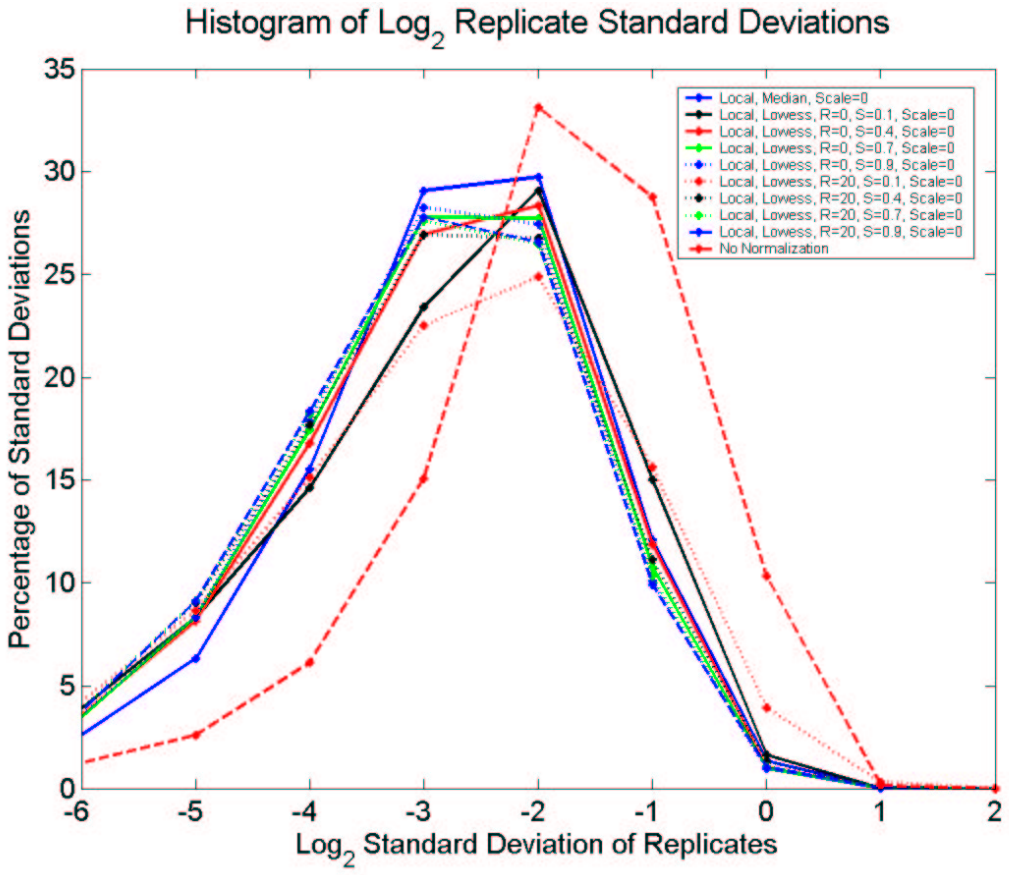


Figure 4 Comparison of Local Versus Global Normalization Techniques

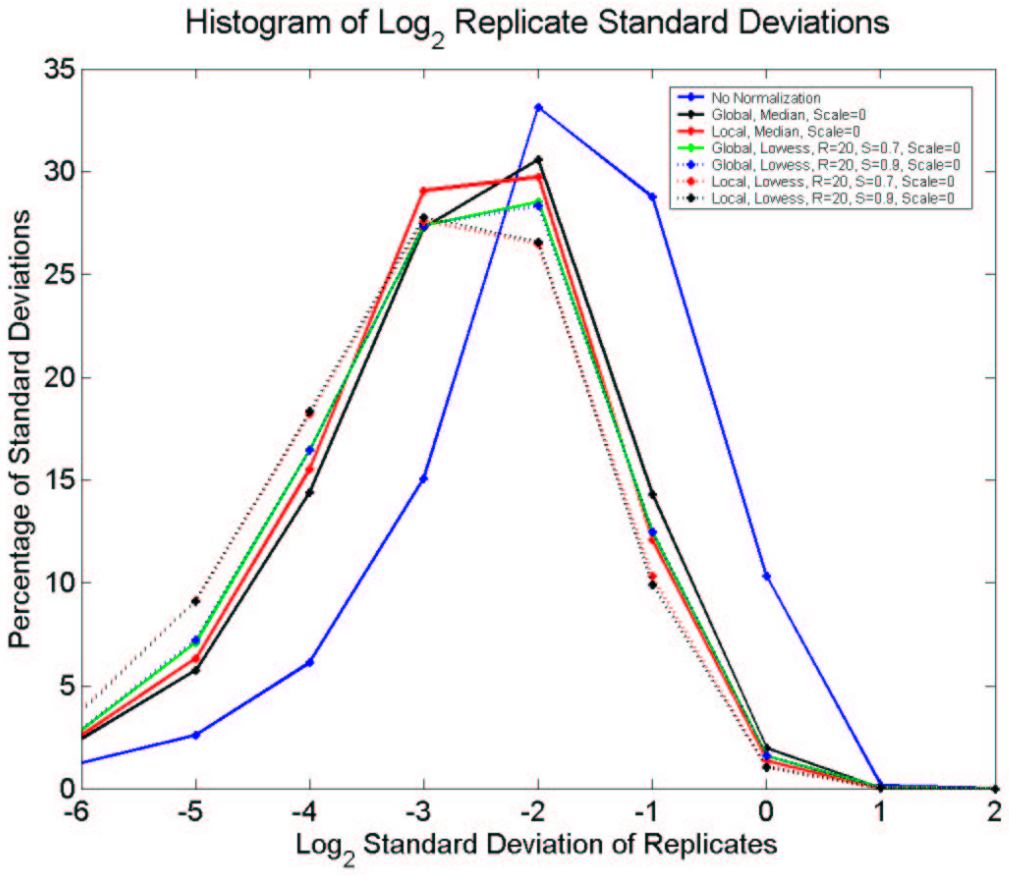
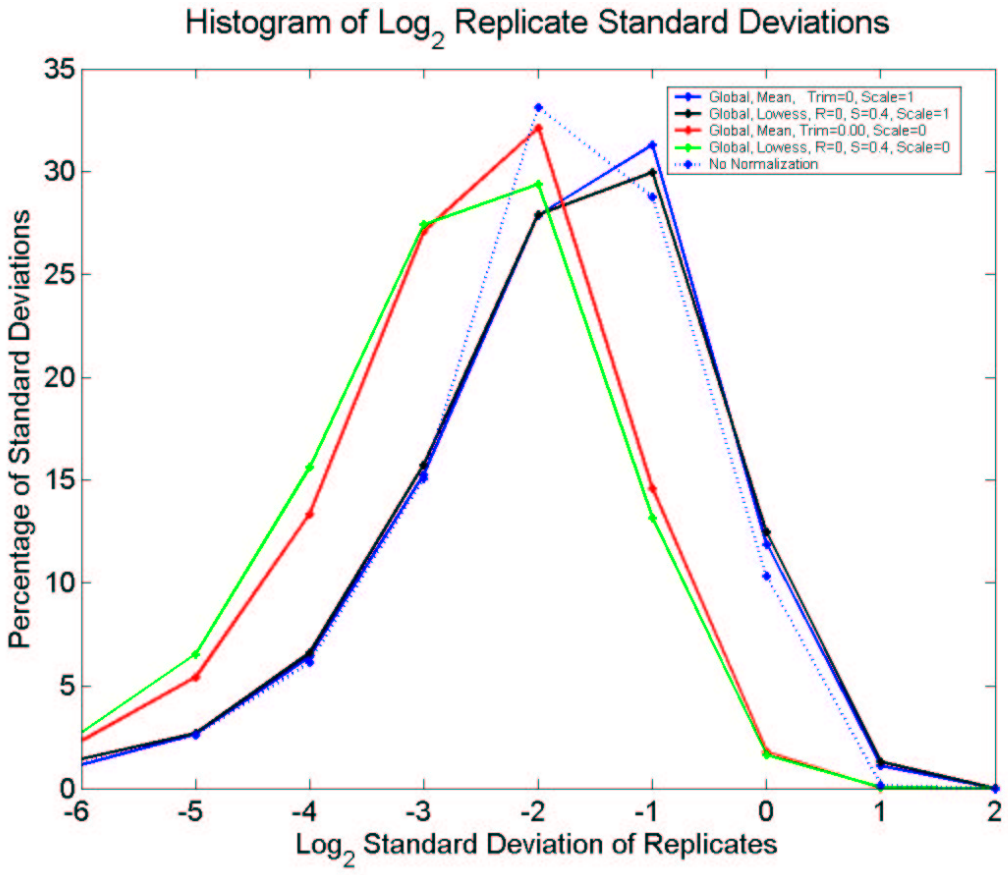


Figure 5 Comparison of Global Scaling versus No Scaling Normalization Techniques



Appendix 1

Interaction of cross-validation with normalisation

Pseudo-code for supervised classification

- construct K validation groups, where a single training case is a single slide.
- for each normalisation scheme, N
    - for each validation group, G
        - hold out the set G, and normalise each member of G individually (ie slide-by-slide, ie training-case-by-training-case)
        - denote the training set as T (everything but G)
        - normalise each member of T individually (ie slide-by-slide, ie training-case-by-training-case)
        - train on the normalised set T
        - count test errors made on the normalised set G

    end
    - report the test accuracy of N made over all validation groups, G.

end

Because we used single-slide normalisation, in practice each slide was only normalised once, outside of both loops. However, if multi-slide normalisation were to be used, then cross-validation would become slightly more complicated. In such a case, normalisation would be dependent on how the validation groups were constructed, since normalisation would be performed on each training set as a whole, and each validation set as a whole.