Assessing the efficiency of *dye-swap* normalization to remove systematic bias from two-color microarray data

Fatima Sanchez-Cabo
Department of Biomolecular Sciences, UMIST
P.O. Box 88, Manchester M60 1QD, U.K.
Institute for Genomics and Bioinformatics
Christian Doppler Laboratory for Genomics and Bioinformatics,
Graz University of Technology,
8010 Graz, Austria

Andreas Prokesch, Gerhard G. Thallinger, Roland Pieler and Zlatko Trajanoski
Institute for Genomics and Bioinformatics
Christian Doppler Laboratory for Genomics and Bioinformatics,
Graz University of Technology, 8010 Graz, Austria

Philip D. Butcher and Jason Hinds
Bacterial Microarray Group, St.George's Hospital Medical School,
Cranmer Terrace, London, U.K.

Leah E. A. Holmes, Susan G. Campbell, Mark P. Ashe and Simon Hubbard
Department of Biomolecular Sciences, UMIST
P.O. Box 88, Manchester M60 1QD, U.K.

Kwang-Hyun Cho
School of Electrical Engineering, University of Ulsan, Ulsan, 680-749, Korea

Olaf Wolkenhauer[*]
Department of Computer Science
University of Rostock, Rostock, Germany
Address: Albert Einstein Str. 21, 18051 Rostock, Germany.
E-mail: wolkenhauer@informatik.uni-rostock.de,
Tel./Fax:+49 (0)381 498 33 35/99.

March 24, 2004

[*] *To whom correspondence should be addressed.*

Abstract

Microarrays are a powerful tool in functional genomics, what allow a simultaneous analysis of the expression level of thousands of genes under different conditions. In order to compare measurements within and across arrays and to correct non-biological variation masking meaningful information, *normalization* is an essential task prior to any further analysis. Among all the available normalization techniques, *LOWESS* has proved useful for the normalization of data generated from the two main microarray platforms (two-color arrays and affymetrix chips) due to its ability to remove intensity dependent effects. However, the use of this robust estimator to correct the data without taking account of biological characteristics is a concern often raised by microarray analysts. In addition, powerful software packages are needed to perform such computationally expensive normalization and several parameters need to be fixed in advance, resulting in differently corrected data sets for the different sets of parameters used.

Reverse labeling designs are common setups in two-color microarray experiments if comparison between the co-hybridized samples is of interest. Using three different data sets, this paper assesses the effectiveness of *dye-swap* normalization, a method that makes use of the intrinsic information provided by this type of experimental design. The results show how *dye-swap* normalization corrects the bias introduced by the different properties of the dyes, removing intensity dependent effects. Furthermore, *dye-swap* normalization corrects the data accounting for gene-dye effects and the transformation of the data is justified on a biological basis. The results present *dye-swap* normalization as a valid alternative to normalize two-color microarray data.

The paper also reviews the assumptions made and the formulas applied to correct the data using *dye-swap* normalization. In addition, a generalization of the *dye swap* normalization formula is implemented to normalize data generated from microarray experiments for which a large proportion of genes are expected to be differentially expressed.[*]

[*] The figures and results presented in this paper have been implemented using several Bioconductor packages from R, MATLAB® (*Mathworks Inc.*) and ArrayNorm. A collection of files and supplementary material is available from http://www.sbi.uni-rostock.de

Keywords: Two-color microarrays, normalization, *LOWESS*, dye-swap, replicates.

1 Introduction

Two-color microarray experiments estimate simultaneously the relative expression level of a set of genes in two biological samples. To allow such a comparison, messenger RNA (mRNA) from the populations of interest is reverse transcribed and labeled using two different fluorescent dyes (usually Cyanine dyes, Cy3 and Cy5)(Schena 2002). Afterwards, both samples (related to the "channels" of the scanner used to read the array) are hybridized onto the microarray, where Polymerase Chain Reaction (PCR) products that represent all or part of the genes in the genome were spotted (Eisen and Brown 1999, Schulze and Downward 2001). The slide is then scanned at two different wavelengths corresponding to the range of the emission spectra of the fluorophore. For each channel a high resolution image is generated, which is then analyzed in a process referred to as "spot finding". The spots are quantified into single intensity values for each channel. These two intensity values are the estimators of the relative expression level of the gene in the two samples. The spotfinding or scanning software (e.g. GenePix, Imagene) also provides an estimator of the background intensity for a given spot, and in both channels. The data analyst then has the option to correct the data by, for example, subtracting the background from the spot intensity.

In microarrays, the process of removing non-biological variation that is masking meaningful information is known as *normalization*. The correction of the data, according to those factors introducing systematic errors, is an essential stage prior to the analysis and biological interpretation of the data. In two-color microarray experiments, the so-called dye effect is one of the most important sources of systematic errors. Several properties are different for the two dyes, i.e. their quantum efficiency and their gene specific incorporation properties (Dobbin et al. 2003, Tseng et al. 2001). They make it necessary to balance the intensities of both channels before further analysis. To compare two measurements that are actually read in different scales, they must be brought in to the same range. There are two non-exclusive strategies that can be employed to normalize microarray data:

- Normalization by self-consistency (Kepler et al. 2002) using all genes: There are three main methods based on the assumption that the overall intensity should be the same for both channels, i.e., most of the genes should be equally expressed in both compared samples. These methods are the global method (Yang et al. 2001, Yang et al. 2002), the use of a *LOWESS* function (Cleveland 1979) correcting intensity-dependent data (Yang et al. 2001, Yang et al. 2002) and the use of the regression line (Quackenbush 2001). From all of them, the use of a *LOWESS* function to normalize within a slide is the most robust and popular.

- Normalization using the quality control elements introduced in the experiment: This refers to the intrinsic and extrinsic controls, the use of replicated genes within the array to correct spatial effects, the use of replicated arrays and the swapping of the dyes for replicated arrays. The latter is a requirement to apply *dye-swap* normalization.

Dye swap experiments are extended and well established in the microarray community (Dobbin et al. 2003, Kerr and Churchill 2001). This paper studies the effectiveness of *dye-swap* normalization, which makes use of the information provided by this type of setup. Data from three different experiments were normalized using *dye-swap* and compared against the standard *LOWESS* normalized data. These experiments were chosen in order to test two different features: (1) A self-self hybridization experiment was conducted to assess the efficiency of *dye-swap* normalization in the correction of intensity dependent systematic bias, i.e. the different properties of the two dyes used to label the co-hybridized samples. (2) A good normalization method should preserve the biological characteristics of the data. For this reason the correlation of technical and biological replicates after normalization was analyzed for two experiments, a growth curve experiment for *M.tuberculosis* and an experiment to study the yeast transcriptional repressors Mig1p and Mig2p (Rolland et al. 2002, Campbell et al. *Manuscript in preparation,* 2004).

The paper is organized as follows. Firstly, the three main self consistency methods are discussed. These are the global approach, *LOWESS* (Yang et al. 2001, Yang et al. 2002) and the linear regressive approach (Quackenbush 2001). In Section 3, the most important quality control elements in microarrays are briefly described and *dye-swap* normalization is explained in detail. The different formulas proposed in the literature and their equivalence are also discussed. To conclude, in Section 4, *LOWESS* and *dye-swap* normalization are applied to three different data sets. The results are discussed in the following section.

2 Within array normalization by self-consistency: *LOWESS* correction

Microarrays allow us to simultaneously measure the response of thousands of genes to specific biological conditions.

Due to the large number of genes spotted onto an array, one might think that, on the whole, most genes will not show a significant change in the expression level between the two samples being compared. Under this premise, differences among the overall intensity of both channels would be the consequence of non-biological variation. An important source of systematic errors in two-color microarray experiments are the different properties of the dyes used to label the two samples (Yang et al. 2001, Dobbin et al. 2003). Under the assumption that most of the genes should be equally expressed in both samples, we ought to correct the data so that the distribution of the expression ratios has a central value of one. Choosing the median as an estimator of the central tendency of the distribution, the data are corrected to accomplish

$$median_{i=1,\ldots,n_g}\left(\frac{R_i}{G_i}\right) \cong 1 \;\Rightarrow\; \log_2\left(median_{i=1,\ldots,n_g}\left(\frac{R_i}{G_i}\right)\right) \cong 0,$$

where R_i represents the intensity of the red channel for gene i, G_i the same for the green one. n_g indicates the number of genes spotted on the array. This transformation can be achieved by estimating an expression ξ (Yang et al. 2001, Yang et al. 2002), as

$$R = \xi \cdot G.$$

The different estimators of ξ will result in the three different within-array normalization methods:

The **global method** looks for a constant which relates the overall intensity of both channels. A common choice is (Yang et al. 2001)

$$\xi = 2^{median_{i=1,\ldots,n_g}\left(\log_2\left(\frac{R_i}{G_i}\right)\right)}.$$

The **linear regression** method (Quackenbush 2001) fits a regression line to the scatter plot (G,R). Under the assumption that most of the genes should be equally expressed for both channels, the regression line should have a slope one. Hence,

$$R = m \cdot G + n \rightarrow \frac{R}{m} - \frac{n}{m} = G\ .$$

From that follows $\xi \simeq m$, where m is the slope of the regression line fitted to the scatter plot and n is the intercept with the ordinate.

The ***LOWESS***[†] function was first introduced by Cleveland (1979). This function is estimated through a locally weighted polynomial regression for a fixed subset of genes in the neighborhood of every gene i. As a tool to normalize microarray data, it first appeared in Yang et al. (2001). From the scatter plot $(A{,}M)$, where

$$M = \log_2 \left(\frac{R}{G} \right) \text{ and}$$
$$A = \frac{1}{2} \cdot (\log_2 G + \log_2 R) \ ,$$

the *LOWESS* function $c(A_i)$ can be calculated:

$$c(A_i) : I \mapsto \mathbb{R},$$

where the set of indexes I denotes all genes spotted on the array. Under the assumption that most of the genes are equally expressed in both channels, A is the overall intensity level measured in the array as it can be observed by

$$\log_2 R \ \simeq \ \log_2 G \Rightarrow A = \frac{1}{2} \cdot (\log_2 G + \log_2 R) \ \simeq \log_2 G \simeq \log_2 R \ .$$

The fitting of the *LOWESS* function $c(A)$ from the $(A{,}M)$ scatterplot leads to:

$$M = \log_2 \left(\frac{R}{G} \right) \cong c(A) \ \Rightarrow \xi = k(A) = 2^{c(A)}.$$

Regardless of the method used to estimate ξ, the data will be corrected as follows:

$$\log_2 \left(\frac{R}{G} \right) \cong \gamma \Rightarrow \log_2 \left(\frac{R}{G} \right) - \gamma \cong 0 \Rightarrow \log_2 \left(\frac{R}{G \cdot \xi} \right) \cong 0,$$

where $\gamma = \log_2 (\xi)$. Denoting the corrected data by the superscript ^{c}, it follows that

$$M_i^c = M_i - \gamma_i, \quad \text{for all } i.$$

This is equivalent to correct both channels intensity values, for every spotted gene i as:

$$R_i^c = R_i,$$
$$G_i^c = G_i \cdot \xi_i.$$

[†] *LOcally WEighted leaSt Squares* (LOWESS)

Because the dye effect appears to be intensity dependent in most of the cases (Yang et al. 2001, Yang et al. 2002, Workman et al. 2002), *LOWESS* has become a popular method for within-array normalization. The global dye correction method transforms all of the genes using a single value for the whole slide, whilst the regression method is highly sensitive to outliers. In consequence, the *LOWESS* approach appears to be the most suitable option among the three in reducing the differential effects of the dyes.

3 Within array normalization using quality control elements: *Dye-swap* normalization

The three self-consistency methods described above provide a general approach to correct the dye effect. However, as shown by van de Peppel et al. (2003) they are not suitable for all those experiments for which a global shift in mRNA expression occurs. In those situations, the intrinsic information of the experiment must be used to normalize the data. To this end, a good experimental design should provide quality control elements, including control spots, replicated genes within the array or replicated arrays for which the dyes are swapped. Different material can be spotted as controls in the microarray, for example, genomic DNA (gDNA), "spiked genes"(van de Peppel et al. 2003, Benes and Muckenthaler 2003), or a Microarray Sample Pool (MSP) (Yang et al. 2002). For the controls to be useful in the normalization, their intensities should cover the whole intensity range. In that case, the *LOWESS* function or any other non-linear function fitted to the controls (using for example the Levenberg-Marquardt algorithm (Marquardt 1963)) can be used to determine the relationship between both channels, and this function can then be used to correct the whole data set.

The use of replicates meets a double target in microarray experiments (Churchill 2002): Biological replicates are a requirement to provide statistical significance measures for differences in gene expression (Black and Doerge 2002) and to average out the differences among individuals. Technical replicates remove random errors introduced in the experiment and replicated slides can be used to normalize the data if the dyes are swapped.

Let us consider a particular gene i for which the expression level in two samples of mRNA is measured. We will refer to the two biological samples to be compared as s and r. Let us suppose that during the reverse transcription of mRNA into cDNA the sample denoted by s was labelled with Cy5 (red) and the sample denoted by r with Cy3 (green). For every spotted gene i the following expression is considered

$$M_i = \log_2\left(\frac{R_i}{G_i}\right).$$

Using the same material, the reverse transcription process and labelling are repeated, but in this case the dyes are swapped so the sample s is labelled with Cy3 (green) and the r with Cy5 (red). For the same gene i we thus have

$$M'_i = \log_2\left(\frac{R'_i}{G'_i}\right).$$

From these two equations, we obtain

$$M_i = \log_2\left(\frac{R_i}{G_i}\right) = \log_2\left(\frac{s_i}{r_i}\cdot k_i\right) = \log_2\left(\frac{s_i}{r_i}\right) + \log_2 k_i = \log_2\left(\frac{s_i}{r_i}\right) + c_i, \tag{1}$$

$$M'_i = \log_2\left(\frac{R'_i}{G'_i}\right) = \log_2\left(\frac{r_i}{s_i}\cdot k'_i\right) = -\log_2\left(\frac{s_i}{r_i}\right) + \log_2 k'_i = -\log_2\left(\frac{s_i}{r_i}\right) + c'_i, \tag{2}$$

where r_i stands for the intensity of the gene i in sample r and s_i for the same value in sample s. The target is to estimate $\log_2(\frac{s_i}{r_i})$ from M_i, M'_i. Hence, it follows that

$$M_i - c_i = \log_2\left(\frac{s_i}{r_i}\right),$$

$$-M'_i + c'_i = \log_2\left(\frac{s_i}{r_i}\right).$$

In these expressions, c_i and c'_i account for the different properties of the dyes. As suggested in Yang et al. (2001) under the name "self normalization", if $c_i \simeq c'_i$ for all i (see Appendix A for an explanation), adding both equations

$$M_i - M'_i \simeq 2\cdot\log_2\left(\frac{s_i}{r_i}\right) \Longrightarrow \frac{1}{2}\cdot(M_i - M'_i) \simeq \log_2\left(\frac{s_i}{r_i}\right). \tag{3}$$

Earlier, Kerr et al. (2000) had proposed a formula to estimate the corrected logged expression ratio for a *dye-swap* experiment, under the assumption of normality of the logged intensities:

$$\log_2\left(\frac{s_i}{r_i}\right) = \frac{1}{2}[(M_i - M'_i) - (mean_i(M_i) - mean_i(M'_i))]. \tag{4}$$

(3) and (4) are equivalent if $c_i \simeq c'_i$, for all i. In that case, $mean_i(M_i) \simeq mean_i(M'_i)$ and (3) and (4) are the same formula. (4) was also used by Tseng et al. (2001) and it

is equivalent to performing a global normalization to correct possible global shifts weakening the reproducibility of the technical replicates. This can be due to different PMT (Photomultiplier Tube) settings, differences in labeling or hybridization between the two replicates, etc. Once these errors are corrected, the conventional *dye-swap* normalization (3) is performed to correct the different properties of the dyes on a gene by gene base. All through the text *dye-swap* normalization was calculated using (4). The main advantage of *dye-swap* normalization is the correction of the data preserving the characteristics of every gene. In addition, it accounts for the different incorporation rate of the two dyes to different sequences. Note also that the implementation of the method is straight forward and the computational cost very low. The main apparent disadvantage is the need for an additional microarray slide to allow dye reversal. However, the inaccuracy of two-color microarrays makes it necessary to provide technical replicates in every condition. Hence, there is no real extra cost in performing the technical replicates with a swap of the dyes.

4 Results

4.1 Self-self hybridization

The scope of the experiment was to test the influence of different surface coatings on the quality of in-house spotted mouse cDNA microarrays. The microarrays comprise several cDNA libraries: NIA (National Institute of Aging), BMAP (Brain Molecular Anatomy Project) and costum libraries. They were PCR amplified and spotted without purification in 3xSSC/1.5M Betaine spotting buffer. Including controls and duplicates, 36672 features were spotted on each array in one spotting run.

Four surfaces with different coatings were compared: Epoxy, amino and aldehyde surfaces from Schott-Nexterion and GAPS II coated slides from Corning. In order to be able to analyze as many spots as possible, RNA had to be chosen from a transcriptionally active source. Therefore, murine mesenchymal stem cells were induced to adipogenesis and eight time point samples were taken over 14 days of differentiation. These samples were pooled together and self versus self hybridizations were performed on each two slides per surface.

Data Preprocessing

The data was scanned with the GenePix 4000B scanner (AxonInstruments 1995-2004) and the image quantification software used was GenePix®4.0.1. After grid alignment, obvious areas of background artifacts were manually flagged and genes were filtered out according to the two following filter criteria: Number of saturated pixels in at least one of the channels greater than 50 and sum of medians smaller than 1500. The latter removes low intensity spots for which expression levels cannot be reliably measured with the scanner. *Dye-swap* and *LOWESS* normalization were performed using the marrayNorm package from Bioconductor. The filtered data was normalized using ArrayNorm (Pieler et al. 2004). The commands used in Bioconductor can be found in the Supplementary Material.

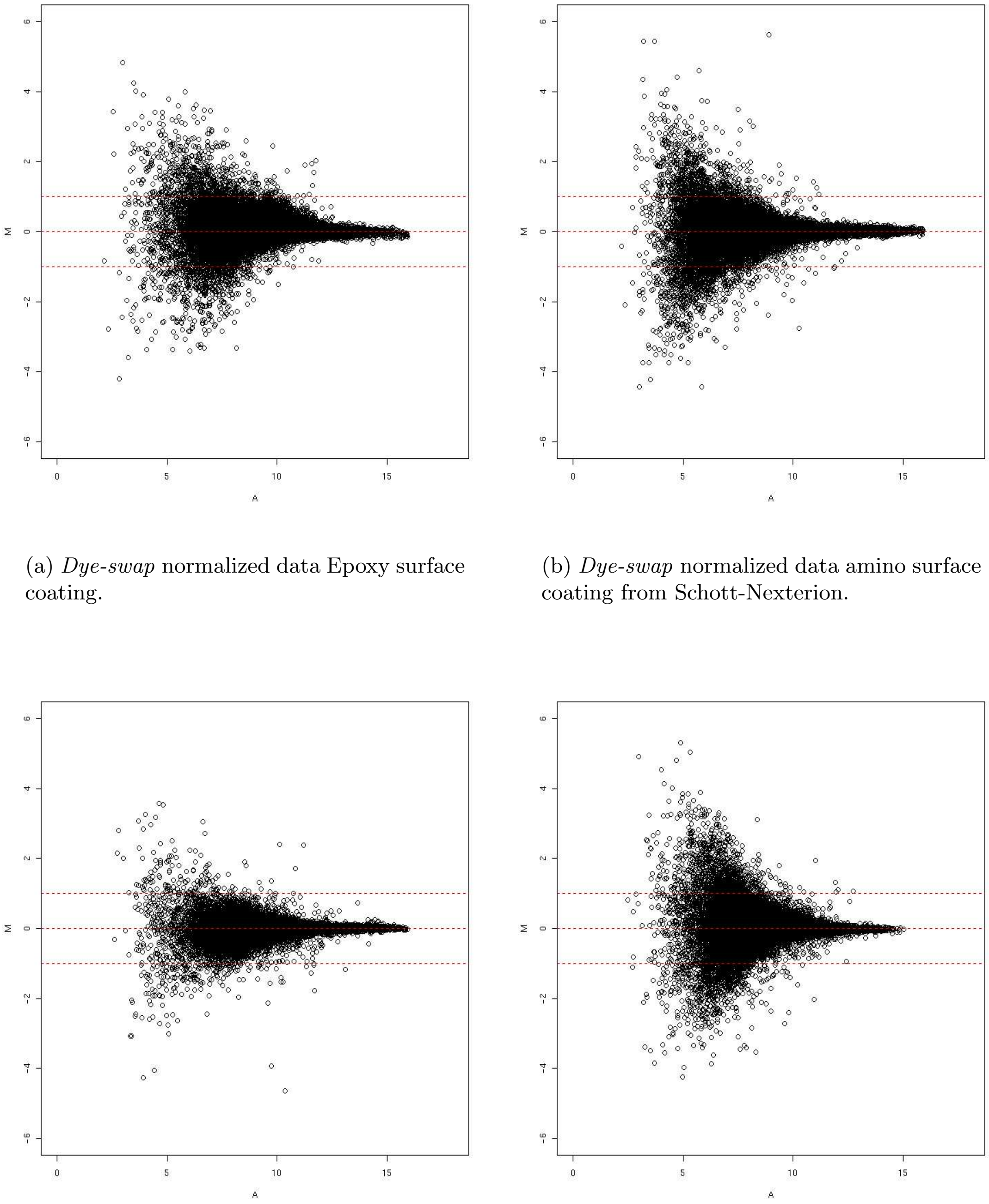
Bias removal

Although no biological replicates were available, the experiment was interesting to test how well *dye-swap* normalization and *LOWESS* normalization are able to remove the bias and if *dye-swap* normalization can correct the intensity dependent effect. Because the same material was labelled with the two dyes and hybridized together onto the same slide type twice, the pair of technical replicates can also be considered as a dye swap pair. The distribution of the corrected log ratios should be centered around 0 with very small dispersion. Figures 1 and 2 show the normalized data after *dye-swap* and *LOWESS* normalization. For *LOWESS* normalization the eight slides were normalized independently and the average of the technical replicates was calculated, resulting in a unique value per surface coating to be tested.

In addition, we calculated the standard deviation of the filtered normalized data. Table 1 shows the results.

4.2 Study of the *M.tuberculosis* growth curve

This experiment aimed to understand the growth curve for *M.tuberculosis*, taking measurements after 6, 14, 20 and 30 days. Four replicated arrays of RNA samples from each time point were hybridized. In total, sixteen arrays were produced, using for the "signal" channel the four samples of RNA extracted from *M.tuberculosis* and using gDNA for the "reference" channel. The advantage of this reference design is that all genes in the



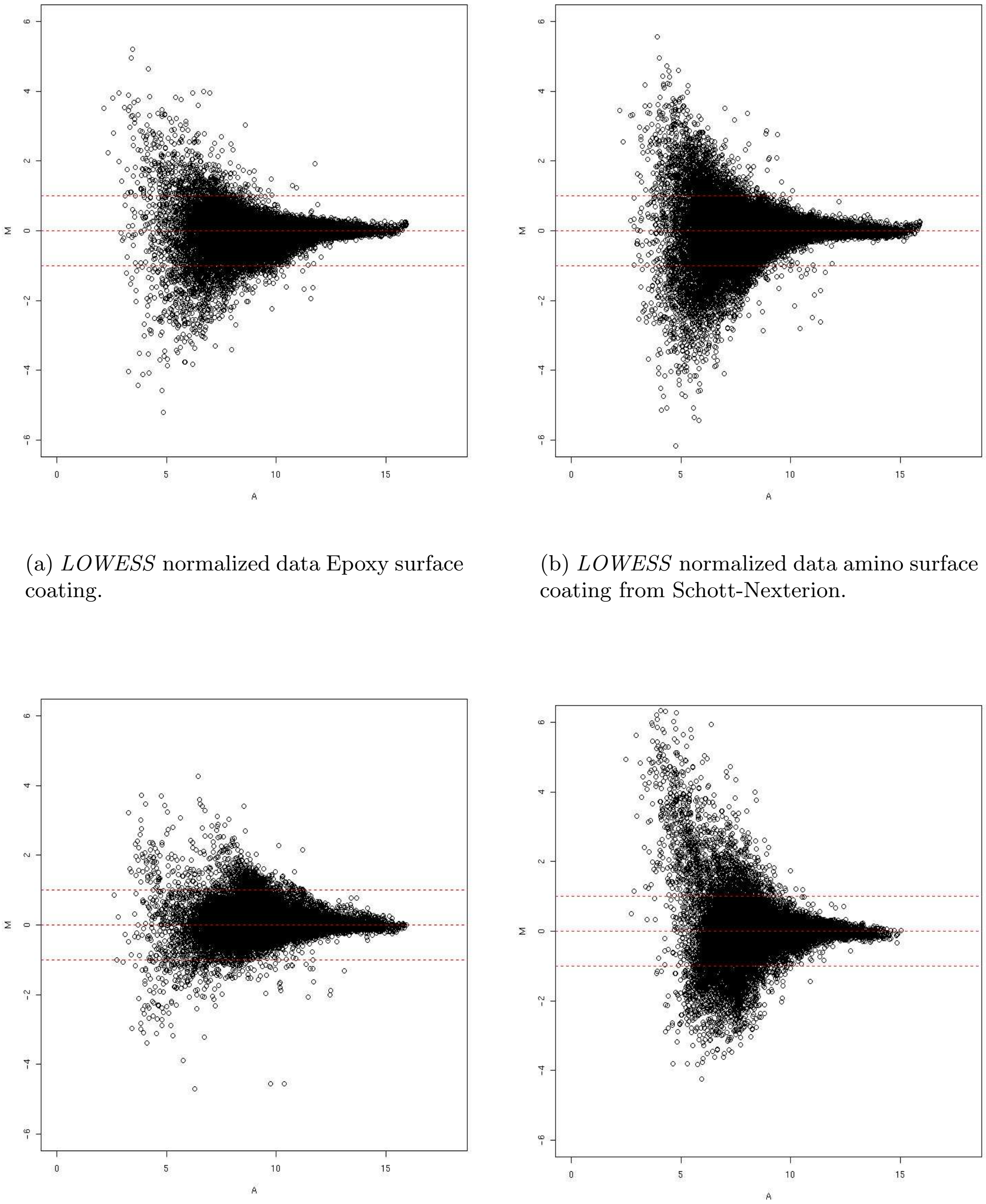
(a) *Dye-swap* normalized data Epoxy surface coating.

(b) *Dye-swap* normalized data amino surface coating from Schott-Nexterion.

(c) *Dye-swap* normalized data aldehyde surfaces from Schott-Nexterion.

(d) *Dye-swap* normalized data GAPS II coated slides from Corning.

Figure 1: *Dye-swap* non-filtered normalized data



(a) *LOWESS* normalized data Epoxy surface coating.

(b) *LOWESS* normalized data amino surface coating from Schott-Nexterion.

(c) *LOWESS* normalized data aldehyde surfaces from Schott-Nexterion.

(d) *LOWESS* normalized data GAPS II coated slides from Corning.

Figure 2: *LOWESS* non-filtered normalized data

Table 1: Standard deviation of the filtered data after *dye-swap* normalization and after *LOWESS* normalization. The standard deviation after *LOWESS* is always approximately double than the standard deviation of the filtered log ratios normalized using *dye-swap*.

condition	**std. filtered log ratios *dye-swap***	**std. filtered log ratios *LOWESS***
1	0.1058109	0.2002904
2	0.0969764	0.1548934
3	0.1015136	0.2620740
4	0.1089701	0.2671140

genome are present in the gDNA. Hence, every gene should give a homogeneous signal for the denominator of the ratio of both channels. A broader discussion on this topic can be found in (Talaat et al. 2002). The labelling reactions were performed independently and the dyes were swapped for one out of the four replicates. Denoting by $a = 1, 2, ..., 16$ the number of the array, the experiment can be summarized as

$$\text{for } a \neq 4, 8, 11, 16 \left\{ \begin{array}{rcl} \text{Green} & : & \text{RNA (signal)}, \\ \text{Red} & : & \text{gDNA (reference)}, \end{array} \right.$$

$$\text{for } a = 4, 8, 11, 16 \left\{ \begin{array}{rcl} \text{Green} & : & \text{gDNA (reference)}, \\ \text{Red} & : & \text{RNA (signal)}. \end{array} \right.$$

PCR products of the 3924 genes of the genome of *M.tuberculosis* strain H37Rv were spotted once in every slide. In addition, different types of controls were printed at different locations. The normalization controls were 5s, 16s and 23s ribosomal RNA genes, printed in every sub-grid. The 16s and 23s rRNA were printed in a three-fold dilution series. Many of the controls gave a saturated signal in the RNA channel. The reason is that whilst gDNA used for the reference has a single copy of the rRNA genes, so equal in abundance to the other genes in the genome, the prokaryotic RNA is total RNA consisting of 98% rRNA and just 2% mRNA. Hence, in the RNA channel a greater proportion of the RNA hybridised to the control spots relative to the rest of the gene spots and so the higher intensities presented by the control spots was not in the same range as the intensities for the other genes. The control spots were excluded from the analysis and all of the results in this paper refer to the 3924 printed genes. Although there were no duplicated genes on the slide, PCR products from the two IS6110 transposase family elements were present. Each of them has sixteen copies. Differences of only a few nucleotides have been detected between the sequenced copies, so we can expect their intensity levels to be very similar after proper normalization of the data.

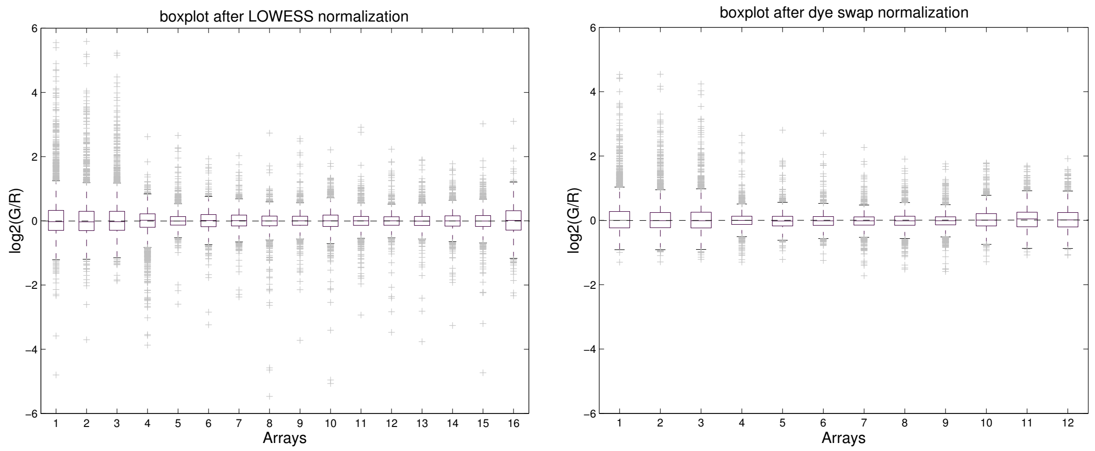
Data preprocessing

The slides were scanned with the Affymetrix 428^{TM} scanner (MWG-Biotech 2004) and the image quantification software used was Imagene ® (BioDiscovery 2004). The use of gDNA reference made the use of all the genes printed in the array feasible because all of them gave a reliable signal in the reference channel. In addition, no gene had to be removed due to high background intensity. Following the analysis of the background intensity, it was decided not to perform background subtraction. There were two reasons: First, the overall background intensity was very small if compared to the foreground intensity. In the second place, we found that the noise patterns that appeared in the background reconstruction were inherited by the foreground after background subtraction. All of this analysis and the *dye-swap* normalization of the data was achieved using the normalization module of the program MADE (Sanchez-Cabo et al. 2003). Because the implementation of *LOWESS* depends on the chosen parameters (i.e. width of the neighborhood, smooth degree, etc.) we normalized the data using the *LOWESS* function from the limma package from Bioconductor (Gentleman et al. 2003). The functions used can be found in the Supplementary Material.

Correlation of replicated measurements

For this particular experiment, relatively few genes are expected to change in expression between the two hybridized samples at every time point. *LOWESS* and *dye-swap* normalization were applied to correct the data. Both removed the bias as presented in Figure 3.

A good normalization method should correct the systematic bias while preserving the biological information in the data. To test the second, we classified the slides using hierarchical clustering (Eisen and Brown 1999). The results obtained support the hypothesis that *dye-swap* normalization preserves the biological information of the data better than



(a) Boxplot after *LOWESS*normalization.

(b) Boxplot after *dye-swap* normalization.

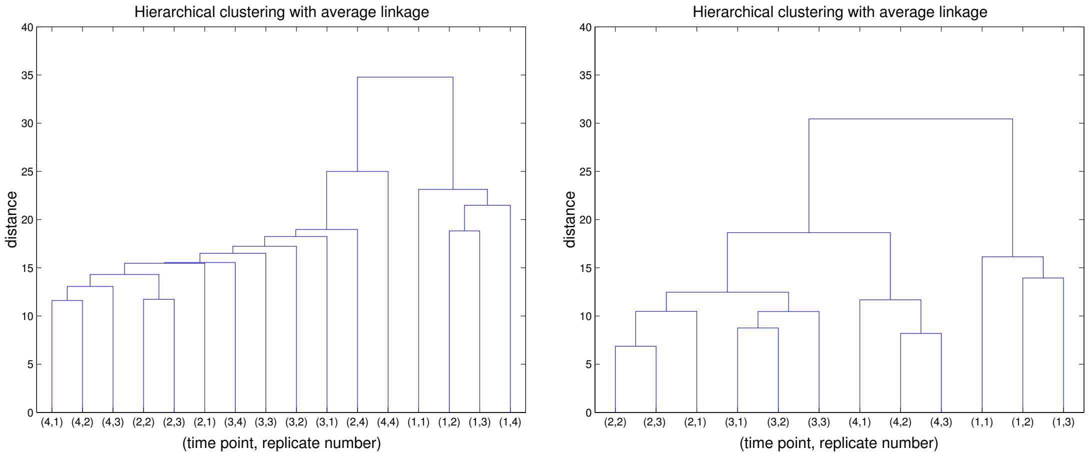
Figure 3: Distribution of the log-ratios for the 16 arrays of the experiment after *LOWESS* and after *dye-swap* normalization. Every four consecutive boxplots (three after *dye-swap* normalization) are the replicates at a particular time point.

LOWESS (see Figure 4). However, the comparison might not be considered fair because using *dye-swap* normalization the three slides classified for every condition have been normalized with a common slide, while *LOWESS* normalized every slide independently.

The study of the correlation among the elements of the IS6110 family would be a fairer comparison. Therefore, we calculated the coefficient of variation (CV) after *dye-swap* and after *LOWESS* normalization. The results are displayed in Table 2. The overall CV was smaller for the elements of the IS6110 family after *dye-swap* normalization than after *LOWESS*. In addition, the CV was larger after *dye-swap* than after *LOWESS* normalization for only one slide (slide 13 (time point four, first replicate)).

4.3 Analysis of the yeast transcriptional repressors Mig1p and Mig2p

Microarray analysis was used to investigate *Saccharomyces cerevisiae* strains deleted for the transcriptional repression genes, *MIG1* and *MIG2*. Mig1p and Mig2p function in the yeast glucose repression/derepression pathway to directly repress genes required for the use of alternate carbon sources (Rolland et al. 2002). In wild type, $mig1\Delta$ and $mig2\Delta$ strains, protein synthesis is inhibited following glucose removal (Ashe et al. 2000). However, the double mutant, $mig1\Delta mig2\Delta$, maintains translation following a switch to no glucose (Campbell et al. *Manuscript in preparation,* 2004). In an attempt to identify changes



(a) Hierarchical clustering of the replicates after *LOWESS* normalization.

(b) Hierarchical clustering of the replicates after *dye-swap* normalization.

Figure 4: Hierarchical clustering of the replicates. Replicates of the same time point cluster perfectly together after *dye-swap* normalization. After *LOWESS* normalization slide (2,1) is closer to slide (3,4) than to the rest of the second time point replicates. Slides (2,4) and (4,4) are missclassified.

Table 2: Dispersion of the IS6110 elements in every slide after *LOWESS* and *dye-swap* normalization. The quality measure used was the Coefficient of Variation (CV).

After *LOWESS* normalization			
(time, replicate)	Mean	STD	CV
(1,1)	0.6966	0.2516	0.3612
(1,2)	0.6881	0.1400	0.2034
(1,3)	0.7930	0.1292	0.1630
(1,4)	1.6077	0.3205	0.1993
(2,1)	0.8928	0.0599	0.0671
(2,2)	0.9898	0.2957	0.2988
(2,3)	0.9407	0.2332	0.2480
(2,4)	1.2124	0.1546	0.1275
(3,1)	0.9113	0.1780	0.1953
(3,2)	0.9365	0.1586	0.1694
(3,3)	1.2248	0.1383	0.1129
(3,4)	1.0549	0.3597	0.3410
(4,1)	0.9005	0.0826	0.0917
(4,2)	0.9370	0.1766	0.1885
(4,3)	0.9793	0.1884	0.1924
(4,4)	1.4126	0.3152	0.2231
overall mean			**0.1989**

After dye-swap normalization			
(time, replicate)	Mean	STD	CV
(1,1)	0.9426	0.1495	0.1586
(1,2)	0.6761	0.1387	0.2051
(1,3)	0.6845	0.0933	0.1363
(2,1)	0.9110	0.0587	0.0644
(2,2)	0.8142	0.1275	0.1566
(2,3)	0.8260	0.1080	0.1307
(3,1)	0.8539	0.0580	0.0679
(3,2)	0.8171	0.0953	0.1167
(3,3)	0.8654	0.0842	0.0973
(4,1)	0.8662	0.1352	0.1560
(4,2)	0.7748	0.1360	0.1755
(4,3)	0.7555	0.1290	0.1707
overall mean			**0.1363**

in gene expression profiles that may be responsible for these translational phenotypes labeled cDNA from each mutant was compared to wild type cDNA using competitive hybridization on spotted arrays representing the whole genome of *S. cerevisiae.* Six arrays were hybridized for each mutant and wild type, of which three experiments used reciprocal labeling of RNA. The experiment was then repeated using duplicate RNA extracts to give a total of thirty-six hybridizations.

Data preprocessing

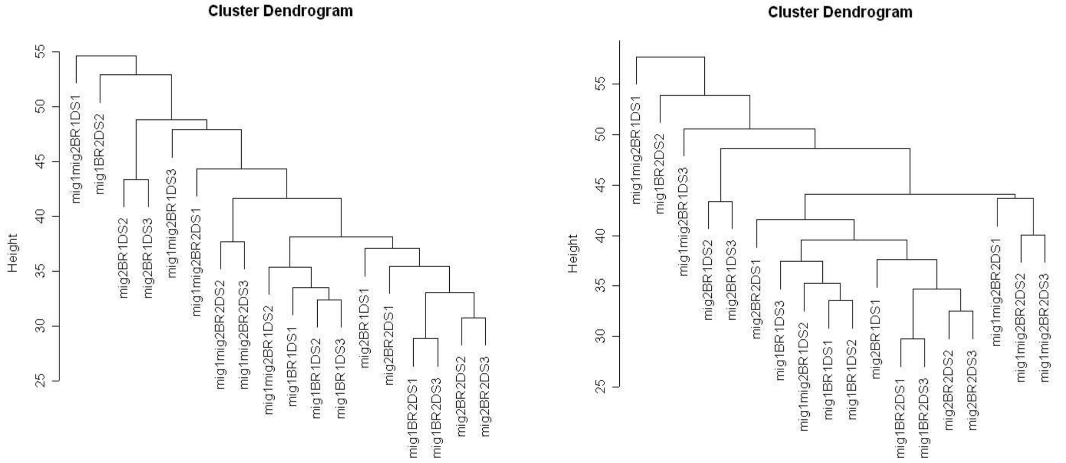
In two-color microarrays designed to study eukaryotic gene expression, the use of an homogeneous reference as gDNA presents practical difficulties due to the high introns-exons ratio. In consequence, the quality of the data is rarely as good as for the previous experiment presented. Regions with high levels of background are often observed and also spatial artifacts. At this stage, flagging of bad spots becomes an issue. At the moment, there is no consensus about the desirable balance between number of spots filtered out and the criteria to be applied to get reliable biological conclusions. For this particular data set, no standard filtering criteria was applied. Areas with obvious artifacts or very high background were flagged. In this case background subtraction was performed. The slides were scanned using the GenePix 4000A scanner (AxonInstruments 1995-2004) and the images were quantified using GenePix$^{®}$.

The data was first normalized using the marrayTools package from Bioconductor (Gentleman et al. 2003). The functions used can be checked in the Supplementary Materials. The normalization of the filtered data was computed using ArrayNorm (Pieler et al. 2004).

Correlation of replicated measurements

For this experiment only those genes regulated by Mig1p and Mig2p were expected to change. Hence, *dye-swap* normalization and *LOWESS* are suitable to normalize the data and they should both remove the bias in a similar manner. In fact, a comparison of both normalization methods showed a very similar list of differentially expressed genes when using the fold change. However, Figure 5 shows how the across replicates variation is not the same for the two compared methods. Consequently, if more biological replicates had been available and statistical methods such as *t-test* had been applied to the data, different sets of differentially expressed genes would have been found.

Figure 5 shows the hierarchical clustering of the replicates after both, *dye-swap* and *LOWESS* normalization using the marrayTools package from Bioconductor. The correlation between the biological and technical replicates was calculated using all the genes without paying attention to the bad quality spots. In both figures, *mig*1Δ, *mig*2Δ and *mig*1Δ*mig*2Δ are the three compared mutants. BR stands for Biological Replicate (1,2) whereas DS 1, 2 and 3 represent the three dye swap pairs.

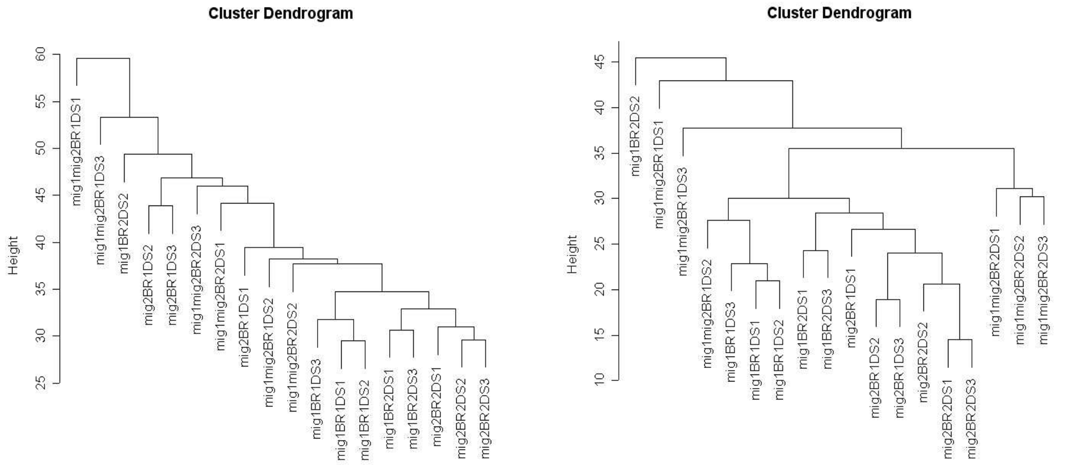


(a) Hierarchical clustering of the replicates after *LOWESS* normalization including all spots.

(b) Hierarchical clustering of the replicates after *dye-swap* normalization including all spots.

Figure 5: Hierarchical clustering of the replicates without filtering the data. Both plots show a similar result when looking at the correlation of the replicates without excluding the flagged spots. In the case of *LOWESS*, all slides were independently normalized and the mean between the technical replicates was calculated.

Because further clustering analysis and detection of differentially expressed genes was performed using only those genes with a reliable measurement, the data were also normalized using ArrayNorm (Pieler et al. 2004) removing the genes that were flagged. The correlation between the replicates was compared for the filtered data normalized using *dye-swap* and *LOWESS*. The results showed a better preservation of the biological meaning of the data after *dye-swap* normalization than after *LOWESS* normalization (see Figure 6).



(a) Hierarchical clustering of the replicates after *LOWESS* normalization for the filtered data.

(b) Hierarchical clustering of the replicates after *dye-swap* normalization for the filtered data.

Figure 6: Hierarchical clustering of the replicates. After *dye-swap* normalization all replicates cluster perfect with the exception of mig1mig2BR1 and mig1BR2DS2. However, after *LOWESS* normalization the biological and technical replicates do not correlate well.

5 Discussion

At present, there is no consensus view regarding either the normalization method that better corrects the systematic bias introduced in microarray experiments or the means to validate a normalization algorithm. Data sets that had been normalized using standard methods such as *LOWESS* have been shown to give different results if normalized using spike-in controls (van de Peppel et al. 2003). In general, it is difficult to know *a priori* which assumptions hold for every particular data set, because it is exactly the behavior on the whole that is unknown to the researcher when planning a microarray experiment. However, it is entirely reasonable that a good normalization method should preserve the biological characteristics of the experiment. Robust normalization methods have often given rise to concerns in this respect (Tsodikov et al. 2002, van de Peppel et al. 2003).

The results from the self-self hybridization pointed out several important facts: Firstly, *dye-swap* normalization calculated using formula (4) removes the intensity dependent effect as efficiently as *LOWESS*. Secondly, conditions 1 and 2 show a tail in the high intensity range (after *LOWESS* normalization), which is probably due to saturation artefacts. This did not appear after *dye-swap* normalization. Thirdly, for both the filtered and the non-filtered data, the spot dispersion was smaller after *dye-swap* normalization than after *LOWESS*.

The other two data sets were chosen to test how well *dye-swap* normalization preserves the correlation among technical and biological replicates. For the growth curve experiment the results were definitive in favor of the *dye-swap* normalization. However, the correlation among replicates could be artificially high because all three replicates were normalized with the same slide. For this reason the same comparison was done for the microarray experiment studying the Mig1p Mig2p yeast transcriptional repressors. Although the results for the non-filtered data did not show much variation between both methods, technical and biological replicates were much better correlated after *dye-swap* than after applying *LOWESS* to the filtered data. The necessity for the filtering of low intensity spots is a source of much debate. In this experiment filtering removed obvious low quality areas and the percentage of poor spots in every slide is shown in the Supplementary Material.

Yet, both methods present the same important drawback: they can only be applied to experiments for which most genes are expected to behave in the same way in the two compared samples. However, *dye-swap* normalization could be easily applied to experiments for which *a priori* information about the number of genes differentially expressed is available. For them, formula (4) should be:

$$\log_2\left(\frac{s_i}{r_i}\right) = \frac{1}{2}(M_i - M_i') - \frac{1}{2}\left[n^{th}perc_{\{i=1,\ldots,n_g\}}(M_i) - (100-n)^{th}perc_{\{i=1,\ldots,n_g\}}(M_i')\right],$$

where $n^{th}perc$ indicates the percentage of genes expected to be changing between the two hybridized conditions.

6 Conclusions

Making use of three different data sets, this paper illustrates the effectiveness of *dye-swap* normalization. *Dye-swap* normalization removes systematic bias in the data accounting for intensity dependent effects and is a natural way to correct the data because each step is justified on a biological basis. The results present *dye-swap* normalization as a valid alternative to *LOWESS* with the advantage of its low computational cost. An apparent disadvantage is the need for an extra slide where the dyes are switched. However, technical replicates must be provided in two-color microarray experiments, therefore a switch of the dyes does not necessarily increase the cost of the experiment. Dobbin et al. (2003) propose a more economic reverse dye design.

Yet, the method as formulated by Kerr et al. (2000) is only valid under the premises that most genes are equally expressed and that the logged intensity values are normally distributed. With exception of the length of the tails (which contain the differentially expressed genes) the symmetry and uni-modality of microarray data is often assumed as true. Having shown that *dye-swap* normalization corrects the data appropriately, its general application to experiments for which the proportion of genes expected to change it is known, is also simple and straight forward.

Nothing is concluded about the necessity of establishing standard criteria to filter out poor quality spots. However microarrays are a quantitative tool that provide numerical information. The resulting measurements should then be as reliable as possible.

Appendix A: Different properties of Cy3 and Cy5

The basic assumption made by Yang et al. (2001) in the *dye-swap* normalization method, is that $c_i \simeq c'_i$. This appendix tries to explain under which conditions this is true, from a theoretical view, without considering random errors affecting the quality of the labeling or hybridization.

The two cyanine dyes differ in several aspects. Some of them are intrinsical to the dyes and independent on the sample or the sequence the dyes are labelling. These are, for example, the different quantum yield, different quenching properties or the different photobleaching properties of the dyes (Tseng et al. 2001). In consequence, they are neither sample- nor gene-dependent, and they are not supposed to change significatively from one array to another, and neither within an array. Formulating this in a mathematically form, we have that:

$$\begin{aligned} \text{Quantum Yield} &: \text{QY(dye,gene,sample)=QY(dye)} \\ \text{Quenching} &: \text{Qn(dye,gene,sample)=Qn(dye)} \\ \text{Photobeaching} &: \text{PH(dye,gene,sample)=PH(dye)} \end{aligned}$$

However, there is another difference between Cy3 and Cy5 that is essential in two-color microarrays. Due to the different size of their molecules, Cy3 and Cy5 incorporate differently to particular sequences. Hence, some genes have been observed to incorporate one dye more efficiently than the other (Dobbin et al. 2003). Kerr et al. (2000) introduced in the ANOVA model proposed in a posterior publication (Kerr and Churchill 2001) the dye × gene effect. Although not originally expected, experimental data showed several examples of the gene-dependent different incorporation properties of the two cyanine dyes. Again, we can formulate this as:

$$\text{Incorporation : In(dye,gene,sample)=In(dye,gene)}$$

Using the same nomenclature as in Section 3, *if the gain set to scan both slides was the same*, the intensity level of a particular gene i measured in the two channels can be expressed as:

$$\begin{aligned} R_i = f(s_i) &= \mathrm{QY}(\mathrm{Cy5}, i, s) \cdot \mathrm{Qn}(\mathrm{Cy5}, i, s) \cdot \mathrm{PH}(\mathrm{Cy5}, i, s) \cdot \mathrm{In}(\mathrm{Cy5}, i, s) \cdot s_i \\ &= \mathrm{QY}(\mathrm{Cy5}) \cdot \mathrm{Qn}(\mathrm{Cy5}) \cdot \mathrm{PH}(\mathrm{Cy5}) \cdot \mathrm{In}(\mathrm{Cy5}, i) \cdot s_i \\ G_i = g(r_i) &= \mathrm{QY}(\mathrm{Cy3}, i, r) \cdot \mathrm{Qn}(\mathrm{Cy3}, i, r) \cdot \mathrm{PH}(\mathrm{Cy3}, i, r) \cdot \mathrm{In}(\mathrm{Cy3}, i, r) \cdot r_i \\ &= \mathrm{QY}(\mathrm{Cy3}) \cdot \mathrm{Qn}(\mathrm{Cy3}) \cdot \mathrm{PH}(\mathrm{Cy3}) \cdot \mathrm{In}(\mathrm{Cy3}, i) \cdot r_i \end{aligned}$$

The same is true for R'_i and G'_i:

$$\begin{aligned} R'_i = f'(r_i) &= \mathrm{QY}(\mathrm{Cy5}, i, r) \cdot \mathrm{Qn}(\mathrm{Cy5}, i, r) \cdot \mathrm{PH}(\mathrm{Cy5}, i, r) \cdot \mathrm{In}(\mathrm{Cy5}, i, r) \cdot r_i \\ &= \mathrm{QY}(\mathrm{Cy5}) \cdot \mathrm{Qn}(\mathrm{Cy5}) \cdot \mathrm{PH}(\mathrm{Cy5}) \cdot \mathrm{In}(\mathrm{Cy5}, i) \cdot r_i \\ G'_i = g'(s_i) &= \mathrm{QY}(\mathrm{Cy3}, i, s) \cdot \mathrm{Qn}(\mathrm{Cy3}, i, s) \cdot \mathrm{PH}(\mathrm{Cy3}, i, s) \cdot \mathrm{In}(\mathrm{Cy3}, i, s) \cdot s_i \\ &= \mathrm{QY}(\mathrm{Cy3}) \cdot \mathrm{Qn}(\mathrm{Cy3}) \cdot \mathrm{PH}(\mathrm{Cy3}) \cdot \mathrm{In}(\mathrm{Cy3}, i) \cdot s_i \end{aligned}$$

Equation (1) and (2) can be then expressed as:

$$\begin{aligned} M_i &= \log_2\left(\frac{R_i}{G_i}\right) = \log_2\left(\frac{s_i}{r_i} \cdot \frac{\mathrm{QY}(\mathrm{Cy5}) \cdot \mathrm{Qn}(\mathrm{Cy5}) \cdot \mathrm{PH}(\mathrm{Cy5})}{\mathrm{QY}(\mathrm{Cy3}) \cdot \mathrm{Qn}(\mathrm{Cy3}) \cdot \mathrm{PH}(\mathrm{Cy3})} \cdot \frac{\mathrm{In}(\mathrm{Cy5}, i)}{\mathrm{In}(\mathrm{Cy3}, i)}\right) = \log_2\left(\frac{s_i}{r_i}\right) + c_i, \\ M'_i &= \log_2\left(\frac{R'_i}{G'_i}\right) = \log_2\left(\frac{r_i}{s_i} \cdot \frac{\mathrm{QY}(\mathrm{Cy5}) \cdot \mathrm{Qn}(\mathrm{Cy5}) \cdot \mathrm{PH}(\mathrm{Cy5})}{\mathrm{QY}(\mathrm{Cy3}) \cdot \mathrm{Qn}(\mathrm{Cy3}) \cdot \mathrm{PH}(\mathrm{Cy3})} \cdot \frac{\mathrm{In}(\mathrm{Cy5}, i)}{\mathrm{In}(\mathrm{Cy3}, i)}\right) = -\log_2\left(\frac{s_i}{r_i}\right) + c'_i, \end{aligned}$$

from which is clear that $c_i \sim c'_i$. Although the functions $f(\bullet)$, $g(\bullet)$ may not be linear and more factors can be influencing the difference between Cy3 and Cy5, the example proposed here proves the assumption that $c_i \sim c'_i$.

If any random error occurred or the PMT settings were not set to the same value the difference in the medians of the two slides in the formula proposed by Kerr et al. (2000) (4) would account for it and *dye-swap* normalization would still work as probed all through the text.

Acknowledgements. F.S.C. was supported by a BBSRC studentship. A.P. was supported by the Austrian GEN-AU project GOLD, grant GZ 200.059/6-VI/2/2002. G.G.T. and R.P. were supported by the Austrian GEN-AU project BIN, grant GZ 200.067/4VI/1/2002. The collaboration between F.S.C. and A.P., G.G.T and R.P. was supported by the EU Marie Curie Training Site Program "Genomics of Lipid Metabolism. The M.tuberculosis microarray studies were supported by grant 062511/Z/00/Z awarded to P.B by The Wellcome Trust under its Functional Genomics Resources Initiative to fund the BμGαS (the Bacterial Microarray Group at St Georges Hospital Medical School) multi-collaborative microbial pathogen microarray facility. The Mig1p/ Mig2p micro-array studies were supported largely by a Wellcome Trust project grant 061867/Z/00/Z to M.P.A. L.E.A.H. and F.S.C. are supported by an MRC studentship. S.G.C. is supported by a Wellcome Trust project grant 067328/Z/02/Z. K.-H. Cho acknowledges the support received by the Korean Ministry of Science and Technology (Korean Systems Biology Research Grant, M10309000006-03B5000-00211). O.W. acknowledges the funding received by the U.K. Department for the Environment, Food and Rural Affairs (DEFRA) as part of the M.Bovis post genomics programme, conducted in collaboration with the Veterinary Laboratories Agency (VLA), Weybridge.

References

Ashe, M., De Long, S. and Sachs, A. 2000. Glucose depletion rapidly inhipbits translation initiation in yeast. *Molecular Biology of the Cell* **11**, 833–848.

AxonInstruments: 1995-2004. http://www.axon.com. Accessed 29 February 2004.

Benes, V. and Muckenthaler, M. 2003. Standardization of protocols in cDNA microarray analysis. *Trends in Biochemical Sciences* **28**, 244–249.

BioDiscovery: 2004. ImaGene website. http://www.biodiscovery.com/imagene.asp. Accessed 29 February 2004.

Black, M. and Doerge, R. 2002. Calculation of the minimum number of replicate spots required to detection of significant gene expression fold change in microarray experiments. *Bioinformatics* **18**, 1609–1616.

Campbell, S., Holmes, L. and Ashe, M.: *Manuscript in preparation,* 2004.

Churchill, G. 2002. Fundamentals of experimental design for cDNA microarrays. *Nature Genetics Supplement* **32**, 490–495.

Cleveland, W. 1979. Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association* **74**, 829–836.

Dobbin, J., Shih, J. and R., S. 2003. Calculation of the minimum number of replicate spots required to detection of significant gene expression fold change in microarray experiments. *Bioinformatics* **19**(7), 803–810.

Eisen, M. and Brown, P. 1999. DNA arrays for analysis of gene expression. *Methods Enzymol.* **303**, 179–205.

Gentleman, R., Rossini, R., Dudoit, S. and Hornik, K.: 2003. The bioconductor FAQ. http://www.bioconductor.org. Accessed 29 February 2004.

Kepler, T., Crosby, L. and Morgan, K. 2002. Normalization and analysis of DNA microarray data by self-consistency and local regression. *Genome Biology* **3**(7), research0037.1–0037.12.

Kerr, K. and Churchill, G. 2001. Experimental design for gene expression microarrays. *Biostatistics* **2**, 183–201.

Kerr, K., Martin, M. and Churchill, G. 2000. Analysis of variance for gene expression microarray data. *Journal of Computational Biology* **7**, 819–837.

Marquardt, D. 1963. An algorithm for least squares-estimation of nonlinear parametrs. *Journal of the Society for Industrial and Applied Mathematics* **11**, 431–441.

MWG-Biotech: 2004. Affymetrix 428^{TM} array scanner. http://www.mwg-biotech.com. Accessed 29 February 2004.

Pieler, R., Sanchez-Cabo, F., Hackl, H., Thallinger, G. and Trajanoski, Z.: 2004. Arraynorm: Comprehensive normalization and analysis of microarray data. In Press. *Bioinformatics.*

Quackenbush, J. 2001. Computational analysis of microarray data. *Nature Reviews Genetics* **2**(6), 418–427.

Rolland, F., Winderickx, J. and Thevelein, J. 2002. Glucose-sensing and signalling mechanisms in yeast. *FEMS Yeast Research* **2**(2), 183–201.

Sanchez-Cabo, F., Cho, K., Trajanoski, Z. and Wolkenhauer, O.: 2003. A graphical user interface to normalize microarray data. *DSC 2003.*

Schena, M.: 2002. *Microarray Analysis.* Wiley-Liss.

Schulze, A. and Downward, J. 2001. Navigating gene expression using microarrays - A technology review. *Nature Cell Biology* **3**, 190–195.

Talaat, A., Howard, S., Hale IV, H., Lyons, R., Garner, H. and Johnston, S. 2002. Genomic DNA standards for gene expression profiling in *Mycobacterium tuberculosis. Nucleic Acids Research* **30**(20), e104.

Tseng, G., Oh, M., Rohlin, L., Liao, J. and Wong, W. 2001. Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Research* **29**(12), 2549–2557.

Tsodikov, A., Szabo, A. and Jones, D. 2002. Adjustments and measures of differential expression for microarray data. *Bioinformatics* **18**, 251–260.

van de Peppel, J., Kemmeren, P., van Bakel, H., Radonjic, M., van Leenen, D. and Holstege, F. 2003. Monitoring global messenger RNA changes in externally controlled microarray experiments. *EMBO Rep* **4**, 387–393.

Workman, C., Jensen, L., Jarmer, H., Berka, R., Gautier, L., Nielsen, H., Saxild, H., Nielsen, C., Brunak, S. and Knudsen, S. 2002. A new non-linear normalization method for reducing variability in DNA microarray experiments. *Genome Biology* **3**(9), research0048.1–0048.16.

Yang, Y., Dudoit, S., Lin, D., Peng, V., Ngai, J. and Speed, T. 2002. Normalization for cDNA microarray data: A robust composite method adressing single and multiple slide systematic variation. *Nucleic Acids Research* **30**(4), e15.1–e15.10.

Yang, Y. H., Dudoit, S., Luu, P. and Speed, T. P.: 2001. Normalization for cDNA microarray data. *SPIE BIOS 2001.*