Is \textit{LOWESS} a Panacea in the Normalization of Microarray Data?

Fatima Sanchez-Cabo  
Department of Biomolecular Sciences, UMIST  
P.O. Box 88, Manchester M60 1QD, U.K.

Institute of Biomedical Engineering and  
Christian Doppler Laboratory for Genomics and Bioinformatics,  
Graz University of Technology,  
8010 Graz, Austria

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

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

Zlatko Trajanoski  
Institute of Biomedical Engineering and  
Christian Doppler Laboratory for Genomics and Bioinformatics,  
Graz University of Technology, 8010 Graz, Austria

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.

July 9, 2003

*To whom correspondence should be addressed.
Abstract

Microarrays are a powerful tool in functional genomics, allowing to monitor the expression level of thousands of genes simultaneously and over time. In order to compare measurements within and across arrays and to correct non-biological variation masking meaningful information, the normalization of the data is an unavoidable task prior to any further analysis. The search for universal normalization techniques, valid for all kind of experiments, is a central problem in the microarray community.

This paper will demonstrate the influence of the normalization in the conclusions of the data analysis. To this end, we compare two well established techniques to correct the bias that the different properties of the dyes introduce in two-color microarrays: LOWESS and dye-swap normalization (also referred to as self-normalization). Although LOWESS has become a default option in several software packages, we demonstrate here the limitations and drawbacks of such normalization method when compared to dye-swap normalization. In the first place, the kind of experiments to which LOWESS can be applied is quite restricted. Furthermore, the replicated measurements showed worse correlation after LOWESS within-array normalization than after dye-swap normalization. The increase in the within replicates variability had an immediate effect in the further analysis, in particular, in the genes detected as differentially expressed.

The paper also discusses in detail the assumptions that a particular experiment must satisfy to apply an appropriate normalization of the data. We provide a guide of the normalization method that should be applied, depending on the experimental design. *

Keywords: Two-color microarrays, normalization, experimental design, LOWESS, dye-swap.

*All the figures and results presented in this paper were implemented using the software MATLAB® (Mathworks Inc.). A collection of files is available from http://www.sbi.uni-rostock.de
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, mRNA (messenger RNA) from the populations of interest is reversely transcribed and labelled using two different fluorescent dyes (usually Cyanine dyes, Cy3 and Cy5). Afterwards, both samples (related to the “channels” of the scanner used to read the array) are hybridized onto the microarray, where PCR (Polymerase Chain Reaction) 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 fluor. This generates for each channel a high resolution image, which is then analyzed in a process referred to as “spot finding”. The spots are quantified into single intensity values for each channel for every gene spotted on the array. 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 has then the option to correct the data by, for example, subtracting the background from the foreground.

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 either systematic or random errors, is an essential stage prior to the analysis and biological interpretation of the data. In two-color microarray experiments, an important source of systematic error is the so-called dye effect. The different properties of both dyes, including their different range of the excitation spectra and their gene specific incorporation properties, 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 to the same range. This process is often denoted as “within array” normalization (Luu et al. 2001, Yang et al. 2002). As defined by Kepler et al. (2002) there are two strategies that can be employed to this end:

- Within array normalization by self-consistency using all the 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 (Luu et al. 2001, Yang et al. 2002), the use of a LOWESS function (Cleveland 1979) correcting intensity-dependent data (Luu 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 slide is the most robust and popular.

- Within array 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, the use of replicated arrays and the swap of the dyes for replicated arrays. The latter is a requirement to apply dye-swap normalization.

In this paper, the two approaches for within-array normalization are contrasted through the comparison of LOWESS and dye-swap normalization. Both methods were applied to the normalization of a growth curve experiment for *M. tuberculosis*. Their effect on the data was assessed according to different criteria: Biological validation, correlation among the replicated measures and genes detected as differentially expressed. Due to the use of gDNA (genomic DNA) in the reference channel, and in the absence of *a priori* knowledge, self-consistency normalization using all the genes in the array could be falsely transforming the data, while dye-swap normalization corrected the dye effect and improved the reliability of our results.

The application of dye-swap normalization was possible owing to an intelligent design that provided not just biological replicates but replicates for which the dyes were swapped. In general, there is a close relationship between normalization and design of experiments. The latter is fortunately gaining importance in microarray studies (Churchill 2002, Kerr and Churchill 2001, Yang and Speed 2002).

The paper is organized as follows. Firstly, the three main self consistency methods are discussed. These are the global method, LOWESS (Luu 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 the dye-swap method is explained in detail. Once the two main approaches to within array normalization have been described, it is discussed in Section 4 the type of experiments for which they are
suitable. A table that summarizes the most appropriate normalization method regarding to the characteristics of the experiment is presented. To conclude, in Section 5, LOWESS and dye-swap normalization are applied to the *M.tuberculosis* growth curve experiment. The correlation of the replicates and the genes detected as differentially expressed after correcting the data using both methods will be used to draw conclusions about their reliability.

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 compared samples. 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 (Luu 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

\[
\text{median}_{i=1,\ldots,n_{g}} \left( \frac{R_{i}}{G_{i}} \right) \cong 1 \Rightarrow \log_{2} \left( \text{median}_{i=1,\ldots,n_{g}} \left( \frac{R_{i}}{G_{i}} \right) \right) \cong 0,
\]

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 \( \xi \) (Luu et al. 2001, Yang et al. 2002), as

\[
R = \xi \cdot G.
\]

The different estimators of \( \xi \) 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

\[ \xi = \text{median}_{i=1,\ldots,n_g} \left( \frac{R_i}{G_i} \right). \]

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**\(^\dagger\) 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 Luu et al. (2001). From the scatter plot \((A,M)\), where

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

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 for 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:

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

\(^\dagger\)LOcally WEighted leaSt Squares (LOWESS)
Regardless to the method used to estimate $\xi$, 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:

$$R_i^c = R_i,$$

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

Because the dye effect appears to be intensity dependent in most of the cases (Luu et al. 2001, Yang et al. 2002, Workman et al. 2002), LOWESS has become a popular method for within-array normalization. Whilst the global dye correction method transforms all the genes using a unique value for every slide and the regression method is highly sensitive to outliers, the LOWESS approach appears as the most suitable option to reduce the effect of the different properties 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. Yet, they are not suitable for all those experiments for which the assumption of most genes being equally expressed in both channels is not valid. 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, gDNA, “spiked genes”, or a Microarray Sample Pool (MSP) (Yang et al. 2002). The latter is becoming a common practice. 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 data (using for example the Levenberg-Marquardt algorithm) 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 is not just a procedure to remove random errors introduced in the experiment, but also a requirement to provide statistical significance of differences in gene expression (Black and Doerge 2002). In addition, an experiment providing replicates for which the dyes have been swapped allows us to easily correct the dye effect. Dye-swap normalization was first suggested in Luu et al. (2001) under the name “self normalization”.

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

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

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

For this expressions, \( c_i \) and \( c'_i \) account for the different properties of the dyes. Because \( c_i \simeq c'_i \) (see Appendix A for an explanation), adding both equations, yields

\[
M_i - M'_i \simeq 2 \cdot \log_2 \left( \frac{s_i}{r_i} \right) \implies \frac{1}{2} \cdot (M_i - M'_i) \simeq \log_2 \left( \frac{s_i}{r_i} \right).
\]
Subtracting them, we have

\[(M_i + M_i') - 2 \cdot c_i = 0 \implies c_i = \frac{1}{2} \cdot (M_i + M_i'),\]

where \(c_i\) can be estimated from the data plotted in the scatter plot \(\frac{1}{2} \cdot (A + A')\) vs. \(\frac{1}{2} \cdot (M + M')\) (Luu et al. 2001).

The main advantage of the dye-swap normalization is that it transforms the data preserving the characteristics of every gene. Note also that the computational cost for the implementation of this method is very low.

4 Normalization and experimental design

Following the description of the two main within-array normalization methods, this section provides a discussion about their application and universality. As already mentioned in Section 3, LOWESS correction will be only suitable for those designs in which most of the genes are expected to be equally expressed in both compared samples. Otherwise, the assumptions made are not valid, and a dramatic transformation of the data will lead to erroneous conclusions.

These are some examples of microarray experiments for which the previous statement does not always hold: Experiments for which a great number of genes is expected to appear differentially expressed, experiments with a reference design (Kerr and Churchill 2001, Yang and Speed 2002) and experiments for which no a priori information is available.

Within the first category, “low-density” microarrays are becoming common due to the advance of microarray technology. In these arrays just target genes detected in previous experiments are spotted. Hence, most of the genes in the array are expected to change in expression and self-consistency methods are not suitable to normalize the data generated from them.

For reference designs, correction through LOWESS is usually not appropriate. For this kind of experiment, a biological bias may appear for those genes that are not expressed for a particular biological condition but that have always a positive reading in the reference channel, often constant across the slide. The self-consistency normalization methods assume that all genes are equally expressed and would correct this real bias as if it was an effect due to dye properties.
There are advantages and disadvantages in the use of a reference design. For instance, the microarray community aims to be able to compare results from different microarray experiments and across many different biological conditions. To this end, the establishment of a common reference for all experiments of the same organism would broaden the possibilities of microarray technology. An additional and common problem in microarrays is the high number of missing values resulting from low signals in one of the hybridized samples. Setting a reference that provides homogeneous and overall strong signals for all spots in the array, increases the quality of the experiment and allows the use of most of the spotted genes for further analysis. See Talaat et al. (2002) for a more extensive discussion of the use of a reference. However, the reference design has also disadvantages. For example, two measurements are needed to provide the same information than one single log ratio gives in the loop design. In these experiments, the two populations of interest are directly compared in the same array (Kerr et al. 2000). In spite of the possible disadvantages, the use of reference designs is becoming common practice and it is therefore important to carefully assess for every particular case if the assumption implicit in the self-consistency methods holds.

Table 1 shows the importance of setting a good experimental design, introducing some of the quality control elements described in this paper (controls, replicated genes within the slide, replicated slides, replicated slides with dyes swapped). Despite the increasing cost of the experiment, it will lead to more reliable results.

Table 1: Summary of the suitable dye correction methods according to the design of the experiment. If dye-swapped replicates are not available, controls, replicated slides or replicated genes within the slide can be used.

<table>
<thead>
<tr>
<th>Expected Results</th>
<th>Most genes equally expressed</th>
<th>Unknown</th>
<th>Most genes differentially expressed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Loop Design</td>
<td><em>LOWESS</em></td>
<td>controls, dye-swap</td>
<td>global method (25% or 75%)</td>
</tr>
<tr>
<td>Reference Design</td>
<td>dye-swap / controls / reference channel</td>
<td></td>
<td>dye-swap</td>
</tr>
</tbody>
</table>
5 Results and discussion

Description of the experiment

Dye-swap normalization and *LOWESS* correction were applied to the same experiment in order to correct the dye effect and to investigate how both methods transform the data. The aim of the experiment was to study 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* (four replicated arrays for each RNA sample of the four time points) and using gDNA for the “reference” channel. The advantage of this reference design is that all genes in the genome are presented in the gDNA. Hence, every gene should give a homogeneous signal for the denominator of the ratio of both channels. A broader discussion 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 $a = 1, 2, \ldots, 16$ the number of the array, the experiment can be summarized as

\[
\begin{align*}
\text{for } a \neq 4, 8, 11, 16 & \quad \{ \text{Green} : \text{RNA (signal)}, \quad \text{Red} : \text{gDNA (reference)}, \\
\text{for } a = 4, 8, 11, 16 & \quad \{ \text{Green} : \text{gDNA (reference)}, \quad \text{Red} : \text{RNA (signal)}. 
\end{align*}
\]

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 copy of rRNA, so equal in abundance to the other genes in the genome, RNA presents 98% rRNA and just 2% mRNA. Hence, much more RNA hybridized to the control spots than to the rest of the gene spots. In addition, the range of intensities presented by the the control spots did not cover the whole intensity...
range in which the rest of the spots was expressed. The control spots were for those reasons excluded from the analysis and all the results in this paper refer to the 3924 printed genes. Although there were not duplicated genes in 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.

The use of gDNA reference made feasible the use of all the genes printed in the array 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 this analysis was done with the normalization module of the program MADE (Sanchez-Cabo et al. 2003).

After the background analysis, the first approach to within-slide-normalization was to use all the genes in the array. Among the self-consistency methods, LOWESS correction was chosen due to the intensity dependence for low values suggested by the \((A,M)\) scatterplot. This is illustrated in Figure 1.

![LOWESS function for array 1](image1.png) ![Corrected scatterplot for array 1](image2.png)

(a) LOWESS function fitted to the raw data. (b) Corrected data using the LOWESS function.

Figure 1: LOWESS correction for the first array.
The data set was transformed according to the preliminary assumption that most of the genes would be equally expressed in both channels, as observed in Figure 3(a). However, this assumption might not be true for all arrays involved in this reference experiment. Dye-swap normalization was then applied, employing the information from the replicates for which the dyes had been swapped. Both normalization methods were then compared according to three assessment criteria: Biological validation, correlation among replicated measures and differentially expressed genes.

**Biological validation**

To study the overall response of the genes at a particular time point, the distribution of the log ratios for every array were compared beforehand (Figure 2) and after the two within array normalization methods (Figure 3). After *LOWESS* normalization, all distributions were perfectly centered around zero (Figure 3(a)). However, after dye-swap normalization, the arrays measuring expression level at the first time point presented a log ratios distribution still centered around a negative value (Figure 3(b)). Because the experiment was analyzing a growth curve, this bias could have a biological meaning. The reference channel provided a positive reading for every gene, while the signal channel could be showing a large number of genes still not expressed.

Without any further information about the proportion of genes expected to be differentially expressed with respect to the gDNA at every time point, we could not assert that the *LOWESS* function corrected the data in the right way. However, if the conditions in which the dye swap was applied were carefully observed (i.e., same amount of initial material, same gain set to scan the slides) the correction using dye-swap normalization would be more reliable since no assumptions were made to apply this second method to the data.

**Correlation among replicated measures**

In the presence of replicated arrays or replicated genes within an array, a unique representative value of the expression level for every gene under a particular biological condition must be obtained at the end of the normalization process. This value will then be used for further analysis (clustering, classification, detection of genes differentially expressed,
Figure 2: Distribution of the log-ratios of the 16 arrays of the *M.tuberculosis*, before within array normalization. Different line-types represent the four replicates at the same time point.

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.
After normalization, the variability of the measurements should be due just to the biological mechanism that we aim to understand. Hence, the difference in the expression level of one gene across technical replicates (replicated slides with the same hybridized material or replicated spots within a slide) should be minimal. The study of the variability of the replicated measures appears as a reliable method to test the effect of the different normalization methods on the data. Tseng et al. (2001) and Huber et al. (2002) refer to the coefficient of variation (CV) as an appropriate method to test the quality of replicated microarray experiments. According to that, we studied the coefficient of variation for the replicated measures at each of the time points after LOWESS and dye-swap normalization. We calculated:

\[
CV_t = \frac{\text{std}(X_{1t}^i, \ldots, X_{nt}^i)}{\text{mean}(X_{1t}^i, \ldots, X_{nt}^i)},
\]

where \( t = 6 \) days, 14 days, 20 days or 30 days, \( n_r \) is the number of replicated slides per time point and \( X_{jt}^i = \frac{R_{jt}^i}{G_{jt}^i} \) (\( j^{th} \) replicated expression level of gene \( i \) at time \( t \)). Figure 4 shows the results. Dye-swap normalization performs better for all the time points. Just 37 genes (0.94% of the arrayed genes) present a CV greater than 0.5. In addition, just four of them have an extremely high CV. However, after LOWESS normalization 323 genes presented a CV greater than 0.5 in at least one time point. This is the 8.23% of the whole data set. The supplementary material, available from http:www.sbi.uni-rostock.de, demonstrates how the genes with poor correlation among replicates, are not necessarily those with low intensity signals.

Another way to test the improvement in correlation among the replicated measurements per time point is looking at the hierarchical clustering of the replicates. Hierarchical clustering among the replicates will give an idea of the closeness of the replicated measurements. As shown in Figure 5, the distance among the replicates after dye-swap normalization is much smaller than after LOWESS normalization. Furthermore, replicates from the same time point appear ordered and very close to each other, while they are all mixed after LOWESS normalization.

To conclude the analysis of the replicates after both normalization methods, we focussed in the two IS6110 transposase family elements. As previously mentioned, there are
Figure 4: Mean-CV scatterplot for each one of the four time points. The mean and standard deviation of the four replicated ratios at every time point were calculated after the two normalization methods.

(a) CV after LOWESS normalization.  
(b) CV after dye-swap normalization.

Figure 5: Hierarchical clustering of the replicates. After dye-swap normalization only three replicates per time point remain. Except the first replicate at the second time point, replicates of the same time point appear very close together. After LOWESS normalization, the overall distance has increased and the replicates of the same time point do not cluster together as after dye-swap normalization.
sixteen copies of each, making a total of 32 PCR products that should be almost identical after normalization of the data. Table 2 shows the mean, standard deviation (STD) and CV of those genes in every slide after normalization of the data using dye-swap normalization and LOWESS normalization. The mean of the CV of the different IS6110 elements across all arrays is better after dye-swap than after LOWESS normalization. It is also striking how large the CV of the IS6110 elements is in the fourth replicate measured at the third time point. In contrast, the coefficient of variation is smaller than 0.4 for all the arrays after dye-swap normalization.

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).

<table>
<thead>
<tr>
<th>(time, replicate)</th>
<th>After LOWESS normalization</th>
<th>Mean</th>
<th>STD</th>
<th>CV</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1,1)</td>
<td></td>
<td>0.7803</td>
<td>0.3484</td>
<td>0.4465</td>
</tr>
<tr>
<td>(1,2)</td>
<td></td>
<td>0.7208</td>
<td>0.1229</td>
<td>0.1703</td>
</tr>
<tr>
<td>(1,3)</td>
<td></td>
<td>0.8858</td>
<td>0.3439</td>
<td>0.3882</td>
</tr>
<tr>
<td>(1,4)</td>
<td></td>
<td>1.72</td>
<td>0.4627</td>
<td>0.2687</td>
</tr>
<tr>
<td>(2,1)</td>
<td></td>
<td>0.8803</td>
<td>0.04</td>
<td>0.0454</td>
</tr>
<tr>
<td>(2,2)</td>
<td></td>
<td>1.0282</td>
<td>0.2956</td>
<td>0.2875</td>
</tr>
<tr>
<td>(2,3)</td>
<td></td>
<td>0.9927</td>
<td>0.3579</td>
<td>0.3579</td>
</tr>
<tr>
<td>(2,4)</td>
<td></td>
<td>1.2688</td>
<td>0.1574</td>
<td>0.1240</td>
</tr>
<tr>
<td>(3,1)</td>
<td></td>
<td>0.9637</td>
<td>0.1400</td>
<td>0.1453</td>
</tr>
<tr>
<td>(3,2)</td>
<td></td>
<td>0.9860</td>
<td>0.1988</td>
<td>0.2016</td>
</tr>
<tr>
<td>(3,3)</td>
<td></td>
<td>1.2529</td>
<td>0.0988</td>
<td>0.0788</td>
</tr>
<tr>
<td>(3,4)</td>
<td></td>
<td>1.1744</td>
<td>0.8472</td>
<td>0.7214</td>
</tr>
<tr>
<td>(4,1)</td>
<td></td>
<td>0.9049</td>
<td>0.0812</td>
<td>0.0897</td>
</tr>
<tr>
<td>(4,2)</td>
<td></td>
<td>0.9714</td>
<td>0.1972</td>
<td>0.2030</td>
</tr>
<tr>
<td>(4,3)</td>
<td></td>
<td>1.0401</td>
<td>0.2259</td>
<td>0.2172</td>
</tr>
<tr>
<td>(4,4)</td>
<td></td>
<td>1.4619</td>
<td>0.2996</td>
<td>0.2049</td>
</tr>
<tr>
<td>overall mean</td>
<td></td>
<td></td>
<td></td>
<td>0.2469</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>(time, replicate)</th>
<th>After dye-swap normalization</th>
<th>Mean</th>
<th>STD</th>
<th>CV</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1,1)</td>
<td></td>
<td>0.3197</td>
<td>0.0737</td>
<td>0.2305</td>
</tr>
<tr>
<td>(1,2)</td>
<td></td>
<td>0.4122</td>
<td>0.0765</td>
<td>0.1855</td>
</tr>
<tr>
<td>(1,3)</td>
<td></td>
<td>0.4188</td>
<td>0.1672</td>
<td>0.3991</td>
</tr>
<tr>
<td>(2,1)</td>
<td></td>
<td>0.7312</td>
<td>0.0757</td>
<td>0.1035</td>
</tr>
<tr>
<td>(2,2)</td>
<td></td>
<td>0.9066</td>
<td>0.1127</td>
<td>0.1243</td>
</tr>
<tr>
<td>(2,3)</td>
<td></td>
<td>0.9776</td>
<td>0.1253</td>
<td>0.1281</td>
</tr>
<tr>
<td>(3,1)</td>
<td></td>
<td>0.8421</td>
<td>0.0718</td>
<td>0.0852</td>
</tr>
<tr>
<td>(3,2)</td>
<td></td>
<td>0.8516</td>
<td>0.0949</td>
<td>0.1114</td>
</tr>
<tr>
<td>(3,3)</td>
<td></td>
<td>0.9034</td>
<td>0.2849</td>
<td>0.3153</td>
</tr>
<tr>
<td>(4,1)</td>
<td></td>
<td>0.7785</td>
<td>0.1251</td>
<td>0.1607</td>
</tr>
<tr>
<td>(4,2)</td>
<td></td>
<td>0.7715</td>
<td>0.1378</td>
<td>0.1787</td>
</tr>
<tr>
<td>(4,3)</td>
<td></td>
<td>0.8344</td>
<td>0.2773</td>
<td>0.3323</td>
</tr>
<tr>
<td>overall mean</td>
<td></td>
<td></td>
<td></td>
<td>0.1962</td>
</tr>
</tbody>
</table>

Differentially expressed genes

Hoffmann et al. (2002), described how after different normalization methods, applied to oligonucleotide arrays, different groups of genes are detected as differentially expressed. Among all the possible aims of microarray experiments, the detection of differentially expressed genes under different biological conditions is one of the simplest and most common objectives (Luu et al. 2001, Dudoit et al. 2002, Kerr et al. 2000, Long et al. 2001).
To compare the expression level of a particular gene across different conditions, a t-statistic can be calculated for every gene $i$ and in two biological conditions $c_1$, $c_2$ (Luu et al. 2001):

$$t_{ic_1c_2} = \frac{\bar{x}_{ic_1} - \bar{x}_{ic_2}}{\sqrt{\frac{s^2_{ic_1}}{n_{c_1}} + \frac{s^2_{ic_2}}{n_{c_2}}}},$$

where

$$\bar{x}_{ic_1} = \frac{1}{n_{c_1}} \sum_{j=1}^{n_{c_1}} x_{ij} = \frac{1}{n_{c_1}} \sum_{j=1}^{n_{c_1}} \log_2 R_{ij} G_{ij},$$

and

$$s^2_{ic_1} = \frac{1}{n_{c_1} - 1} \sum_{j=1}^{n_{c_1}} (x_{ij} - \bar{x}_{ic_1})^2.$$

From (3) it is immediate to conclude that the standard error of the replicated measurements is essential to detect differentially expressed genes. The lower the correlation of the replicated measurements for every gene at every biological condition, the higher the value for $s_{ic_1}$ and $s_{ic_2}$. According to (3), large values for $s_{ic_1}$ and $s_{ic_2}$ will result in a small value of $t_{ic_1c_2}$, independently on the difference of means $(\bar{x}_{ic_1} - \bar{x}_{ic_2})$. In consequence, some genes that present a significant difference among their mean values will have a small t-statistic and will not be detected as differentially expressed due to the large across replicates variability.

The precise subset of genes differentially expressed between two biological conditions in an experiment must be obtained calculating the p-values of the t-statistics. Parametric methods are often not accurate to this end since microarray data do not usually follow a normal or symmetric distribution. Dudoit et al. (2002) suggest several methods to adjust the p-values. However, because the aim of this paper is not to detect reliable genes differentially expressed but to prove the effect of the two compared normalization methods, we tried just the visual method proposed in (Dudoit et al. 2002). Since one of the targets of the experiment was to identify genes that change significatively from the first to the last measured time point, we compared the QQ-plots of the sample of $t_{14}$-statistic values. As expected from the results of the correlation analysis, the number of genes that present odd t-statistics (i.e. potential differentially expressed genes) was much larger after dye-swap normalization than after LOWESS normalization (see Figure 6). This supports the thesis that genes that might present a significant difference between the mean of their expression
levels are not detected as differentially expressed after *LOWESS* normalization due to the poor correlation among replicates.

![QQ-plot of the t-statistics after LOWESS normalization.](image)

(a) QQ-plot of the t-statistics after *LOWESS* normalization.

![QQ-plot of the t-statistics after dye-swap normalization.](image)

(b) QQ-plot of the t-statistics after dye-swap normalization.

Figure 6: QQ-plot displaying the quantiles of the sample of t-statistics against the quantiles of a standard normal distribution. The sample of t-statistics was calculated to estimate the change in the expression level of every gene from time point 1 to time point 4.

### 6 Conclusions

The use of the *LOWESS* function to correct the dye effect has been lately emphasized, appearing as the default option in different software analysis packages (e.g. Gene Spring (SiliconGenetics Gene Spring website. 2001)). In this paper it was argued that there are disadvantages associated with this approach and how the kind of experiment conducted and the experimental design are two important considerations when choosing a normalization method.

*LOWESS* normalization implies fairly strict assumptions that make it difficult to become a widespread method, applicable to all kind of microarray experiments. We described the experiments for which *LOWESS* could be applied and showed that it is a high risk strategy in experiments for which no *a priori* knowledge about the expected proportion of differentially expressed genes is available. For “low-density” microarrays it is unlikely to be suitable and for a reference design experiment, the experimental conditions must be carefully observed before application of this method.
Furthermore, the effect of LOWESS on the correlation of the replicated measurements is an important consideration. For the data set presented in this paper, almost ten percent of the data showed poor reproducibility in at least one time point after the application of LOWESS. It would be desirable to have a normalization method that reduces the across replicates variability, increasing the reliability of the results of the analysis of the data. For example, genes that present a very large variability for their replicated values will not be detected as differentially expressed even when the difference between the mean value across the compared biological conditions is significant.

For the \textit{M.tuberculosis} experiment, the use of replicates for which the dyes had been swapped allowed a proper normalization of the data. Just one percent of the genes showed poor correlation of the replicated slides after the application of this method. Hierarchical clustering on the replicates and the study of the replicated spots within the slide demonstrated also a better performance of the dye-swap normalization method.

In general, the use of quality control elements in the experiment can improve the normalization of the data compared to the normalization of the data by self-consistency. In spite of the limitations of material and slides, the conclusions inferred in this paper should encourage the experimenter to invest more resources in experiments that provide reliable data. For that, it is essential to provide the elements that are necessary for a proper normalization.

\textbf{Appendix A: Different properties of Cy3 and Cy5}

The basic assumption made in the dye-swap normalization method, is that $c_i \simeq c_i'$. This can be explained as follows.

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:

Quantum Yield : \( QY(dye, gene, sample) = QY(dye) \)

Quenching : \( Qn(dye, gene, sample) = Qn(dye) \)

Photobleaching : \( PH(dye, gene, sample) = PH(dye) \)

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 \( \times \) 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:

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:

\[
R_i = f(s_i) = QY(Cy5, i, s) \cdot Qn(Cy5, i, s) \cdot PH(Cy5, i, s) \cdot In(Cy5, i, s) \cdot s_i \\
= QY(Cy5) \cdot Qn(Cy5) \cdot PH(Cy5) \cdot In(Cy5, i) \cdot s_i
\]

\[
G_i = g(r_i) = QY(Cy3, i, r) \cdot Qn(Cy3, i, r) \cdot PH(Cy3, i, r) \cdot In(Cy3, i, r) \cdot r_i \\
= QY(Cy3) \cdot Qn(Cy3) \cdot PH(Cy3) \cdot In(Cy3, i) \cdot r_i
\]
The same is true for $R'_i$ and $G'_i$:

\[
R'_i = f'(r_i) = QY(Cy5, i, r) \cdot Qn(Cy5, i, r) \cdot PH(Cy5, i, r) \cdot In(Cy5, i, r) \cdot r_i
\]

\[
= QY(Cy5) \cdot Qn(Cy5) \cdot PH(Cy5) \cdot In(Cy5) \cdot r_i
\]

\[
G'_i = g'(s_i) = QY(Cy3, i, s) \cdot Qn(Cy3, i, s) \cdot PH(Cy3, i, s) \cdot In(Cy3, i, s) \cdot s_i
\]

\[
= QY(Cy3) \cdot Qn(Cy3) \cdot PH(Cy3) \cdot In(Cy3) \cdot s_i
\]

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

\[
M_i = \log_2 \left( \frac{R_i}{G_i} \right) = \log_2 \left( \frac{s_i}{r_i} \cdot \frac{QY(Cy5) \cdot Qn(Cy5) \cdot PH(Cy5)}{QY(Cy3) \cdot Qn(Cy3) \cdot PH(Cy3)} \cdot \frac{In(Cy5, i)}{In(Cy3, i)} \right) = \log_2 \left( \frac{s_i}{r_i} \right) + c_i,
\]

\[
M'_i = \log_2 \left( \frac{R'_i}{G'_i} \right) = \log_2 \left( \frac{r_i}{s_i} \cdot \frac{QY(Cy5) \cdot Qn(Cy5) \cdot PH(Cy5)}{QY(Cy3) \cdot Qn(Cy3) \cdot PH(Cy3)} \cdot \frac{In(Cy5, i)}{In(Cy3, i)} \right) = -\log_2 \left( \frac{s_i}{r_i} \right) + c'_i,
\]

from which is clear that $c_i \sim c'_i$. Not much work has been published in this direction. 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 $c_i \sim c'_i$. A work with real data to prove this thesis is in progress.

**Acknowledgements.** The authors would like to thank the Welcome Trust funded Bacterial Microarray Group at St. George’s Hospital Medical School in London, the Streptomyces group at UMIST, and the Bioinformatics group in the Biomedical Engineering Institute at the Technical University of Graz, Austria, in particular to Gerhard Thallinger for proof-reading. This collaboration was supported by the EU Marie Curie Training Site Program. We would also like to thank the reviewers for their useful comments.

**References**


