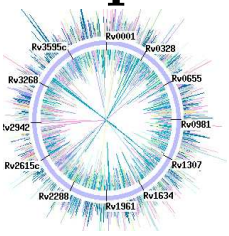
Normalisation and Quality Control for DNA Microarray Data:

Analysis of a *M.tuberculosis* growth curve experiment



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

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

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.

June 17, 2003

* *To whom correspondence should be addressed.*

Contents

1 Introduction

M. tuberculosis is the world leading cause of death by a single infectious factor. It is estimated that in the next 20 years 1 billion people will be newly infected, 200 million will get sick and *M.tuberculosis* will cause 35 million deaths.

The complicated genesis of *M.tuberculosis* makes developing effective treatment for it a continued challenge for researchers. In addition it has been recently discovered the link between tuberculosis and the human immunodeficiency virus.

For those reasons, the sequence of the 4,411,529 base pairs of the *M. tuberculosis* genome was an important step in the understanding of the mechanisms that control this bacteria. The project was developed by the researches of the Wellcome Trust Genome Campus of the Sanger Centre, in collaboration with the Institut Pasteur in Paris. The complete annotation can be found in *Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence* Nature 393, 537-544 (1998).

In an interview by Jennifer Fisher Wilson to Julian Parkhill, a member of the pathogen sequencing group at the Sanger Centre and coauthor of the Nature paper, he said: "We hope that our success in sequencing the tubercle bacillus will lead to better treatment for tuberculosis. Genome sequencing of pathogenic bacteria is basic research (...) it provides a data set that will underpin all future work on these organisms"

His hope is already a reality. With the advance of microarray technology, the possibility of monitoring the response of all genes in the genome of an organism is already a fact. Although microarrays focusses in the transcription level and tells little about the proteins that will be later encoded, they have been already a key player in many discoveries all around the world.

This report focusses in the different methods developed by the Systems Biology Group to extract valuable information from microarray data. More in particular, we are trying to establish a normalisation protocol as universal as possible. We have showed how some of the available normalisation methods hold very restrictive assumptions that make them impossible to be applied to all different types of microarray experiments. Applying these methods to the data generated by the Bacterial Microarray Group at St George's Hospital Medical School (BμG@S) was very useful to study the effect of the normalisation in the replicated profiles and how this will affect the number of genes detected as differentially expressed. The Bacterial Microarray Group at St George's Hospital Medical School (BμG@S) has been funded by The Wellcome Trust to generate whole genome microarrays for twelve bacterial pathogens, among them *M.tuberculosis*. We hope these methods will be valuable in the future for the achievement of important biological conclusions about the mechanisms that regulate the smaller particle of life: the cell.

The report is structured as follows: In the Section number 2 we describe the flow of the normalisation process, which are the main sources of non-biological variability introduced in microarray experiments and which possibilities are readily available for their correction. This methods were applied to the three biological replicates performed by BμG@S to achieve some conclusions about the growth curve of *M.tuberculosis*. The results obtained are reported in Sections 3,4 and 5. For any of the three biological replicates *LOWESS* and dye-swap normalisation were applied to correct the data. The replicates before and after any of the methods were studied Finally, some conclusions about the detection of genes differentially expressed are also presented.

2 Normalisation methods

2.1 Introduction

Normalisation of Microarray Data is a very important issue unavoidable previous to further analysis. In this pages we will try to summarize all the different possibilities that can be chosen at every stage of the normalisation process. We will see how different combinations of methods provide a very different time series profile for the normalised data. It is important to choose a suitable method to normalise our data reasoning in which way every of the approaches corrects our data and which are the assumptions made for all of them.

The following pages are structured as following:

- In the next section a sketch of the possibilities that can be chosen at every stage of the normalisation process is described.
- Afterwards, we will apply the theoretical approach to normalise the growth curve experiment carried out by the Bacterial Group at St. George's Medical School, London. This experiment has three different biological replicates that were performed in completely independent conditions and that are named according to the period when they were performed. The three time series have a different experimental design, what made not possible to apply every one of the normalisation methods to every one of them.
- For everyone of the time series and according to the different combinations of normalisation methods at every stage, we show the resulting time series.
- To conclude, we show the genes that appear differentially expressed performing a student-t test for the three biological replicates. The error model for the data is presented and the problems that to use a t-test for this data presents are described.

2.2 Sequential method: Options implemented in MADE

2.2.1 Dye correction

After background correction, systematic errors must be corrected. The most important of all of them is the one introduced by the different properties of both fluorescent dyes labelling the two RNA pools. We have detected four properties that are different for both dyes. The most important of them is the lower incorporation rate of Cy5, but as well the quantum yield, the photobleatching and the quenching properties are different. All these differences distort the real intensity values of both channels so they must be balanced previous any further analysis of the data. However, we must be careful at this stage. The most popular dye correction methods are based on the idea that the majority of the genes are equally expressed in both channels. But this is not going to be the case of all experiments. For this reason, two different approaches were implemented in our interface. The correction of the dye effect (as well known as within-array normalisation) can be performed:

- using the whole data set to normalise the data, as well known as self-consistency.
- using the quality control elements provided in the experiment. This includes the dye-swap normalisation, the use of spotted controls or the use of a reference channel.

Dye-effect correction by self-consistency

Assuming that most of the genes are going to be equally expressed in both channels, an expression ξ is estimated to force the overall intensity of both channels to be the same. Both channels intensities would be then related according to the expression:

$$R = \xi \cdot G,$$

Effects corrected		***Options***
Background effect		1. Background subtraction 2. No subtraction
Spatial effect		$p_i = \frac{r_i}{g_i}$
Dye effect	*Using all genes*	1. Global constant 2. Linear regression 3. LOWESS function 4. LOWESS for print-tips
	Quality control elements	1. Dye-swap normalisation 2. Use of spotted controls
Array effect: Across replicates normalisation		1. Against median all arrays 2. Against median value reference channel
Average experimental replicates (slides/spots)		
Array effect: Across samples normalisation		1. Against all arrays 2. Against arrays in J 3. Against median value reference channel
Transformation of the data		1.$\log_2(\bullet)$ transformation 2. $\sqrt{\bullet}$ transformation 3. lin-$\log_2$ transformation 4. arsinh($\bullet$) transformation

where $R \equiv red$ and $G \equiv green$. The estimation of this expression ξ is going to result in many different methods to correct the different properties of the dyes. Four of them can be selected in our interface:

Global normalisation In this case, we assume that the systematical bias due to the different dyes properties is affecting all spotted genes in the array in the same amount. A constant k relating both channels is estimated. If most of the genes are expected to be equally expressed, then a good representative value of the distribution of the ratios is:

$$k = med_i \frac{R_i}{G_i}$$

and $\xi = k$. For experiments for which a high percentage of genes is differentially expressed comparing both channels, the use of the first or third quartiles are more suitable options. The three choices are implemented in our interface.

Linear regression normalisation A regression line is fitted 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.

LOWESS normalisation Looking at the (A,M) plot it can be detected if the distribution of the log ratios depends on the intensity. In this case, it is not appropriated to correct every spot by the same amount as the global method does. At the same time, the linear regression method is very sensitive to outliers, so a more robust alternative is required. For these reasons the use of

a LOWESS function to correct the dye bias is becoming more important in the normalisation of microarray data. (A,M) scatter plot will show:

$$M = \log_2 \frac{S_i}{R_i} \ ,$$
$$A = \frac{1}{2} \cdot (\log_2 S_i + \log_2 R_i) \ .$$

The *LOWESS* function $c(A_i) : \mathrm{I} \mapsto \mathrm{R}$ can be calculated from this plot, where the set of indexes I denotes all genes spotted on the array. The fitting of the *LOWESS* function $c(A)$ from the (A,M) scatterplot leads to:

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

To estimate this function in MATLAB takes extremely long, but we can improve its efficiency using a C function implemented by the Jackson's laboratory. LOWESS is computationally efficient also in R.

LOWESS for different print tips During the spotting process, the spots located in the same "grid" are printed by the same print tip. Some authors suggest that different LOWESS functions should be fitted for the different print tip subgroups. In our interface we have implemented the scatter plots that show the genes ordered as in the array to detect print tip effects and correct them if necessary. However, and although the option is implemented in our interface, we would expect this effect to cancel with the ratios in two color-microarrays.

Regardless to the method used to estimate ξ, any of them corrects the data to get,

$$\frac{R_i}{G_i} \cong 1 \Rightarrow M = \log_2 \frac{R_i}{G_i} \cong 0$$

For this reason, to look at the scatter plots, boxplots and kernel fitted functions before and after the correction is essential.

Dye-effect correction using the quality elements provided in the experiment

In general, there are many experiments for which the assumption of most genes equally expressed cannot be known "a priori" or for which a very different number of genes is expected to be differentially expressed in both channels. In those cases we would rely on the quality control elements. We have implemented two methods:

Dye-swap normalisation Given two arrays for which the same material was labelled with a different dye each time, for every spotted gene i the following expressions are considered

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

From these two equations, we obtain

$$M_i = \log_2 \left(\frac{R_i}{G_i} \right) = \log_2 \left(\frac{s_i}{r_i} \cdot k_i \right) = \log_2 \left(\frac{s_i}{r_i} \right) + \log_2 k_i = \log_2 \left(\frac{s_i}{r_i} \right) + c_i,$$
$$M'_i = \log_2 \left(\frac{R'_i}{G'_i} \right) = \log_2 \left(\frac{r_i}{s_i} \cdot k'_i \right) = -\log_2 \left(\frac{s_i}{r_i} \right) + \log_2 k'_i = -\log_2 \left(\frac{s_i}{r_i} \right) + c'_i,$$

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 expression, c_i and c_i' depend on the properties of the dyes. It is important to mention at this point, that the scan gains must have been set at the same numbers from array to array (no necessarily from channel to channel). Otherwise c_i and c_i' are not similar and the dye swap normalisation method cannot be applied to the data. Considering that $c_i \simeq c_i'$, adding both equations, we have

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

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

Using the controls In order to normalise a particular Microarray data set, we can spot a predefined set of genes as controls. They have some particular and known characteristics and this information can be used to reduce the non-biological variability introduced in the experiment. But not every subset of genes in the array can be considered as a set of controls. Some requirements must be fulfilled:

- The subset of genes should *cover the whole range of intensities* expected for the array. Otherwise, if the effects that want to be removed are intensity-dependent (as we know is the case of most of the dye effects) we cannot know the effect of the systematical error in a particular range just by looking at the controls.
- The subset of genes that are used as controls should be *plotted all over the array.* If plotted in very specific areas their values can be affected by spatial effects in the array. It happens the same in the case of duplicated genes within an array.
- It would be desirable that the controls were *constant across different biological conditions*, or at least that the response of the control genes to different biological conditions was known for those genes.
- Those genes should give a *good reading* in both channels.

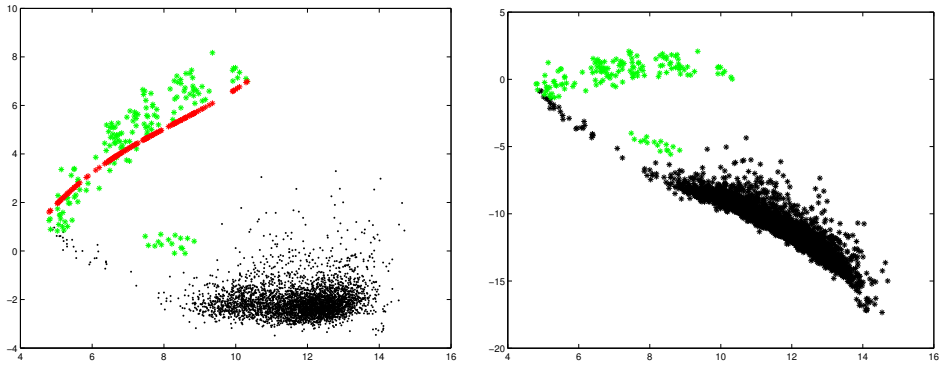
If controls covering the whole intensity range are available, we can normalise our data according to them. For controls for which the expression level in both channels is expected to be the same, a non-linear function can be fitted to the (A,M) plot of the controls and used to correct the entire data set. However, because the number of controls available per slide is usually not very large, we do not recommend to fit a LOWESS function but a more general method such as Levenberg-Marquardt. The model used will be in most of the cases a quadratic function.

Normalisation controls for the TB data set. Those are 5s rRNA?(*), 16s rRNA?(*) and 23s rRNA?(*). Ribosomal RNA molecules are components of the ribosomes, the large multimolecular structures that act as factories for protein synthesis. During translation, ribosomes attach to mRNA molecules and migrate along them as they go, analogous in a way to the role of RNA polymerase in transcription. Ribosomes are made up of rRNA molecules and proteins, and are extremely numerous in most cells. Each ribosome contains one copy of each of the

different rRNA molecules (3 for prokaryotes and 4 for eukaryotes). The most efficient system would be for the cell to produce the same number of each of these molecules. Because the gDNA contains a copy of every gene in the genome, it contains as well a copy of the rRNA. However, the total RNA used for the signal channel contains 98 % of rRNA against just 2 % of mRNA. While the abundance of gDNA (reference channel) in the controls should be similar to the abundance of gDNA in the rest of the spots, the signal channel is expected to present a much higher intensity value than the signal intensity for the rest of the spots. Both channels log ratio for the controls will be much higher than the same for the rest of the spots. If we correct all the data according to the function that we obtain from centering the ratio of both channels for the controls around 1, we are bringing really down the rest of the ratios. This can be seen in Figure 2.1.

In consequence, many controls appear saturated. The spotted controls were printed in a three-fold dilution series. As we go down the dilution series there is less printed material, so the intensity values should decrease. We won't use in the analysis those saturated measurements.

Furthermore, we should check if our controls fulfill the conditions previously described. The range of intensities is for most of the arrays between 4 and 15.7. The use of controls to normalise the data set would be appropriated if the genes in the subset of controls cover all the range of intensities. Otherwise we would be just correcting some genes. As shown in Figure 2.1 for most of the arrays our controls won't fulfill this requirement.



(a) MA plot for the controls (green) and the genes (black). Levenberg-Marquardt function (red) was fitted to the controls.

(b) Corrected data using the function fitted to the controls.

Fig. 2.1: Use of controls to correct the TB data set.

2.2.2 Array effect

Across replicates normalisation

If the overall intensity from one array to the other is observed to be different, this can be due to real biological variability (in one biological condition most genes are higher expressed than in the other) or it can be due to non-biological variation. If we expect most of our genes to be equally expressed for both channels in all the arrays, we may want yet to correct small variations from the mean value of 1 from array to array. Besides that, replicated arrays should have all the same or a very similar mean value. For this reason it would be useful to normalise across replicates. There are two possibilities:

- Dividing every gene i expression level for the median of the expression level of this gene across different replicated slides. This will bring all our slides around a value of one for the ratios or

zero for the log ratios. We would expect,

$$q_{ij} \simeq k \simeq 1 \quad \forall\ j \in \text{set replicates}$$

where,

$$q_{ij} = \frac{\text{signal}_{ij}}{\text{reference}_{ij}}.$$

Signal and reference have been already normalised within array. We correct then our data according to:

$$t_{ij} = \frac{q_{ij}}{med_j(q_{ij})}$$

where med_j is the median of the gene expression level across the different replicates.

- Dividing the signal value for the median of the reference channel across replicated slides.

$$t_{ij} = \frac{\text{signal}_{ij}}{med_j(\text{reference}_{ij})}$$

where med_j is the median across the different replicates and signal, reference have been already within-array normalised.

It is important to remark that with the first approach we are bringing the log ratio distribution of every array around zero. We must be careful in the observance of this hypothesis for every data set. The replicates must be similar among them, but may be not centered around zero.

2.2.3 Replicate handling

At this stage of the normalisation process we are in conditions to take the average among replicated spots within an array or replicated arrays within a biological class. The reason for that is to remove the random error that could not be remove in the previous normalisation steps for not being systematic or for not affecting both channels in the same amount.

Imagine that we have two replicated measurements of a particular gene i in a given biological condition. These are $\frac{R_i^1}{G_i^1}$, and $\frac{R_i^2}{G_i^2}$.

The ratio of the raw intensity values measured for every spot in the two channels (Red and Green) can be modelled as follows:

$$\frac{R_1}{G_1} = N_1 \cdot \frac{A_R^1}{A_G^1} + \text{error}_1$$
$$\frac{R_2}{G_2} = N_2 \cdot \frac{A_R^2}{A_G^2} + \text{error}_2$$

G stands for the intensity of material labelled with the green dye that binds to the spot i and R is the same for the material labelled with the red dye. A_1, A_2 are the abundance of RNA in the samples and 1, 2 indicate the replicate number. It is important to mention that N_1, N_2 are summarizing the systematic error due to the dyes together with the spatial and print-tip effects that affect both channels in the same amount.

Because the dye effect is the most important of the systematic errors in microarrays, after within array normalisation the ratio of the corrected data ^{c} will appear as follows:

$$\left(\frac{R_1}{G_1}\right)^c = \frac{A_R^1}{A_G^1} + \text{error'}_1 \tag{2.1}$$
$$\left(\frac{R_2}{G_2}\right)^c = \frac{A_R^2}{A_G^2} + \text{error'}_2 \tag{2.2}$$

Where error'$_1$ and error'$_2$ are summarizing the random errors after within array normalisation.

The aim of replicating a measurement is to minimize the random errors introduce in the estimator of the relative expression level of a gene in the two studied samples. According to (2.1), (2.2) it follows:

$$\frac{1}{2}\cdot\left[\left(\frac{R_1}{G_1}\right)^c+\left(\frac{R_2}{G_2}\right)^c\right]=\frac{1}{2}\cdot\left[\frac{A_R^1}{A_G^1}+\frac{A_R^2}{A_G^2}\right]+\frac{1}{2}\cdot(\text{error'}_1+\text{error'}_2).$$

Random errors are assumed to be independent[1] and identically distributed as:

$$\text{error'}_1\sim N(0,\sigma^2)$$
$$\text{error'}_2\sim N(0,\sigma^2)$$

From the properties of variance:

$$V\left[\frac{1}{2}\cdot(\text{error'}_1+\text{error'}_2)\right]=\frac{1}{4}\cdot V\left[\text{error'}_1+\text{error'}_2\right]=\frac{1}{4}\cdot 2\cdot\sigma^2=\frac{\sigma^2}{2}.$$

Hence, since the random and systematic errors have been reduced, we would be in conditions to calculate any transformation of the final corrected value.

2.2.4 Across samples normalisation

After averaging the data, we will get for every gene i a unique ratio t_{is} where s denotes the biological condition.

$$t_{is}=\frac{1}{n_r}\sum_{j=1}^{n_r}t_{ij}\ \text{ for every biological condition s for which } n_r \text{ replicates were performed.}$$

Before getting the final log-ratio that will be used for the analysis of the data, we need to make the measurements comparable across different biological conditions. To achieve this objective, we have three possibilities:

Option 1. Dividing the expression level of the gene in every biological condition by the median value of the profile. Mathematically we express it as:

$$e_{is}=\frac{t_{is}}{med_s(t_{is})}$$

Where med_s is the median of the gene expression level across the different biological samples. This approach approximates every overall array expression level even more close to one.

Option 2. Dividing the ratio of every gene by the ratio in the first array. This is very useful for time series analysis. It is more than a way to normalise the data, a way to visualize the data, but must be taken into account for any further study and will affect, for example, in the clustering method chosen.

$$e_{is}=\frac{t_{is}}{t_{i1}}$$

Option 3. If we average corrected signal and corrected reference independently we obtain, for every biological condition s:

$$\text{signal}_{is}=\frac{1}{n_r}\sum_{j=1}^{n_r}\text{signal}_{ij}$$
$$\text{reference}_{is}=\frac{1}{n_r}\sum_{j=1}^{n_r}\text{reference}_{ij}$$

[1]We know that the error between slides are just independent for biological replicates for which the original RNA was pooled from independent cell populations

We can then normalise our data like:

$$e_{is} = \frac{\text{signal}_{is}}{med_s \text{reference}_{is}}$$