5 TB 01-01

5.1 General comments

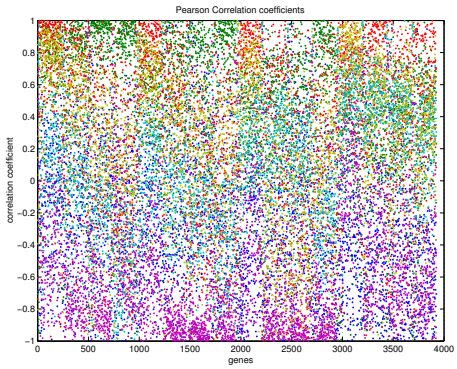
This is the last time series performed. Measurements were taken after 7, 14, 20 and 30 days. Four replicated arrays of RNA samples from each time point were hybridized. In total, sixteen arrays were produced, using for the "signal" channel the four samples of RNA extracted from *M.tuberculosis* (four replicated arrays for each RNA sample of the four time points) and using gDNA for the "reference" channel. For this time series, the dyes were not swapped. The signal channel (RNA) was always labelled with Cy3 and gDNA was Cy5. That is the reason why we just analyze the different results obtained after LOWESS normalisation. Because the array number 14 doesn't correspond to this experiment, this array was set to the results in array 13 in order to keep a regular number of replicates that eases the computational calculations.

5.2 Before normalisation

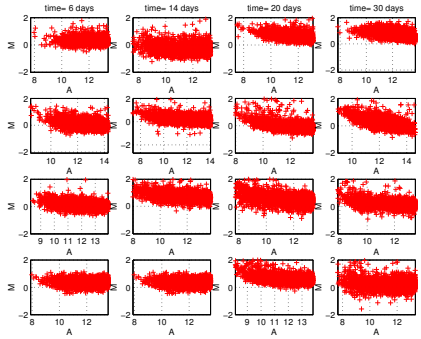
Figure 5.1 shows the data before normalisation. According to plot (a), we can see how the quality of the replicates is since the beginning not very good.

Figure 5.1(b) shows that the intensity dependent effect is not as obvious as for the other data sets. In that case, LOWESS would not be necessary. Should be enough just with global normalisation if we don't have the appropriate computational tools to compute LOWESS. Otherwise, both should show a very similar result.

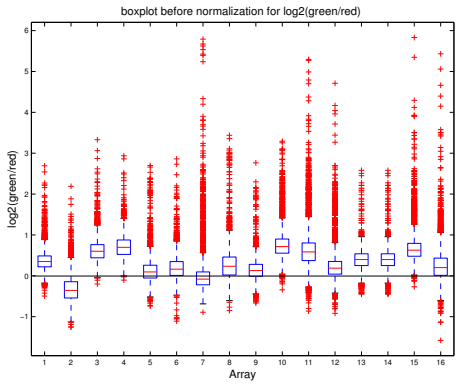
From Figure 5.1 (c) and (d) it can be observed how the second array shows a completely different profile than its partners. For every time point the overall distribution of the replicated arrays is very different. This is in concordance with our curtain plot.



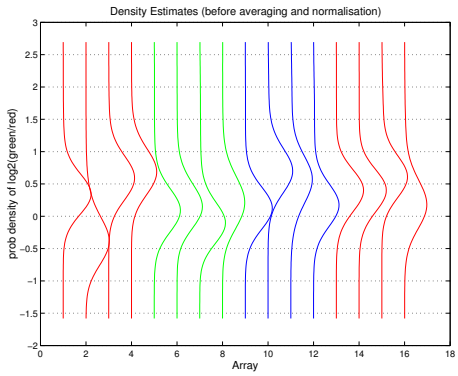
(a) Curtain plot showing the correlation coefficient for the different replicates for every gene.



(b) MA plots for all the arrays before normalisation.



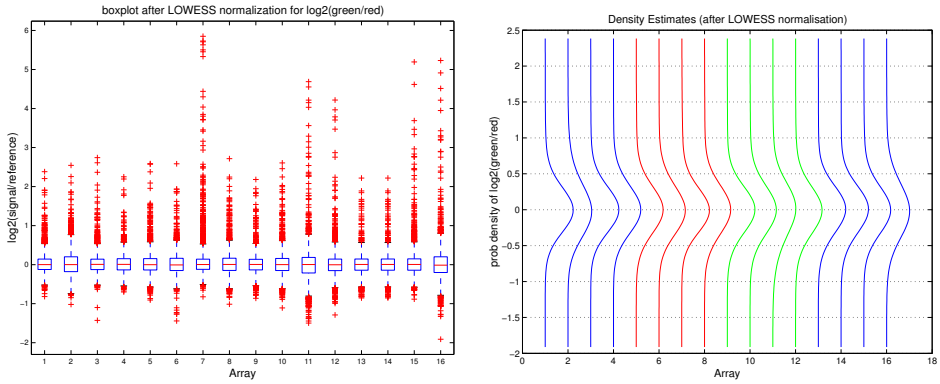
(c) Box plot all arrays before normalisation.



(d) Distributions all arrays before normalisation.

Fig. 5.1: Distributions of the arrays before normalisation. Every four consecutive boxplots/distributions are technical replicates of the same time point.

5.3 After LOWESS normalisation



(a) Box plot all arrays after LOWESS normalisation.

(b) Distributions all arrays after LOWESS normalisation.

Fig. 5.2: Distributions of the arrays after normalisation.

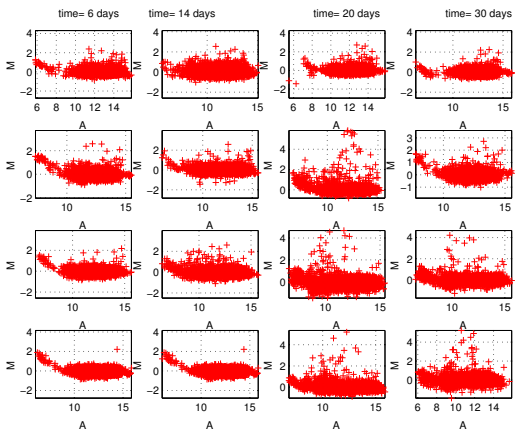
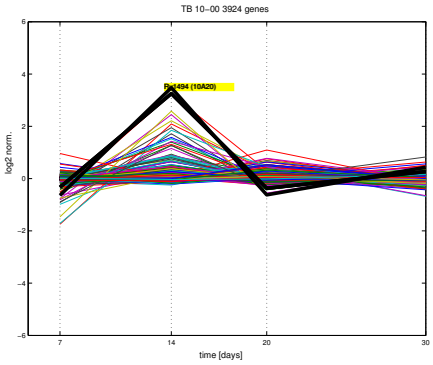
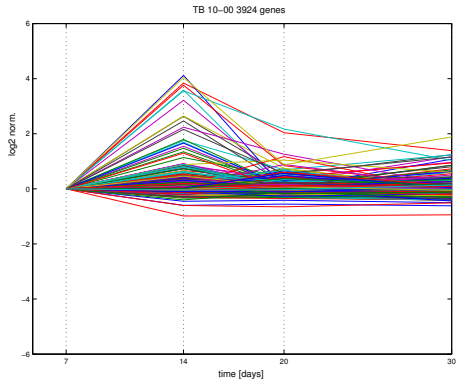


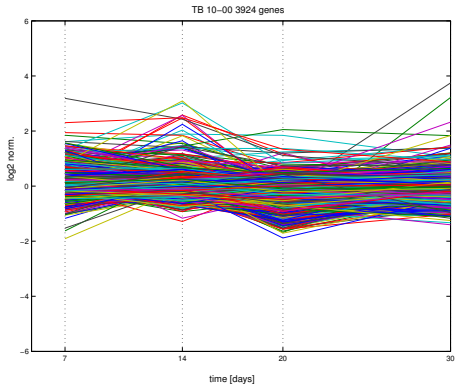
Fig. 5.3: MA plots for all the arrays after LOWESS normalisation



(a) Time series after LOWESS normalisation and across samples normalisation option 1.

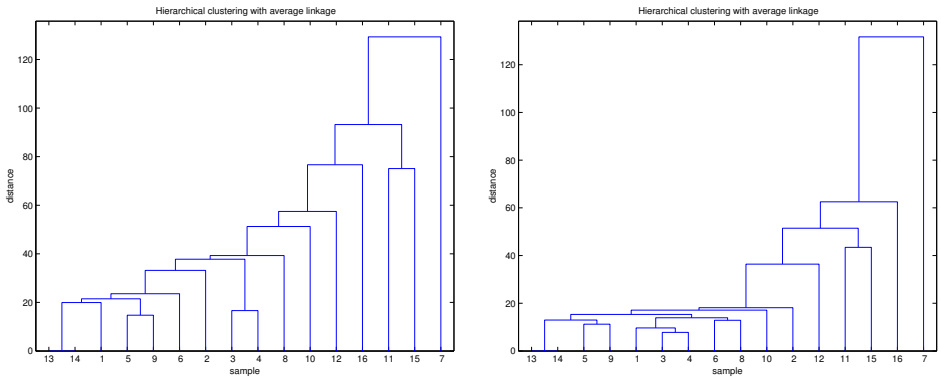


(b) Time series after LOWESS normalisation and after across samples normalisation option 2.



(c) Time series after LOWESS normalisation and across samples normalisation option 3.

Fig. 5.4: Time series after LOWESS normalisation and different across samples options



(a) Hierarchical clustering on the replicates before normalisation.

(b) Hierarchical clustering on the replicates after LOWESS normalisation.

Fig. 5.5: Comparing the two plots we can see the bad quality of the raw replicates (the same was visualized with the curtain plot) and how after LOWESS the distance among replicates has improved. Still replicates from the same time points appear not to correlate. We can't know how dye swap normalisation would have corrected the results.

6 Detection of genes differentially expressed

Student-t distribution

For microarray experiments, the student-t distribution is one of the traditional methods to detect genes differentially expressed. The reason is its computational simplicity and its intuitive meaning. However, since the t-test must be applied for random variables independent and identically distributed following a $N(\mu, \sigma)$ distribution, and because the performance of thousands of t-test at the same time increases the ETI, there are particular problems when applying the student-t test to find genes differentially expressed in microarray experiments:

Low correlation among biological replicates Because the values must be independent, a proper t-test must be applied to the biological replicates and not to the technical ones, since the last were not performed in independent conditions. The problem is that biological replicates try to summarize the properties of the whole population, giving a very low correlation due to the different individuals that they stand for. The correlation among biological replicates may be as low as 0.3 in some cases.

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:

$$t_{ic_1c_2} = \frac{\bar{x}_{ic_1} - \bar{x}_{ic_2}}{\sqrt{\frac{s^2_{ic_1}}{n_{c_1}} + \frac{s^2_{ic_2}}{n_{c_2}}}}, \tag{6.1}$$

where

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

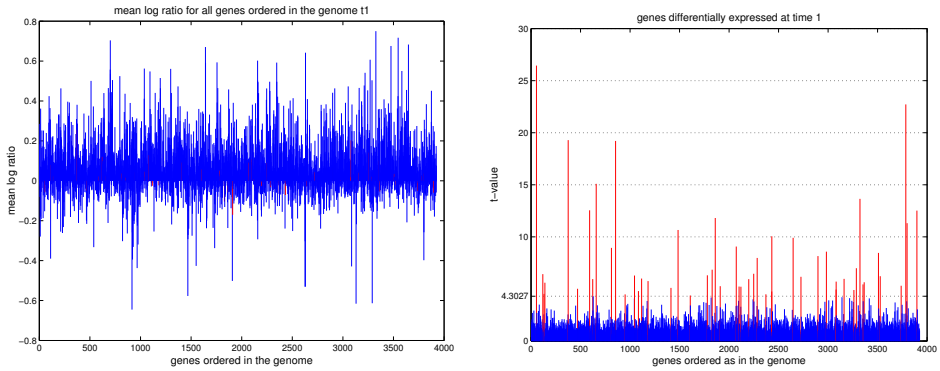
$$s^2_{ic_1} = \frac{1}{n_{c_1} - 1} \sum_{j=1}^{n_{c_1}} (x_{ij} - \bar{x}_{ic_1})^2.$$

From (6.1) 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 (6.1), 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.

For the TB data set, we applied a t-test to detect genes differentially expressed among consecutive time points. As shown in Figure 6.1, the genes detected as differentially expressed were not those with the highest difference among the two consecutive time points, but those for which the difference between biological replicates was very small.

To solve the problem of the poor correlation among biological replicates used to detect genes differentially expressed, the experimenters usually apply a simple fold change detector or a student-t test to every one of the biological replicates. The reliability of the genes detected as differentially expressed will depend on its robustness, i.e. in how many of the three biological replicates they were found to be differentially expressed.

Normality of microarray data If we observe the distribution of the spots within a particular slide (for any of the channels or for the ratio of both), we can see in Figure 6.2 how the data often doesn't follow a normal distribution. The fact that the tails are much longer for the data



(a) Expression level (log-ratio measurement) for all the genes at the first time point after averaging the three biological replicates. The red lines show the expression level of the genes detected as differentially expressed by the t-test.

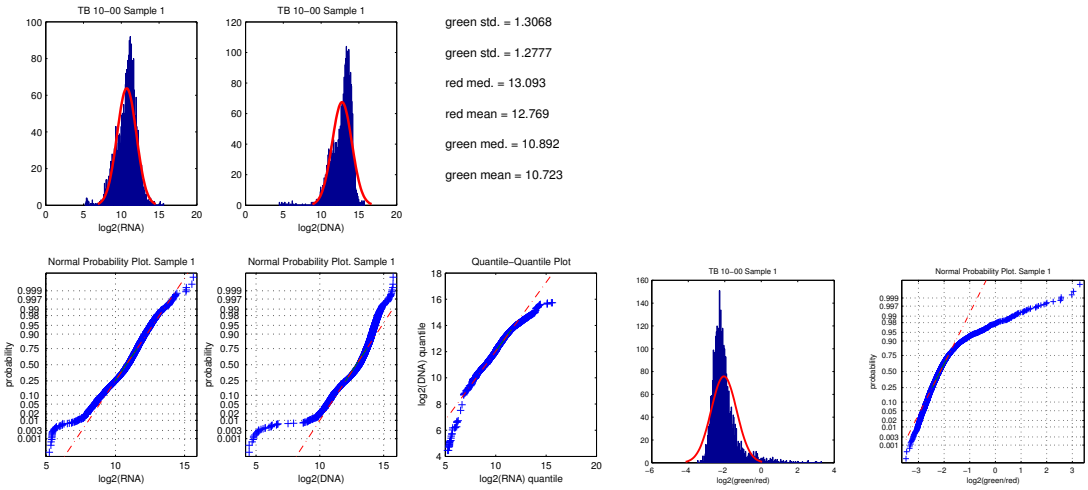
(b) t-values of all the genes at the first time point after averaging the three biological replicates. The red lines show the expression level of the genes detected as differentially expressed by the t-test. As expected, the t-value of the genes differentially expressed is higher than the threshold.

Fig. 6.1: Expression level and t-value for the all genes at the first time point after averaging the three biological replicates. We can see how (due to the dispersion of their replicates) those genes with higher expression level are not those detected as differentially expressed using the t-test.

than for the normally distributed data leads to the conclusion that more genes will be selected as differentially expressed than if we had real normally distributed data.

Due to these reason the use of non-parametrical test is essential. In our toolbox, we have implemented the Mann-Whitney test.

Error Type I It could be said that in the detection of genes differentially expressed we are testing n_g null hypothesis at the same time. For this reason, if any of the H_0 is selected with a $\alpha = 0.05$ in the whole we would have $0.05 \times n_g$ false discoveries. For this reason, the p-values are usually corrected using Bonferroni. Some other not so exigent approaches suggest to rank the p-values and considering as differentially expressed just those genes with their p-value among the 5 percent of higher p-values.



(a) Distribution of the log intensity for every channel in the first array of the second biological replicate. The same not-normal shape appear for most of the slides.

(b) Distribution of the log ratios in the first array of the second biological replicate.

Fig. 6.2: Distribution of the log-transformed data for the first slide of the second biological replicate.

7 Conclusions

This report tried to extract some conclusion about the reliability of two of the most important normalisation methods used in microarray: Dye-swap and LOWESS. Comparing the results obtained for the three biological replicates we could conclude something about the way the two methods behave. Unfortunately, we would have needed replicated slides for which the dyes were swapped and at least three replicates per time point. This was just the case of TB-10-00. For this data set a more detailed work is presented in a paper submitted for publication. However, important conclusions were still extracted from the analysis of all this data:

Assumptions of the self consistency methods: The self consistency methods make the assumption that most gene in a slide will be equally expressed in both hybridized samples. Without non a priori information it is risky to normalize the data according to those methods.

Requirements of dye-swap normalisation: The basic assumption under which the dye-swap normalisation make sense is that the properties of the dyes are stable from slide to slide. This will be just true if the two slides were scanned at the same gain and using some standards to fix a particular gain to the scan. If the data is scanned without paying attention to these requirements, it can be again dramatically transformed. Whilst we can go back to the original data if we think that the normalisation applied was not correct, it is a harder task to recover the original data after scanning.

Across replicates normalisation: This method will center the distribution of the ratios of all slides around 1. This is, as in the case of the self consistency methods, a operation that can falsely transform the data hiding the real biological information of the results. However, we understand that the different overall intensity between technical replicates must be due to some random or systematic error. For this reason, we are inspecting some other methods that preserve the real value of the data but diminish intensity variability from slide to slide due to non-biological variation.

Which method to choose? The time series for all three biological replicates shown similar profiles, although the dispersion among gene profiles is greater after dye-swap normalisation than after LOWESS normalisation. The question is, are we looking for normalisation methods that transform our data very nicely? or, are we looking for methods that transform the data removing the non-biological normalisation and preserving the biological information? A good normalisation method should not transform extremely the data, should reduce the across replicates variability and, as consequence, should provide a reliable number of genes as differentially expressed.

8 Publications related with this project

- F. Sanchez-Cabo, K.Y.Cho, Z.Trajanoski and O.Wolkenhauer.
  A Graphical User Interface to Normalise Microarray Data, DSC 2003
  Available from http://www.sbi.uni-rostock.de

- F. Sanchez-Cabo, K.Y.Cho, P. Butcher, J.Hinds and O.Wolkenhauer.
  Is *LOWESS* a panacea in the normalisation of microarray data?
  Available from http://www.sbi.uni-rostock.de

9 Further information

- Systems Biology Group webpage, http://www.sbi.uni-rostock.de
- BμG@S webpage, http://bugs.sghms.ac.uk/index.php

10 Acknowledgements

The authors would like to thank the Wellcome Trust funded BμG@S group at St.Georges Hospital Medical School and the Streptomyces group at UMIST for their unconditional help. They would also like to thank the Bioinformatics Group from the Biomedical Engineering Department of the Technical University of Graz, Austria. This collaboration was possible thanks to the Marie Curie Fellowship Program. To all of them, thanks for useful discussions and continuous feedback that made possible the elaboration of this report.