Comparison of Microarray Pre-Processing Methods

K. Shakya, H. J. Ruskin, G. Kerr, M. Crane, J. Becker

Dublin City University, Dublin 9, Ireland

Abstract

Data pre-processing in microarray technology is a crucial initial step before data analysis is performed. Many pre-processing methods have been proposed but none has proved ideal to date. Frequently, datasets are limited by laboratory constraints so that the need is for guidelines on quality and robustness, to inform further experimentation while data are yet restricted. In this paper, we compared the performance of 4 popular methods, namely *MAS5*, *Li & Wong pmonly* (*LWPM*), *Li & Wong subtractMM* (*LWMM*) and *RMA*. The comparison is based on analysis carried out on sets of laboratory-generated data from the Bioinformatics Lab, National Institute of Cellular Biotechnology (NICB), Dublin City University, Ireland. These experiments were designed to examine the effect of Bromodeoxyuridine (5-bromo-2-deoxyuridine, *BrdU*) treatment in deep lamellar keratoplasty (*DLKP*) cells. The methodology employed is to assess dispersion across the replicates and analyze the false discovery rate. From the dispersion analysis, we found that variability is reduced more effectively by *LWPM* and *RMA* methods. From the false positive analysis, and for both *parametric* and *non-parametric* approaches, *LWMM* is found to perform best. Based on a complementary *q*-value analysis, *LWMM* approach again is the strongest candidate. The indications are that, while *LWMM* is marginally less effective than *LWPM* and *RMA* in terms of variance reduction, it has considerably improved discrimination overall.

1 Introduction

Microarray technology allows the monitoring of expression levels of thousands of genes, simultaneously, which in turn helps to explore gene sequence information and ultimately gene function(s). Since microarray gene expression data are charac-

terized by high dimensionality and noisiness, the initial steps of microarray experiments are very crucial in terms of feeding 'clean' data to the downstream analysis, namely, identification of gene expression patterns.

An important initial step, in microarray technology, is data pre-processing. Preprocessing removes systematic errors between arrays, introduced by labeling, hybridization and scanning. Several such methods have been developed so far (Mastrogianni 2007), but, this paper focuses on 4 popular methods which comprise basic tools for much experimental work. These are: *MAS5* (Microarray Suite version 5, core to the Affymetrix system providing instrument control, data acquisition and analysis for the entire genechip platform), two alternatives of the *Li & Wong* method, *Li & Wong pmonly* (*LWPM*) and *Li & Wong subtractMM* (*LWMM*) and *RMA* (Robust Multichip Average). *RMA* consists of 3 steps: a background adjustment, quantile normalization and finally summarization (Bolstad et al. 2003). *Li & Wong pmonly* (*LWPM*) method ignores the *MM* (Mismatch probe intensities) while *Li & Wong subtractMM* (*LWMM*) uses the *PM* (Perfect Match) – *MM* (Mismatch) value to adjust the non-specific binding (NSB) during background adjustment (Wu et al. 2004). These methods are compared on the basis of dispersion across replicates (Bolstad et al. 2003, Novak et al. 2006), distinguished, as explained in Sect. 2.1 and also in terms of false discovery rates.

2 Materials and Methods

The dataset used for this comparison is laboratory-generated (experiments performed: Bioinformatics Lab, National Institute of Cellular Biotechnology (NICB), Ireland). The experiments were performed on Affymetrix GenechipTM, Human Genome U133 set (HG-U133A) and were designed to investigate patterns of gene expression changes in *DLKP* cells treated with thymidine analogue (5-bromo-2-deoxyuridine, *BrdU*), during three different periods, 0 (control), 3 and 7 days. For each time point, 3 microarrays were used (McMorrow 2006). The dataset is thus modest, but typical of experiments targeted to exploratory analysis.

2.1 Dispersion Analysis

Dispersion analysis is used to assess the ability of each pre-processing method to reduce systematic error, introduced during treatment stage. A method giving large dispersion implies that, at the analysis stage, some genes are falsely declared to be differentially expressed, and vice versa. The approach is commonly based on two precision criteria:

(i) The ability to minimize differences in pairwise comparisons between arrays of replicates: Theoretically, genes are not differentially expressed across replicates, but should produce similar values. *MA*-plots are used here for pairwise compari-

sons as these conveniently illustrate the distribution of intensity values and log ratios and can give a quick overview of the data. In *MA*-plots, methods which minimize the distance between the loess curve and the $M = 0$ line, are considered optimal, as a gene is less likely to be falsely declared as differentially expressed in such cases. The 4 pre-processing methods were applied to the datasets for comparison using this criterion. For each of the 4 methods, nine *MA*-plots were produced, (for 3 replicates at 3 time points). The absolute distance, between the line $M = 0$ and the loess curve was measured for every intensity. Finally, the mean was calculated for each intensity across the nine comparisons. The loess function becomes prohibitively time consuming for datasets corresponding to more than 20,000 probes and it was not possible to include all 506,944 probes on the array used in experiment. Hence, we chose to apply the process for ten random samples of 5,000 probes each, and then to calculate the mean of the ten vectors. The final mean vector was found to be almost the same for the 10 random samples, which argues for good consistency. Averaging over a number of samples also improved reliability even though excessive smoothing can obscure finer details.

(ii) The precision of the expression measures, estimated by the standard deviation across the replicates: After pre-processing, most noise would be removed from the data, and the biological replicates should have similar values. Thus, for a given method, high residual standard deviation across the replicates (for each gene, and/or time point), implies poor reliability. The following process was applied: for each method and gene, at each time point, the standard deviation and the mean were calculated across the replicates. To investigate the behavior of the standard deviation of the mean, a loess curve based on these calculated values was fitted in order to visualize the trend in the data.

2.2 False Positive Analysis

Microarray experiments measure expression values of thousands of genes simultaneously. Many of these are not in fact expressed but repeated statistical testing at levels of significance ($\alpha \sim 0.05$) can lead to a large number of false positives. The number of false positives, generated after pre-processing, is used as a second criterion for comparison and measures the specificity of the pre-processing technique.

Procedures of testing for similar levels of expression between any two genes may be either *parametric* or *non-parametric* with use typically dependent on sample size. With a smaller number of replicates, the assumption of normality is less robust. A *non-parametric* approach such as *Wilcoxon* or *Mann-Whitney*, while not reliant on distributional assumptions of sample measurements, has less ability to distinguish between methods and a larger number of replicates are desirable. In a typical laboratory investigation, practical restrictions can apply, so that frequently the approach is to perform complementary analysis as an internal checking procedure. Consequently, some reliance is placed on the ability of pre-processing techniques to detect poor quality results.

In the work presented here, false positive analysis is based on *FWER* and *FDR* measures. The family-wise error rate (*FWER*) (Novak et al. 2006), is defined as the probability of occurrence of at least one false positive (V) over all true null hypotheses (corresponds to no relationship between gene expression measurement and response variable). Thus,

$$FWER = pr\,(V \geq 0) \qquad \ldots \qquad (1)$$

The one-step *Bonferroni* method (Gordon et al. 2007), together with the sequential *Westfall & Young* (*maxT*) adjusted *p*-value method (Westfall et al. 1993), were used to estimate the *FWER* statistics. The advantage of the latter is that it takes the dependence structure between genes into account, giving improved power in many cases (e.g. where there is positive dependence between genes).

The false discovery rate (*FDR*), (Benjamini et al. 1995), is an alternative and less stringent concept of error control, which is defined to be the expected proportion of erroneously rejected null hypotheses (Type *I* errors), amongst all those rejected.

$$FDR = E\,[V/R \mid R>0].\,Pr\,(R>0) \qquad \ldots \qquad (2)$$

Controlling the *FDR* is desirable since the quantity controlled is more directly relevant than that for the *FWER* (i.e. statistical power is improved for the former).

False-positive analysis was carried out, using libraries available with the *R* '*Bioconductor*' software: *multtest* and *q*-value. The *multtest* package implements multiple testing procedures for controlling different Type *I* error rates, for *FWER* and *FDR*. For our analysis, the procedures used to control *FWER* and *FDR* respectively were, (a)*Bonferroni* and *maxT*, and (b)*Benjamini & Hochberg* (1995) and *proc-2* correction (Benjamini et al. 2001, Speed 2003).

The *pFDR* (Storey 2001), is the expected proportion of Type *I* error among the rejected hypotheses when the number of rejected hypotheses is strictly positive.

$$pFDR = E(\,V/R \mid R > 0) \qquad \ldots \qquad (3)$$

The *pFDR* can be interpreted in Bayesian terms. If we wish to perform m identical tests of a null hypothesis versus an alternative hypothesis based on the statistics $T_1, T_2, \ldots, \ldots, T_m$, for a given significance region Γ,

$$pFDR(\Gamma) = E\,[V(\Gamma)/R(\Gamma) \mid R(\Gamma)>0] \qquad \ldots \qquad (4)$$

Associated with the *pFDR* an adjusted *p*-value, known as the *q*-value, is defined. The *q*-value gives each hypothesis test a significance in terms of a given error rate. It measures the minimum false discovery rate that is incurred when the test result is deemed significant. For an observed statistic $T= t$, the *q*-value of t is:

$$q - value(t) = \inf_{\{\Gamma_\alpha : t \in \Gamma_\alpha\}} pFDR(\Gamma_\alpha) \qquad \ldots \qquad (5)$$

Where $\{\Gamma_\alpha\}$ is the set of nested rejection regions.

Equation (5) indicates that the *q*-value is a measure of the strength of an observed statistic, with respect to *pFDR*; it is the minimum *pFDR* that can occur when rejecting a statistic with value t for the set of nested significance regions. Thus, it relates to false discovery rate.

The *q*-value module of *Bioconductor* has been used here for *pFDR* analysis, for both *parametric* and *non-parametric* estimation. The *q*-value package functions permit computation and presentation of *q*-values for multiple comparison features.

3 Results

3.1 Dispersion Analysis

Fig. I.(a) illustrates the results of analysis for minimization of differences across replicates (criterion (i), Sect. 2.1). Both the *RMA* and *LWPM* methods perform better than *MAS*5 and *LWMM* in that they minimize variability around the $M = 0$ line and show potential for relatively few genes to be falsely declared as differentially expressed. The same hierarchy of performance across these methods is found in Fig. I.(b), corresponding to analysis under criterion (ii), Sect. 2.1 and also in Fig. II. Better performing methods both use the *PM* Correction, which may explain the effect observed.

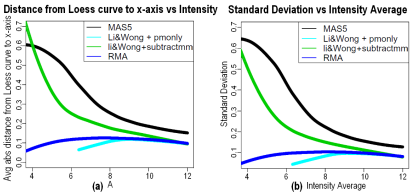

Fig. I.(a) and (b) Comparison of the 4 preprocessing methods based on the two dispersion criteria

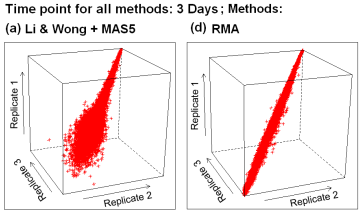

Fig. II. Plot of the 3 replicates on the x, y, z axis for the 4 methods of Table I. Note the high variability of differentially expressed genes for (a) and to a lesser extent for (d)

To obtain a numerical estimate of the relative precision, linear models were fitted between replicates, (3 pairwise comparisons between the 3 replicates), for each time point and for each method. From the 36 R^2 values obtained, an average was taken of the 3 values for each method and each time point. In Table I, these values are summarized by averaging over time points, then over time points and methods, (mean row). Again, *RMA* and *LWPM* have the highest R^2 values, indicating that the amount of variation explained is higher, compared to other methods, and implying better reproducibility and precision. This is a crude measure of goodness of fit, but acts as our indicator of relative reliability.

Table I: Average R^2 associated with each time point and each method

Sample	***MAS***	***LWPM***	***LWMM***	***RMA***
Time 1	0.9265326	0.9904766	0.9724816	0.9952248
Time 3	0.9165260	0.9857380	0.9603595	0.9930388
Time 7	0.9080216	0.9649691	0.9434947	0.9850423
Mean	0.9170267	0.9803946	0.9587786	0.9911020

3.2 False Positive Analysis

3.2.1 *Multtest* Results

Using the *multtest* library, analysis was performed on the number of rejected hypotheses as a function of the adjusted *p*−value, (calculated according to equations, Sect. 2.2). The objective is to highlight those pre-processing methods that have a minimum number of false positives for a given number of differentially expressed genes, i.e. to find those methods, which have a high number of differentially expressed genes for a large adjusted *p*−value range.

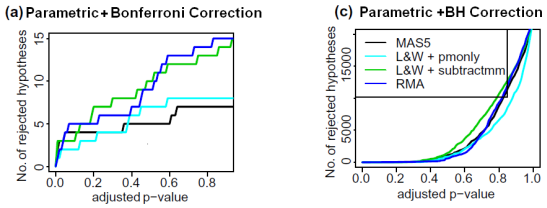

Fig. III. *Parametric* approach

Number of rejected hypotheses is larger for the *FDR* method as compared to the *FWER* (Fig. III.), in agreement with the view that the *FWER* criterion is more conservative than the *FDR*. Nevertheless, better discrimination between pre-processing methods is achieved for the *FWER* criterion, with *RMA* and *LWMM* able to identify considerably more differentially expressed genes than *MAS*5 and *LWPM*, for a given adjusted *p*-value. For *FDR* procedures, *LWMM* is slightly improved, whereas *LWPM* performs slightly worse. Curves for *MAS*5 and *RMA* overlap over partial ranges, so that real differences between methods are small.
Comparing results of *parametric* and *non-parametric* approaches showed that *Wilcoxon +MaxT* correction and *Wilcoson + Bonferroni* correction (*FWER* criterion) produced the same shape curves. Similarly, *Wilcoxon +BH* and *Wilcoxon +Proc_2* (*FDR* criterion). In general, the *non-parametric* approach is unsatisfactory in terms of discrimination, especially for the *FWER* criterion, as all pre-processing methods produce similar curves. For the two *FDR* procedures, *LWMM* is slightly improved, whereas *RMA*, which performed well in *parametric* analyses, is slightly worse. Given the closeness of the curves, however, the difference is not marked.

Summary Results of false positive analyses:
Fig. IV. summarizes results for the false positive correction procedures, Sect. 3.2.1. For each of the four false positive correction procedures, (two for each of *FWER* and *FDR*), for each *p*-value, the four values (number of rejected hypotheses, corresponding to the 4 pre-processing methods), is divided by the largest of these, so that hierarchy and relative variation between the four are preserved,

(normalization). Then for each *p*-value, and for each of the 4 pre-processing methods, the normalized values (for four positive-correction approaches) are averaged. This mean value is used to plot mean percentage of the best value for the 4 false positive correction procedures versus adjusted *p*-value.

Summary results are shown in Fig. IV.(a) for *parametric* and Fig. IV.(b) for *non-parametric* approaches. Based on these, *LWMM* outperforms the other pre-processing methods for $0.2 \leq p < 1$ and, somewhat surprisingly, for very low *p*-value $0 < p \leq 0.045$ for the *parametric* approach, and for $0.06 \leq p < 1$ for *non-parametric* approach: (*LWMM* performance is distinctly improved here). For mid-range $0.045 \leq p \leq 0.2$, *RMA* outperformed *LWMM* in the *parametric* case. The best pre-processing method is indeterminate, however, for $0 \leq p \leq 0.06$ (*non-parametric*). The 0.06 threshold here is possibly due to the low number of replicates for the *Wilcoxon* test, however, and the approach is more robust under assumptions that apply.

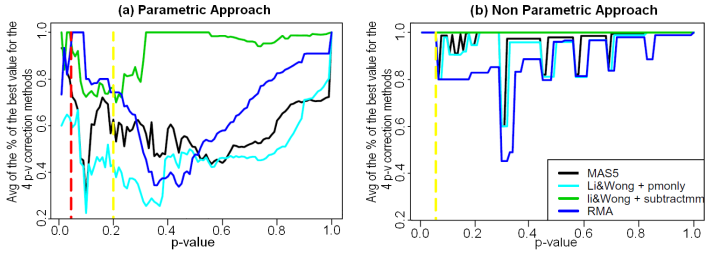

Fig. IV.(a) and (b) Summarization of false positive correction procedures with the *parametric* approach (a) and the *non-parametric* approach (b)

3.2.2 Results based on *q*-value Library

The *q*-value library, of *Bioconductor*, was used to estimate *pFDR* (Storey 2008). Results are similar to those for the *multtest* library in both *parametric* and *non-parametric* cases. All 4 pre-processing methods give similar shape curves; however *LWMM* and *RMA* performed best for $0 < q \leq 0.6$ and $q \geq 0.6$, respectively, (*parametric* case). For the *non-parametric*, *LWMM* was best across all *q*-values, implying that a higher number of differentially expressed genes are captured.

4 Conclusion

An analysis of 4 different microarray pre-processing methods, namely *MAS5*, *LWPM*, *LWMM* and *RMA*, was performed with respect to their dispersion and false positive analysis. Dispersion comparisons indicate that technical variability is addressed more effectively by *LWPM* and *RMA* methods: (supported by results of the fitted linear model and R^2, Coefficient of determination measures).

Comparison, of false positive rates, in ranges $0.045 \leq p \leq 0.2$ and $p > 0.2$, indicate that *RMA* and *LWMM*, respectively, performed best, (*parametric* case). *LWMM* also outperformed other methods, $0 < p \leq 0.045$, (possibly due to small sample size). For $0 < p \leq 0.06$, (*non-parametric*), all methods performed equivalently, with *LWMM* best for $p \geq 0.06$. In q-value tests for *pFDR* analysis, *LWMM* outperformed *RMA*, (for *non-parametric* and $0 < q \leq 0.06$, *parametric*), supporting previous analyses.
Given that sample size is relatively small for these data, i.e. methods are less robust for small p, results of *false positive* analysis indicate *LWMM* to be the *best pre-processing method*. Based on *dispersion* analysis results, *LWPM outperformed* other methods. While choice of method may depend on analysis purpose, these control for two very relevant experimental criteria, (not necessarily of equal importance in an investigation), so that one is typically preferred. Consequently, results presented here, even for modest sized dataset, do provide some distinct guidelines on pre-processing method choice for microarray data.

5 References

Benjamini, Y., Krieger, A., Yekutieli, D.: Two-staged Linear Step-Up FDR Controlling Procedure, Technical Report, Tel-Aviv University and Department of Statistics, Wharton School, University of Pennsylvania (2001)

Bolstad, B. M., Irizarry, R. A., Astrand, M., Speed, T. P.: A comparison of normalization methods for high density oligonucleotide array data based on variance and bias, Bioinformatics. **19(2)**, 185–193 (2003)

Cleveland, W.S.: Visualizing Data, Summit, New Jersey: Hobart Press (1993)

Cope, L. M., Irizarry, R. A., Jaffee, H. A., Wu, Z., Speed, T. P.: A benchmark for affymetrix genechip expression measures. Bioinformatics (Oxford, England), **20**(**3**), 323–331(2004)

Gordon, A., Glazko, G., Qiu, X., Yakovlev, A.: Control of the mean number of false discoveries, Bonferroni and Stability of multiple testing. Ann. Appl. Statist., **1**(**1**), 179-190 (2007)

Hubbell, E., Liu, W. M., Mei, R.: Robust estimators for expression analysis. Bioinformatics (Oxford, England), **18**(**12**), 1585–1592 (2002)

Irizarry, R. A., Hobbs, B., Speed, T. P. et al.: Exploration, normalization and summaries of high density oligonucleotide array probe level data. Biostatistics, **4**(**2**), 249–264 (2003)

Mastrogianni A., Dermatas E., Bezerianos A.:Robust Pre-processing and Noise Reduction in Microarray Images. Proceeding (555), Biomedical Engineering (2007)

McMorrow, J.: Ph.D. thesis. Dublin City University, Ireland (2006)

Novak, J. P., Kim, S. Y., Xu, J., Modlich, O. et al. : Generalization of DNA microarray dispersion properties: microarray equivalent of *t*-distribution. Biol Direct, **1**(**27**) (2006)

Speed, T. P.: Statistical Analysis of Gene Expression Microarray Data. Published by CRC, ISBN 1584883278, 9781584883272

Stafford, P., Editor: Methods in Microarray Normalization (Drug Discovery Series), USA, CRC press, ISBN 1420052780, 9781420052787

Storey, J. D.: The positive false discovery rate: A Bayesian interpretation and the q-value. Ann. Statist. **31**(**6**), 2013–2035 (2001)

Westfall, P. H., Young S. S.: Resampling based multiple testing: Examples and methods for p-value adjustment. Wiley, England (1993)

Wu, Z., Irizarry, R. A., Gentleman, R., Martinez-Murillo, F., Spencer, F.: A Model-Based Background Adjustment for Oligonucleotide Expression Arrays. Journal of American Statistical Association, **99**(**468**), 909-917 (2004)