Comparing the effects of background, normalization and summarization on gene expression estimates

B. M. Bolstad

December 7, 2002

Introduction

We have placed the computation of probeset expression measures into a three step process: Background, Normalization and Summarization. For each of these steps, numerous alternatives had been proposed, but a study of the effects of combining the different proposals together and examining the effect of these on gene expression has not been done. We will combine together five different background methods, with two normalization methods and three summarization methods. Since normalization has been discussed before in Bolstad et al (2003) we will contrain ourselves to testing quantile normalization versus not using normalization at all. At the summarization step we shall try to quantify the effects of robustness and using information from multiple chips.

We will attempt to examine the effects that each combination of processing steps has on detecting differential expression and on predicting fold change.

Methods

The methods discussed are shown in table 1. We have tried to examine evey combination, except that we have used the RMA background only with the median polish.

Background

We define background correction as the process of correcting probe intensities on an array using information only on that array.

RMA Background

The RMA background model is that we observe $O = S + N$. Where S is signal and N is background. We assume S is distributed $\exp(\alpha)$ and that N is distributed $\mathrm{N}\left(\mu, \sigma^2\right)$. The corrected intensities

Background	Normalization	Summarization
None	None	Median Polish
RMA Background	Quantile	Tukey Biweight (1 step)
MAS 5.0 Background		Average of Logs
Ideal Mismatch		
MAS 5.0 + Ideal Mismatch		

Table 1: Methods compared

are given by

$$E\left(s|O=o\right) = a + b\frac{\phi\left(\frac{a}{b}\right) - \phi\left(\frac{o-a}{b}\right)}{\Phi\left(\frac{a}{b}\right) + \Phi\left(\frac{o-a}{b}\right) - 1}$$

where $a = s - \mu - \sigma^2\alpha$ and $b = \sigma$. Estimating the parameters has proved troublesome, the parameters are currently estimated in an ad-hoc way. We generally only correct PM probes.

MAS 5.0 Background

Documented in Affymetrix (2001), the MAS 5.0 background breaks the array into regions, within each grid picks a background and noise value for that grid. The background/noise adjustment for each probe intensities is gven by taking a weighted average of the grid background/noise values. The weights are dependent on distance from the probe location to the center of each of the grids.

Ideal Mismatch

This procedure is discussed in Affymetrix (2001), uses the MM probes to correct the PM probes. Originally PMs were corrected by substracting MMs. However, many MM's are greater than the corresponding PM and so simple subtraction creates creat difficulties. To sidestep this problem the MAS 5.0 algorithm uses the Ideal Mismatch, which is defined as the MM when physically possible and something close but smaller than the PM when it is not.

Normalization

Normalization is the process of removing non-biological variability between arrays. A comparision of normalization methods has been considered before in Bolstad et al (2003). To remove the possible effects that normalization might have on the final gene expression (and confounding our comparision of background methods) estimates we have computed expression estimates in each case both with and without normalization.

Summarization

Summarization is the process of combining the preprocessed PM probes together to compute an expression measure for each probeset on the array.

Median Polish

In the RMA method we fit a robust model to the log of the preprocessed PM probes for a particular probeset using the median polish algorithm with probe and chip effects and use the fitted model to predict chip expression. This method allows us to combine information across chips.

Tukey Biweight (1 Step)

The MAS 5.0 algorithm uses the Tukey Biweight 1 step alogrithm to combine probe values for a particular probeset from a single array to compute the expression value for that array.

Average of logs

This is an unrobust single chip analog of the other two methods. That is we take the mean of the log preprocessed PM probes for a particular probeset as the expression value for that probeset on that array.

Computation of expression measures as a three-step process

We can consider the procedure for generating expression measures as a three step process. Let X be raw probe intensities across all arrays and E be probeset expression measures. Let B be the operation which background corrects probes on each array, N be the operation which normalizes across arrays and S be the operation which combines probes together to compute an expression measure. The the process of computing measures of expression can be written as

$$E = S\left(N\left(B\left(X\right)\right)\right)$$

In the case of RMA, B is the RMA background process described above, N is the quantile normalization and S is an operation which takes $\log_2$ of the probes and fits a robust linear model. For MAS 5.0, B is the background/noise correction followed by subtracting the ideal mismatch from the PM probes, N is an operation which leaves the data unchanged (in the MAS 5.0 framework normalization takes place after summarization) and S is the process which takes $\log_2$ of the data and then uses the 1-Step Tukey biweight.

Data

The data used is a spikein latin square experiment carried out by Affymetrix. This dataset was used in the creation of the MAS 5.0 algorithm. It consists of 59 chips, where 14 probesets have been spiked in a known concentrations. The concentration of the spike-ins are 0, 0.25, 0.5, 1, 2, 4, 8, 16, 32, 64, 128, 256, 512, and 1024pM. In most cases there are three replicate chips haveing the same concentration profile, however there are also two concentration profiles that are replicated 12 times each and one concentration profile is represented only twice.

Results

We will compare methods by looking at the effect of the processing steps on the variability of non spike-in probe sets, ability to predict true fold change using observed data and ability to detect differential expression using observed fold change.

Given data from 59 arrays one may look at all pairwise combinations of single chips (a total of 1711 comparisons) or at pairwise comparisons between groups of arrays with same spike in concentrations. In that case there are 14 spikein groups making for a total of 91 comparisons. We will use the aggregate comparisons, but analysis with individual comparisons has yielded similar results.

Expression differences M are computed by differencing the log scale expression measures. That is for probeset i the difference in expression between chips j and k given by $M_i = E_{ik} - E_{ij}$.

For every combination of background, normalization and summarization method given in table 1 we compute expression measures. It is on the basis of these expression measures that we will compare methods.

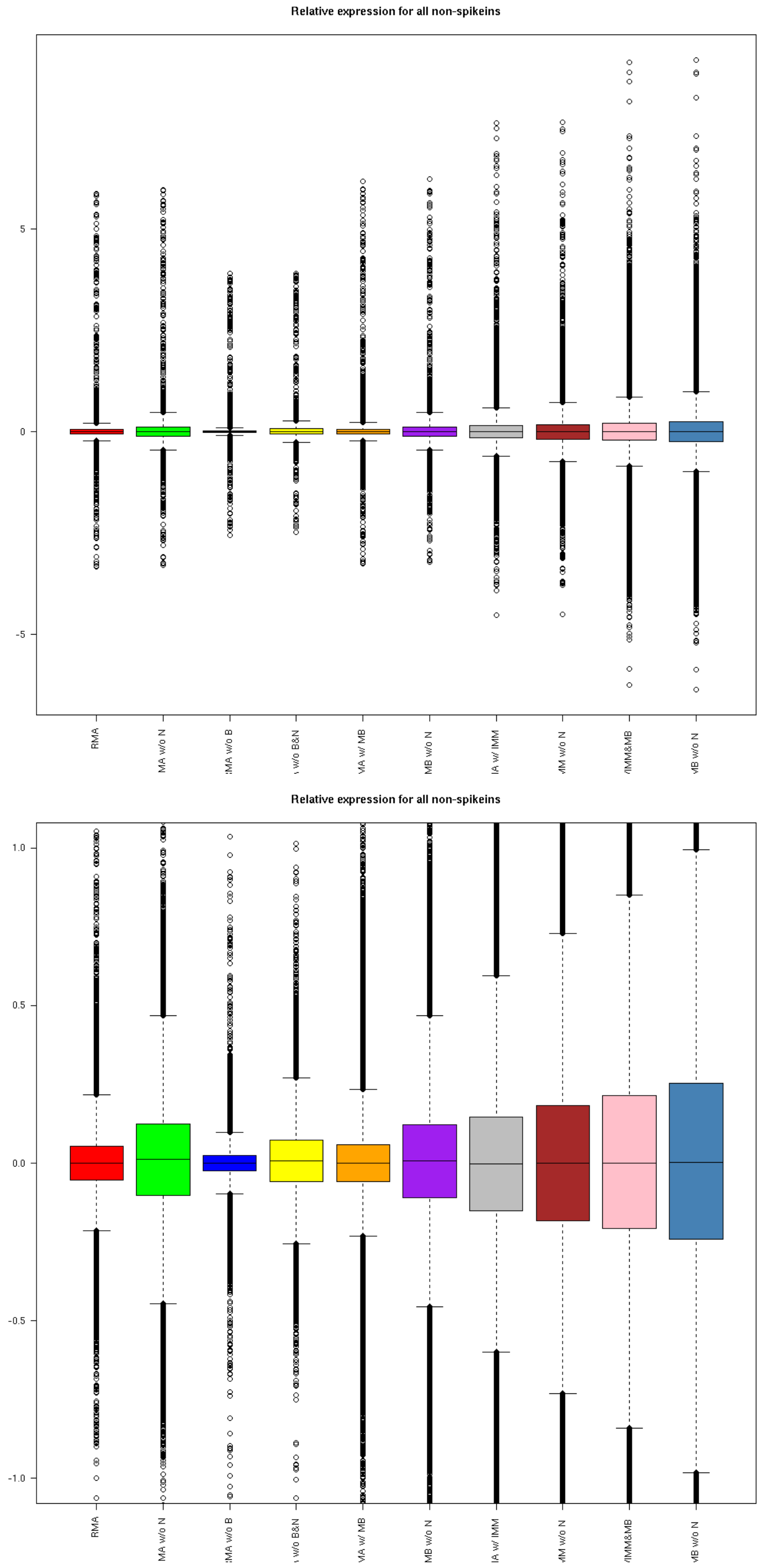

Figure 1: Boxplot of non differential expression differences by preprocessing method using median polish expression summary

We look first at the variability of the expression differences in non-spike in probesets. Ideally we would like non-differential probesets to be flagged as non-differential by which ever selection method we choose. Thus, it would be desirable to have the measure of differential expression to be less variable and close to zero. Figure 1 shows the variability of the expression measure for the non differential probesets. The boxes are, in order left to right, Complete RMA method, RMA without normalization, RMA without background, RMA without both background and normalization, RMA summarization with both normalization and MAS5.0 background, RMA summarization without normalization and with MAS5.0 background, RMA summarization with both normalization and ideal mismatch correction, RMA summarization without normalization and with ideal mismatch correction, RMA with normalization and both MAS5.0 and then IMM correction and finally RMA without normalization and both MAS5.0 and then IMM correction. One can see that as expected normalization reduces general variability by comparing each pair with and without normalization. One also sees that background correction adds more variability, with the ideal mismatch correction increasing the variabilty the most. The case where the data was not background processed but was normalized using quantile normalization produced the least variable plot.

Creating boxplots using the same background correction and normalization settings with the Tukey biweight and Average Log summary methods show similar results, background adds variability and normalization removes it. Figure 2 plots the non differential expression measures for the four background methods with normalization. In most cases the median polish is slightly less variable than the other two methods. However, in the case when there was no background correction, only normalization, Average Log PM was the least variable.

We now turn our attention to how well the observed expression differences match up with the true expression differences. We plot the observed verus the truth and fit a linear regression. Ideally a linear regression between the two would have slope 1. Table 2 contains the slope estimates. We see that normalization has little effect on the slope, but makes the R^2 slightly larger. The background adjustment procedures increase the slope with the ideal mismatch procedure increasing the slopes the most. In general the slope for the biweight was higher than the corresponding slope for the median polish, the slope when using AvgLog PM was always lower than the other two methods.

We thus far have two conflicting conclusions: the backgrounds are bad for increasing the variability of non differential probesets, but good for accurately predicting true fold change. To reconcile the two issues we look at a third issue: how is our ability to detect fold-change affected by the different processing options.

We compare the ability of expression measures to detect differential expression by looking at Receiver Operating Characteristic (ROC) curves. ROC curves for the expression measures computed using the median polish summary measure are shown in figure 3. We see that the more agressive backgrounds are less able to detect differential expression than using no background correction, in each case normalization helped detect differential expression.

The three summarization methods are compared using ROC curves in figure 4. In almost all cases of the different backgrounds, the Average Log summary performed better. The best method was to use no background correction, normalize and then summarize with the Average Log PM.

Conclusions

We found that the background adjustments added variability to the measures of differential expression for non differential probesets. However, by using background adjustments we were more accurately able to predict the true expression difference. By using the ROC curves we were able to reconcile these two and found that using no background correction gave us the best results when

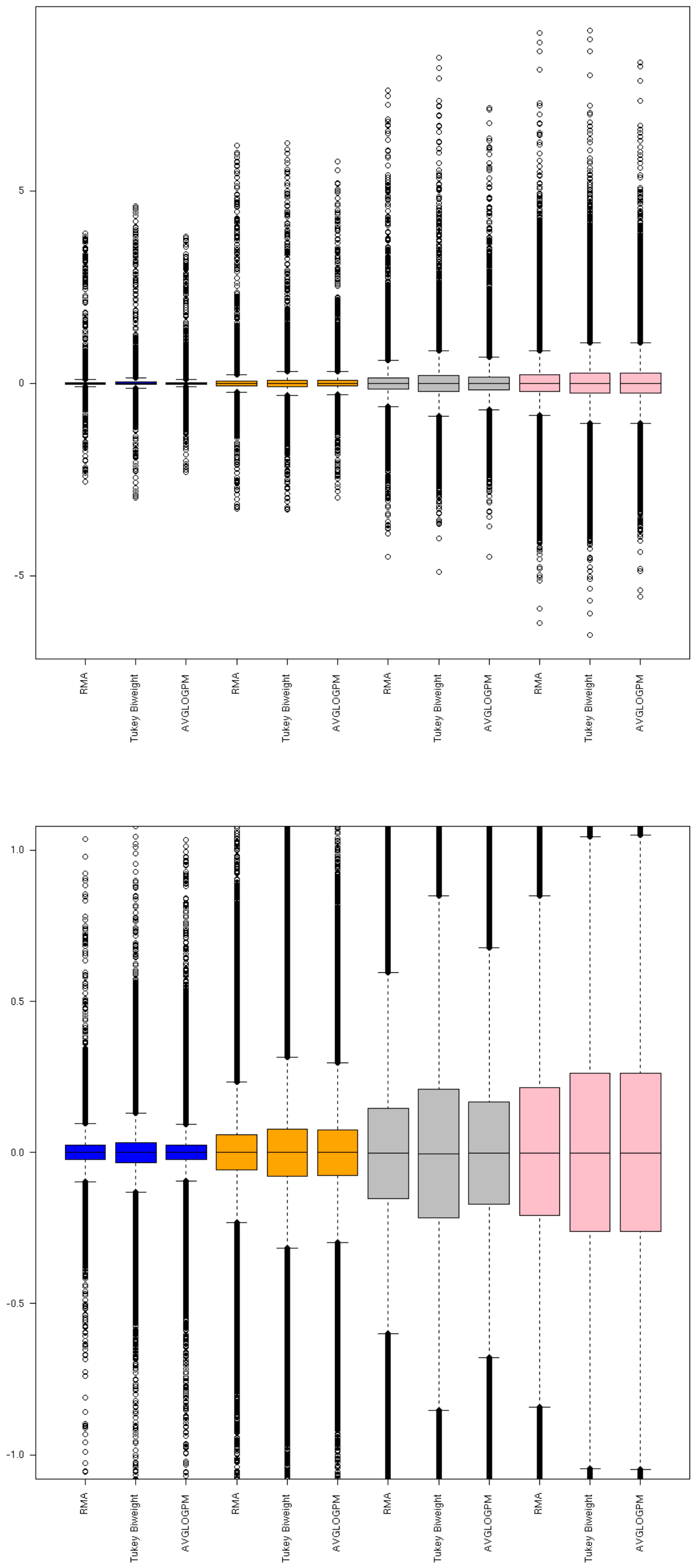

Figure 2: Boxplot of non differential expression differences by preprocessing method using different expression summaries

Summary	Background	Normalization	Slope	R-squared
Median Polish	RMA background	Quantile	0.607	0.966
Median Polish	RMA background	none	0.607	0.964
Median Polish	none	Quantile	0.484	0.955
Median Polish	none	none	0.484	0.952
Median Polish	Mas5.0	Quantile	0.591	0.967
Median Polish	Mas5.0	none	0.591	0.965
Median Polish	IMM	Quantile	0.683	0.974
Median Polish	IMM	none	0.683	0.972
Median Polish	Mas5.0 + IMM	Quantile	0.694	0.967
Median Polish	Mas5.0 + IMM	none	0.694	0.967
Biweight	none	Quantile	0.509	0.962
Biweight	none	none	0.509	0.960
Biweight	Mas5.0	Quantile	0.611	0.975
Biweight	Mas5.0	none	0.611	0.972
Biweight	IMM	Quantile	0.719	0.97
Biweight	IMM	none	0.719	0.969
Biweight	Mas5.0+IMM	Quantile	0.717	0.967
Biweight	Mas5.0+IMM	none	0.717	0.965
AvgLog	none	Quantile	0.453	0.953
AvgLog	none	none	0.453	0.950
AvgLog	Mas5.0	Quantile	0.567	0.965
AvgLog	Mas5.0	none	0.568	0.962
AvgLog	IMM	Quantile	0.664	0.971
AvgLog	IMM	none	0.664	0.970
AvgLog	Mas5.0 + IMM	Quantile	0.678	0.965
AvgLog	Mas5.0 + IMM	none	0.678	0.963

Table 2: Regression Slopes and R^2 comparing observed to truth using spike-ins

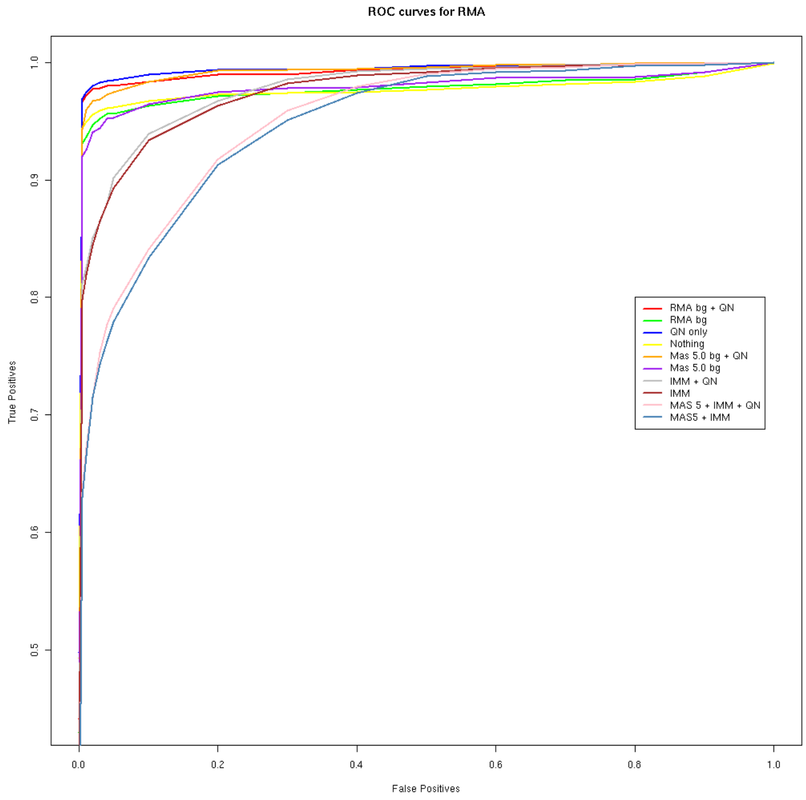

Figure 3: ROC curves for the Median Polish Summary Measure across different pre-processing steps

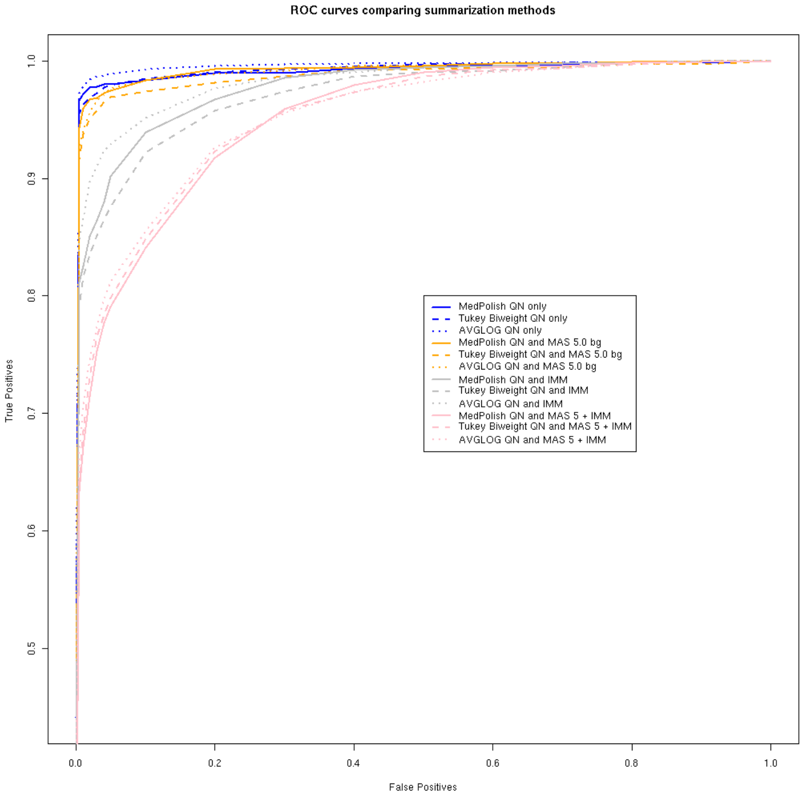

Figure 4: ROC curves for different summary methods

picking differential genes.

The collection of methods used allows us to capture features of both the RMA and MAS 5.0 algorithm. It is clear from our results that it is the background correction methodologies that are driving the differences between the two methods.

References

[1] Bolstad, B.M., Irizarry R. A., Astrand, M., and Speed, T.P. (2003), A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Bias and Variance. To appear in Bioinformatics

[2] Affymetrix (2002) Statistical Algorithms Description Document `http://www.affymetrix.com/support/technical/whitepapers/sadd_whitepaper.pdf`

[3] Irizarry, RA, Hobbs, B, Collin, F, Beazer-Barclay, YD, Antonellis, KJ, Scherf, U, Speed, TP (2003) Exploration, Normalization, and Summaries of High Density Oligonucleotide Array Probe Level Data. To appear in Biostatistics

[4] Rafael A. Irizarry, Benjamin M. Bolstad, Francois Collin, Leslie M. Cope, Bridget Hobbs, and Terence P. Speed (2002) Summaries of Affymetrix GeneChip Probe Level Data. Accepted to Nucleic Acids Research