Effects of Pre-processing on Expression Estimates: Background and Normalization

Ben Bolstad <bolstad@stat.berkeley.edu>, Division of Biostatistics, University of California, Berkeley

Introduction

Background adjustment and cross chip normalization are important steps when computing an expression measure. The choice of method can have drastic effect on the accuracy (low bias) and precision (low variance) of gene expression estimates. We will show the effect that several of the proposed preprocessing methods have on gene expression estimates.

The process of computing expression estimates will be viewed as a three-step process: background correction, normalization and probeset summarization. We will favor RMA (median polish) for our summarization step, but vary both background and normalization steps in our analysis. For our comparison of background methods we will use quantile normalization followed by RMA (median polish) summarization. We will compare normalization methods in the absence of background correction.

Note that this is not an exhaustive comparison of all the different methods that have been proposed.

Methods

Background Correction

A background correction method does one or more of the following:

- **Corrects for background noise and processing effects**
- **Adjusts for cross hybridization**
- **Adjusts estimated expression values to fall on proper scale**

We shall compare the following methods:

- **No background**
- **RMA convolution background**
  - Observed intensity is modeled as sum of normal background and exponential signal
- **MAS 5.0 location dependent**
  - Array is broken into grids and lowest 2% of intensities is used to compute background
- **Ideal mismatch**
  - A subtraction based upon the Mismatch probes
- **MAS 5.0 location dependent + Ideal mismatch**
  - A Combination of the two methods above
- **Standard curve adjustment**
  - Motivated by Figure 1. Concentration dependent adjustment based on model for PM and MM

Normalization

Non-biological factors can contribute greatly to the variability of data from different arrays. Normalization is a process of reducing unwanted variation across chips. It may use information from multiple chips.

We shall compare:

- **No normalization**
- **Quantile normalization**
- **Scaling**
  - Median chip is used as baseline

Data

We will use the Affymetrix HGU_95A spike-in data set (http://www.affymetrix.com/support/datasets.affx). It consists of 59 chips, with 14 probesets spiked-in at 14 different concentrations using a Latin square design. By using spike-in information we have "truth" to measure against.

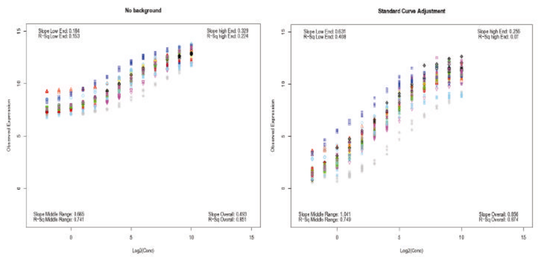
Results

Observed Expression vs Concentration

Table 1. Regression slopes of observed expression versus concentration (both $\log_2$ scale). Higher slopes are better, the ideal slope is 1. Low is less than 2, middle is 2 - 8, and high is 8 - 10.

Method	All	Low	Mid	High
No Background	0.49	0.18	0.67	**0.33**
RMA	0.63	0.38	0.78	0.33
MAS 5.0	0.59	0.32	0.75	0.33
Ideal Mismatch	0.69	0.52	0.82	0.30
MAS 5.0/IMM	0.70	0.56	0.82	0.29
Standard Curve Adjustment	**0.86**	**0.63**	**1.04**	0.26

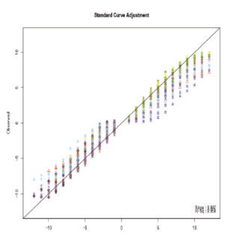
Figure 1. Plots of observed expression versus concentration. The plot for no background correction shows a concentration dependent curve. This motivates the standard curve adjustment method, where we estimate concentration and then linearize the curve. The second plot shows the results when standard curve adjustment has been applied.



Observed vs Expected Fold Change

Figure 2. Plot of observed fold change against expected fold change for the Standard Curve Adjustment method. The 45 degree line is indicated.

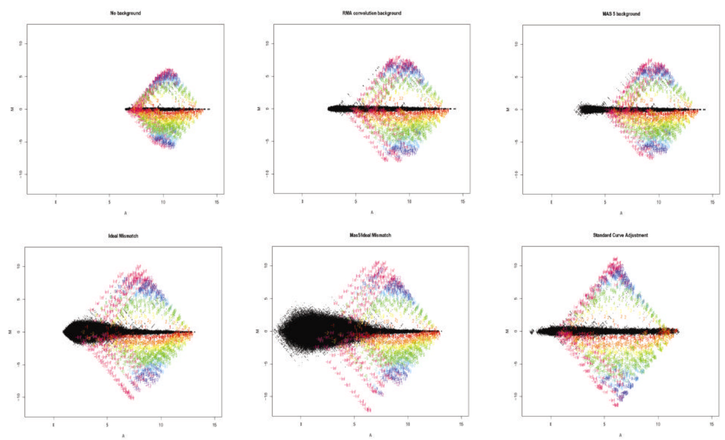
Table 2. Slope estimates for the regression of observed fold change against expected fold change. A higher slope is better. A slope estimate near 1 is desirable.



Method	Slope
No Background	0.48
RMA	0.62
MAS 5.0	0.58
Ideal Mismatch	0.68
MAS 5.0/IMM	0.69
Standard Curve Adjustment	**0.84**

Composite M vs A Plot

Figure 3. M vs A plots for all 91 different pair wise comparisons. Left to right: No background, RMA convolution background, MAS5.0, Ideal Mismatch, MAS5.0/IdealMM and Standard Curve Adjustment. Ideally, we would like low noise for non-differential probesets at the low end and the spike-ins to accurately estimate the true fold change.



Background Methods Have Different Tradeoffs

Table 3. It is important to be able to detect differential genes and have accurate estimates of fold change. The Standard Curve Adjustment did well in both comparisons.

Background Method	Detect Differential Genes	Accurate estimates of Fold change
No Background	**Good**	**Poor**
RMA	**Good**	**Poor**
MAS 5.0	**Good**	**Poor**
Ideal Mismatch	**Poor**	**Good**
MAS5.0/IdealMM	**Poor**	**Good**
Standard Curve Adjustment	**Good**	**Good**

Normalization Has Little Effect on Spike-ins

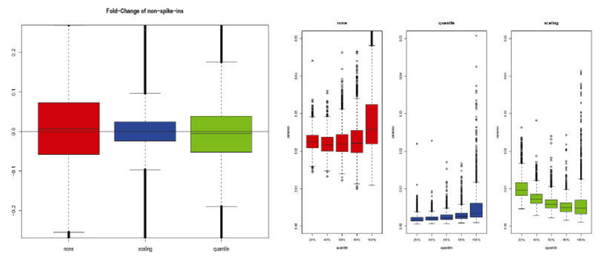
Table 4. Slope estimates for the regression of observed expression against spike-in concentration show little change after normalization. There are slightly higher R^2 values after normalization.

Method	All	Low	Mid	High	FC
No Normalization	0.493 (0.845)	0.185 (0.148)	0.664 (0.733)	0.328 (0.207)	0.484 (0.952)
Quantile	0.493 (0.851)	0.184 (0.153)	0.665 (0.741)	0.329 (0.224)	0.484 (0.955)
Scaling	0.493 (0.852)	0.186 (0.156)	0.663 (0.742)	0.33 (0.225)	0.484 (0.954)

Normalization Reduces Variability

Figure 4. (Left) Boxplot of fold change estimates for non-spike-in probesets. The quantile method (blue) was least variable, followed by scaling (green) and not normalizing (red).

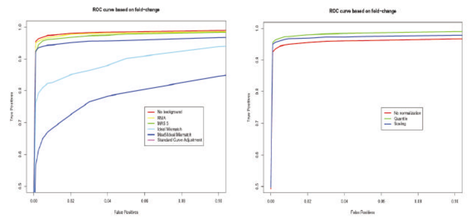
(Right) Boxplots of variability of expression estimates across all 59 arrays for non-spike-ins, broken into 5 blocks each of 20% of the data (low to high). The quantile method (blue) had the lowest variability.



ROC Curves

Figure 5. ROC curves when picking probesets using fold change. Using no background correction gives the fewest false positives. The methods using the ideal mismatch give more false positives.

Normalization improved things, but did not have as large an effect as changing the background method.



Discussion

We found that the choice of preprocessing method can have drastic effect on the computed expression estimate.

The choice of background correction method is particularly important. We found that some methods could more accurately predict the true fold change, but would trade this off for much higher variability in low intensity signals. Other methods would have low variability in the low intensity signals but more poorly predicted fold change.

Normalization is good for reducing variability. It had minimal effect on the magnitude of the computed fold changes for the spike-in. By normalizing we reduced the variability of non-differential genes.

We have not examined how different background and normalization methods interact with each other, but we expect that our general conclusions hold true.

For supplementary information and additional figures see http://www.stat.berkeley.edu/users/bolstad