Assessing Low-Level Processing of Microarray Data

Alexander Ploner[1], Lance D. Miller[2], Per Hall[1], Jonas Bergh[3], and Yudi Pawitan[1]

[1]Medical Epidemiology and Biostatistics, Karolinska Institutet [2]Genome Institute of Singapore [3]Cancer Center Karolinska

1. Background

Photolithographically synthesized high-density oligonucleotide arrays represent a stable and highly standardized platform for measuring the expression profiles of tens of thousands of genes simultaneously. The pre-processing however that is required to convert the raw measurement data into values representative of relative mRNA abundance is still open to debate and comprises two essential choices: that of expression measure, which summarizes the multiple measurements per gene on each chip, and that of normalization procedure, which makes measurements comparable between chips. Numerous approaches to both steps have been suggested. Evaluations on artificial reference data indicate that there is currently no combination that is optimal for all data sets and purposes, even though it has been demonstrated that these choices have a severe impact on any subsequent data analysis.

2. Working Proposition

Our hypothesis is that given a modern large-scale chip covering a large percentage of a species' genome, randomly selected pairs of genes will be *on average* uncorrelated.

Note that we do *not* claim the absence of all correlation between genes, but rather that (a) the number of biologically meaningful relationships between genes in regulatory pathways is small compared to all possible combinations of genes, and (b) positive and negative correlations will tend to cancel out when averaged.

A low-level analysis strategy will be deemed suitable for a given data set, if the resulting normalized expression values are on average uncorrelated for randomly chosen pairs of genes.

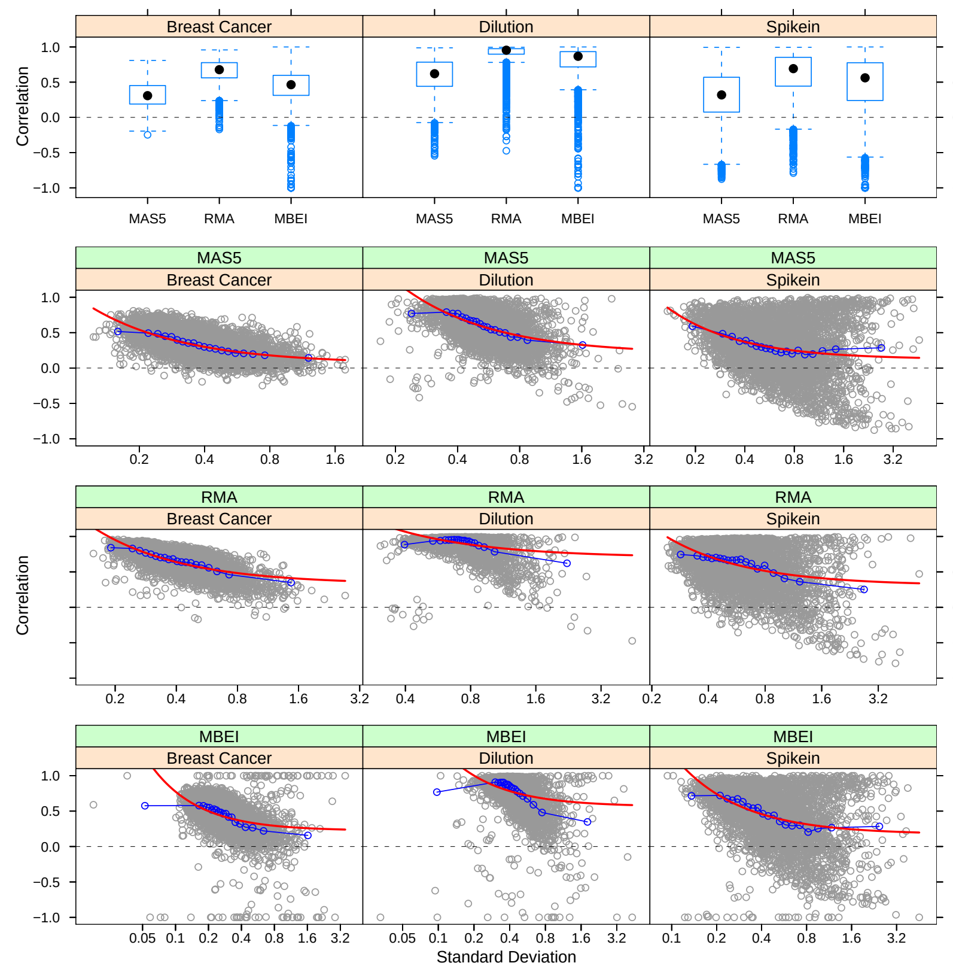

Figure 1: *Correlations for unnormalized expression values of 5000 randomly selected pairs of genes. Lower part: correlations vs. standard deviations; blue = averages, red = model.*

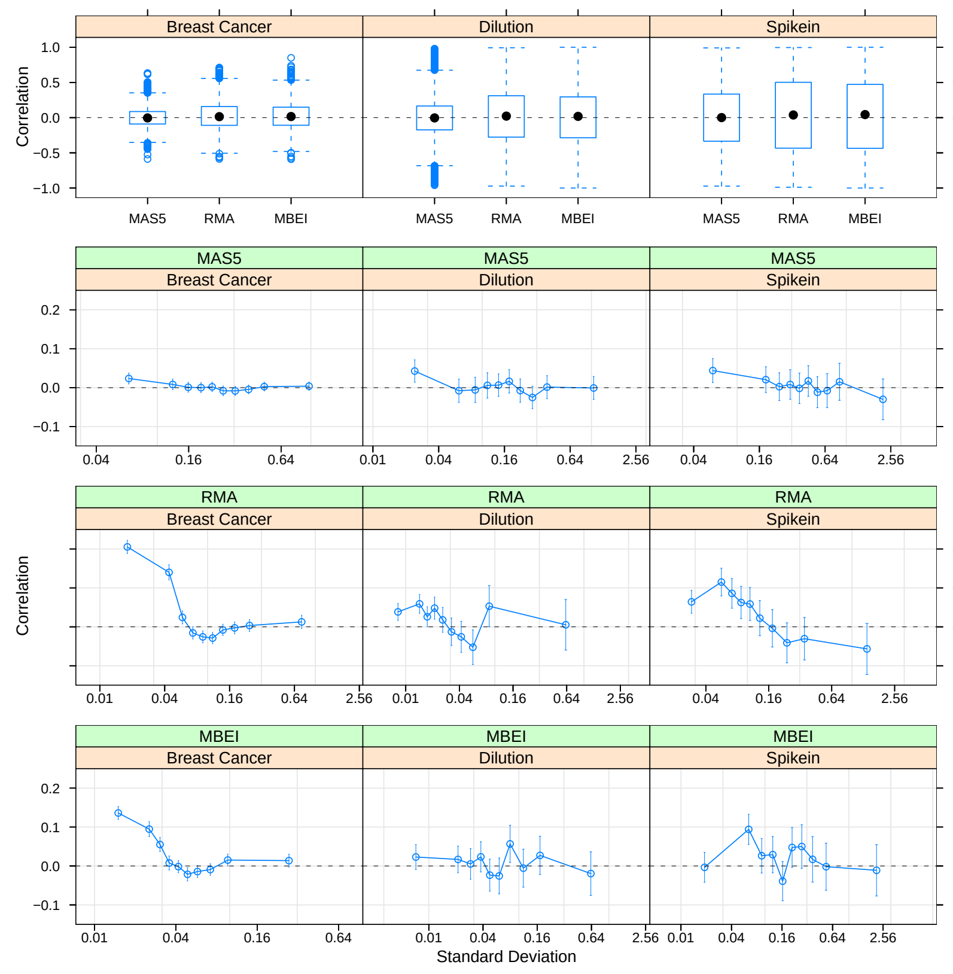

Figure 2: *Correlations between the normalized expression values of the 5000 randomly selected pairs of genes, using the default normalization procedures. Lower part: only averages are shown; 95% confidence intervals are indicated by vertical bars.*

3. Results

Lack of normalization causes correlation. The boxplots in Figure 1 are centered far from zero for all expression measures and data sets. The scatterplots of correlations vs. products of standard deviations show that correlation decreases with increasing variability. The average correlation (in blue) follows roughly the simple model (in red) described below.

Default normalization removes excess correlation. The boxplots in Figure 2 are nicely centered around zero.

Genes with low variability are poorly normalized by RMA and MBEI. The average correlations shown in the lower part of Figure 2 are far from zero at the lower end of variability for these expression measures.

Normalization on housekeeping genes fails to remove excess correlation. The boxplots in Figure 3 are all centered away from zero. Additionally, the curves of average correlations are well above those for the default normalization procedures.

Measured expression of absent genes is poorly normalized in RMA and MBEI. Using additional quality control information, gene expression can be classified as absent or present on each chip. Figure 4 shows that independent of the mean intensity level, the average correlation between pairs of genes is linked to the average number of absent/present calls for each pair, most strongly for RMA and MBEI. Pairs with both genes predominantly present (green) do equally well for all expression measures.

4. Conclusions

- In all three data sets, RMA and MBEI with default normalization are not suitable for genes with low-variability .
- For the Breast Cancer data, the set of generic housekeeping genes is not suitable for normalization.

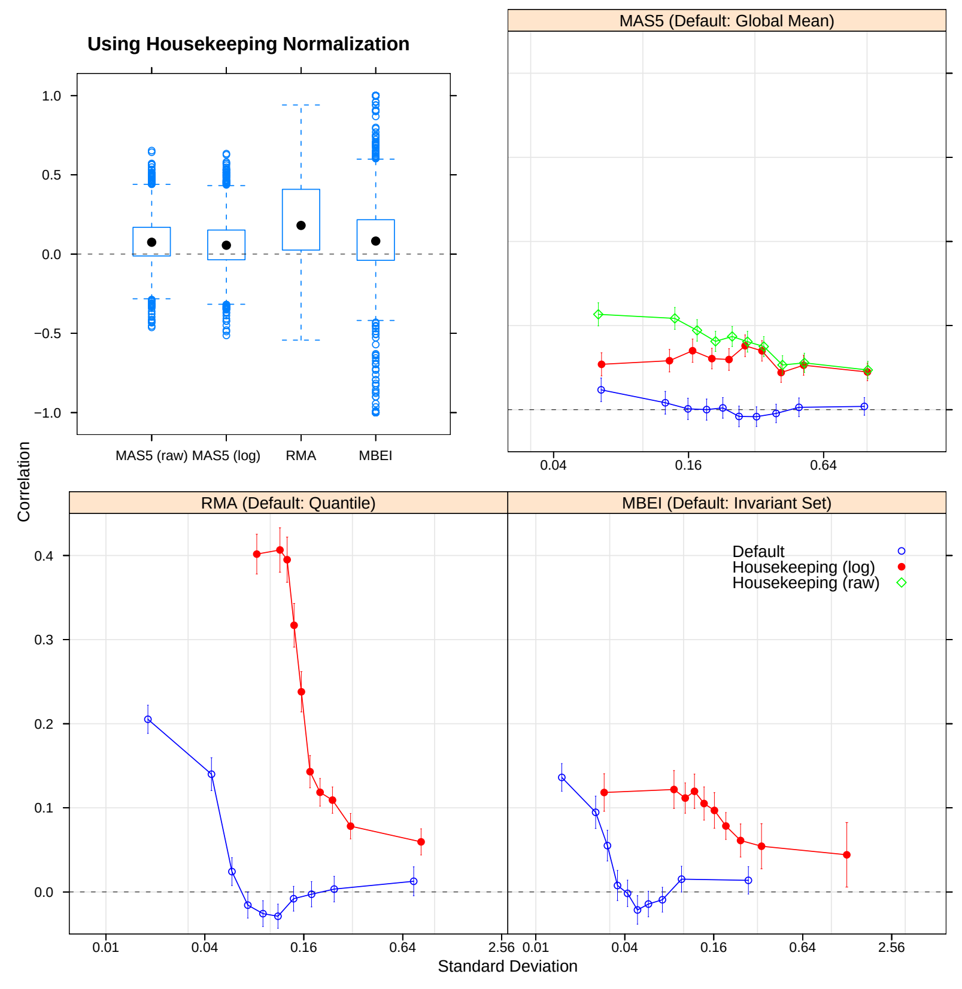

Figure 3: *Correlations for the Breast Cancer data set, comparing normalization to the mean of a predefined set of 100 houskeeping genes with the default normalization procedure.*

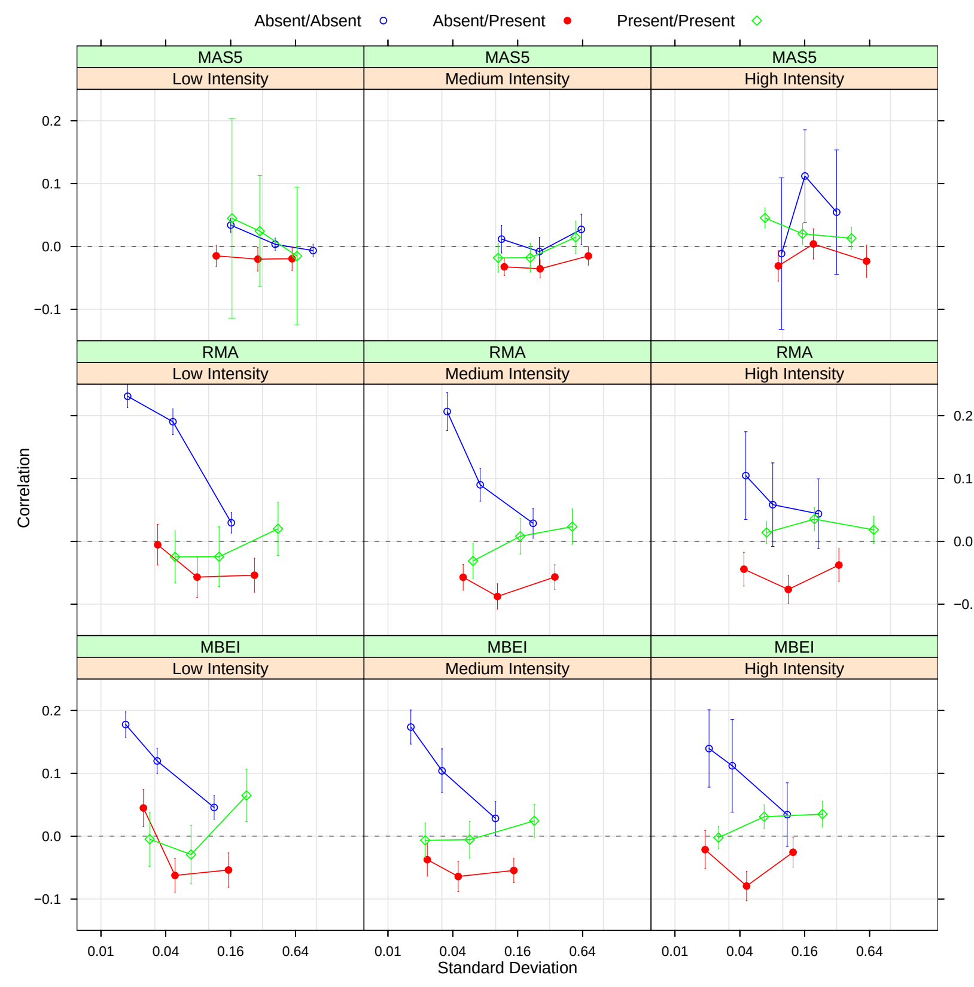

Figure 4: *Correlations of the randomly sampled pairs of genes for the Breast Cancer data set, grouped by mean intensity (columns of plots) and percentage of present calls (multiple curves).*

- For the Breast Cancer data, RMA and MBEI require removal of absent genes in order to be suitable.

Correlations between random pairs of genes allow the direct comparison of different low-level processing strategies for a specific and real data set.

5. Model

We consider as experimental unit one microarray chip with the associated sample from the biological population under study. Each chip yields observation y_i for gene i specified by the array design. We write this as a random variable

$$Y_i = \theta + \psi_i + \epsilon_i$$

with θ a random array effect, ψ_i a random gene effect, and ϵ_i the gene-specific measurement error, all uncorrelated. This can be manipulated to yield

$$Corr(Y_i, Y_j) \approx \frac{\sigma_\theta^2}{\sigma_i \sigma_j} + \text{error},$$

where σ_i^2 is the variance of Y_i. This model can be fitted to the data as in Figure 1 or just serve as a motivation for plotting correlation against the product of standard deviations when investigating lack of fit (Figures 2-4).

6. Methods

MAS5 expression values were used with global mean normalization as default procedure. For RMA, quantile normalization was applied before computing expression values. Similarly, computation of MBEI values followed normalization to a baseline array of average intensities. Housekeeping gene normalization was based on the probes with suffix `2000_` on the Affymetrix HGU133A chip.

The Dilution and Spike-in data sets are artifical reference data that have been widely used for low-level processing. The Breast Cancer data set was collected from a population-based breast cancer cohort at Karolinska Hospital, Stockholm.

Contact: `Alexander.Ploner@meb.ki.se`