BIOS 7659 Journal Club:

A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. (Bolstad et al., 2003)

Tim Vigers

9/22/2020

Introduction

- The goal of normalization is to separate the interesting biological variation from the variation that is a result of sample preparation, array production and processing, etc.
- Affymetrix proposes scaling the arrays so that each one has the same mean expression (based on a summary measure).
  - This does not work well when there are non-linear relationships between arrays.

Alternatives

- Other approaches, such as non-linear smooth curves or transforming data to standardize the distribution of intensities across arrays, rely on picking a "baseline" array.
- Bolstad et al. compare three different approaches, all of which combine data from every single array (complete data).

Cyclic loess

- Basically an extension of the M vs. A plots discussed in class, but applied to pairwise combinations of Affymetrix arrays.
- M is the difference in log expression values and A is the average (a Bland-Altman plot).

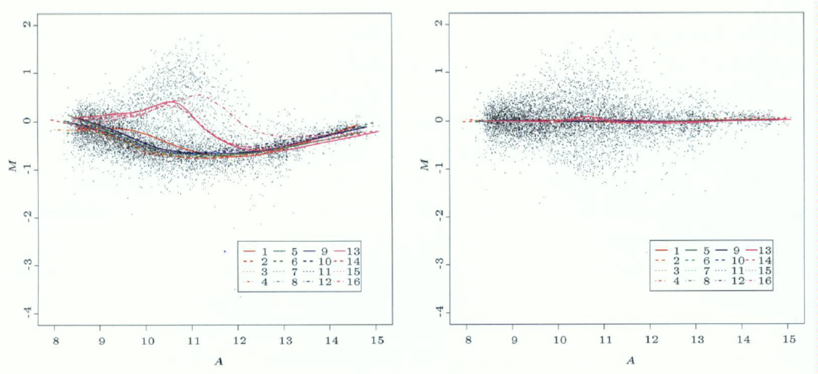

Figure 1: Dudoit et al., 2002

Cyclic loess

1. Take two arrays i and j, each with probes $k = 1, ..., p$.
2. Create an MA plot for these two arrays, and fit a loess curve through these data:

$$M_k = log_2(\frac{x_{ki}}{x_{kj}}), A_k = \frac{log_2(x_{ki}x_{kj})}{2}$$

3. Subtract the normalization curve fits $M'_k = M_k - \hat{M}_k$ and obtain adjusted probe intensities:

$$x'_{ki} = 2^{A_k + \frac{M'_k}{2}}, x'_{kj} = 2^{A_k - \frac{M'_k}{2}}$$

4. Take each of these adjustments (one for each pairwise comparison between arrays) and weight them equally across the set of arrays.

Contrast method

Very similar to the cyclic loess method, because it's another way of normalizing based on M vs. A:

1. Data is converted to the log scale and the basis is transformed (this is just a fancy linear algebra step).
2. $n - 1$ normalizing curves are fit to the transformed basis as in cyclic loess.
3. Data is transformed again so that the normalizing curves lie on the horizontal, this time using a smooth function.
4. This normalized data is transformed back to the original basis and exponentiated.

This is slightly faster than cyclic loess but fitting the curves can still be slow.

Contrast method

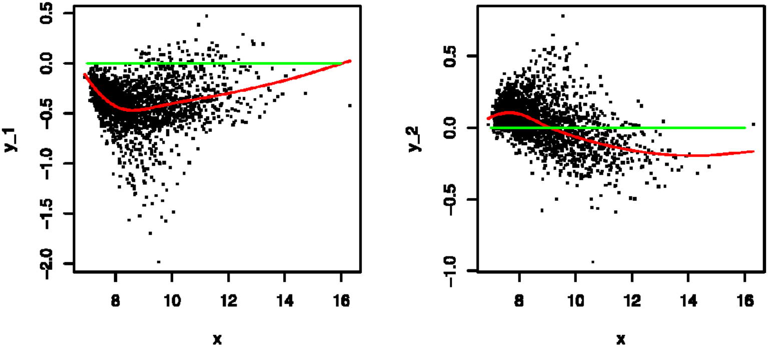

Figure 2: Scatter plots of the 2 contrasts against the mean for 3 arrays, A, B, and C, prior to normalizing. The red curve is the fitted normalizing curve, and the green line is the reference line. (Åstrand, 2003)

The quantile method

- The goal of this method is to standardize the distribution of probe intensities across all arrays.
- The approach is an *n*-dimensional extension of the fact that, given a quantile-quantile plot where all of the points are on a straight diagonal line, you can be fairly sure that the two data vectors have the same distribution.
- We want to project our data onto the unit vector $(\frac{1}{\sqrt{n}}, ..., \frac{1}{\sqrt{n}})$.
- Let $\mathbf{q}_k = (q_{k1}, ..., q_{kn})$ be the vector of k^{th} quantiles for $k = 1, ..., p$. Then:

$$proj_{\mathbf{d}}\mathbf{q}_k = (\frac{1}{n}\sum_{j=1}^{n} q_{kj}, ..., \frac{1}{n}\sum_{j=1}^{n} q_{kj})$$

Quantile normalization algorithm

1. Given n arrays with p probe intensity measurements, make the $p \times n$ matrix X, where each column has all the data from a single array.
2. Sort each column of X to produce X_{sort}. So, each row in X_{sort} is a quantile.
3. Take the mean of each row, and replace every value in the row with the mean to produce X'_{sort}
4. Put each column of X'_{sort} back in the original ordering from X to produce $X_{\text{normalized}}$

This approach could theoretically be a problem for probes that have the same value across all arrays, but in practice this isn't an issue.

Scaling method

Based on the approach suggested by Affymetrix, but this paper uses a probe-level version.

1. Choose a baseline array x_{base}: Usually this is the median array, but doesn't necessarily have to be.
2. For each other array, calculate the mean trimmed intensity $\tilde{x}_i$ and find
$$\beta_i = \frac{\tilde{x}_{\text{base}}}{\tilde{x}_i}$$
3. Normalized intensities are $x_i' = \beta_i x_i$

Non-linear method

- The scaling method is the same as fitting a straight line with intercept 0 between x_{base} and x_i.
- This can be extended to non-linear methods, usually a loess curve such that:

$$x_i' = \hat{f}_i(x_i)$$

where $\hat{f}_i(\cdot)$ is the curve mapping from array i to baseline.

Data

Bolstad et al. tested these techniques on two datasets:

1. 30 arrays each from liver and central nervous system cell lines, plus 15 arrays with mixtures of cell lines in 75:25, 50:50, and 25:75 proportions.
2. A dilution series of 27 arrays, plus two sets of triplicates (6 arrays total) from a Latin square experiment.

Results

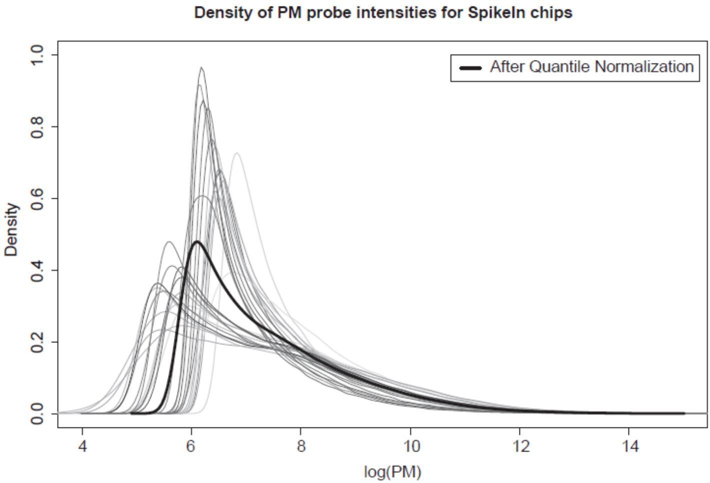

Figure 3: A plot of the densities for PM for each of the 27 spike-in datasets, with distribution after quantile normalization superimposed. (Bolstad et al., 2003)

Results

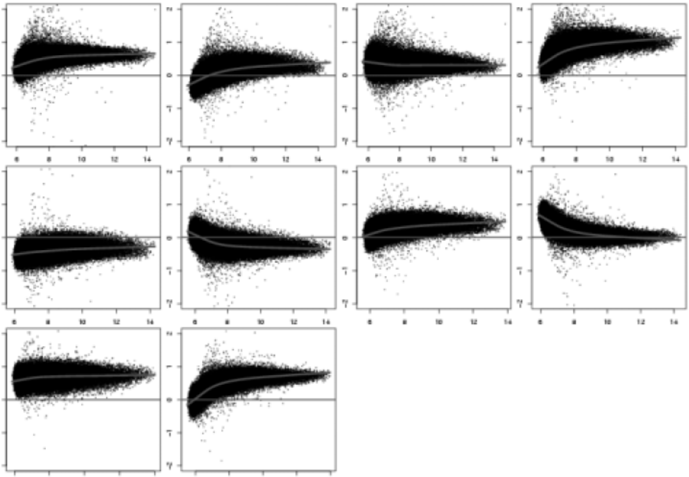

Figure 4: 10 pairwise M versus A plots using liver (at concentration 10) dilution series data for unadjusted data. (Bolstad et al., 2003)

Results

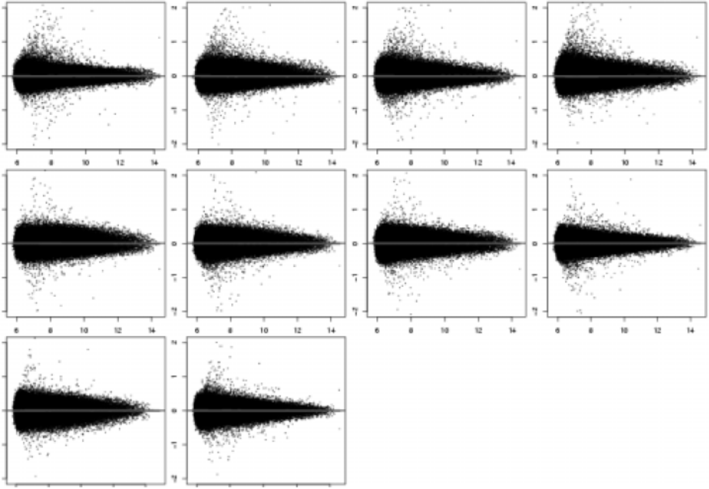

Figure 5: 10 pairwise M versus A plots using liver (at concentration 10) dilution series data after quantile normalization. (Bolstad et al., 2003)

Results

- To compare methods at the probeset level, this paper uses the Robust Multichip Average (RMA).
- Plotting the log variance ratio against the log mean between two methods allows you to examine “differences in the between array variations and intensity dependent trends.”
- When the loess line is below the x-axis, the first method has lower variance, and vice versa when the line is above the horizontal axis.

Results

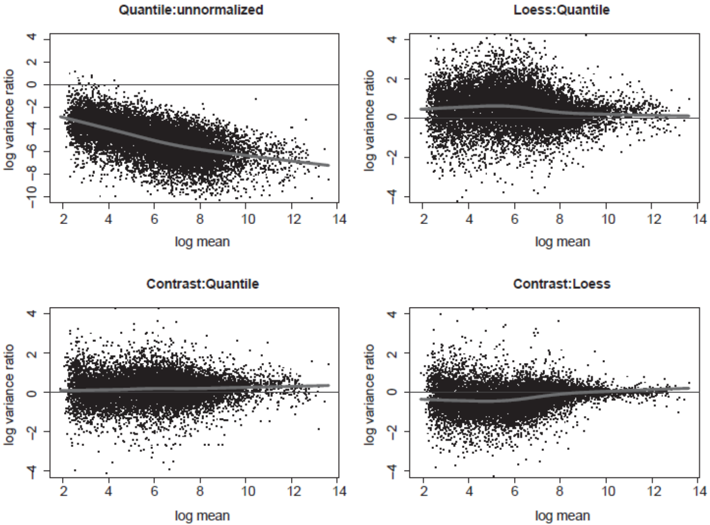

Figure 6: log_2 variance ratio versus average log_2 mean across 5 arrays for liver dilution data at concentration 10. (Bolstad et al., 2003)

Results

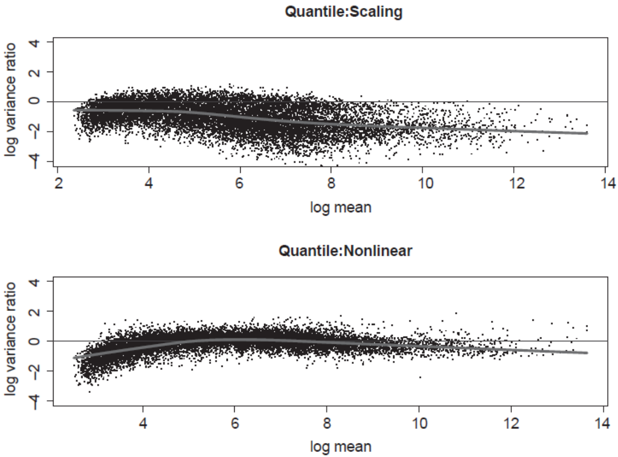

Figure 7: *log_2* variance ratio versus average *log_2* mean using the spike-in data. Comparing the baseline methods with the quantile method. (Bolstad et al., 2003)

Results

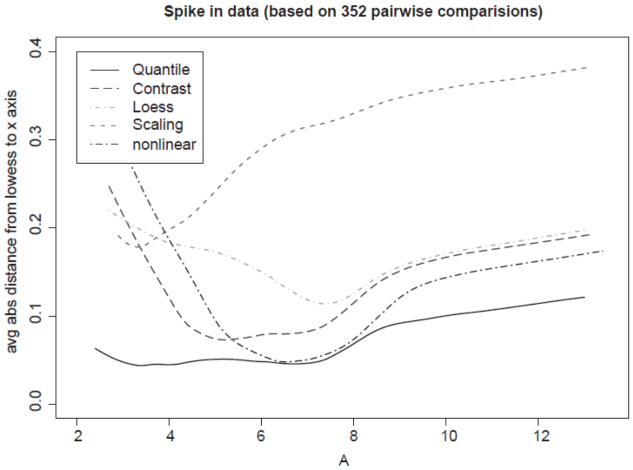

Figure 8: Comparing the ability of methods to reduce pairwise differences between arrays by using average absolute distance from loess smoother to x-axis in pairwise M versus A plots using spike-in dataset. Smaller distances are favorable. (Bolstad et al., 2003)

Results

- The spike-in data can be used to assess bias for these techniques by fitting a regression model that includes the known concentration:

$$\log_2(\text{Expression}) = \beta_0 + \beta_1 \log_2(\text{concentration}) + \epsilon$$

- Ideally you would see $\beta_1 = 1$ for the spike-in probesets and $\beta_0 = 0$ for the non-spike-in sets.
- However, none of the median slopes for the non-spike-in probesets are exactly 0, which suggests that spike in concentration affects intensity.
- The authors recommend adjusting the slopes of the normalization procedures but subtracting the median slope of the non-spike-ins (e.g. $0.845 + 0.005 = 0.850$).
- There is likely some error due to "pipette" effect, so concentrations may not be exactly correct.

Results

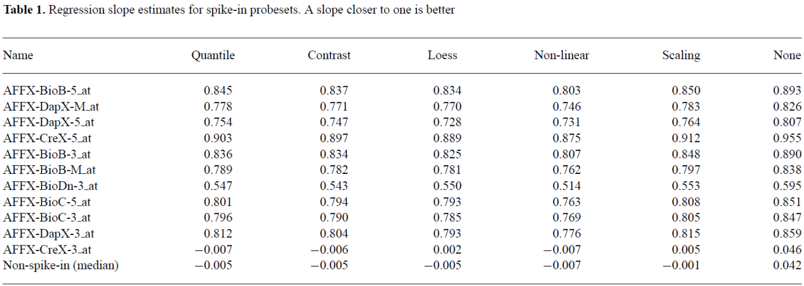

Table 1. Regression slope estimates for spike-in probesets. A slope closer to one is better

Name	Quantile	Contrast	Loess	Non-linear	Scaling	None
AFFX-BioB-5_at	0.845	0.837	0.834	0.803	0.850	0.893
AFFX-DapX-M_at	0.778	0.771	0.770	0.746	0.783	0.826
AFFX-DapX-5_at	0.754	0.747	0.728	0.731	0.764	0.807
AFFX-CreX-5_at	0.903	0.897	0.889	0.875	0.912	0.955
AFFX-BioB-3_at	0.836	0.834	0.825	0.807	0.848	0.890
AFFX-BioB-M_at	0.789	0.782	0.781	0.762	0.797	0.838
AFFX-BioDn-3_at	0.547	0.543	0.550	0.514	0.553	0.595
AFFX-BioC-5_at	0.801	0.794	0.793	0.763	0.808	0.851
AFFX-BioC-3_at	0.796	0.790	0.785	0.769	0.805	0.847
AFFX-DapX-3_at	0.812	0.804	0.793	0.776	0.815	0.859
AFFX-CreX-3_at	−0.007	−0.006	0.002	−0.007	0.005	0.046
Non-spike-in (median)	−0.005	−0.005	−0.005	−0.007	−0.001	0.042

Figure 9: (Bolstad et al., 2003)

Results

- Next they compared the means produced by quantile normalization and the non-linear method.
- Used two sets of triplicates (6 arrays) where the fold change between the spike-in concentrations is high.
- About half of the spike-in probesets are high in one triplicate and low in the other, and the opposite for the remaining probesets.
- For the non-linear method, they changed which array was used as baseline, including two “synthetic” baseline arrays made by taking probe-wise means and medians.

Results

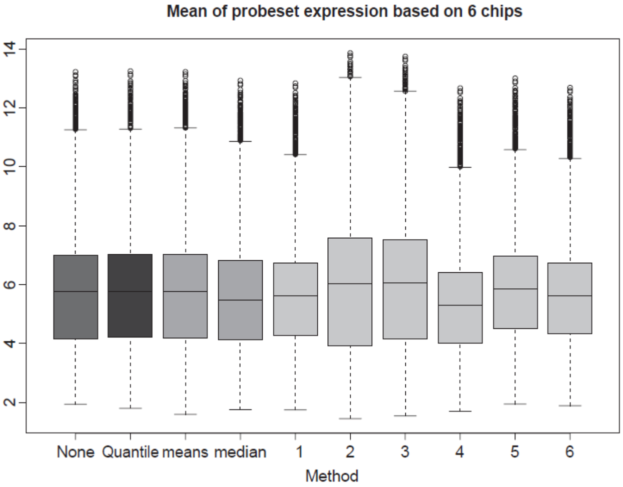

Figure 10: Distribution of average (over 6 chips) of a probeset expression measure using different baseline normalizations. (Bolstad et al., 2003)

Results

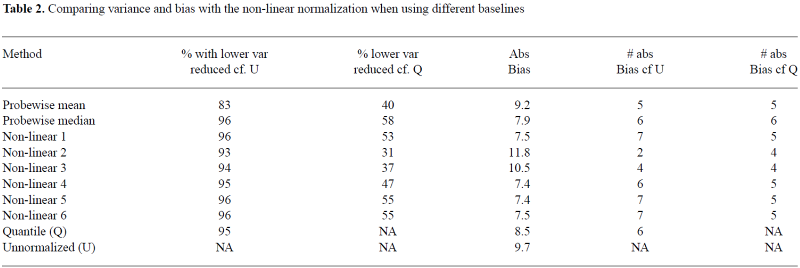

Table 2. Comparing variance and bias with the non-linear normalization when using different baselines

Method	% with lower var reduced cf. U	% lower var reduced cf. Q	Abs Bias	# abs Bias cf U	# abs Bias cf Q
Probewise mean	83	40	9.2	5	5
Probewise median	96	58	7.9	6	6
Non-linear 1	96	53	7.5	7	5
Non-linear 2	93	31	11.8	2	4
Non-linear 3	94	37	10.5	4	4
Non-linear 4	95	47	7.4	6	5
Non-linear 5	96	55	7.4	7	5
Non-linear 6	96	55	7.5	7	5
Quantile (Q)	95	NA	8.5	6	NA
Unnormalized (U)	NA	NA	9.7	NA	NA

Figure 11: (Bolstad et al., 2003)

- For all approaches, about 95% of probesets have reduced variance.
- Compared to the quantile approach, over 50% of the probesets have reduced variance for 4 approaches (median, NL1, NL5, and NL6).

Results

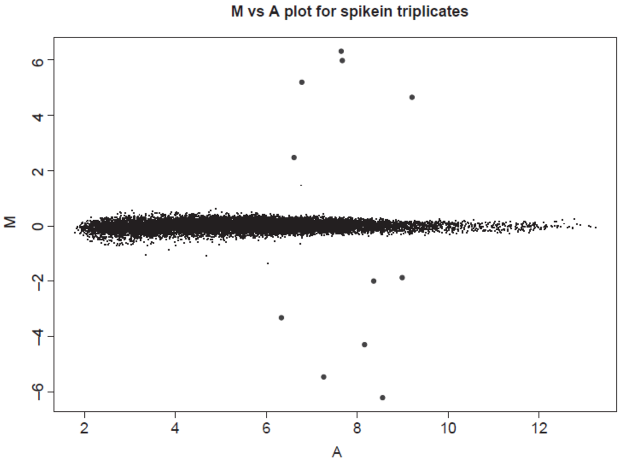

Figure 12: M versus A plot for spike-in triplicate data normalized using quantile normalization. Spike-ins are clearly identified. (Bolstad et al., 2003)

Conclusions

- All three complete data methods (cyclic loess, contrast, and quantile) reduced variability across arrays more than scaling. The non-linear approached performed about as well as all three.
- The complete data methods performed comparably in terms of bias. The non-linear approach did not perform well in spike-in regressions. The slopes for scaling were closer to 1 but more variable than other methods.
- Choice of baseline can affect downstream analysis, so complete data methods are preferable.
- Quantile normalization is the fastest of the complete data methods, and therefore the best approach.

Questions

1. What are some potential disadvantages of the quantile approach?
2. What about advantages of the Affymetrix-recommended approach?
3. What are some potential problems with this paper's methodology?
4. Would you draw the same conclusions as the authors based on these results?

References

1. Dudoit, S., Yang, Y. H., Callow, M. J., & Speed, T. P. (2002). Statistical Methods for Identifying Differentially Expressed Genes in Replicated cDNA Microarray Experiments. Statistica Sinica, 12(1), 111–139. JSTOR.
2. Åstrand, M. (2003). Contrast Normalization of Oligonucleotide Arrays. Journal of Computational Biology, 10(1), 95–102. https://doi.org/10.1089/106652703763255697
3. Bolstad, B. M., Irizarry, R. A., Åstrand, M., & Speed, T. P. (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics, 19(2), 185–193. https://doi.org/10.1093/bioinformatics/19.2.185