Normalization of Microarray Data

Paul Gauthier, Michael Ringenburg
gauthier@cs.washington.edu, miker@cs.washington.edu

December 17, 2003

1 Sources of Error in Microarray Results

DNA microarrays are a powerful technology for analysis of gene expression levels within cells. Both the cDNA and oligonucleotide technologies for microarrays are fairly young and are prone to a broad collection of errors and inaccuracies [KYML02]. In this paper we briefly discuss some sources of error and detail a variety of methods for normalizing the data with respect to some of these error sources.

Essentially microarrays operate by attaching fragments of single-stranded DNA to a small glass plate or chip in a grid pattern. Each of the thousands of spots on the grid contains many copies of a unique sequence. A sample containing unknown quantities of mRNA sequences is deposited on each spot on the plate. mRNA sequences in the sample hybridize with the single-strand sequences attached to the chip. The sample mRNA is tagged with a dye or marker which will be visible on an image of the chip. Genes with higher expression levels appear as spots with higher intensities on the scanned slide image. With cDNA technology red and green dyes are used to run two experiments simultaneously on the same slide. With the oligonucleotide technology each slide runs a single experiment using only a single marker.

These technologies rely on hybridization between the sample strands and the strands affixed to the slide. One class of potential inaccuracies occurs because hybridization can occur without a perfect match between the strands. Sequences which are *mostly* similar may still bond to some extent, confusing the results. Isoforms of very similar sequence can often easily hybridize and mask the measurement of the target gene. The sequences fixed to the slides are normally quite short and are often systematically selected from the $3'$ end of the target gene. Other genes of substantially different sequence, but with a similar $3'$ end may therefore bond to a given spot.

The oligonucleotide arrays use *perfect match* (PM) and *mismatch* (MM) pairs of spots to combat these probe specificity issues. The MM spot has one base changed as compared to its matching PM spot. If a result shows high expression levels on both a PM and an MM spot, the PM result should be discounted as it may indicate the target gene is being overwhelmed by another similar sequence. Unfortunately, this calibration can distort legitimate PM matches which also happen to successfully hybridize to the MM spot.

At a higher level there are larger questions about the accuracy of microarrays. There is evidence that slides often contain misprinted spots with incorrect probe sequences attached. One analysis found over 20% of the spots on a slide contained incorrect sequences. Further, comparing cDNA, oligonucleotide and more traditional Northern blot analysis has shown wide discrepancies. The same sample analyzed by all there technologies produced results that varied over almost two orders of magnitude [KYML02].

The remainder of this paper is concerned with addressing a specific subset of error sources.

- Dye color variation – The intensity of the red and green dyes used in cDNA microarrays may not be directly comparable due to chemical differences in the dyes.
- Scanning variation – Results from different slides may be incomparable because of differences in the scanning process.

- Print-tip effects – The mRNA samples are spotted onto the slide with a grid of print tips. Results from spots printed with different print tips may not be directly comparable due to differences in the tip opening or accumulated wear and tear.
- Slide preparation and wet-lab variables – Differences in the process leading up to the actual microarray experiment may introduce variations in the results. Slight temperature variations in the sample cultures or differences in how the cultures are prepared for each slide are examples of this type of inacuracy.
- Variance increases with intensity – The variance of measurements appears to increase with the overall expression level of a gene [HvHS$^+$02]. A given increase in expression level is less significant for a highly expressed gene making it hard to ascertain which results are indeed significant.

2 Correcting for Experimental Differences

The raw output of a cDNA microarray is the set of $(\log R_i, \log G_i)$ tuples of red-green spot intensities scanned from the slide. Usually these values have been background-corrected by substracting the intensity of the nearby slide background. Given those values, we can define $M_i = \log \frac{R_i}{G_i}$ and $A_i = \frac{1}{2}\log(R_i G_i)$ for each of the genes on the slide. In this section we discuss different methods for obtaining M_i^* the normalized values of M_i as covered in [YDLS01, PYK$^+$03].

All of these normalization methods are based on the assumption that some of the genes have nearly constant expression levels. For these constantly expressed genes we would expect $M_i = \log \frac{R_i}{G_i} \approx 0$ and any observed deviation from $M_i \approx 0$ is the result of some experimental difference such as a dye bias. Ideally, one should only use the constantly expressed genes to determine the normalization adjustments for the whole collection. In practice, there are a range of options available some of which are listed below. The best method for identifying constantly expressed genes may depend on the specifics of the experiment. The normalization methods we will discuss vary in their robustness when their inputs contain some differentially expressed genes.

- All genes - The method for determining the normalization adjustments should be robust to outliers (highly differentially expressed genes).
- Control genes - The experimental setup may include genes specially intended to be constantly expressed. There may also be an expectation that certain genes will be constantly expressed due to biological constraints (housekeeping genes).
- Rank invariant genes - If genes are rank ordered based on their $\log R_i$ and $\log G_i$ values, use that set of genes whose rank is stable, or nearly stable, for normalization.

To normalize M_i, we need to estimate some normalization factor c such that $M_i^* = M_i - c \approx 0$ for constantly expressed genes. The normalization factor c will then be used to compute $M_i* = M_i - c$ for all the (possibly differentially expressed) remaining genes.

2.1 Global Normalization

Global normalization assumes that the red and green dye intensities are related by a constant factor. That is, $M_i \approx \alpha$ for the constantly expressed genes. Typically the constant α is estimated by taking the median of the control genes. The normalized intensities are therefore $M_i^* = M_i - \alpha$. Figure 1(A) shows a raw microarray dataset without normalization and Figure 1(B) shows the same data after global normalization.

2.2 Linear Normalization

Linear normalization assumes that the relationship between the dyes depends on the overall intensity of the dyes, A_i, in a linear fashion. So for constantly expressed genes $M_i \approx \beta_0 + \beta_1 A_i$ for appropriate constants β_0

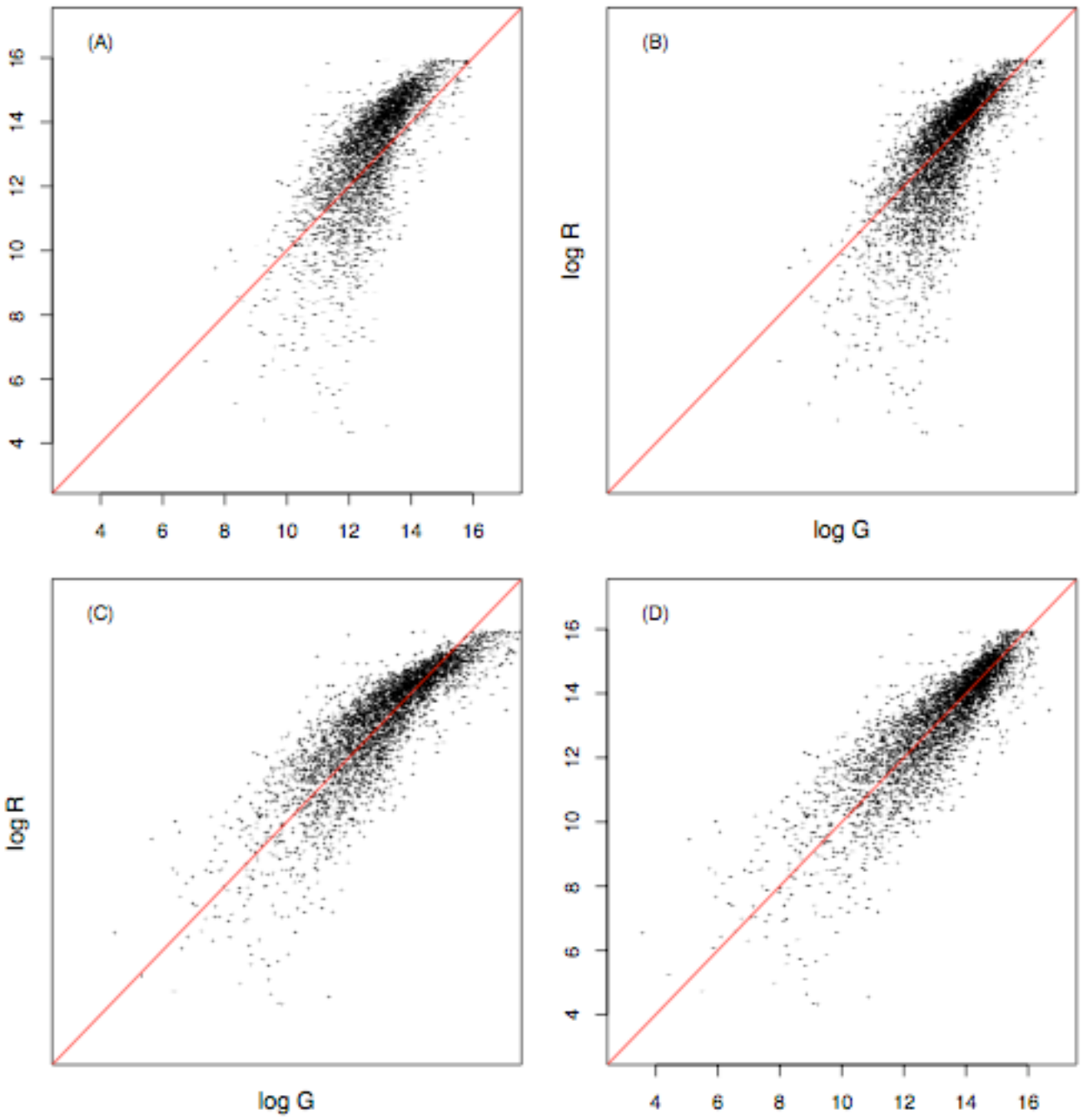

Figure 1: Scatter plots from [PYK$^+$03] of $(\log G_i, \log R_i)$ for **(A)** an unnormalized cortical stem rat cell microarray data set (see Section 2.7); and that data **(B)** globally, **(C)** linearly, and **(D)** non-linearly normalized.

and β_1. These constants are typically estimated by a least-squares fit through the M_i vs. A_i plot of all the control genes. The normalized intensities are therefore $M_i^* = M_i - \beta_1 A_i - \beta_0$. Figure 1(C) shows the results of linear normalization.

2.3 Non-Linear Normalization

Non-linear normalization also assumes that the dye relationship varies with intensity. Rather than fitting a line through the data, the *lowess* fit of the data is used. Lowess produces a robust locally-linear fit of the data. Since it is robust, it will tolerate some differentially expressed genes in the control group.

With this method we have $M_i^* = M_i - c(A_i)$ where $c(A_i)$ is the result of the lowess fit through the M_i vs. A_i plot of all the control genes. Figure 1(D) shows the results of non-linear normalization.

2.4 Dye Swap Experiments

In a dye swap experiment the same pair of mRNA samples is hybridized against two microarrays with the dye assignments reversed. This results in $(\log R_i, \log G_i)$ results from one slide and $(\log R_i', \log G_i')$ from the other. Given this, we have M_i, A_i as before as well as $M_i' = \log \frac{R_i'}{G_i'}$ and $A_i' = \frac{1}{2}\log(R_i' G_i')$ from the dye-swapped slide.

If $M_i^* = M_i - c$ and $M_i^{*'} = M_i' - c'$ where c and c' are determined using any of the above methods, then we should expect $M_i - c \approx -(M_i' - c')$. Since this is a dye swap experiment, we also expect $c \approx c'$. Using these assumptions Park, et al. derive

$c \approx \frac{1}{2}\left[\log \frac{R_i}{G_i} + \log \frac{R_i'}{G_i'}\right] = \frac{1}{2}(M_i + M_i')$

Any of the normalization methods (global, linear or nonlinear) can be applied to a dye-swap experiment by using $M_i'' = \frac{1}{2}(M_i + M_i')$ and $A_i'' = \frac{1}{2}(A_i + A_i')$ in place of M_i and A_i.

2.5 Print Tip Effects

The sample mRNA is applied to each of the spots on a slide using a print tip. There are normally far fewer print tips than the total number of genes, so sections of a slide are each spotted by a different tip. The openings on the ends of these tips may be of different sizes or shapes, or may wear differently over time. As such, the results from spots printed with different tips may not be comparable. The normalization techniques discussed above can be used separately on each set of genes printed by a single print tip. Each print tip group of genes should have it's own normalization factor c estimated separately by whichever method is chosen.

2.6 Oligonucleotide Microarrays

The above normalization techniques are equally applicable to a series of d single color oligonucleotide slides. Say each slide $k = 1..d$ produces a set of measured intensities y_{ki} for each gene i. Each slide $k = 2..d$ can be separately normalized against slide 1 using $M_{ki} = \log \frac{y_{ki}}{y_{1i}}$ and $A_{ki} = \frac{1}{2}\log y_{ki} y_{1i}$. After all the slides have been normalized in this fashion their results should be comparable to each other.

2.7 Evaluation of Methods

Park, et al. [PYK+03] analyzed a microarray data set from a study of cortical stem rat cells. In this experiment $I = 2$ different (very closely related) tissue samples were hybridized against a cDNA microarray at $J = 6$ different time points. As well, each microarray hybridization was repeated $K = 3$ times for a total

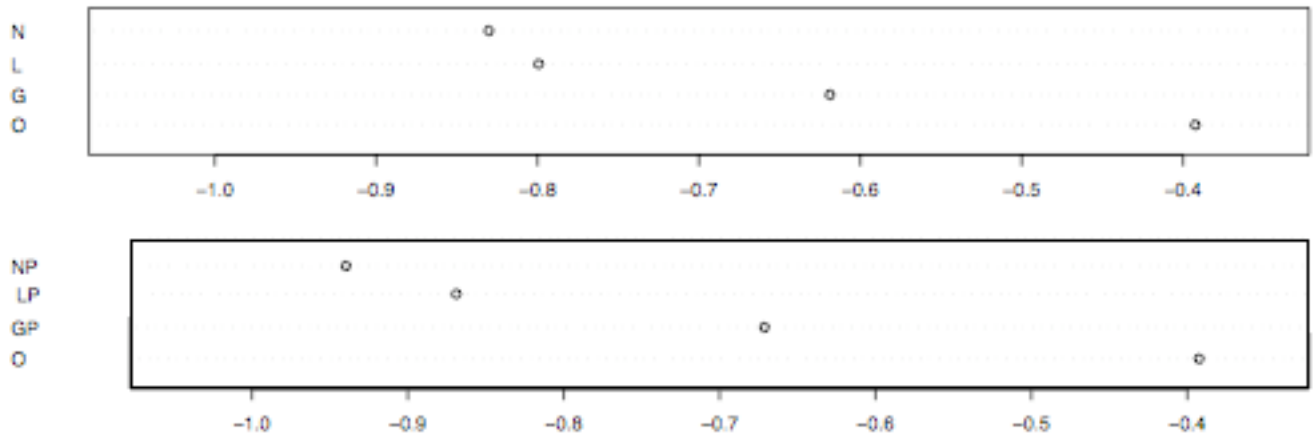

Figure 2: Results from the Park, et al. [PYK+03] variance analysis showing the mean value of $\log \sigma_l^2$ variance estimates for various normalization strategies. **O, G, L, N** refer the original (not normalized) data and the globally, linearly and non-linearly normalized data respectively. **GP, LP, NP** refer to those normalization methods applied separately to the genes from each print tip group.

of 36 slide results sets. Each microarray contained spots for $N = 3,840$ genes. The results data from this experiment are y_{ijkl}, the logarithm of the red to green intensity ratio from group $i = 1..I$, at time $j = 1..J$, replication $k = 1..K$ for gene $l = 1..N$.

The variance for each gene l can be estimated by $\sigma_l^2 = \frac{1}{IJ(K-1)} \Sigma_i \Sigma_j \Sigma_k (y_{ijkl} - \bar{y}_{ij\cdot l})$, where $\bar{y}_{ij\cdot l} = \frac{1}{K} \Sigma_{k=1}^{K} y_{ijkl}$. We can examine the distributions of the $l = 1..N$ variances σ_l^2 of each gene after the data has been normalized by the various methods. Better normalizations will result in smaller variance estimates. Park, et. al, used this model and a more flexible ANOVA model to estimate variance. The results shown in Figure 2 are from the ANOVA model, but they are substantially similar to those from the above variance estimates.

Park, et al. conclude that the intensity dependant normalization methods (linear and non-lineaer) outperform the global method. It is not clear that the non-linear method provides significant additional benefit beyond the linear approach. They also conclude that within print tip normalization seems to provide enough benefit to make it worth considering.

3 Variance Stabilization

The previously described algorithms focus on the goal of normalizing the intensities of microarray data. In this section, we introduce techniques which have the additional goal of variance stabilization. These algorithms transform the intensity data and replace the standard log intensity ratio metric (M) with a new metric Δh such that the variance $v(h_k)$ for a gene k is not dependent on the gene's mean intensity u_k. In this section we describe a method due to Huber *et al.* [HvHS+02] from 2002. Similar work was done by Geller *et al.* [GGHR03] in 2003. We begin by addressing the motivation for variance stabilization. We then present the model due to Huber *et al.* [HvHS+02], and close with a discussion of the results.

3.1 Motivation

As we can see in Figure 3 (taken from Huber *et al.* [HvHS+02]), there is a strong dependence between the variance and the mean. The variance has a non-zero y-interecept, and tends to increase approximately quadratically with the mean. Huber *et al.* note that the same pattern was also visible in other experiments, and with different array types (Figure 3 displays cDNA array data).

Variance stabilized data is desirable for microarray analysis because it allows for easier comparison between

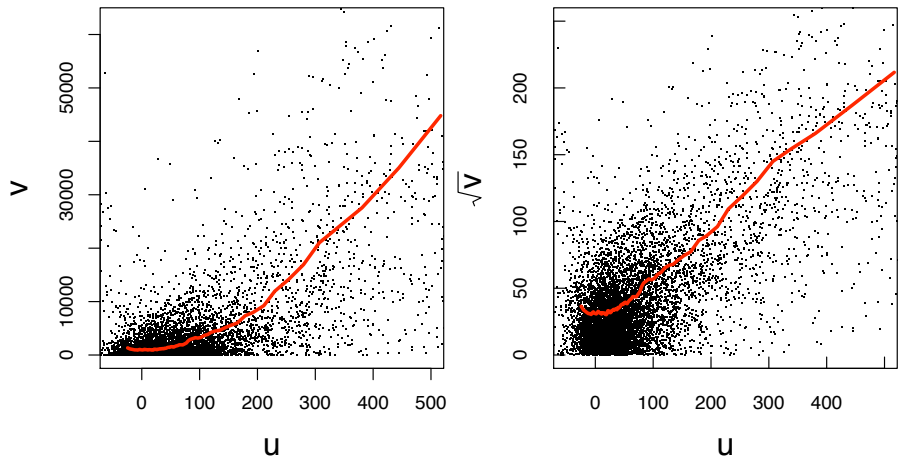

Figure 3: Two plots taken from Huber *et al* [HvHS+02]. The left plot shows the variance (y-axis) versus the mean (x-axis) from an 8400 element cDNA slide. The right plot shows the standard deviation rather than the variance. The dots represent single genes, and the solid line shows a moving average.

genes. Without stabilization, a large differential expression for a high intensity gene could potentially be *less* significant than a small differential expression for a low intensity gene. After stabilization, however, if we view expression differences in terms of Δh, we are guaranteed that a larger difference corresponds to a greater likelihood of significance.

3.2 The Model

Huber *et al.* [HvHS+02] use linear normalization to calibrate the slide and dye intensities. Their method is similar to that described by Yang *et al.* [YDLS01], but is generalized to work with an arbitrary number of slides or dyes. Instead of normalizing one color to the other, they normalize each slide to the first slide. All slides are thus effectively transitively normalized to each other. For each slide (other than the first) $i = 2, \ldots, d$, we have a scaling factor s_i and an offset o_i. Their normalization equations are then

$$y_{ki} \longmapsto \tilde{y}_{ki} = o_i + s_i y_{ki}$$

where k is the gene and i is the slide or dye number. As Park *et al.* [PYK+03] showed, nonlinear normalization typically only leads to small gains over linear normalization, so the choice of using linear normalization is likely sufficient.

Huber *et al.* [HvHS+02] next model the variance-mean dependence quadratically. This choice is backed up by Figure 3, which shows a roughly quadratic curve. Figure 3 also shows a non-zero y-intercept, so the quadratic model also contains an independent constant term c_3:

$$v_k = v(u_k) = (c_1 u_k + c_2)^2 + c_3 \, .$$

Huber *et al.* [HvHS+02] then apply the variance stabilization method of Tibshirani [Tib88] to the variance equation, and obtain the transform:

$$h(y) = \int^y 1/\sqrt{v(u)} du \, .$$

Solving the integral gives:

$$h(y) = \gamma \, \mathrm{arcsinh}(a + by) \, .$$

Combining with the normalization equation and leaving off the overall scaling factor γ gives:

$$h(\tilde{y}) = \mathrm{arcsinh}(a + b(o_i + s_i y_{ki})) \, .$$

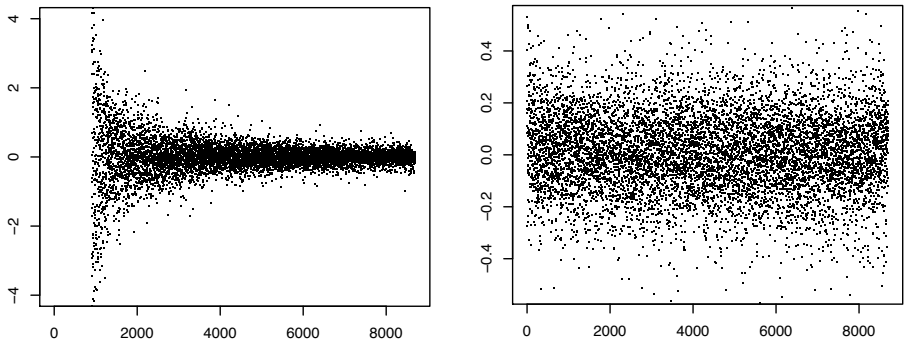

Figure 4: Two plots taken from Huber *et al* [HvHS$^+$02]. The left plot shows the variance-mean dependence of lowess-normalized data. The right plot shows the variance-mean dependence of data normalized with the method of Huber *et al.* [HvHS$^+$02]. The x-axis indicates the rank of the mean of the gene. In the lowess plot, the y-axis tracks the log ratio. In the Huber *et al.* plot, the y-axis measures the Δh statistic.

Finally, setting $a_i = a + bo_i$ and $b_i = bs_i$ gives:

$$h(\tilde{y}) = \text{arcsinh}(a_i + b_i y_{ki}) \, .$$

In order to use the derived transform, we first need to estimate the parameters. We would like to estimate the parameters using the genes which are not differentially expressed. However, we do not know which genes are or are not differentially expressed. If we knew the parameters, though, we could estimate which genes were differentially expressed. Thus, Huber *et al.* [HvHS$^+$02] suggest using Maximum Likelihood Estimation by Expectation Maximization. They iteratively estimate the parameters form the constantly expressed genes, and estimate the constantly expressed genes from the parameters, until they converge to a local likelihood maxima.

3.3 Results

Figure 4 (taken from Huber *et al.* [HvHS$^+$02]) compares the variance distribution of data normalized with lowess nonlinear normalization with the variance distribution of data normalized with the variance stabilization method of Huber *et al.* [HvHS$^+$02]. The data is from a cDNA microarray measuring neighboring regions of a kidney tumor. In both graphs, the x-axis indicates the rank of the mean of the gene. In the lowess graph, the y-axis measures the log expression ratio, which, as we can see, overcorrects for the variance-mean dependency. In the variance stabilization graph, the y-axis measures the Δh statistic, which seems to remove the variance-mean dependency.

4 Other Normalization Methods

A number of other normalization methods have been proposed. Workman *et al.* [WJJ$^+$02] suggest using cubic splines for nonlinear normalization. Schadt *et al.* [SLEW01] propose a normalization method based on changes in the ranks of the intensity values. Kepler *et al.* [KCM02] describe a local regression based approach to normalization. Luck [Luc01] and Munson [Mun01] also suggest other, alternative normalization techniques.

5 Conclusions

We have discussed normalization techniques for microarray data which can solve some of the problems cited by Kothapalli *et al* [KYML02]. In particular, the techniques presented by Yang *et al.* [YDLS01] address the problems with dye color variation, print-tip effects, scanning variation, and slide-preparation and wet-lab variables. The techniques presented by Huber *et al.* [HvHS$^+$02] and Geller *et al.* [GGHR03] also attack the problem of mean-variance dependency.

All of these papers, however, fail to address the question of what effect their normalization techniques have on the next stage of processing. Typically, normalization is run as a preprocessing step, prior to clustering, classification, feature selection, et cetera. We would be interested in seeing a future study comparing the efficacy of various postprocessing algorithms after these normalization techniques have been applied.

References

[GGHR03] S. C. Geller, J. P. Gregg, P. Hagerman, and D. M. Rocke. Transformation and normalization of oligonucleotide microarray data. *Bioinformatics*, 19(14):1817–1823, 2003.

[HvHS$^+$02] W. Huber, A. von Heydebreck, H. Sültmann, A. Poustka, and M. Vingron. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, 18(Supplement 1):S96–S104, 2002.

[KCM02] T. B. Kepler, L. Crosby, and K. T. Morgan. Normalization and analysis of DNA microarray data by self-consistency and local regression. *Genome Biology*, 3(7):research0037.1–research0037.12, 2002.

[KYML02] R. Kothapalli, S. Yoder, S. Mane, and T. Loughran. Microarray results: how accurate are they? *BMC Bioinformatics*, 3(1):22, 2002.

[Luc01] S.D. Luck. Normalization and error estimation for biomolecular expression patterns. In *Proceedings of SPIE BiOS*, volume 4266, San Jose, CA, USA, Jan. 2001.

[Mun01] P. Munson. A 'consistency' test for determining the significance of gene expression changes on replicate samples and two convenient variance-stabilizing transformations. In *GeneLogic Workshop on Low Level Analysis of Affymetrix GeneChip Data*, 2001.

[PYK$^+$03] T. Park, S.-G. Yi, S.-H. Kang, S.Y. Lee, Y.-S. Lee, and R. Simon. Evaluation of normalization methods for microarray data. *BMC Bioinformatics*, 4(1):33, 2003.

[SLEW01] E. E. Schadt, C. Li, B. Ellis, and W. H. Wong. Feature extraction and normalization algorithms for high-density oligonucleotide gene expression array data. *Journal of Cellular Biochemistry*, Supplement 37:120–125, 2001.

[Tib88] R. Tibshirani. Estimating transformation for regression via additivity and variance stabilization. *J. American Statistical Association*, 83:394–405, 1988.

[WJJ$^+$02] C. Workman, L. J. Jensen, H. Jarmer, R. Berka, L. Gautier, H. B. Nielsen, H.-H. Saxlid, C. Nielsen, S. Brunak, and S. Knudsen. A new non-linear normalization method for reducing variability in DNA microarray experiments. *Genome Biology*, 3(9):research0048.1–0048.16, 2002.

[YDLS01] Y. H. Yang, S. Dudoit, P. Luu, and T. P. Speed. Normalization for cDNA microarray data. In *Proceedings of SPIE BiOS*, volume 4266, San Jose, CA, USA, Jan. 2001.