Normalization for cDNA Microarray Data

Yee Hwa Yang[a*], Sandrine Dudoit[b*], Percy Luu[c] and Terence P. Speed[a,d]

[a] Department of Statistics, University of California at Berkeley
[b] Department of Biochemistry, Stanford University
[c] Department of Molecular and Cell Biology, University of California at Berkeley
[d] Division of Genetics and Bioinformatics, The Walter and Eliza Hall Institute, Australia.

ABSTRACT

There are many sources of systematic variation in microarray experiments which affect the measured gene expression levels. Normalization is the term used to describe the process of removing such variation, e.g. for differences in labeling efficiency between the two fluorescent dyes. In this case, a constant adjustment is commonly used to force the distribution of the log-ratios to have a median of zero for each slide. However, such global normalization approaches are not adequate in situations where dye biases can depend on spot overall intensity and location on the array (print-tip effects). This paper describes normalization methods that account for intensity and spatial dependence in the dye biases for different types of cDNA microarray experiments, including dye-swap experiments. In addition, the choice of the subset of genes to use for normalization is discussed. The subset selected may be different for experiments where only a few genes are expected to be differentially expressed and those where a majority of genes are expected to change. The proposed approaches are illustrated using gene expression data from a study of lipid metabolism in mice.

Keywords: cDNA microarray, normalization, dye bias, robust smoother, scaling, dye-swap.

1. INTRODUCTION

DNA microarrays are part of a new class of biotechnologies which allow the monitoring of expression levels for thousands of genes simultaneously. Applications of microarrays range from the study of gene expression in yeast under different environmental stress conditions to the comparison of gene expression profiles for tumors from cancer patients. In addition to the enormous scientific potential of DNA microarrays to help in understanding gene regulation and interactions, microarrays have very important applications in pharmaceutical and clinical research. By comparing gene expression in normal and disease cells, microarrays may be used to identify disease genes and targets for therapeutic drugs.

For cDNA microarrays, the purpose of dye normalization is to balance the fluorescence intensities of the two dyes (green Cy3 and red Cy5 dye) as well as to allow the comparison of expression levels across experiments (slides). Dye bias can be most obviously seen in an experiment where two identical mRNA samples are labeled with different dyes and subsequently hybridized to the same slide. In this situation, it is rare to have the dye intensities equal on average and often the intensities are higher for the green dye. This bias can stem from a variety of factors including physical properties of the dyes (heat and light sensitivity, relative half-life), efficiency of dye incorporation, experimental variability in probe coupling and processing procedures, and scanner settings at the data collection step. Many of these factors, whether internal or external to the sample, present unique difficulties to a global normalization procedure. Furthermore, the relative gene expression levels (as measured by log ratios) from replicate experiments may have different spreads due to differences in experimental conditions. Some scale adjustment may then be required so that the relative expression levels from one particular experiment do not dominate the average relative expression levels across replicate experiments.

This paper describes normalization methods for different types of cDNA microarray experiments. We illustrate the different approaches using gene expression data from a study of lipid metabolism in mice (Callow *et al.*[1]). The

Send correspondence to Yee Hwa Yang
E-mail: `yeehwa@stat.berkeley.edu`

* *These authors contributed equally to this work.*

goal of these experiments was to identify genes with altered expression in apolipoprotein AI knock-out (apo AI ko) mice with very low HDL cholesterol levels (treatment groups) compared to inbred C57Bl/6 control mice.

The paper is organized as follows. Section 2 describes the different subsets of genes commonly used for normalization purposes. The data used to illustrate the strengths and weaknesses of the various normalization methods are described in Section 3. In Section 4, we briefly review a number of existing normalization methods and discuss new methods we have developed for different types of microarray experiments. The results are presented in Section 5. Finally, Section 6 summarizes our findings and outlines open questions.

2. WHICH GENES TO USE

Normalization can be done in a number of ways, depending on the experimental set-up. We distinguish between three situations: (i) within-slide normalization, (ii) paired-slides normalization for dye-swap experiments, and (iii) multiple slide normalization (see Section 4). In each of these situations, a decision must be made as to the set of genes to use for the normalization. A number of considerations influence this decision, such as the proportion of genes that are expected to be differentially expressed in the red and green samples, and the availability of control DNA sequences. Three types of approaches are described next.

All genes on the array. Frequently, biological comparisons made on microarrays are very specific in nature-*i.e.*, only a small proportion of genes are expected to be differentially expressed. Therefore, the remaining genes are expected to have constant expression and so can be used as indicators of the relative intensities of the two dyes. In other words almost all genes on the array may be used for normalization when there are good reasons to expect that (i) only a relatively small proportion of the genes will vary significantly in expression between the two mRNA samples (see exception with self-normalization), or (ii) there is symmetry in the expression levels of the up/down-regulated genes.

Constantly expressed genes. Instead of using all genes on the array for normalization, one may use a smaller subset of genes, often called housekeeping genes, that are believed to have constant expression across a variety of conditions (*e.g.* β actin). Although it is very hard to identify a set of housekeeping genes that do not change significantly under any conditions, it may be possible to find sets of "temporary" housekeeping genes for particular experimental conditions. A limitation of housekeeping genes is that they tend to be highly expressed and hence may not be representative of other genes of interest.

Controls. An alternative to normalization by housekeeping genes is to use spiked controls or a titration series of control sequences. In the spiked controls method, synthetic DNA sequences or DNA sequences from an organism different from the one being studied are spotted on the array (with possible replication) and included in the two different mRNA samples at equal amount. These spotted control sequences should thus have equal red and green intensities and could be used for normalization. In the titration series approach, spots consisting of different concentrations of the same gene or EST are printed on the array. These spots are expected to have equal red and green intensities across the range of intensities. Genomic DNA, which is supposed to have constant expression levels across various conditions, may be used in the titration series. In practice, however, genomic DNA is often too complex to exhibit much signal and setting a titration series that spans the range of intensities for different experiments is technically very challenging. We are aware of efforts to achieve the same result with what might be called *pseudo-genomic DNA*, (J. Ngai, pers. comm.).

3. DATA

A) Apo AI experiment

The apo AI experiment was carried out as part of a study of lipid metabolism and atherosclerosis susceptibility in mice. Apolipoprotein AI (apo AI) is a gene known to play a pivotal role in HDL metabolism. Mice with the apo AI gene knocked-out have very low HDL cholesterol levels and the goal of the apo AI experiment was to identify genes with altered expression in the livers of these knock-out mice compared to inbred control mice.

The treatment group consisted of eight mice with the apo AI gene knocked-out and the control group consisted of eight "normal" C57Bl/6 mice. For each of these 16 mice, target cDNA was obtained from mRNA by reverse transcription and labeled using a red-fluorescent dye, Cy5. The reference sample used in all hybridizations was prepared by pooling cDNA from the eight control mice and was labeled with a green-fluorescent dye, Cy3. In this experiment, target cDNA was hybridized to microarrays containing 6,384 cDNA probes, including 200 related to lipid metabolism. Note that we call the spotted DNA sequences "genes", whether they correspond to actual genes, ESTs (expressed sequence tags), or DNA sequences from other sources.

Each of the 16 hybridizations produced a pair of 16-bit images, which were processed using the software package $\mathcal{S}$pot.[2] The main quantities of interest produced by the image analysis methods (segmentation and background correction) are the (R, G) fluorescence intensity pairs for each gene on each array. After image processing and normalization (see Section 4 below) the gene expression data can be summarized by a matrix X of log-intensity ratios $\log_2 R/G$, with p rows corresponding to the genes being studied and $n = n_1 + n_2$ columns corresponding to the n_1 control hybridizations (C57Bl/6) and n_2 treatment hybridizations (apo AI knock-out). In the experiment considered here $n_1 = n_2 = 8$ and $p = 5,548$.

Differentially expressed genes were identified by computing t-statistics. For gene j, the *t-statistic* comparing gene expression in the control and treatment groups is

$$t_j = \frac{\bar{x}_{2j} - \bar{x}_{1j}}{\sqrt{\frac{s_{1j}^2}{n_1} + \frac{s_{2j}^2}{n_2}}},$$

where $\bar{x}_{1j}$ and $\bar{x}_{2j}$ denote the average background corrected and normalized expression level of gene j in the n_1 control and n_2 treatment hybridizations, respectively. Similarly, s_{1j}^2 and s_{2j}^2 denote the variances of gene j's expression level in the control and treatment hybridizations, respectively. Large absolute t-statistics suggest that the corresponding genes have different expression levels in the control and treatment groups. The statistical significance of the results was assessed based on p-values adjusted for multiple comparisons. These adjusted p-values were estimated using Westfall and Young's step-down adjusted p-value algorithm 4.1.[3] The analysis of the data-set is described in detail in Dudoit *et al.*.[4]

B) Follow-up experiment

The 20 "clones" with the largest absolute t-statistics in the apo AI experiment were selected and spotted on a mini-array. Some clones actually comprised more than one clone and these were purified and re-checked. Each of the approximately 50 distinct clones from the top 20 clones were spotted eight times on the mini-array down a single column in the same print-tip group. Anticipating that most genes on the mini-array were differentially expressed, a dye-swap experiment was done to allow normalization of the red and green fluorescence intensities. In the first hybridization, named `C3K5`, the treatment (apo AI ko) mRNA is labeled red and the control mRNA is labeled green. In the second hybridization, named `C5K3`, the original labeling is reversed, with treatment labeled green and control labeled red.

In addition to the 50 distinct clones that were spotted on the mini-array, another 72 genes were spotted in the same pattern for normalization purposes. These genes were studied in another experiment and are not expected to be differentially expressed in the apo AI knock-out mice. They are treated as proxies for housekeeping genes (for this experiment only) and used for normalization purposes in order to (re-)examine differential expression of the approximately 20 genes from the original apo AI experiment. Without this set of genes, verification of the correctness of the self-normalization procedure described below would have been difficult, as most of the genes from the apo AI experiment were expected to be differentially expressed.

4. METHODS

4.1. Single-slide data displays

Notation. For a spot j, $j = 1, \ldots, p$, let R_j and G_j denote the measured fluorescence intensities (after background correction) for the red and green dyes, respectively.

Single-slide expression data are typically displayed by plotting the log-intensity $\log_2 R$ of the red dye *vs.* the log-intensity $\log_2 G$ of the green dye. We find that such plots give an unrealistic sense of concordance and we prefer to plot the log intensity ratio $M = \log_2 R/G$ *vs.* the mean log-intensity $A = \log_2 \sqrt{RG}$. An M *vs.* A plot amounts to a 45^o counterclockwise rotation of the $(\log_2 G, \log_2 R)$-coordinate system, followed by scaling of the coordinates. If M' and A' denote the rotated coordinates, then $A = A'/\sqrt{2}$ and $M = M'\sqrt{2}$.

An M *vs.* A plot is thus another representation of the (R, G) data in terms of the log-intensity ratios M which are the quantities of interest to most investigators. We have found M *vs.* A plots to be more revealing than their $\log_2 R$ *vs.* $\log_2 G$ counterparts in terms of identifying spot artifacts, detecting intensity dependent patterns in the log-ratios etc. They are also very useful for the purpose of normalization as illustrated next.

4.2. Within-slide normalization : Location

In this case, the normalization is done separately for each slide, using only the red and green intensities for this slide. Several approaches are described below.

4.2.1. Global normalization

Global methods assume that the red and green intensities are related by a constant factor. That is, $R = k \cdot G$, and in practice, the center of the distribution of log-ratios is shifted to zero:

$$\log_2 R/G \to \log_2 R/G - c = \log_2 R/(kG).$$

A common choice for the location parameter $c = \log_2 k$ is the median or mean of the log-intensity ratios for a particular gene set. Global normalization methods are mentioned as pre-processing steps in a number of papers on the identification of differentially expressed genes in single-slide cDNA microarray experiments. In one of the first treatments of the problem, Chen *et al.*[5] assume that $R = k \cdot G$ and propose an iterative method for estimating the constant normalization factor k and cut-offs for the red and green intensity ratio R/G. In some software packages (*e.g.* GenePix[6]), a constant normalization factor is estimated such that the arithmetic mean of the intensity ratios of all the genes on a given microarray is one. Global normalization methods are still the most widely used methods in spite of the evidence of spatial or intensity dependent dye biases in numerous experiments.

4.2.2. Intensity dependent normalization

In many cases, the dye bias appears to be dependent on spot intensity, as revealed by plots of the log-ratio M *vs.* overall spot intensity A. An intensity or A-dependent dye normalization method may thus be preferable to global methods.

We use the robust scatter-plot smoother `lowess` from the statistical software package `R`[7] to perform a local A-dependent normalization:

$$\log_2 R/G \to \log_2 R/G - c(A) = \log_2 R/(k(A)G),$$

where $c(A)$ is the `lowess` fit to the M *vs.* A plot. The `lowess()` function is a scatter-plot smoother which performs robust locally linear fits. In particular, the `lowess()` function will not be affected by a small percentage of differentially expressed genes which will appear as outliers in the M *vs.* A plot. The user defined parameter f is the fraction of the data used for smoothing at each point; the larger the f value, the smoother the fit. We typically use $f = 20\%$.

Sapir and Churchill[8] suggest using the orthogonal residuals from the robust regression of $\log R$ *vs.* $\log G$ as normalized log-ratios. Since an M *vs.* A plot amounts to a 45^o counterclockwise rotation of the ($\log G$,$\log R$) - coordinate system (up to multiplicative constants), their method is similar to fitting a robust regression line through the M *vs.* A plot instead of a `lowess` curve. One can view this linear normalization as a more constrained version of intensity dependent normalization. Kepler[9] proposes a more general intensity dependent normalization approach which uses a different local regression method instead of the `lowess` smoother function.

4.2.3. Within-print-tip-group normalization

Every grid in an array is printed using the same print-tip. Different experiments may be done using different printing set-ups depending on the layout of the tips in the print-head of the arrayer (*e.g.* 4 by 4 or 2 by 2 print-heads). Some systematic differences may exist between the print-tips, such as slight differences in the length or in the opening of the tips, and deformation after many hours of printing. Alternatively, print-tip groups are proxies for spatial effects on the slide. Within-print-tip-group normalization is simply a (print-tip + A)-dependent normalization, that is,

$$\log_2 R/G \to \log_2 R/G - c_i(A) = \log_2 R/(k_i(A)G),$$

where $c_i(A)$ is the `lowess` fit to the M *vs.* A plot for the ith grid only, $i = 1, \ldots, I$, and I represents the number of print-tips.

4.3. Within-slide normalization: Scale

After within-print-tip-group normalization, all the normalized log-ratios from the different print-tip groups will be centered around zero. However, it is possible that the log-ratios from the various print-tip groups have different spread and some scale adjustment is required.

One approach that we have found to work is to assume that all log-ratios from the ith print-tip group follow a normal distribution with mean zero and variance $a_i^2\sigma^2$, where σ^2 is the variance of the true log-ratios and a_i^2 is the scale factor for the ith print-tip group. In order to perform scale normalization, the scale factors a_i for the different print-tip groups must be estimated. Enforcing the natural constraint $\sum_{i=1}^{I} \log a_i^2 = 0$, with I denoting the total number of print-tips on the array, the maximum likelihood estimate for a_i is

$$\hat{a}_i = \frac{\sum_{j=1}^{n_i} M_{ij}^2}{\sqrt[I]{\prod_{k=1}^{I} \sum_{j=1}^{n_i} M_{kj}^2}},$$

where M_{ij} denotes the jth log-ratio in the ith print-tip group, $j = 1, \ldots, n_i$. A robust alternative to this estimate, which we find preferable, is

$$\hat{a}_i = \frac{MAD_i}{\sqrt[I]{\prod_{i=1}^{I} MAD_i}},$$

where the median absolute deviation MAD is defined by

$$MAD_i = \text{median}_j\{\mid M_{ij} - \text{median}_j(M_{ij}) \mid\}.$$

This procedure assumes that a relatively small proportion of the genes will vary significantly in expression between the two mRNA samples. In addition, it assumes that the spread of the distribution of the log-ratios should be roughly the same for all print-tip groups. The robust statistic MAD, like the robust `lowess` smoother, will not be affected by a small percentage of differentially expressed genes which will appear as outliers in the M *vs.* A plots.

4.4. Paired-slides normalization (dye-swap)

Paired-slides normalization applies to dye-swap experiments: two hybridizations for two mRNA samples, with dye assignment reversed in the second hybridization.

Denote the normalized log-ratios for the first slide by $\log_2 R/G - c$ and those for the second slide by $\log_2 R'/G' - c'$. Here, c and c' denote the normalization functions for the two slides; these could be obtained by any of the within-slide normalization methods described above. If $c \approx c'$, then

$$\frac{1}{2}\left[\log_2 R/G - c - \left(\log_2 R'/G' - c'\right)\right] \approx \frac{1}{2}\left[\log_2 R/G + \log_2 G'/R'\right] = \frac{1}{2}\log_2 \frac{RG'}{GR'} = \frac{1}{2}(M - M').$$

Thus, we may combine the relative expression levels for the two slides without explicit normalization. We refer to this procedure as *self-normalization*. The main assumption here is that $c \approx c'$ and this method can be applied to all genes, even if they are differentially expressed. With this approach, the genes that are not supposed to change should have $(M - M')/2 = \frac{1}{2}\log_2(RG'/GR') \approx 0$. The validity of this assumption may be checked using a set of genes expected to have constant expression levels (*e.g.* housekeeping genes or genomic DNA), if such as set is available.

Given that the dye assignments are reversed in the two experiments, one expects that the normalized log-ratios on the two slides are of equal magnitude and opposite sign, that is,

$$\log_2 R/G - c \approx -(\log_2 R'/G' - c').$$

Therefore, rearranging the equation and assuming again that $c \approx c'$, we can estimate the normalization function c by

$$c \approx \frac{1}{2}\left[\log_2 R/G + \log_2 R'/G'\right] = \frac{1}{2}(M + M').$$

In practice, $c = c(A)$ is estimated by the `lowess` fit to the plot of $\frac{1}{2}(M + M') = \frac{1}{2}\log_2 RR'/GG'$ *vs.* $\frac{1}{2}(A + A')$, where this time all the genes are used.

Note that the normalization method just described adjusts for location only and assumes that the spread of the log-ratios is roughly the same for the two slides.

4.5. Multiple slide normalization

After within-slide normalization, all normalized log-ratios will be centered around zero, regardless of the normalization method. Multiple slide normalization methods, which aim to allow experiment to experiment comparisons, may also need to be adjusted for scale when the different slides have substantially different spreads in their log-ratios. Failing to perform a scale normalization could lead to one or more slides having undue weight when averaging log-ratios across experiments.

The within-slide scale normalization method described in Section 4.3 may also be used for multiple slide scale adjustment. We are currently evaluating this approach with experiments where a scale normalization seems called for.

4.6. Comparison between different normalization methods

In order to compare the different within-slide normalization methods, we consider their effect on the location and scale of the log-ratios. We produce density plots of the log-ratios for each of the normalization methods using a gaussian kernel density estimator (`density()` function of the statistical software package R, bandwith size of 0.17).

The different methods are also evaluated based on their ability to identify genes which are known to be differentially expressed. A good method should enable a clear distinction between differentially and constantly expressed genes, as reflected by the t-statistics and the adjusted p-values. That, is one expects a large jump in the t-statistics and adjusted p-values between the least extreme of the differentially expressed genes and the most extreme of the remaining genes. For experiment (A), the apo A1 gene is knocked-out in the eight treatment mice, so one expects the t-statistics to take on very large negative values for this gene. In order to compare the different methods, we produce truncated plots of the extreme t-statistics for each of the methods.

5. RESULTS

5.1. Within-slide normalization

Global normalization amounts to a vertical translation in an M *vs.* A plot and does not allow for spatial or intensity dependent dye biases. This may not be the best strategy as suggested by the M *vs.* A plot in Figure 1. The 16 within-print-tip-group `lowess` curves clearly illustrate the dependence of the log-ratio M on the overall spot intensity A. Furthermore, four within-print-tip-group `lowess` curves stand out from the remaining twelve curves, indicating strong print-tip or spatial effects. These 4 curves correspond to the last row of print-tips in the 4 by 4 print-head (print-tips 13, 14, 15, and 16). This pattern was visible in the images, where the bottom 4 grids tended to have high red signal. Hence, Figure 1 suggests that within-print-tip-group intensity dependent normalization methods may be preferable to global methods.

Figure 2 displays the spatial distribution on the array of the largest 5% absolute log-ratios, after within-print-tip-group location normalization (panel (a)) and after within-print-tip-group location and scale normalization (panel (b)). In panel (a), there is a disproportionately large number of extreme log-ratios in the lower four grids. This pattern is also noticeable in Figure 3, which shows boxplots of the log-ratios for the different print-tip groups and suggests that the spread of the log-ratios for the last four grids is larger than that in the remaining twelve grids. After scale normalization the extreme log-ratios seem to be evenly distributed on the array (panel (b) of Figure 2).

5.2. Paired-slides normalization (dye-swap)

For the dye-swap experiment (B), Figure 4 displays an M *vs.* A plot of the expression levels for both slides. The solid cyan curve represents the `lowess` fit using the constantly expressed genes in the `C3K5` slide and the dotted black curve represents the `lowess` fit using the constantly expressed genes in the `C5K3` slide. The two normalization lines for the dye-swap experiment are very similar, suggesting that the dye bias is similar in the two slides and hence that self-normalization is appropriate.

Figure 5 is a plot $\frac{1}{2}(M - M') = \frac{1}{2}\log_2(RG')/(GR')$ *vs.* $\frac{1}{2}(A + A')$ for the self-normalization procedure applied to experiment (B). The solid black curve represents the `lowess` fit ($f = 0.5$) through the constantly expressed genes (blue solid dots). This curve is very close to the dotted black line corresponding to log-ratios of zero and this again confirms that a self-normalization procedure is appropriate.

5.3. Multiple slide normalization

Figure 6 displays boxplots of the log-ratios for each of the 16 slides in experiment (A), after within-print-tip-group location and scale normalization. The boxplots are centered at zero and have fairly similar spreads. Although the slides corresponding to knock-out mice 1, 5, and 6 seem to have larger spread than others, the noise introduced by a scale normalization of the different slides may be more detrimental than a small difference in scale. We thus chose not to adjust for scale in this case.

5.4. Comparison between different normalization methods

Figure 7 shows density plots of the log-ratios for different normalization methods. Without normalization (black curve), the log-ratios are centered around -1 indicating a bias towards the green (Cy3) dye. A global median normalization (red curve) shifts the center of the log-ratio distribution to zero but does not affect the spread. The dependence of the log-ratio M on the overall intensity A is also still present (Figure 1). Both the intensity dependent (green curve) and within-print-tip-group (blue curve) location normalization methods reduce the spread of the log-ratios compared to a global normalization. A within-print-tip-group scale normalization (cyan curve) further reduces the spread slightly.

For experiment (A), Figure 8 shows a plot of the extreme t-statistics for different normalization methods. The global median, intensity dependent and within-print-tip-group location normalization methods seem to perform best in terms of their ability to detect the three knocked-out apo AI genes. Table 1 shows the corresponding p-values adjusted for multiple comparisons. The largest jump in p-values is observed for the within-print-tip-group location normalization. For within-print-tip-group scale normalization, the t-statistics are less extreme and the gap in p-values between the differentially expressed genes and the remaining genes decreases in comparison with a location normalization only.

6. DISCUSSION

This paper introduced location and scale normalization methods for different types of cDNA microarray experiments. The different normalization methods were compared using gene expression data from two sets of experiments: the apo A1 experiment (data-set (A)), with replicated treatment and control slides, and the follow-up dye-swap experiment (data-set (B)), with replicated spots on a slide.

For within-slide normalization, we found that the standard global median normalization can often be inadequate due to spatial and intensity dependent dyes biases. We propose instead a within-print-tip-group location normalization method which consists of applying a robust smoother to a plot of the log-ratios M against overall spot intensities A. Compared to other normalization methods, this approach provided a clearer distinction between the differentially and constantly expressed genes in experiment (A).

The spatial plots in Figure 2 and the boxplots of the location normalized log-ratios in each print-tip group in Figure 3 suggest that some scale adjustment may also be required. However, within-print-tip-group scale normalization seems to have decreased our ability to identify the differentially expressed genes in experiment (A). We believe that this is due to an increase in the variability (the denominator of the t-statistic) of the log-ratios for the eight differentially expressed genes compared to the rest of the genes. In general there is a trade-off between the gains achieved by scale normalization and the possible increase in variability introduced by this additional step. In cases where the scale differences are fairly small, it may thus be preferable to perform only a location normalization. A similar approach to that described in Section 4.3 for within-slide scale normalization may also be extended to perform scale normalization across experiments. Further investigations are underway to develop an improved procedure for scale adjustment and identify better comparison criteria to assess the effectiveness of various normalization procedures.

In order to apply any of the location normalization methods discussed above one must identify a set of genes that satisfy the following: (i) only a relatively small proportion of the genes vary significantly in expression between the two mRNA samples, or (ii) there is symmetry in the expression levels of the up/down-regulated genes. In general, the set of genes to be used in normalization depends on the nature of the experiment. For experiments such as the apo A1 knock-out experiment (A), where only a small proportion of the genes are expected to be differentially expressed, a robust procedure based on all the genes is appropriate. For experiments where a large fraction of the genes are expected to change, it is possible to use a set of constantly expressed genes (housekeeping genes or control sequences). However, these usually represent a small fraction of all the genes on the array and the resulting normalization is likely to be noisy. Alternatively, for dye swap experiments where the normalization curve is expected to be the same for both slides, a self-normalization procedure may be used.

Acknowledgments

We would like to acknowledge Matthew J. Callow from the Lawrence Berkeley National Laboratory for providing the data we used to develop the various normalization approaches. We would also like to thank the members of the Ngai Lab at UC Berkeley for helpful discussions on the biology background.

This work was supported in part by the NIH through grants 5R01MH61665-02 (YHY, PL) and 8R1GM59506A (TPS), and by an MSRI and a PMMB postdoctoral fellowship (SD).

REFERENCES

1. M. J. Callow, S. Dudoit, E. L. Gong, T. P. Speed, and E. M. Rubin, "Microarray expression profiling identifies genes with altered expression in hdl deficient mice," *Genome Research* , 2000. Submitted.
2. M. J. Buckley, *The Spot user's guide.* CSIRO Mathematical and Information Sciences, August 2000. `http://www.cmis.csiro.au/IAP/Spot/spotmanual.htm`.
3. P. H. Westfall and S. S. Young, *Resampling-based multiple testing: examples and methods for p-value adjustment*, Wiley series in probability and mathematical statistics, Wiley, 1993.
4. S. Dudoit, Y. H. Yang, M. J. Callow, and T. P. Speed, "Statistical methods for identifying genes with differential expression in replicated cDNA microarray experiments." (Submitted), 2000.
5. Y. Chen, E. R. Dougherty, and M. L. Bittner, "Ratio-based decisions and the quantitative analysis of cdna microarray images," *Journal of Biomedical Optics* **2**, pp. 364–374, 1997.
6. Axon Instruments, Inc., *GenePix 4000A User's Guide*, 1999.
7. R. Ihaka and R. Gentleman, "R: A language for data analysis and graphics," *Journal of Computational and Graphical Statistics* **5**, pp. 299–314, 1996.
8. M. Sapir and G. A. Churchill, *Estimating the posterior probability of differential gene expression from microarray data.* Poster, The Jackson Laboratory, 2000. `http://www.jax.org/research/churchill/`.
9. T. Kepler, "Normalization and analysis of dna microarray data by self-consistency and local regression." `http://www.ipam.ucla.edu/programs/fg2000/abstracts/fgsn_tkepler.html`.

Table 1. Names and adjusted p-values for the 9 genes with the largest absolute t-statistics for each of the normalization methods described in Section 4. The first column gives the method name, and columns 2 to 10 give the names and adjusted p-values of the top 9 genes for each of the methods. For example, column 2, with the header "1", gives the adjusted p-values for the gene with the most extreme t-statistic. The adjusted p-value calculation is based on an algorithm of Westfall and Young described in Dudoit *et al.*.[4] The symbols "A1", "A3", "SD", "E" and "O" denote apo A1, apo CIII, sterol desaturase, a novel EST and other genes, respectively. The first four genes were confirmed by RT-PCR (Callow *et al.*[1]).

	1		2		3		4		5		6		7		8		9	
None	A1	0	A1	0	A3	0	SD	0	A1	.01	E	.01	O	.02	SD	.07	A3	.10
Median	A1	0	A1	0	A3	0	SD	0	A1	0	A3	0	E	.01	SD	.01	O	.46
Global Lowess	A1	0	SD	0	A1	0	A1	0	A3	0	E	0	A3	.01	SD	.01	O	.41
Print-tip Lowess	A1	0	A1	0	SD	0	A3	0	A1	0	A3	0	E	0	SD	0	O	.71
Print-tip Scale	A1	0	A3	0	A1	0	E	0	SD	0	A1	.02	SD	.06	A3	.12	O	.44

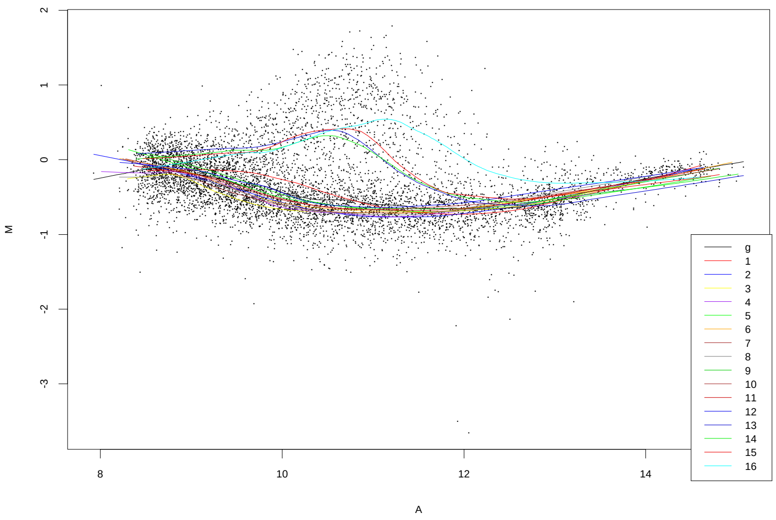
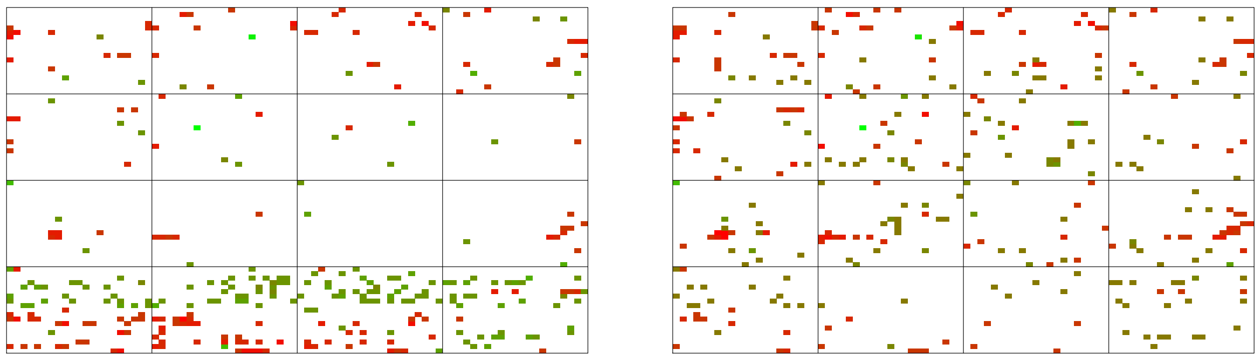


Figure 1. *Within-slide normalization. M vs. A* plot for within-print-tip-group location normalization displaying the `lowess` lines ($f = 20\%$) for each of the 16 print-tips. The curve labeled by "g" corresponds to the `lowess` fit for the entire data-set (data from apo A1 knock-out mouse #8 in experiment (A)).



(a) (b)

Figure 2. *Within-slide normalization.* Spatial plot of the array highlighting the spots with the largest 5% absolute log-ratios. The different shades of red represent positive log-ratios and the different shades of green represent negative log-ratios. The plot is divided into 16 grids representing the 16 different print-tip groups. Each small rectangular cell represents the log-ratio of a spot on the array. (a) Extreme log-ratios after within-print-tip-group location normalization but before scale adjustment. (b) Extreme log-ratios after within-print-tip-group location and scale normalization (data from apo A1 knock-out mouse #8 in experiment (A)).

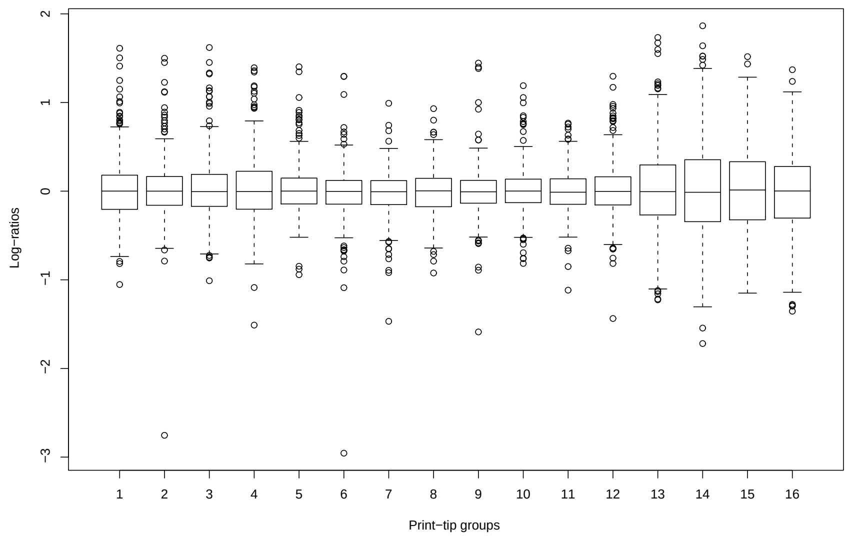


Figure 3. *Within-slide normalization.* Boxplot displaying the log-ratio distribution after within-print-tip-group location normalization for each of the 16 print-tip groups. The array was printed using a 4 by 4 print-head and the print-tip groups are numbered first from left to right then from top to bottom starting from the top left corner (data from apo A1 knock-out mouse #8 in experiment (A)).

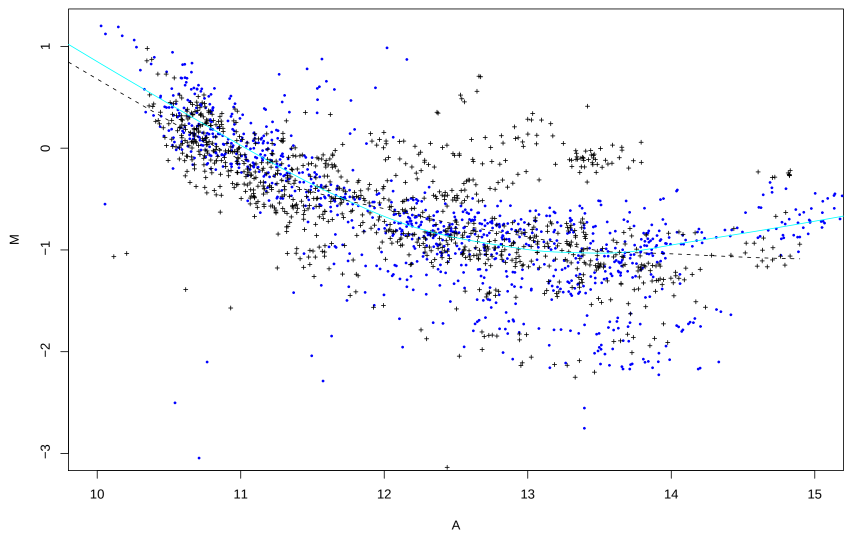


Figure 4. *Paired-slide normalization.* *M vs. A* plot showing the within-slide normalization curves for experiment (B). The blue dots represent the log-ratios for slide C3K5 and the black crosses represent the log-ratios for slide C5K3. The solid cyan curve is the `lowess` fit ($f = 0.5$) through the constantly expressed genes for slide C3K5 and the dotted black curve is the `lowess` fit through the constantly expressed genes for slide C5K3.

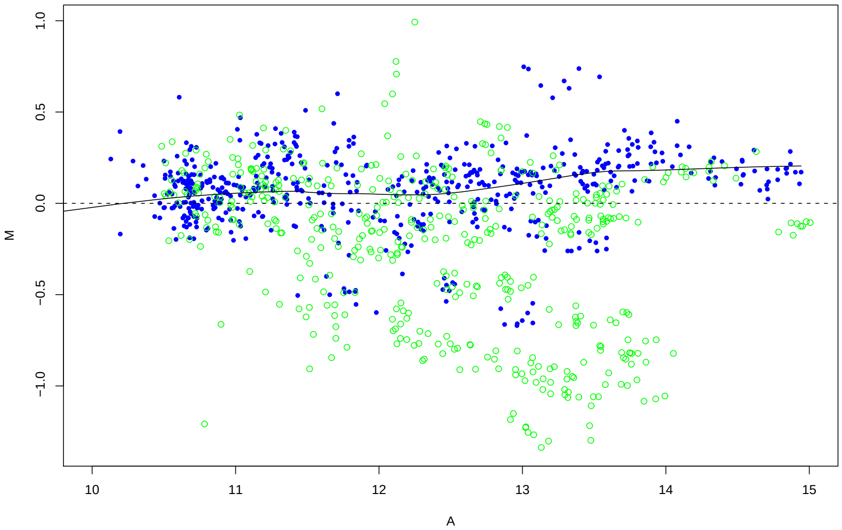


Figure 5. *Paired-slide normalization.* Plot of $\frac{1}{2}(M - M') = \frac{1}{2}\log_2(RG')/(GR')$ *vs.* $\frac{1}{2}(A + A')$ for the self-normalization procedure applied to experiment (B). The green open circles represent the clones related to the apo A1 experiment; these were selected because they had large absolute t-statistics in experiment (A). The blue solid dots represent constantly expressed clones which were selected from another experiment. A solid `lowess` curve ($f = 0.5$) is fitted through the blue solid dots. This `lowess` curve is very close to the horizontal dotted line corresponding to a log-ratio of zero.

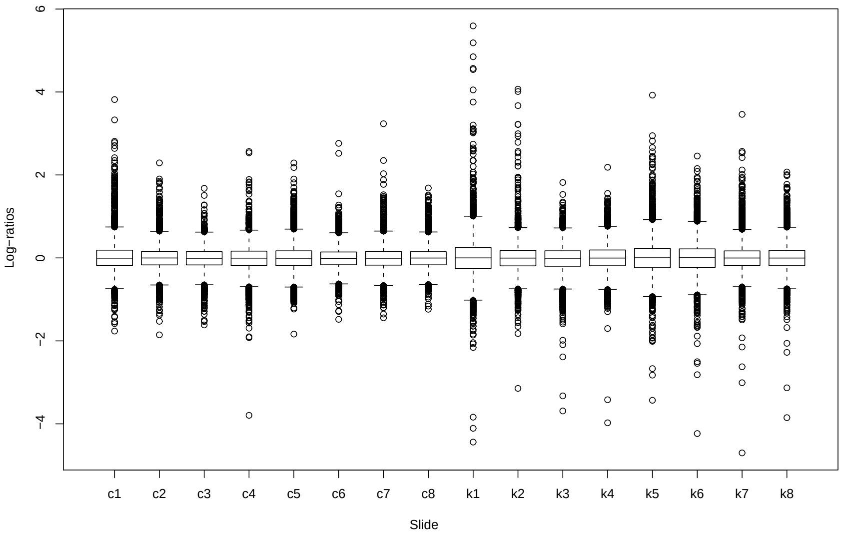


Figure 6. *Multiple slide normalization.* Boxplots displaying the log-ratio distribution for different slides/mice for experiment (A), after within-print-tip-group location and scale normalization. The first 8 boxplots represent the data for the 8 control mice and the last 8 boxplots represent the data for the 8 apo A1 knock-out mice.

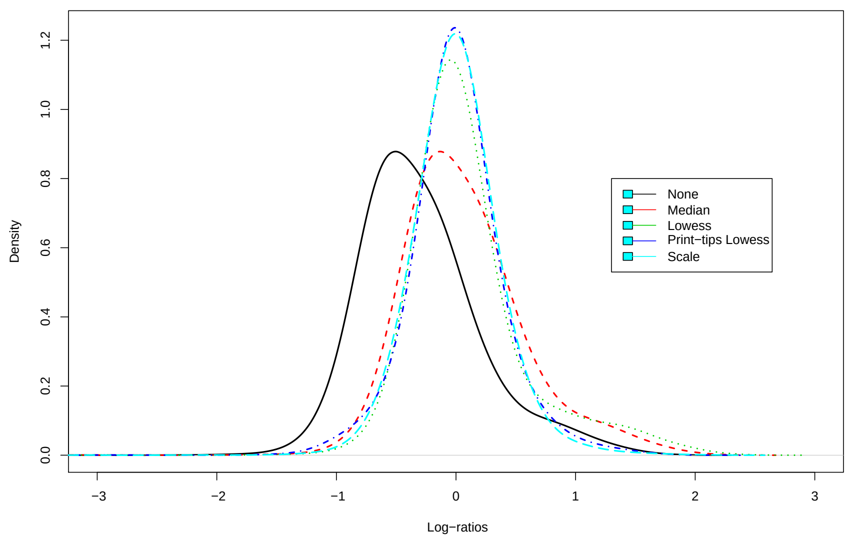


Figure 7. *Within-slide normalization.* Density plots of the log-ratios before and after different normalization procedures. The solid black curve represents the density of the log-ratios before normalization. The red, green, blue, and cyan curves represent the densities after global median normalization, intensity dependent location normalization, within-print-tip-group location normalization, and within-print-tip-group scale normalization, respectively (data from apo A1 knock-out mouse #8 in experiment (A)).

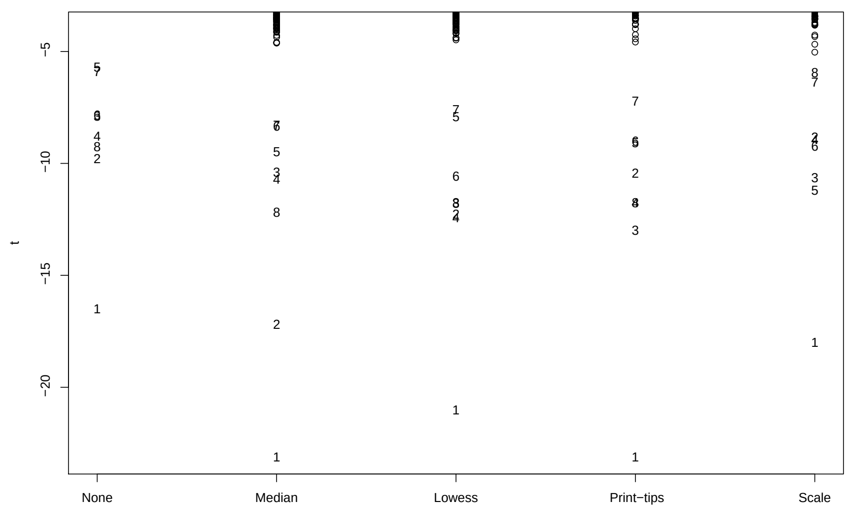


Figure 8. *Within-slide normalization.* Plot of t-statistics for different normalization methods. The numbers from 1 to 8 represent the differentially expressed genes identified in Dudoit *et al.*[4] and confirmed using RT-PCR: indices 1 to 3 represent the three apo A1 genes. Empty circles represent the remaining 6,376 genes where no effect is expected. Only t-values less than -4 are shown.