To appear in: D. Carter (ed.), *METHODS: Selecting Candidate Genes from DNA Array Screens: Application to Neuroscience*.

Normalization of cDNA Microarray Data

Gordon K. Smyth[1] and Terry Speed[12]
1. Walter and Eliza Hall Institute of Medical Research
2. Department of Statistics, University of California, Berkeley

April 4, 2003

Address for correspondence: Dr Gordon K. Smyth, Walter and Eliza Hall Institute, 1G Royal Parade, Parkville, Victoria 3050, Australia, smyth@wehi.edu.au

Abstract

Normalization means to adjust microarray data for effects which arise from variation in the technology rather than from biological differences between the RNA samples or between the printed probes. This article describes normalization methods based on the fact that dye balance typically varies with spot intensity and with spatial position on the array. Print-tip loess normalization provides a well-tested general purpose normalization method which has given good results on a wide range of arrays. The method may be refined by using quality weights for individual spots. The method is best combined with diagnostic plots of the data which display the spatial and intensity trends. When diagnostic plots show that biases still remain in the data after normalization, further normalization steps such as plate-order normalization or scale-normalization between the arrays may be undertaken. Composite normalization may be used when control spots are available which are known to be not differentially expressed. Variations on loess normalization include global loess normalization and 2D normalization. Detailed commands are given to implement the normalization techniques using freely available software.

1 Introduction

In this article we suppose that an experiment has been conducted using a series of two-color cDNA microarrays. Each microarray has been hybridized with RNA from two sources labeled with different fluors. The two color channels will be referred to by convention as red and green. We suppose that the arrays have been scanned to produce images and that these images have been further processed by an image analysis program to produce measured red and green foreground and background intensities for each spot on each array. Before the gene expression profiles of the RNA samples can be analyzed and interpreted, the red and green intensities must be normalized relative to one another so that the red/green ratios are as far as possible an unbiased representation of relative expression.

The purpose of normalization is to adjust for effects which arise from variation in the microarray technology rather than from biological differences between the RNA samples or between the printed probes. Imbalances between the red and green dyes may arise from differences between the labeling efficiencies or scanning properties of the two fluors complicated perhaps by the use of different scanner settings. If the imbalance is more complicated than a simple scaling of one channel relative to the other, as it usually will be, then the dye bias is a function of intensity and normalization will need to be intensity dependent. The dye-bias will also generally vary with spatial position on the slide. Positions on a slide may differ because of differences between the

print-tips on the array printer, variation over the course of the print-run, non-uniformity in the hybridization or from artifacts on the surface of the array which affect one color more than the other. Finally, differences between arrays may arise from differences in print quality, from differences in ambient conditions when the plates were processed or simply from changes in the scanner settings. Therefore normalization between as well as within arrays will need to be considered.

Write R and G for the background-corrected red and green intensities for each spot. Normalization is usually applied to the log-ratios of expression, which will be written $M = \log_2 R - \log_2 G$. The log-intensity of each spot will be written $A = (\log_2 R + \log_2 G)/2$, a measure of the overall brightness of the spot. (The letter M is a mnemonic for *m*inus while A is a mnemonic for *a*dd.) It is convenient to use base-2 logarithms for M and A so that M is units of 2-fold change and A is in units of 2-fold increase in brightness. On this scale, $M = 0$ represents equal expression, $M = 1$ represents a 2-fold change between the RNA samples, $M = 2$ represents a 4-fold change, and so on.

Any negative values for R or G will yield missing values for M and A and the corresponding spots will be excluded from subsequent analysis including normalization. The frequency of negative values depends very much on the image analysis program and the background estimation method used. SPOT [1] for example using a "morph" background gives very few negative intensities while other programs such as GenePix [2] using a "median" background may often result in 30% or more negative values. The loss of information which results from omitting such spots from the analysis is usually not great because spots with negative values for either R or G are usually too faint to show good evidence of differential expression. In any case, the relative merits of the different background correction methods is beyond the scope of this article.

The plan of this article is as follows. Section 2 describes diagnostic plots to visualize intensity and spatial trends. Section 3 describes the basic normalization method, print-tip loess normalization, designed to adjust for intensity and spatial trends. Section 4 describes composite loess normalization in which use is made of control spots known to be not differentially expressed. Section 5 considers normalization for other trends, in particular, correcting for print-order effects. Section 6 describes scale normalization between arrays. Section 7 describes the use of spot quality weights and Section 8 gives detailed commands to implement the normalization techniques using freely available software.

2 Visualization of Intensity and Spatial Trends

The sub-array loess normalization methods described in this article are based on the fact that dye balance typically varies with spot intensity and with spatial position on the array. It is a useful trouble-shooting step to display these trends visually as part of the normalization process.

The relationship between dye-bias and intensity can be seen best in an MA-plot, which is a scatterplot of the M-values against the A-values for an array [3]. Figure 1 shows an MA-plot for an array showing three different trend lines. The horizontal blue line shows the median of the M-values. The orange curve shows the overall trend line as estimated by loess regression, a robust smoother based on local polynomial regression [4]. The yellow line shows the loess curve through a set of control spots known to be not differentially expressed. This array shows a gradual trend from green-bias at low intensity to red-bias at high intensity. Other arrays will

show different trends, perhaps even a reverse of the trend seen here, but a trend of some sort is almost always present.

Spatial variation can be seen in several ways, the most direct being a spatial image plot as in Figure 2. The microarray for which data is displayed here was printed using a 48 tip print-head with tips in a 12×4 arrangement, so there is a 12×4 pattern of print-tip groups on the array. Each spot on the array corresponds to one small square region on the plot. Spatial plots of the background intensities or un-normalized M-values are particularly useful. Figure 2 shows green background values as estimated by the "morphG" measure from the SPOT image analysis program. The image shows that the array tends to be more green around the edges in the four corners. There is also a green region in tip rows 8 and 9 and columns 3 and 4. A plot of the M-values shows a similar pattern to that seen from the green background values, that the array is relatively more green near the edges and near the corners.

A simpler method of describing spatial patterns is to focus attention on the print tip groups. There may be slight physical differences between the print tips, perhaps differences in length or in the size of the opening or deformations after many hours of printing. Even in the absence of differences between the pins, the print tip groups can be used as a surrogate for more general spatial variation across the array. Figure 3 shows side-by-side boxplots of the M-values for each of the 48 print tip groups of the same array as in Figures 1 and 2. The M-values are higher (more red) in the middle of each sequence of four and in the middle of the overall sequence corresponding to tip rows 7, 8 and 9. (A boxplot displays graphically the so-called 5-number summary of a set of numbers, the three quartiles and the maximum and minimum. The central box of the plot extends from the first to the third quartile and therefore encompasses the middle 50% the data.) The boxplots provide a different view of the spatial pattern seen in the spatial image plot.

The ideas of spatial and intensity trends may be combined by considering separate loess curves for each of the print tip groups. Figure 4 shows a 12 x 4 grid of MA-plots and loess curves for the individual print tip groups. For this array, the slope and shape of the curves is broadly consistent over the print-tip groups although the height varies. The height of the curves varies between tip groups in a similar way to the height of the boxplots in Figure 3.

3 Print-tip Loess Normalization

The idea of print-tip loess normalization can be visualized in Figure 4. Each M-value is normalized by subtracting from it the corresponding value of the tip group loess curve. The normalized log-ratios N are the residuals from the tip group loess regressions, i.e.,

$$N = M - \text{loess}_i(A)$$

where $\text{loess}_i(A)$ is the loess curve as a function of A for the ith tip group. Each loess curve is constructed by performing a series of local regressions, one local regression for each point in the scatterplot. Technically, the local regressions are linear (degree=1), are based on the 40% of the spots which are closest in terms of A-value to the spot being predicted (span=0.4) and are estimated using re-descending M estimation with Tukey's biweight function (family="symmetric") [4]. This method was proposed by Yang *et al* [5]. We recommend this

method as a routine normalization method for cDNA arrays. It corrects the M-values both for sub-array spatial variation and for intensity-based trends.

A simpler form of loess normalization is global loess normalization

$$N = M - \text{loess}(A)$$

where $\text{loess}(A)$ is the global loess curve plotted in Figure 1. This variation does not take into account sub-array variation and we do not recommend it as a routine method. It could be used if careful examination of spatial plots shows that spatial variation is negligible.

Another way to model spatial variation is to use a two-dimensional loess curve. This can be combined with intensity-based loess normalization to give the 2D normalization strategy,

$$N = M - \text{loess}(r, c) - \text{loess}(A)$$

where $\text{loess}(r, c)$ is a two-dimensional loess curve which is a function of the overall row position r and the column position c of the spot on the array. This method models spatial variation using a smooth two-dimensional surface instead of via step-changes at the print-tip groups as does print-tip loess. The intensity-based trend is assumed to be global rather than varying across the array as for print-tip loess normalization. We do not use 2D normalization as a routine normalization strategy because of concern that imperfections on the array may present sudden rather than smooth changes and concern that the 2D loess curve may confuse local clusters of differential expression on the array with the spatial trend to be removed. Further research is being conducted actively on two-dimensional normalization strategies so it is possible that recommendations will change in the future.

There are many additional features which can be added to the normalization process. However, further normalization should be applied only when diagnostic plots show strong evidence of the need for such normalization, as unnecessary estimation and removal of trends adds noise to the data. One more complicated variation on print-tip loess normalization is to further standardize the M-values in each print tip to have the same scale

$$N_s = N / \text{mad}_i$$

where mad_i is the median absolute deviation of the N for the ith tip group and N_s is the print-tip scale-normalized log-ratio. This method was compared with ordinary print-tip loess normalization by Yang *et al* [5]. We do not recommend this as a routine procedure but it can be useful for particularly noisy arrays.

4 Composite Loess Normalization

It is usual to use all or most of the genes on the array in the normalization methods described above. It can be useful to modify this strategy if a suitable set of control spots is available which are known not to be differentially expressed. To be of most use in loess normalization, the control spots should span as wide a range of intensities as possible. A satisfactory set of controls for this purpose is a specially designed microarray sample pool (MSP) titration series in which the entire clone library is pooled and then titrated at a series of different concentrations. Theoretically all labeled cDNA sequences should hybridize to this mixed probe sample, so it should be minimally subject to any sample specific biases.

The loess curve through the control spots offers security that the curve is not biased by differentially expressed genes. On the other hand, the use of all genes for normalization offers the most stability in terms of numbers of spots and the most flexibility in terms of estimating tip group specific trends. In some cases it can be beneficial to use a compromise between the sub-array loess curves and the global titration series curve.

Figure 1 shows the loess curve through a series MSP titration spots. Yang *et al* [6] propose the normalization

$$N = M - p(A)\text{loess}_{\text{MSP}}(A) - \{1 - p(A)\}\text{loess}_i(A)$$

where $\text{loess}_{\text{MSP}}(A)$ is the loess curve through the MSP spots and $p(A)$ is the proportion of spots on the array with A-values less than A. The idea of this proposal is that normalization will be increasingly based on the global MSP curve rather than the individual tip-group curves at higher intensities where the individual curves are less reliable due to the smaller number of spots. In the composite normalization procedure, it is best to use constant local regressions (degree=0) to construct the MSP loess curve so that any necessary extrapolation of the MSP curve outside of the intensity range of the control spots will also be constant, this being the most conservative extrapolation policy.

5 Correcting for Other Trends

There are many other trends which could be estimated and adjusted for in the normalization step, although normally these are of less importance than the intensity and spatial trends already considered. For example, there can be differences between the purity of DNA from different amplification batches or from different clone libraries. This can mean that different spots on the microarray contain different effective quantities of DNA. Different amplification batches and different clone libraries are typically associated with different well plates used to hold DNA during the printing of the array. Differences in DNA purity show up on a scanned array in that the intensities and possibly log-ratios of the spots can vary with the well plate used to hold DNA during printing. Figure 5 shows the M-values of the first array from the ApoAI knock-out experiment reported by Callow *et al* [7]. The horizontal axis gives print-order, i.e., the numerical order in which the spots were laid down during the printing of the array. This array was printed with a 4 x 4 arrangement of print-tips and with 19 rows and 21 columns in each tip group. This means that the print-order index goes from 1 to 19 x 21 = 399, and that 4 x 4 = 16 spots share each print-order index. The solid line on the plot shows the median M-value for each group of 16 consecutive print-order indices. The spots in each of these groups were printed with DNA from the same plate. The plot shows that a series of plates starting around print-order 169 have higher median M-values than the rest of the array. Indeed it turns out that spots with print orders between 169 and 252 were printed with DNA from a different library to the other spots. One can normalize for this effect by subtracting from the M-values the medians shown in the plot. One would then proceed on to print-tip loess normalization as described earlier.

Print-order normalization should be used when, as in this case, exploratory plots of the data reveal a substantial print-order effect in the M-values. Print-order plots of data from the array shown in Figures 1 to 4 show that there are clear differences between the intensities of spots printed from different plates but also show that these differences are not reflected to any

substantial degree on the log-ratios. Print-tip normalization is therefore probably not required for this array.

6 Between Array Normalization

Sometimes there are substantial scale differences between microarrays, because of changes in the photomultiplier tube settings of the scanner or for other reasons. In these circumstances it is useful to scale-normalize between arrays. Scale-normalization is a simple scaling of the M-values from a series of arrays so that each array has the same median absolute deviation.

Figure 6 displays side-by-side boxplots of the normalized M-values for a series of six replicate arrays including slide 0924 displayed in Figures 1 to 4. The much longer box for slide 0936 shows that the spread of the M-values is much larger for this array than the others. The different slides appear to be on varying scales. Some re-scaling seems to be called for to make the arrays more comparable and so that slide 0936 is not overly influential. Scale-normalization can be omitted when the widths of the boxplots are reasonably consistent.

7 Weighting for Spot Quality

Most image analysis programs routinely record a variety of descriptive information about each spot apart from the foreground and background intensities. If this information is used to construct a numeric quality measure for each spot then lower quality spots can be down-weighted in the normalization process.

Information which is recorded on each spot typically includes morphological details such as area, perimeter and location plus heterogeneity measures such as standard deviations or inter-quartile ranges across the pixels used to construct the foreground and background estimates. The meaning and relevance of these measures depends on the image analysis program used and there is no universally recognized measure of spot quality. In general, spots can be expected to be unreliable if they are very small or very large relative to the bulk of spots on the array, if they are markedly non-circular, if the background intensities are high, if the signal to noise ratio is low, if the foreground or background regions are very heterogeneous or if the spot is shifted from its expected location. Many image analysis programs flag low quality spots although these flags are not usually quantitative.

When using the SPOT image analysis program, we have found it useful to weight spots according to their area in lieu of a more comprehensive measure of spot quality. Spot area is defined as the number of pixels in the segmented foreground region of the spot. Inspection of the TIFF images of arrays used in the examples in the article suggests that the area in pixels of an ideal circular spot on these arrays is about 165 pixels. Therefore we have weighted spots so that spots with area equal to 165 pixels get full weight while those which are smaller or larger receive less weight. The weight function given in Figure 7 has been used. This weight function gives weights between 0 and 1 for each spot. It fits in with the principle that a spot of zero size should also get zero weight and with the elementary statistical principle that the reliability of the average intensity of a number of pixels should increase linearly with the number of pixels averaged if the correlation between the pixels is constant. Spots which are larger than the ideal size are also down-weighted because these are likely to include pixels not representative of the printed cDNA. Spots which are more than twice as large as an ideal spot are given zero weight.

The values from the weight function are used as relative weights in all the loess regressions used in the normalizations. This weight function is a simple function of only one of the morphological characteristics of the spot and more complex quality measures can easily be imagined. Other measures of spot quality computed from the image analysis output could be used in the same way to provide weights in the normalization. The prerequisite for weights to be useful is that they should be numerical and inversely proportional to the variances of the M-values.

8 Software

Software to carry out the normalization methods described in this article is freely available from the Bioconductor project site http://www.bioconductor.org. The Bioconductor packages use the free statistical programming environment R. For normalization of cDNA arrays, the relevant packages are marrayNorm [8, 9] and limma. Here we give commands from the limma package.

The first step is data input. If all the SPOT output files are in the working directory of the R session and are named after the slide numbers, then the commands

```
slides <- c("0924","0925","0928","0929","0936","0937")
RG <- read.maimages(slides,ext="spot",wt.fun=wtarea(165))
```

will read data from the output files into a data object in the R session. Note the use of the weight function "wtarea" which creates spot quality weights based on the function in Figure 7. This particular weight function is appropriate only for SPOT image output. If another image analysis program has been used then the spot weight function should be omitted or replaced with one appropriate for that program.

The print-tip loess normalized M and A values can be obtained by the commands

```
layout <- list(ngrid.r=12,ngrid.c=4,nspot.r=26,nspot.c=26)
MA <- normalize(RG,layout)
```

Note the "layout" list which specifies the arrangement of the tips on the print-head and the number of rows and columns printed with each tip. The data object "MA" is a list containing the M-values and A-values for all the arrays as components.

Before doing composite normalization, one needs to know which spots are control spots and which are known to be not differentially expressed. The arrays used in the above example are printed with the National Institute of Aging 15k mouse clone library and with an MSP titration series constructed from a pool of all the clones in the library. The following commands will read in the GAL (Genepix Array List) file from the working directory and will undertake the composition normalization described in Section 4:

```
gal <- readGAL()
nonDE <- grep("cDNA",gal[,"ID"])
MA <- normalize(RG,layout,method="composite",controlspots=nonDE)
```

The commands to identify non-differentially expressed control spots will depend on the naming conventions in the allocation list as well as on the detailed controls printed on the array, so it is impossible to give universally applicable commands. For the arrays used in our example, the titration series spots can be identified by the letters "cDNA" in the library ID column of the GAL file.

Finally,

```
M <- norm.scale(MA$M)
```

will scale normalize the M-values across the six arrays as shown in Figure 6. The operator "$" here picks out the component of the MA-list containing the M-values.

Differential expression can be judged by continuing with the commands

```
fit <- gls.series(M,design=c(-1,1,-1,1,1,-1),ndups=2,correlation=0.65,
                  weights=RG$weights)
eb <- ebayes(fit)
```

The meaning of the argument list is as follows. "Design" specifies the dye-swap pattern, "ndups=2" specifies that each gene is printed twice side-by-side on each array and "correlation=0.65" specifies the spatial correlation between the duplicate spots. Note that the spot quality weights have been used again. This is optional but is recommended. The function gls.series estimates the average fold change and a standard deviation for each gene using generalized least squares and taking into account the pattern of dye-swaps, duplicate spots and quality weights. If the array did not have duplicate spots, i.e., if each gene appeared once only on the array, then the function lm.series could be used in place of gls.series and the arguments ndups and correlation would not be required. The function ebayes computes moderated t-statistics and B-statistics [10] which can be used to rank the genes in order of evidence for differential expression. Finally a list of the top selected genes can be output by

```
toptable(fit=fit,eb=eb,genelist=uniquegenelist(gal,ndups=2))
```

The syntax of these commands is refined from time to time, so readers should check the current documentation of the limma or marrayNorm packages before using them.

9 Conclusion

Normalization methods for cDNA microarrays will no doubt see further development in the future, but print-tip loess normalization provides a well-tested general purpose normalization method which gives good results on a wide variety of arrays. The method may be refined by using quality weights for individual spots. The method is best combined with diagnostic plots of the data. When diagnostic plots show that biases still remain in the data after normalization, further normalization steps such as plate-order normalization or scale-normalization between the arrays may be undertaken.

10 Acknowledgements

The authors are grateful to Dr Lynn Corcoran for permission to use unpublished data from her laboratory at the WEHI and to Henrik Bengtsson for helpful discussions on plate-order normalization.

11 References

1. Buckley, M. J. (2000). *Spot User's Guide*. CSIRO Mathematical and Information Sciences, Sydney, Australia. http://www.cmis.csiro.au/iap/Spot/spotmanual.htm.

2. GenePix Pro microarray and array analysis software, Axon Instruments Inc. http://www.axon.com.

3. Dudoit, S., Yang, Y. H., Callow, M. J., and Speed, T. P. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica* 12, 111-140.

4. Cleveland, W. S., Grosse, E., and Shyu, W. M. (1992). Local regression models. In: J. M. Chambers and T. J. Hastie (eds.), *Statistical Models in S*, Wadsworth & Brooks/Cole.

5. Yang, Y. H., Dudoit, S., Luu, P., and Speed, T. P. (2001). Normalization for cDNA microarray data. In M. L. Bittner, Y. Chen, A. N. Dorsel, and E. R. Dougherty (eds.), *Microarrays: Optical Technologies and Informatics*, Volume 4266 of Proceedings of SPIE.

6. Yang, Y. H., Dudoit, S., Luu, P., Lin, D. M., Peng, V., Ngai, J., and Speed, T. P. (2002). Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research* 30(4):e15.

7. Callow, M. J., Dudoit, S., Gong, E. L., Speed, T. P., and Rubin, E. M. (2000). Microarray expression profiling identifies genes with altered expression in HDL deficient mice. *Genome Research* **10**, 2022-2029.

8. Dudoit, S., Yang, Y. H., and Bolstad, B. (2002). Using R for the analysis of DNA microarray data. *R News* **2** (1), 24-32.

9. Dudoit, S., and Yang, Y. H. (2002). Bioconductor R packages for exploratory analysis and normalization of cDNA microarray data. In G. Parmigiani, E. S. Garrett, R. A. Irizarry and S. L. Zeger, editors, *The Analysis of Gene Expression Data: Methods and Software*, Springer, New York. To appear.

10. Lönnstedt, I. and Speed, T. P. (2002). Replicated microarray data. *Statistica Sinica* 12, 31-46.

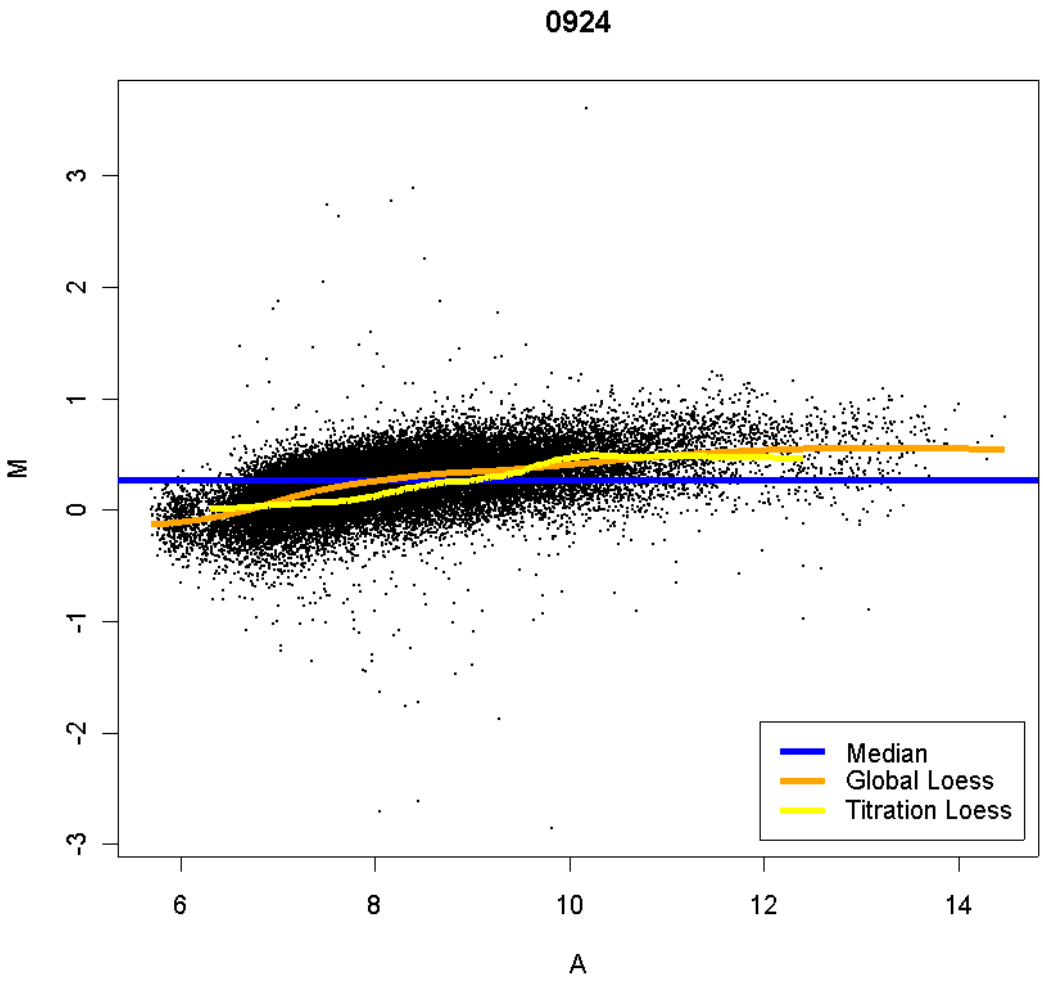

Figure 1. MA-plot showing three different trend lines. The horizontal blue line shows the median of the M-values. The continuous orange curve shows the overall trend line as estimated by loess regression. The yellow curve shows the loess curve through a set of control spots known to be not differentially expressed.

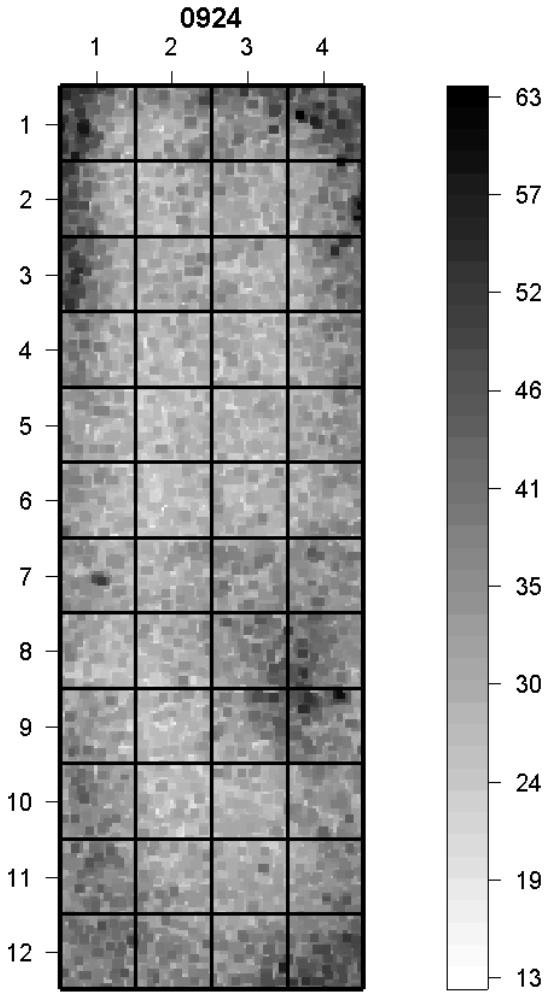

Figure 2. Spatial plot of green background values. The array was printed using a 12 x 4 pattern of print tips. The image shows that the array tends to be more green around the edges in the four corners. There is also a green patch in tip rows 8 and 9 and columns 3 and 4.

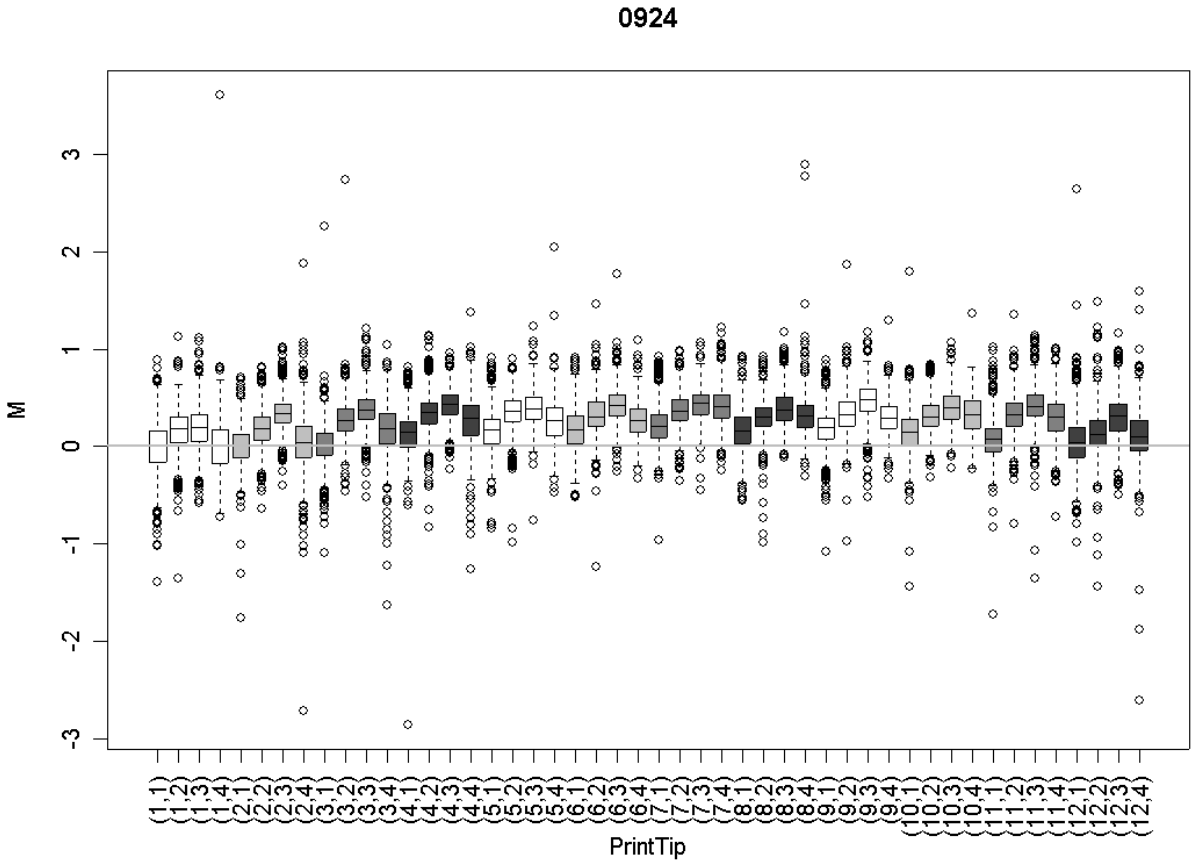

Figure 3.Boxplots of the M-values for each print tip group. The M-values are higher (more red) in the middle of each sequence of four and in the middle of the overall sequence.

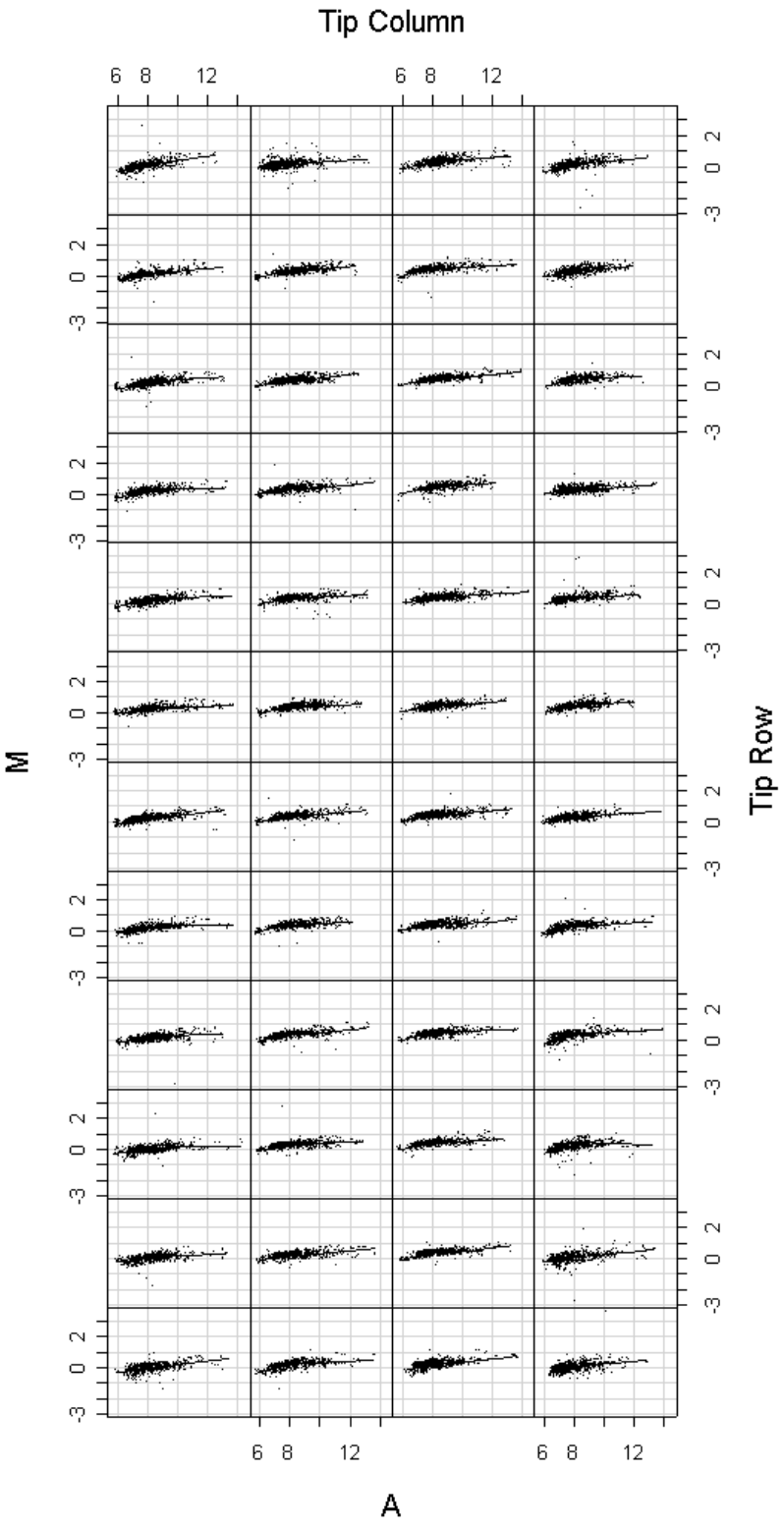

Figure 4. Individual loess curves for the 48 print tip groups for the same array as in Figures 1 to 3.

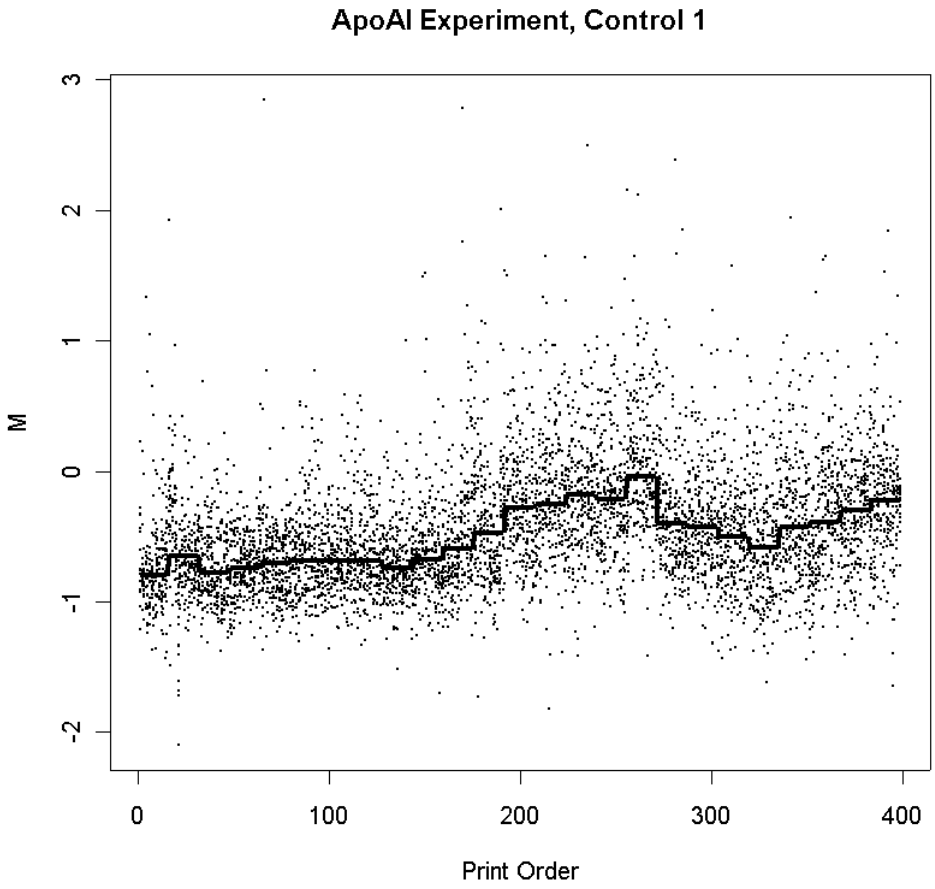

Figure 5. Plate or print-order effects for the first slide in the ApoAI knock-out experiment.

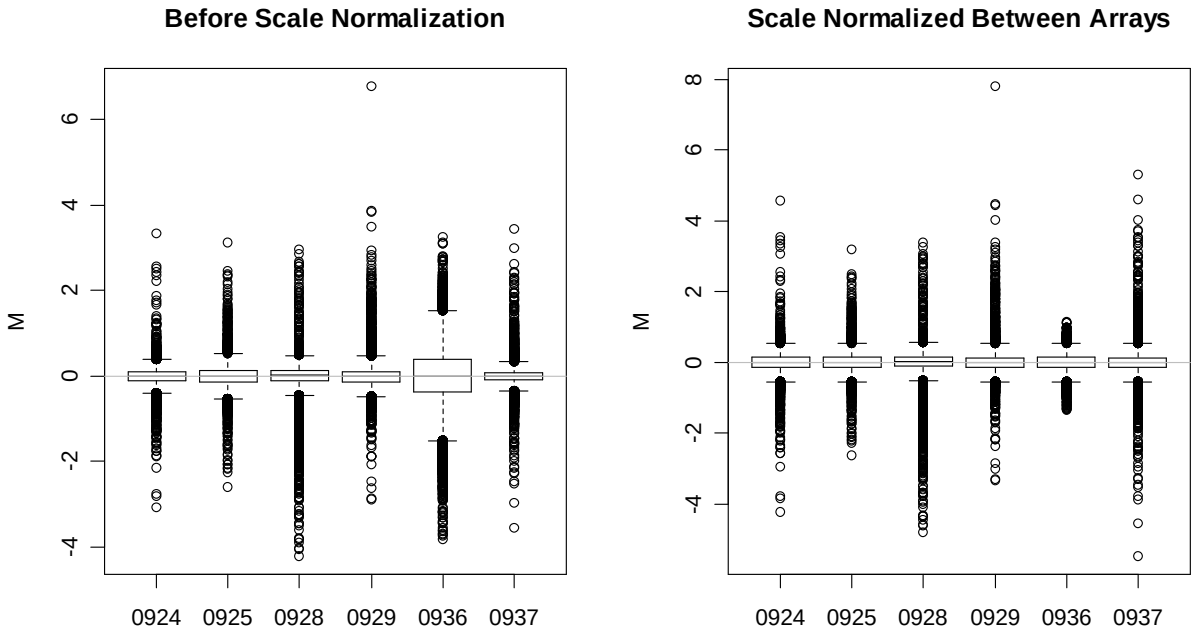

Figure 6. Between array scale normalization for a series of six arrays. (a) Side-by-side boxplots of the M-values from 6 arrays. The arrays are replicates except that three are dye-swap pairs of the others. Array 5 has a much larger spread than the others. (b) Boxplots of the same arrays after scale-normalization to equalize the median absolute value for each array. Data from the Corcoran Lab, WEHI.

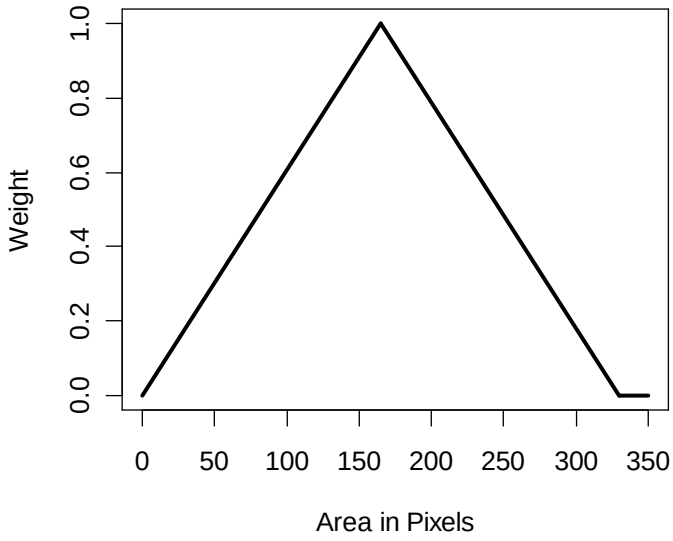

Figure 7. Weight function used to down-weight spots smaller and larger than 165 pixels in the examples in this article.