Analysis of cDNA microarray images

Yee Hwa Yang, Michael J. Buckley and Terence P. Speed

Date received (in revised form): 4th September 2001

Yee Hwa Yang
is a PhD student at the University of California, Berkeley. Her current research is on design and analysis of gene expression data.

Michael J. Buckley
has a PhD in nonparametric regression. His interests are in applied image analysis, statistics and in various aspects of bioinformatics.

Terence P. Speed
is Professor of Statistics at the University of California, Berkeley, and joint head of the Division of Genetics and Bioinformatics at the Walter and Eliza Hall Institute of Medical Research. His research concerns the application of statistics to problems in genetics and molecular biology.

Keywords: *image processing, segmentation, background correction, gene expression*

Abstract

Microarrays are part of a new class of biotechnologies that allow the monitoring of expression levels for thousands of genes simultaneously. Image analysis is an important aspect of microarray experiments, one that can have a potentially large impact on subsequent analyses, such as clustering or the identification of differentially expressed genes. This paper reviews a number of existing image analysis methods used on cDNA microarray data. In particular, it describes and discusses the different segmentation and background adjustment methods. It was found that in some cases background adjustment can substantially reduce the precision – that is, increase the variability of low-intensity spot values. In contrast, the choice of segmentation procedure seems to have a smaller impact.

INTRODUCTION

Image analysis is an important aspect of microarray experiments. It can have a potentially large impact on subsequent analysis such as clustering or the identification of differentially expressed genes. In microarray experiments, hybridised arrays are imaged in a microarray scanner to produce red and green fluorescence intensity measurements at each of a large collection of pixels which together cover the array. These fluorescence intensities correspond to the levels of hybridisation of the two samples to the DNA sequences spotted on the slide. Fluorescence intensities are usually stored as 16-bit images which we view as 'raw' data.

Over the last four years, a number of cDNA microarray image analysis packages for glass slides, both commercial software and freeware, have become available. Some of these packages are variants of those used to analyse radioactive signals from arrays spotted onto nylon membranes. Others are designed specifically for glass slide arrays. These specifically designed packages take advantage of the rigid layout of the spots in their spot-finding algorithm, as well as utilising information from the two channels. The processing of scanned microarray images can generally be separated into three tasks.

- *Addressing* or *gridding* is the process of assigning coordinates to each of the spots. Automating this part of the procedure permits high-throughput analysis.
- *Segmentation* allows the classification of pixels either as foreground – that is, within printed DNA spot – or as background.
- The *intensity extraction* step includes calculating, for each spot on the array, red and green foreground fluorescence intensity pairs (R,G), background intensities and, possibly, quality measures.

An additional aspect associated with image processing is the visualisation of array data. The input to the image analysis procedure consists of a pair of unsigned 16-bit images which are stored as TIFF (tagged image file format) files. These images are named 'R' and 'G', for 'red' and 'green', with R corresponding to the dye Cy5 and G to Cy3. Often images R and G are overlaid for addressing and

Terence P. Speed,
Department of Statistics,
University of California,
Berkeley,
367 Evans Hall,
Berkeley,
CA 94720-3860,
USA

Tel: +1 (510) 642-0613
Fax: +1 (510) 642-7892
E-mail: terry@stat.berkeley.edu

visualisation purposes. The two 16-bit TIFF images (scanned output from the Cy3 and Cy5 channels) are compressed into 8-bit images using a square root transformation. The objective of this transformation is to display fluorescence intensities for both wavelengths using a 24-bit composite RGB overlay image. In this RGB image, blue values (B) are set to zero, red values (R) are used for Cy5 intensities, and green values (G) are used for Cy3 intensities. Alternatives for reducing 16-bit TIFF images into 8-bit images include selecting only 8 bits of each 16-bit image. For example, the software GenePix offers the option of displaying only high intensities, while the bottom 8 bits are ignored. This results in the display of intensities from 256 to 65,535.

An issue prior to the addressing and segmentation stages is whether the pair of input images should be processed separately or simultaneously. Most software packages form a combined image before the addressing stage. Analysing the two fluorescence images separately has the benefit of removing concerns over misregistration between the two images. With nylon filters where only one sample is hybridised onto a membrane, addressing is usually done separately, and warping usually occurs during the image acquisition stage. With glass slide arrays, both input images are generated based on scanning the same rigid glass slide and the two images can often be combined. Such combinations allow addressing and segmentation algorithm to take advantage of signal information from both channels. In our software *Spot*, a combined image is formed with properties that the two inputs – that is, raw images R and G – contribute equally in combination. In addition, very high pixel values are damped in the combined image to prevent very bright pixels from dominating in both the addressing and segmentation phases. Furthermore the combined image is reduced to an 8-bit image for ease of computation. The automatic addressing and segmentation procedures are performed on this 8-bit combined image. The segmentation method produces a spot mask which is used together with the original 16-bit images for extraction of spot foreground and background intensities. Details of this can be found in Yang *et al.*[1]

Automatic addressing permits high throughput analysis

In this paper we review existing image analysis methods, with an emphasis on segmentation and background adjustment. The goal is to extract for each spotted DNA sequence a measure of transcript abundance in the two labelled mRNA samples, as well as to obtain background estimates and quality measures. This section is not meant to be a survey of all microarray image analysis software packages available, but, rather, different packages, proprietary and non-proprietary, are mentioned mainly as examples of implementations of certain methods and algorithms.

ADDRESSING

The basic structure of a microarray image is determined by the arrayer and is therefore known. That is, it is known that there are, say, four rows and four columns of grids, and that within each grid there are 19 rows and 21 columns of spots. However, to address the spots in an image – that is, to match an idealised model of the array with the scanned image data – a number of parameters need to be estimated. These parameters include:

- separation between rows and columns of grids;
- individual translation of grids (caused by slight variations in print-tip positions);
- separation between rows and columns of spots within each grid;
- small individual translations of spots; and
- overall position of the array in the image.

Within a batch of microarray images produced together, the last of these is usually the most highly variable. Other parameters that may in some cases need to be estimated include misregistration of the red and green channels, rotation of the array in the image, and skewness in the array. The last two parameters are important issues for automated gridding algorithms, but a lesser problem if manual grid placement is used. In addition, with the improvement of printing and scanning technologies, some of these parameters such as misregistration between the two channels and small individual translations of spots are likely to decrease in importance.

User intervention in addressing has the potential to increase reliability but it can also be time consuming

To achieve higher levels of accuracy in the measurement process, it is desirable for the addressing procedure to be as reliable as possible. Reliability of the addressing stage can be enhanced by allowing user intervention. However, this has the potential to make the process very slow. Ideally we seek reliability while attempting to minimise user intervention to maximise efficiency. The addressing steps are often referred to as 'gridding' in the microarray literature. Most software systems now provide for both manual and automatic gridding procedures. These are very varied and mostly have not been publicly documented, thus we will not attempt to describe them here. Instead, we focus on the various segmentation and background methods.

SEGMENTATION

Segmentation of a microarray image involves classifying pixels into foreground (within printed cDNA spots) or background

Generally, *segmentation* of an image can be defined as the process of partitioning an image into different regions, each having certain properties.[2] In a microarray experiment, segmentation allows the classification of pixels as foreground (ie within a spot) or background, so that fluorescence intensities can be calculated for each spotted DNA sequence as measures of transcript abundance. Any segmentation method produces a *spot mask*, which consists of the set of foreground pixels for each given spot. Existing segmentation methods for microarray images can be categorised into four groups, according to the geometry of the spots they produce:

- fixed circle segmentation;
- adaptive circle segmentation;
- adaptive shape segmentation; and
- histogram segmentation.

Table 1 lists different segmentation approaches and examples of software implementations. In general, most software packages implement a number of segmentation methods.

Fixed circle segmentation

Fixed circle segmentation fits a circle with a constant diameter to all the spots in the image. This method is easy to implement and works nicely when all the spots are circular and of the same size. It was probably first implemented in the *ScanAlyze* software written by Eisen[3] and it is usually provided as an option in most software. Figure 1 contains a small portion of an array, with spots ranging from 5 to 10 pixels in diameter. A fixed diameter segmentation may not be satisfactory to detect the exact shape for all the spots.

Theoretically, if the background affects the foreground values additively and the background value can be reliably estimated, one could use a very large fixed diameter for segmentation such that the entire spot is covered for all spots. That is, any segmentation that is too large can yield perfectly good (unbiased) estimates if the background contribution can be removed. On the other hand, an ability to detect the exact shape for all spots limits the amount of irregular noise within the spot mask (for example, bright pixels due to dust, scratch or contribution from neighbouring spots).

Adaptive circle segmentation

In this kind of segmentation, the circle's diameter is estimated separately for each spot. The software *GenePix* for the Axon

Table 1: Segmentation methods and examples of algorithms and software implementation

Fixed circle	ScanAlyze, GenePix, QuantArray
Adaptive circle	GenePix, Dapple
Adaptive shape	Spot, region growing and watershed
Histogram method	ImaGene, QuantArray, DeArray and adaptive thresholding

Seeded region growing is an example of adaptive segmentation

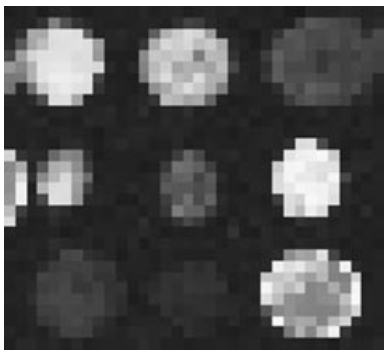

Figure 1: Small portion of the scanned image from the green (Cy3) channel for knock-out mouse #4 from an experiment studying lipid metabolism in mice.[4] This image displays nine spots on gray-scale, where the dark black pixels represent low pixel values and the bright white pixels represent high pixel values. Note the different sizes and shapes of the spots

scanner implements such an algorithm.[5] Note that GenePix and other software provide the user with the option to adjust the circle diameter spot by spot. This practice can be very time consuming, since each array contains thousands of spots. The software *Dapple*[6] finds spots by detecting edges of spots. Briefly, Dapple calculates the negative second derivative of the image (Laplacian). Pixels with high values in the Laplacian image correspond to edges of a spot. In addition, Dapple enforces a circularity constraint by finding the brightest ring (circle) in the Laplacian images.

Some microarray images contain spots of different sizes and shapes

Adaptive circle segmentation methods will work rather well as circular spots are probably typical of most commercially produced arrays. However, spots printed from non-commercial arrayers are rarely perfectly circular and can exhibit oval or doughnut shapes.[7] A circular spot mask can thus provide a poor fit as shown in Figure 2 for a non-circular shaped spot. Sources of non-circularity include the printing process (eg features of the print-tips, uneven solute deposition) or the post-processing of the slides after printing (eg insufficient time of rehydration). Again, segmentation algorithms that do not place restrictions on the shape of the spots are thus more desirable if one is attempting to determine the exact spot shape.

Adaptive shape segmentation

Two commonly used methods for adaptive segmentation in image analysis are the watershed[8,9] and seeded region growing (SRG).[10] These methods are beginning to be applied in microarray analysis, although not in the most widely-used software packages.

Both watershed and SRG segmentation require the specification of starting points, or seeds. A weakness of segmentation procedures using these methods can be the selection of the number and location of the seed points. In microarray image analysis, however, we are in the rather unusual situation where the number of features (spots) is known exactly *a priori*,

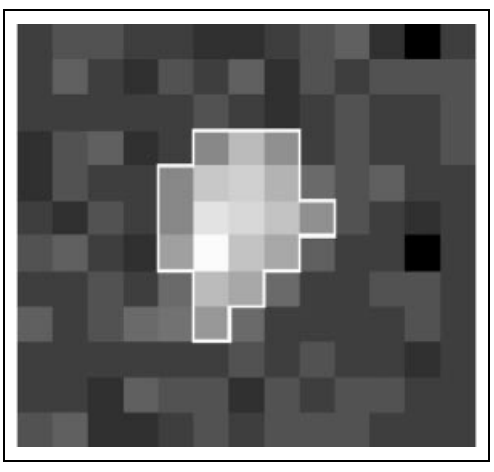

Figure 2: An example of a non-circular shaped spot. The thick white line shows the result of the SRG segmentation. The pixels inside the thick white line are classified as foreground and the other pixels are classified as background

and the approximate locations of the spot centres are determined at the addressing stage. Microarray images are therefore well suited to such methods. The SRG algorithm is implemented in spot. Details regarding the placement of foreground and background seeds can be found in Yang *et al.*[1]

Histogram segmentation

This method uses a *target mask* chosen to be larger than any other spot. For each spot, foreground and background intensity estimates are determined in some fashion from the histogram of pixel values for pixels within the masked area. For example, *QuantArray* uses a square target mask and defines foreground and background as the mean intensities between some predefined percentile values. By default, these are the 5th and 20th percentiles for the background and the 80th and 95th percentiles for the foreground. These methods therefore do not use any local spatial information.

Intensity extraction involves calculating red and green foreground intensity pairs and background intensities for each spot on the array

Another example of this class of methods is described by Chen *et al.*[11] This method uses a circular target mask and computes a threshold value based on a Mann–Whitney test. Pixels are classified as foreground if their value is greater than the threshold, and as background otherwise. This method is implemented in the *QuantArray* software for the GSI Lumonics scanner[12] and *DeArray* by Scanalytics.

It is likely that fluorescence from regions of the slide not occupied by DNA is different from regions occupied by DNA

Simplicity is the main advantage of this method. However, a major disadvantage is that quantification is unstable when a large target mask is set to compensate for variation in spot size. Furthermore, the resulting spot masks are not necessarily connected.

Simplicity is the main advantage of histogram segmentation

INFORMATION EXTRACTION

Spot intensity

Each pixel value in a scanned image represents the level of hybridisation at a specific location on the slide. The total amount of hybridisation for a particular spotted DNA sequence is proportional to the *total fluorescence* at the spot. The natural measure of spot intensity is therefore the sum of pixel intensities within the spot mask. Since later calculations are based on the ratio of fluorescence intensities, we compute the average pixel value over the spot mask. This yields identical results, as the ratio of averages is equal to the ratio of sums. An alternative measure used is ratio of medians, where the median pixel value over the spot mask is computed. This measure is not associated with any biological meaning but can be seen as a robust variant of the ratio of means.

Background intensity

The motivation for background adjustment is the belief that a spot's measured intensity includes a contribution not specifically due to the hybridisation of the target to the probe, for example, non-specific hybridisation and other chemicals on the glass. If such a contribution is indeed present, we would like to measure and remove it to obtain a more accurate quantification of hybridisation. The glass slides are treated chemically so that the spotted cDNA fragments will bind to them. After the cDNA spots are printed, the slides are treated again so that target DNA does not bind to them. Nevertheless, some binding of the target to the slide may occur. Furthermore, there may be some fluorescence away from the spots due to the slide's surface treatment and the glass. It seems likely that the fluorescence from regions of the slide not occupied by DNA is different from that from regions occupied by DNA. It follows that measuring the intensity in some region near a spot and subtracting it may *not* be the best way to correct for this extra contribution, even though this is what many people are doing. It would be interesting to compare the morphological and local background estimates to ones based on local negative controls (ie nearby spotted cDNA sequences which should have no hybridisation signal).

Apart from histogram-based methods,

the segmentation procedures described above produce local background regions as well as segmented spots. We can broadly classify the various background methods implemented in software packages into four categories.

Estimate 'local' background intensities from regions around printed cDNA spots

Local background

Background intensities are estimated by focusing on small regions surrounding the spot mask. Usually, the background estimate is the median of pixel values within these specific regions. Most software packages we have encountered implement such an approach.

The *ScanAlyze* package considers as background all pixels that are not within the spot mask but are within a square centred at the spot centre. This is represented by the dotted square in Figure 3. The median value of these pixels is used as an estimate of the local background intensity. One of the background adjustment methods implemented in QuantArray and ArrayVision considers the area between two concentric circles, such as the area between the two larger circles in Figure 3. By not considering the pixels immediately surrounding the spots, the background estimate is less sensitive to the performance of the segmentation procedure. An alternate set of pixels to be considered as background (implemented in Spot) is shown as the four dashed diamond-shaped areas in Figure 3. These regions are referred to as the *valleys* of the array and have the furthest distance from all four surrounding spots. The local background for each spot can be estimated by the median of values from the four surrounding valleys. Depending on the software, the local valley regions are different, but this method of background estimation is somewhat independent of the segmentation results. The background method implemented by GenePix effectively calculates the median intensity from local valley regions.

Using valley pixels which are very distant from all spots ensures to a large degree that the background estimate is not corrupted by pixels belonging to a spot. Such corruption by bright pixels may occur in the other methods, particularly in the ScanAlyze method, introducing an upward bias into the background estimate. Using remote pixels reduces this bias effectively but entails the use of a smaller number of pixels and therefore increases the variance of the estimate. This is an example of the bias–variance trade-off. Most software packages allow users to choose their preferred version of local background method.

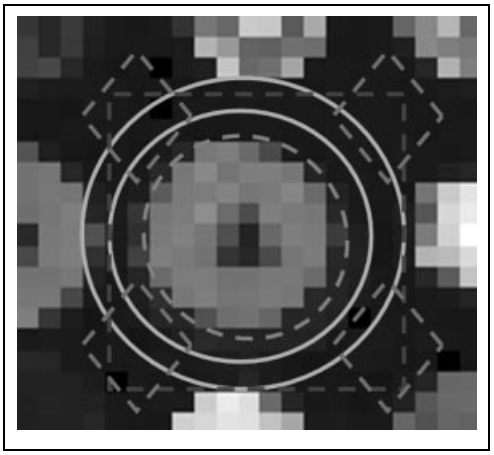

Figure 3: Image illustrating different local background adjustment methods. The region inside the dashed circle represents the spot mask and the other regions bounded by lines represent regions used for local background calculation by different methods. Solid circles: used in QuantArray; dotted square: used in ScanAlyze; and dashed diamond shapes: used in Spot. This image is from KO mouse #8 in an experiment studying lipid metabolism in mice[4]

***Morphological opening* gives background estimates that are lower and less, variable than other estimates**

Morphological opening

This approach to background adjustment relies on a non-linear filter called *morphological opening*. This filter is obtained by computing a form of local minimum filter (an *erosion*) followed by a form of local maximum filter (a *dilation*) with the same window. In a microarray image, the effect of such non-linear filtering using a window that is larger than any of the spots is to remove all spots, replacing them by nearby background values. See Soille[2] for a detailed description of these filters.

In Spot, morphological opening is applied to the original images R and G using a square structuring element with side length at least twice as large as the spot separation distance. This operation removes all the spots and generates an image that is an estimate of the background for the entire slide. For individual spots, the background is estimated by sampling this background image at the nominal centre of the spot. We simply chose to sample this image rather than take an average over a 'background region' because very similar results are expected from both methods. A large window was used to create the morphological background image, hence it is expected to have slow spatial variation.

Morphological opening results in lower background estimates than other simpler methods. More importantly, though, morphological background estimation is expected to be less variable than the other methods, because spot background estimates are based on pixel values in a large local window, and yet are not corrupted (ie biased upwards) by brighter pixels belonging to or on the edge of the spots.

Most programs view quality assessment as 'flagging' spots in one or the other channel

Constant background

This is a global method which subtracts a constant background for all spots. The approaches previously described assume that the non-specific binding to a spot can be estimated by the surrounding area. However, some findings[13] suggest that the binding of fluorescent dyes to 'negative control spots' (eg spots corresponding to plant genes that should not hybridise with human mRNA samples) is lower than the binding to the glass slide. If this is the case, it may be more meaningful to estimate background based on a set of negative control spots. When there are no negative control spots, one could approximate the average background by, for example, the third percentile of all the spot foreground values.

Compared to different background adjustments, choice of segmentation procedure seems to have a smaller impact on downstream analysis

No adjustment

Finally, we also consider the possibility of no background adjustment at all.

Calculating log-ratios with no background adjustment is an option

Quality measures

In addition to the actual spot foreground and background intensities, it is also desirable to collect statistics describing the quality of these measurements. Examples of quality measures provided in most software include variability measures in pixel values within each spot mask, spot size (area in pixels), a circularity measure and relative signal to background intensity. Most software packages provide a reject and accept assessment on spot quality. Dapple defines two measures: *b-score* measures the fraction of background intensities less than the median foreground intensity while *p-score* measures the extent to which the position of a spot deviates from a rigid rectangular grid. A classifier is built based on these two measures to accept, reject or flag any spots. Flagged spots need to be manually accepted or rejected.

Most programs have yet to make fuller use of these measures in their analysis, as relating them to more common statistical concepts such as reproducibility seems to be difficult. Research along these lines is being carried out.

SUMMARY AND RECOMMENDATIONS

Our comparison of different methods[1] found that the choice of background correction method has a larger impact on the log-intensity ratios than the segmentation method used. Thus, finding better segmentation methods may not be as important as choosing a stable and accurate background estimation method. In the estimation of background contribution, our study suggests that morphological opening provides a better estimate of background than other methods. The log-ratios $\log_2 R/G$ computed after morphological background correction tended to exhibit low within- and between-slide variability. In addition, this method did not seem to compromise on accuracy.

Figure 4 displays an *MA* plot where we plot the log-intensity ratio $M = \log_2 R/G$ versus the mean log

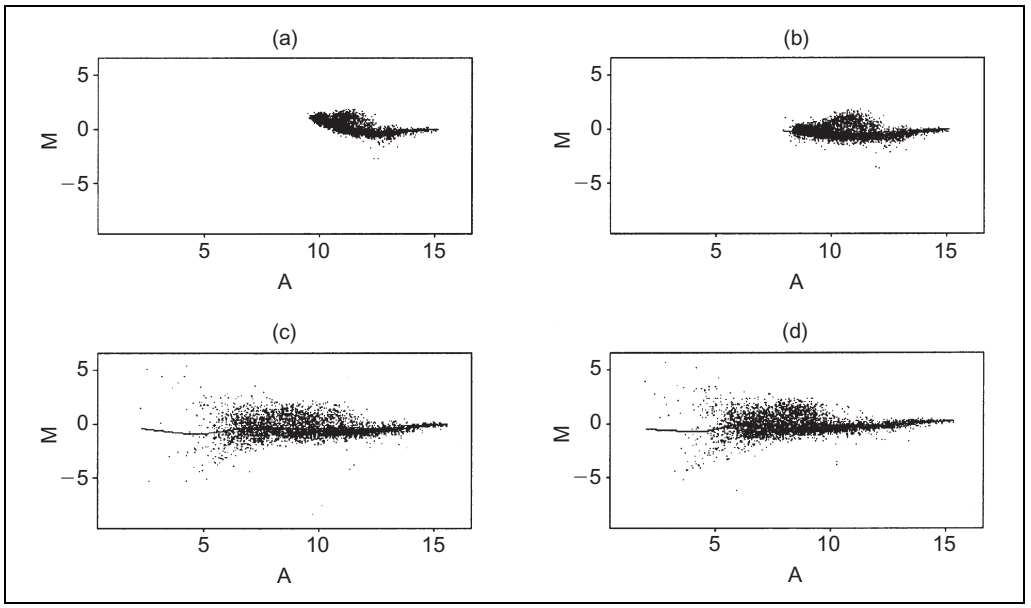

Figure 4: *MA* plot for methods (a) Spot with no background subtraction; (b) Spot with morphological opening; (c) GenePix; and (d) ScanAlyze (data from KO mouse #8 in an experiment studying lipid metabolism in mice[4]

A dilution experiment will provide a more conclusive study

Background adjustment has a larger impact on low intensity spots

intensity $A = \log_2\sqrt{RG}$. The different panels show the *MA* plots from the same image quantified by different image analysis methods. A good image analysis method should permit a clear distinction to be made between differentially expressed genes and noise. Notice that methods using local background adjustment (panels (c) and (d)) show greater variability around the low-intensity spots than methods not using any background subtraction (panel (a)) or or the morphological background adjustment (panel (b)). Local background adjustment tends to blur the distinction between differentially expressed genes and noise.

For a more conclusive study of the statistical properties of different image processing methods, one would need a more rigorous assessment of bias, such as one based on an external measure of truth. One might verify estimated expression levels via Northern blot or reverse transcriptase polymerase chain reaction (RT-PCR). However, it is difficult to compare the quantifications from RT-PCR with those from microarrays.[14] Alternatively, to address the bias issue fully, one could imagine specially created benchmark data sets with some 'ground truth'. One example is a series of dilution experiments, which would bypass the need to know the true fold changes. A dilution experiment refers to a series of hybridisations where the same pair of mRNA samples are competitively hybridised at different concentrations across different arrays. More specifically, for a dilution experiment that aims to compare mRNA samples A and B, hybridisations of A versus B are repeated with different amounts of starting material for each array, for example, 50 μg, 25 μg, 20 μg, 10 μg and 5 μg for A and B in each of five arrays, or possibly different amounts of A and B for the same array. For any given gene, one can deduce from the different concentrations of starting material on different arrays the true value of ratios of ratios. One could then study the behaviour of log-ratios for thousands of genes across the different dilutions and the best image analysis method would be that leading to the smallest overall mean square error (or some similar measure) between estimated and known values.

The comparison of different background correction methods indicates that estimates based on means or medians over local neighbourhoods tend to be quite noisy and can potentially double the standard deviation of the log-ratios. At the other extreme, no background adjustment seems to reduce the ability to identify differentially expressed genes, as shown by the decrease in the magnitude of the *t*-statistics in our study.[1] Therefore, we recommend performing an intermediate background adjustment, which provides less variable estimates than local background methods and more accurate estimates than raw intensities (no background correction at all). Morphological opening seems to provide a good balance in the bias–variance trade-off. In software packages where morphological opening is unavailable, calculating log-ratios without background subtraction can be better than subtraction of a local background estimate.

Acknowledgements
We thank anonymous referees for detailed reading and constructive comments on an earlier version of this paper.

References

1. Yang, Y. H., Buckley, M. J., Dudoit, S. and Speed, T. P. (2001), 'Comparisons of methods for image analysis on cDNA microarray data', Technical report #584, Department of Statistics, University of California, Berkeley.
2. Soille, P. (1999), 'Morphological Image Analysis: Principles and Applications', Springer-Verlag Berlin Heidelberg.
3. Eisen, M. B. (1999). URL: http://rana.lbl.gov/manuals/ScanAlyzeDoc.pdf
4. Callow, M. J., Dudoit, S., Gong, E. L. *et al.* (2000), 'Microarray expression profiling identifies genes with altered expression in HDL-deficient mice', *Genome Res.*, Vol. 10, pp. 2022–2029.
5. GenePix 4000 A User's Guide (1999), Axon Instruments, Inc. URL: http://www.axon.com/GN_Genomics.html#software
6. Buhler, J., Ideker, T. and Haynor, D. (2000), 'Improved techniques for finding spots on cDNA microarrays', University of Washington.
7. Eisen, M. B. and Brown, P. O. (1999), 'DNA arrays for analysis of gene expression', *Methods Enzymol.*, Vol. 303, pp. 179–205.
8. Beucher, S. and Meyer, F. (1993), 'The morphological approach to segmentation: The watershed transformation', Vol. 34, 'Optical Engineering', pp. 433–481.
9. Vincent, L. and Soille, P. (1991), 'Watersheds in digital spaces: An efficient algorithm based on immersion simulations', *IEEE Trans. Pat. Anal. Machine Intell.*, Vol. 13, pp. 583–598.
10. Adams, R. and Bischof, L. (1994), 'Seeded region growing', *IEEE Trans. Pattern Anal. Machine Intell.*, Vol. 16, pp. 641–647.
11. Chen, Y., Dougherty, E. R. and Bittner, M. L. (1997), 'Ratio-based decisions and the quantitative analysis of cDNA microarray images', *J. Biomed. Optics*, Vol. 2, pp. 364–374.
12. QuantArray Analysis Software (1999), GSI Lumonics. URL: http://www.packardbioscience.com/pdf/product_old/quantarray.pdf
13. Lou, X. J. (1999), 'Human primary cell gene expression monitoring using cDNA microarrays', in 'Microarray Algorithms and Statistical Analysis: Methods and Standards', poster at Lake Tahoe Center conference.
14. Bartosiewicz, M., Trounstine, M., Barker, D. *et al.* (2000), 'Development of a toxicological gene array and quantitative assessment of this technology', *Arch. Biochem. Biophys.*, Vol. 376, pp. 66–73.