Biostatistics (2006), **7**, 1, *pp.* 85–99
doi:10.1093/biostatistics/kxi042
Advance Access publication on July 27, 2005

Probabilistic segmentation and intensity estimation for microarray images

RAPHAEL GOTTARDO*, JULIAN BESAG, MATTHEW STEPHENS, ALEJANDRO MURUA
Department of Statistics, University of Washington, Box 354322, Seattle, WA 98195-4322, USA
raph@stat.washington.edu

SUMMARY

We describe a probabilistic approach to simultaneous image segmentation and intensity estimation for complementary DNA microarray experiments. The approach overcomes several limitations of existing methods. In particular, it (a) uses a flexible Markov random field approach to segmentation that allows for a wider range of spot shapes than existing methods, including relatively common 'doughnut-shaped' spots; (b) models the image directly as background plus hybridization intensity, and estimates the two quantities simultaneously, avoiding the common logical error that estimates of foreground may be less than those of the corresponding background if the two are estimated separately; and (c) uses a probabilistic modeling approach to simultaneously perform segmentation and intensity estimation, and to compute spot quality measures. We describe two approaches to parameter estimation: a fast algorithm, based on the expectation-maximization and the iterated conditional modes algorithms, and a fully Bayesian framework. These approaches produce comparable results, and both appear to offer some advantages over other methods. We use an HIV experiment to compare our approach to two commercial software products: Spot and Arrayvision.

Keywords: Bayesian estimation; cDNA microarrays; Expectation-maximization; Gene expression; Hierarchical-t; Image analysis; Iterated conditional modes; Markov chain Monte Carlo; Markov random fields; Quality measures; Segmentation; Spatial statistics.

1. INTRODUCTION

Development of complementary DNA (cDNA) microarray technology allows investigators to measure the expression levels of thousands of genes in tens or hundreds of samples. These measurements have many potential applications, including characterizing and classifying diseases, studying the response to new drugs, and so on. The probes on a cDNA microarray (Schena *et al.*, 1995) are the different DNA sequences that are spotted on a pretreated glass slide using a robotic arrayer. High-density cDNA arrays can contain tens of thousands of spots (probes) for different genes. Microarrays exploit the ability of a single-strand nucleic acid molecule to hybridize to a complementary sequence. When an RNA sample is labeled and hybridized to the array, the amount of labeled RNA that is hybridized to each probe can be measured. Hence, researchers can use a single experiment to measure the expression levels of thousands of genes within a cell.

*To whom correspondence should be addressed.

© The Author 2005. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oupjournals.org.

In a typical application of cDNA arrays, gene expression patterns between two samples (e.g. a treatment and a control) are compared. The RNA is extracted from both samples and each is labeled with a different fluorescent dye. Generally, one dye is red, the other green. Next, the RNA samples are mixed and cohybridized to the probes on the cDNA array, which is then scanned to provide a 16-bit gray-scale image for each dye. The relative intensity of the dyes in each spot measures the relative abundance of that particular RNA type in the sample. Several factors, such as the hydrophobicity of the pretreated glass surface, the humidity as the probe dries, and the speed of drying, induce unequal distribution of probe material in the spot (Hedge *et al.*, 2000) and can result in spots having irregular shape and size. Figure 1 shows an image from one of the data sets discussed later in the paper; note the evidence of doughnut shapes in some spots.

Image analysis is required to produce estimates of the foreground and background intensities for both the red and green dyes for each gene. These estimates are the starting point of any statistical analysis such as testing for differential expression (Tusher *et al.*, 2001; Efron *et al.*, 2001; Newton *et al.*, 2001; Dudoit *et al.*, 2002; Gottardo *et al.*, 2003), discriminant analysis (Golub *et al.*, 1999; Tibshirani *et al.*, 2002), and clustering (Eisen *et al.*, 1998; Tamayo *et al.*, 1999; Yeung *et al.*, 2001). To estimate the intensities, one first needs to locate the spots on the images and then to classify each pixel either as part of a spot or as background. Chen *et al.* (1997) provide an early statistical treatment of this task and Yang *et al.* (2002) discuss the effects of different approaches.

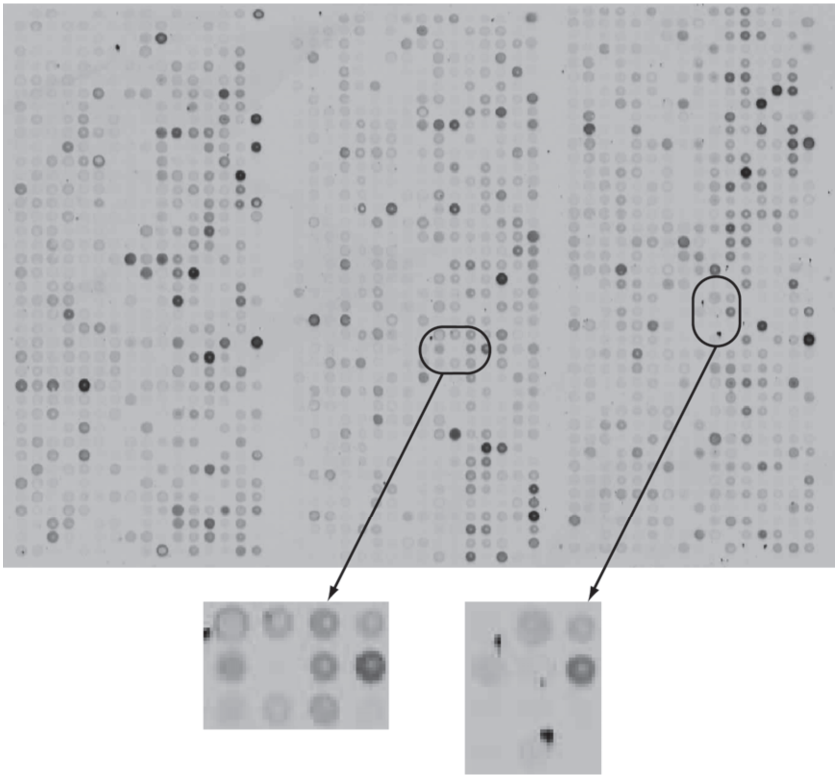


Fig. 1. Three blocks of one of the HIV raw images. The whole image contains 12 blocks. Each block is formed by 16 × 40 spots. At the bottom, we have enlarged two portions of the image containing several artifacts not caused by hybridization of the probes to the slide. Some spots are doughnut shaped with larger intensity on the perimeter of the spot.

There are three main issues in the analysis of microarray images:

(1) Addressing or gridding, which consists of locating the spots on the array;
(2) Segmentation, which consists of classifying the pixels either as foreground (spot) or as background; and
(3) Intensity estimation, which consists of estimating the foreground and background intensities of each spot on the array in each sample. Estimation of the background intensity is usually considered necessary in order to accurately estimate the amount of hybridized cDNA. This is motivated by the fact that the observed intensity of a spot includes a contribution that is not due to the hybridization of the RNA samples to the spotted DNA.

In this paper, we are mainly concerned with segmentation and estimation; we use the simple gridding procedure described in Section 2. Gridding could be a more difficult task for some images and, in this case, we refer the reader to Angulo and Serra (2003) and Katzer *et al.* (2003) for more sophisticated procedures.

Most methods perform segmentation and estimation separately. For the segmentation, some methods fit circles of fixed (Eisen, 1999) or variable radii (Buhler *et al.*, 2000; Axon Instruments Inc., 2003) to the spots but spots are neither of constant size nor strictly circular. Methods based on histograms (GSI Lumonics, 1999; Li *et al.*, 2005) have the disadvantage of not using any spatial information. Though more adaptive, the seeded region growing method (Adams and Bischof, 1994) and mathematical morphology (Angulo and Serra, 2003) do not correctly segment the doughnut-shaped spots in Figure 1.

Once the segmentation has been completed, most methods estimate the foreground and background intensities separately. First, the foreground is estimated by computing either the mean or median intensities of the corresponding pixels. Then, the background is usually estimated locally for each spot. Some methods use the median intensity value of neighboring pixels (Eisen, 1999; GSI Lumonics, 1999; Imaging Research Inc., 2001; Axon Instruments Inc., 2003), while others form background images (Yang *et al.*, 2002; Brändle *et al.*, 2003) from the data and use these to extract background information for each spot. One difficulty with such approaches is that they can produce background estimates larger than the foreground estimates, which must be incorrect since the background-corrected intensity estimates would be negative. This can be problematic because researchers often use the log transformation (Tusher *et al.*, 2001; Efron *et al.*, 2001; Dudoit *et al.*, 2002).

In this paper, we present a probabilistic model for the analysis of microarray images, where segmentation, estimation, and the associated errors are all modeled simultaneously. The model is more flexible than existing approaches, allowing the proper segmentation of a wide range of shapes and sizes of spot. Estimation is robust and the estimated background-corrected intensities are guaranteed to be positive. In addition, our model allows us to compute quality measures and these can be useful in filtering out low-quality spots, for example.

The paper is organized as follows. Section 2 presents our basic gridding algorithm. Section 3 describes the probabilistic model we use to segment the images and estimate the intensities, an expectation-maximization (EM)/iterated conditional modes (ICM) algorithm used for estimation, and related quality measures. In Section 4, we use an HIV experiment to compare the EM/ICM algorithm to a fully Bayesian implementation of our model and to two computer packages. Finally, in Section 5, we discuss our results, some possible extensions, and the current limitations of our methodology.

2. Gridding

A microarray slide typically consists of several gene blocks, containing 100–1000 spots: those considered here contain 12 blocks of 16 × 40 spots. The blocks are far enough apart that we can analyze them separately (Figure 1). For each block, we first obtain a rough estimate of the position of each spot in the block by defining a rectangular grid, such that each rectangle is of about the same size and each contains

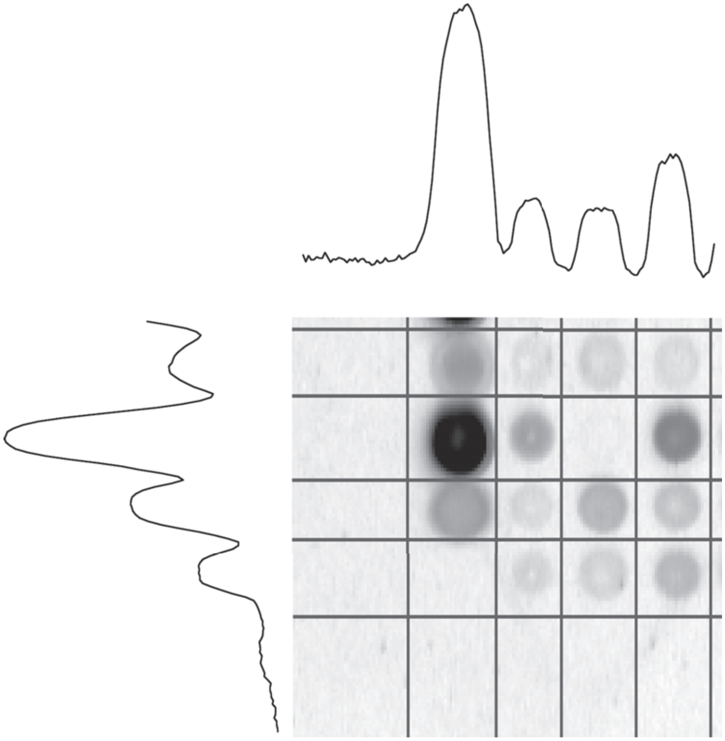


Fig. 2. Illustration of our basic gridding algorithm. The image is projected onto the x-axis and y-axis. The troughs in the two projections define the lines of the grid.

one spot. The method we use is similar to that of Yang *et al.* (2002). Gridding is done independently for each block, by first locating the upper left-hand corner and lower right-hand corner of the block. This only needs to be approximate, and in our case is done manually. Then the image portion representing the block is projected onto the x-axis and the y-axis. The projections form a series of peaks (representing the spots), separated by troughs (region of low intensities between spots). The grid is defined by plotting a line in each trough. Figure 2 illustrates the algorithm.

After gridding, the data take the form $\{y_{rsp}: r = 1, \dots, R;\ s = 1, 2;\ p = 1, \dots, P_r\}$, where y_{rsp} is the intensity of pixel p from the rth rectangle in sample s. Henceforth, by spot r we mean the spot in rectangle r. This is illustrated in Figure 3 with a block of size 2 × 2.

3. IMAGE SEGMENTATION AND INTENSITY ESTIMATION

Segmentation and intensity estimation are carried out concurrently via probabilistic modeling. From now on, $\mathcal{G}a(a, b)$ denotes a gamma distribution with mean a/b and variance a/b^2, $N(a, b)$ a Gaussian distribution with mean a and variance b, and iid means identical and independently distributed. We denote by $(x|y)$ the conditional distribution of x given y.

3.1 *The model*

We assume that the intensity of a pixel in rectangle r and sample s can be described as the sum of a background effect β_{rs}, a hybridization effect ϕ_{rs} if the pixel is classified as foreground, and an additive noise component (3.1). The classification of a given pixel x_{rp} is a random variable, independent of the

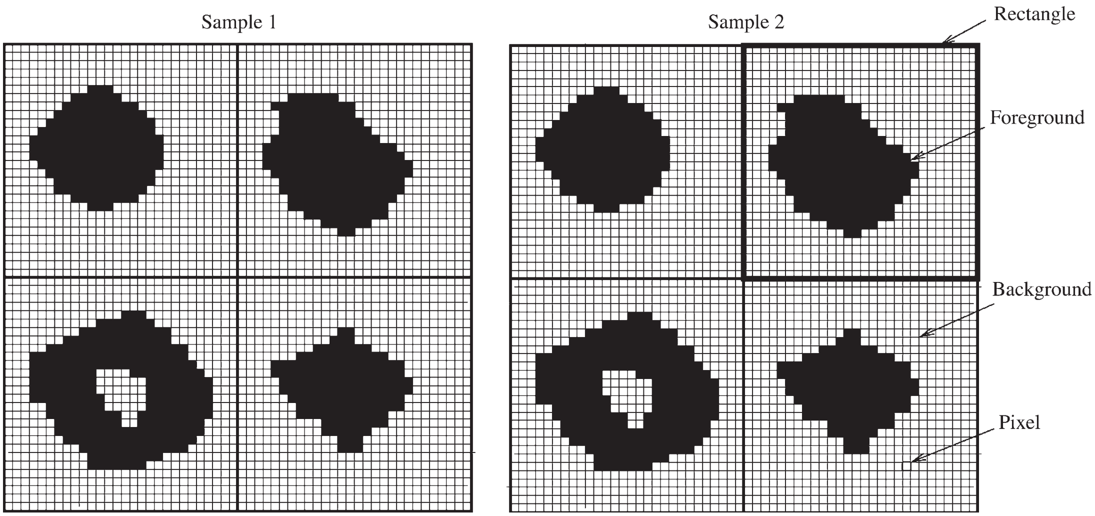


Fig. 3. Example of a block of 2 × 2 genes. We have two images, one for each sample. Each rectangle of the rectangular grid contains one spot. The foreground (resp. background) is represented by the black (resp. white) pixels.

Table 1. *Table of parameters and their descriptions. The subscripts r, s, and p correspond to rectangle, sample, and pixel, respectively*

Parameter	Description
ϕ_{rs}	Hybridization effect
β_{rs}	Background effect
λ_s^β	Background precision parameter
x_{rp}	Pixel classification label
γ	Interaction parameter of the Ising model
λ_{rs}^ϵ	Error precision parameter
ν	Degrees of freedom of the t-distribution

sample index s, taking the value 0 (background) or 1 (foreground). The parameters and their descriptions are summarized in Table 1. We assume the (y_{r1p}, y_{r2p}) are independent conditional on the parameters $(\boldsymbol{\beta}, \boldsymbol{\phi}, \mathbf{x})$ and can be written as,

$$\begin{pmatrix} y_{r1p} \\ y_{r2p} \end{pmatrix} = \begin{pmatrix} \beta_{r1} \\ \beta_{r2} \end{pmatrix} + \begin{pmatrix} \phi_{r1} \\ \phi_{r2} \end{pmatrix} x_{rp} + \begin{pmatrix} \epsilon_{r1p} \\ \epsilon_{r2p} \end{pmatrix} \Big/ \sqrt{w_{rp}}, \tag{3.1}$$

where $\epsilon_{rsp} \overset{\text{iid}}{\sim} N(0, 1/\lambda_{rs}^\epsilon)$ and $w_{rp} \overset{\text{iid}}{\sim} \mathcal{G}a(\nu/2, \nu/2)$ with w_{rp} independent of ϵ_{rsp}. Thus, $\epsilon_{rsp}/\sqrt{w_{rp}}$ follows a bivariate t-distribution with ν degrees of freedom and covariance matrix $\operatorname{diag}(1/\lambda_{r1}^\epsilon, 1/\lambda_{r2}^\epsilon)$. The advantage of this parametrization is that, conditioning on the w_{rp}, the sampling errors are again Gaussian, but with different precisions. The parameter ν is fixed, and we discuss its value in the next section.

We model the background intensity in each rectangle r as a first-order Gaussian intrinsic autoregression (Besag and Kooperberg, 1995) with

$$(\beta_{rs} | \boldsymbol{\beta}_{\partial rs}, \lambda_s^\beta) \sim N\left(\frac{\sum_{r' \in \partial r} \beta_{r's}}{n_r}, \frac{4}{n_r \lambda_s^\beta} \right), \tag{3.2}$$

where ∂r corresponds to the rectangles r' immediately adjacent to r, and $n_r = 2, 3, 4$ is the cardinality of ∂r. The parameters λ_s^{β} are assumed known and we discuss their value in the next section.

The Ising model is often used in low-level image analysis (e.g. Besag, 1986) to encourage neighboring pixels to have the same class. Here, we use a modified symmetric first-order Ising model for the pixel classification label x_{rp} values as follows:

$$(x_{rp}|\mathbf{x}_{r\partial p}, c_r) \propto \exp\left(\gamma \sum_{p' \in \partial p} \mathbf{1}[x_{rp'} = x_{rp}]\right) \mathbf{1}[\| p - c_r \| < 0.5d], \qquad (3.3)$$

where $\mathbf{1}[E]$ is an indicator function equal to 1 if E is true and 0 otherwise, γ is the interaction parameter of the Ising model and ∂p denotes the adjacent pixels to p. We assume that boundary pixels along the perimeter of each rectangle are part of the background and we fix the corresponding x_{rp} values to zero. The number of adjacent pixels for all other x_{rp}s is four. The second indicator function in (3.3) forces the foreground pixels to be contained in a circle of fixed diameter d and center c_r, which reduces the influence of artifacts. We use $d = 22$ pixels; other values may be appropriate for other technologies. The centers c_r are unknown and are estimated with the other parameters; see Section 3.2, where we also discuss the choice of γ, which controls the extent to which neighboring pixels are of the same class.

3.2 *Model fitting by EM/ICM*

ALGORITHM 1 EM/ICM algorithm

```
for $r = 1$ to $R$ do
    Start with initial estimates of $\mathbf{x_r}$, $\boldsymbol{\beta_r}$, $\boldsymbol{\phi_r}$, $\boldsymbol{\lambda_r^{\epsilon}}$ and $\mathbf{w_r}$.
    repeat
        E step
        for $p = 1$ to $P_r$ do
            $w_{rp} = \dfrac{\nu + 2}{\nu + \sum_s \lambda_{rs}^{\epsilon}(y_{rsp} - \beta_{rs} - x_{rp}\phi_{rs})^2}$
        end for
        for $p = 1$ to $P_r$ do
            Update the classification by ICM
            $x_{rp} = \begin{cases} \text{argmax}_x \exp\big(-\frac{1}{2}\sum_s \lambda_{rs}^{\epsilon} w_{rp}(y_{rsp} - \beta_{rs} - x\phi_{rs})^2 & \text{if } \| p - c_r \| < 0.5d \\ \qquad + \gamma \sum_{p' \in \partial p} \mathbf{1}[x_{rp'} = x]\big) & \\ 0 & \text{otherwise} \end{cases}$
        end for
        for $s = 1$ to $2$ do
            Update the background parameter by ICM
            $\beta_{rs} = \dfrac{\lambda_{rs}^{\beta} \sum_{r' \in \partial_r} \beta_{r's} + 4\lambda_{rs}^{\epsilon} \sum_p w_{rp}(y_{rsp} - \phi_{rs}x_{rp})}{n_r \lambda_{rs}^{\beta} + 4\lambda_{rs}^{\epsilon} \sum_p w_{rp}}$
            *Update the foreground parameter*

                $\phi_{rs} = \max\left(0, \frac{\sum_p w_{rp}x_{rp}(y_{rsp} - \beta_{rs})}{\sum_p w_{rp}x_{rp}}\right)$

            *Update the precision parameter*

                $\lambda^{\epsilon}_{rs} = \frac{P_r}{\sum_p w_{rp}(y_{rsp} - \beta_{rs} - x_{rp}\phi_{rs})^2}$

        **end for**
    **until** Convergence
**end for**
```

The EM algorithm (Dempster *et al.*, 1977) can be used for maximum likelihood estimation in multivariate t models (Meng and van Dyk, 1997). Exact computation with Markov random fields, such as in (3.2) and (3.3), is computationally very demanding, but a good approximation can be obtained quickly by the ICM algorithm (Besag, 1986). Here, we use a combination of the EM and the ICM algorithms to fit the model described in Section 3.1. The update for each parameter is given in Algorithm 1, and requires values of λ^{β}_1, λ^{β}_2, ν, and γ. It is possible but computationally intensive to estimate these as part of the algorithm. For example, the degrees of freedom ν can be estimated in the EM framework (Meng and van Dyk, 1997), but each evaluation of the function to be maximized and its gradient involves one sweep through the image.

The parameters λ^{β}_s act as smoothing parameters for each background and we have found the value 0.005 to work well in practice. We set $\nu = 2$, which seems successful in avoiding the influence of bright artifacts during segmentation. Finally, we repeat Algorithm 1 with γ equals 0.2, 0.6, and 0.8, though the exact values are not crucial. This type of strategy is often used in image analysis to avoid fixing the classification of pixels on the basis of unreliable initial estimates (Besag, 1986). In our experience, the above values are satisfactory for a range of data sets generated from the same laboratory. This is supported by our fully Bayesian analysis in Appendix B, which can be used to estimate the parameters on small subsets of the data. It takes about 10 min to fit Algorithm 1 to a single image with 7680 spots using a single Intel Xeon processor at 3.06 GHz. An R package implementing the EM/ICM algorithm will be made freely available for download at http://www.stat.washington.edu/raph/software/.

3.3 *Quality measures*

Our model can be used to derive spot quality measures Q_r (see Appendix A), which we define by

$$Q_r^2 = \frac{\sum_s [(\sum_p \hat{w}_{rp}(1-\hat{x}_{rp}))^{-1} + (\sum_p \hat{w}_{rp}\hat{x}_{rp})^{-1}]/(\hat{\lambda}^{\epsilon}_{rs}(\hat{\phi}_{rs}\log 2)^2)\mathbf{1}(\hat{\phi}_{rs} > 0)}{\sum_s \mathbf{1}(\hat{\phi}_{rs} > 0)}. \tag{3.4}$$

This quantity is defined only if $\hat{\phi}_{r1} > 0$ or $\hat{\phi}_{r2} > 0$, at least 1 pixel is classified as foreground ($\sum_p \hat{x}_p > 0$) and at least 1 pixel is classified as background ($\sum_p(1 - \hat{x}_p) > 0$). For each rectangle, we constrain the segmented region to be contained in a circle of fixed radius and there are always background pixels. If no pixels are classified as foreground, the rectangle is blank and there is no need to compute a quality measure. Similarly, if both $\hat{\phi}_{r1} = 0$ and $\hat{\phi}_{r2} = 0$, we define the rectangle to be blank and set the corresponding x_{rp} values to zero.

For each spot, Q_r is mainly affected by three things:

(1) The number of pixels classified as foreground; a small number of pixels will make the quantity $1/(\sum_p \hat{\omega}_p \hat{x}_p)$ large.
(2) The coefficient of variation in each sample, $(\hat{\lambda}^{\epsilon}_{rs}\hat{\phi}^2_{rs})^{-0.5}$.
(3) The value of the estimated weights $\hat{\omega}_{rp}$. Small weights, which are usually associated with artifacts, will be associated with larger values of Q_r.

Here, we recommend filtering out a spot if its quality measure Q_r is greater than 0.1, which seems to work well in practice.

Algorithm 1 can also lead to hybridization estimates equal to 0. In general, these estimates are associated with low-quality spots and are filtered out using our quality measures. However, they can also correspond to valid spots. Without further information, it is impossible to know if the true expression level for the spot in the corresponding sample is zero or the intensity is below the detection level of the scanner, so we recommend flagging such spots.

4. APPLICATION TO EXPERIMENTAL DATA

In this section, we compare different methods on an HIV experiment in which the expression levels of 7680 cellular RNA transcripts had been assessed in CD4-T-cell lines at time $t = 24$ h after infection with HIV virus type 1. The data set contains 12 HIV-1 genes used as positive controls. Further details are given by van't Wout *et al.* (2003). The raw images are available at http://expression.microslu.washington.edu/expression/index.html. To ease comparisons between methods, we only use the first block of 640 genes, which contains 2 of the 12 HIV control genes. We have applied our method to other blocks and other images and the results were similar.

4.1 *Methods to be compared*

Besides our own EM/ICM and fully Bayesian implementations (as described in Appendix B), we consider two other methods for cDNA microarray image analysis: Spot and Arrayvision. A summary follows:

(1) *Probabilistic approach via EM/ICM*: Segmentation and estimation of the background-corrected intensities are carried out simultaneously. To display the segmented region, we use the estimated x_{rp}s. To estimate the hybridization effects (background-corrected intensities), we use $\hat{\boldsymbol{\phi}}$ and estimate the log-ratio for each spot by $\log_2(\hat{\phi}_{r1}/\hat{\phi}_{r2})$.
(2) *Fully Bayesian approach*: Using the approach described in Appendix B, we obtain a sample from the overall posterior distribution. We estimate the segmented region and hybridization parameters by the posterior means of x_{rp} and ϕ_{rs}.
(3) *Spot*: Segmentation is done using a seeded growing region algorithm (Adams and Bischof, 1994). For a given spot, the foreground intensity is computed as the median intensity of the pixels within the spot. The background intensity is calculated using morphological opening (Serra, 1982; Soille, 1999). The nonlinear filter is applied to the original images using a square structuring element with sides of length at least twice as large as the spot separation distance. The background image is estimated by first replacing each pixel by the minimum local intensity in the square region and then performing a similar operation on the resulting image using the local maximum. If S_i denotes the square centered at pixel i, the background intensity z_i of pixel i is given by $z_i = \max_{j \in S_i} y'_j$, where $y'_j = \min_{k \in S_j} y_k$ with y denoting the original pixel values. This operation removes all the spots and produces an image that is an estimate of the background image. For individual spots, background is estimated by sampling this background image at the nominal center of the spot.

(4) *Arrayvision*: Arrayvision (Imaging Research Inc., 2001) is a commercial software developed for the quantification of gene expression arrays in which spots are allowed to vary in shape and size; exact details are not available in the public domain. Arrayvision offers several methods for background estimation and automatically subtracts it from the foreground intensity value. In the examples explored here, we use the background estimates from circular regions around each spot. Intensities are estimated by the median of all the pixels in each region.

4.2 *Segmentation*

Figure 4 shows the segmentation results using the four methods described in Section 4.1 for a portion of one of the HIV images. Spot and Arrayvision are more flexible than fixed circle segmentation algorithms in allowing noncircular shapes but they still fail to properly segment doughnut-shaped spots. In addition, they both output almost circular regions for the spots that do not show any hybridization on the raw images. On the other hand, our approach with EM/ICM is flexible in allowing all sorts of shapes including doughnuts and is robust to artifacts because of the t-distributed errors. This is not the case with Spot: see, for example, row 3 and column 14 of Figure 4(c). The results from the fully Bayesian approach are comparable to the EM/ICM implementation except for a few spots. However, these spots have quality measures greater than 0.1 (spots colored in gray) and are filtered out by the EM/ICM implementation.

4.3 *Estimation*

The log-ratio estimates from the EM/ICM implementation are almost identical to the ones from the fully Bayesian approach above an overall intensity of about five, except for two spots: one corresponding to an artifact and the other to one of the HIV genes with estimated log-ratio equal to infinity, which cannot be displayed. Below this, estimates can be quite different. The intensities from the EM/ICM implementation can be exactly zero, whereas the fully Bayesian estimates are strictly positive. As a consequence, some of the estimates cannot be displayed in the EM/ICM case. However, these estimates would be associated with large measures of uncertainty computed with the fully Bayesian approach, and would likely be discarded (Gottardo, 2005).

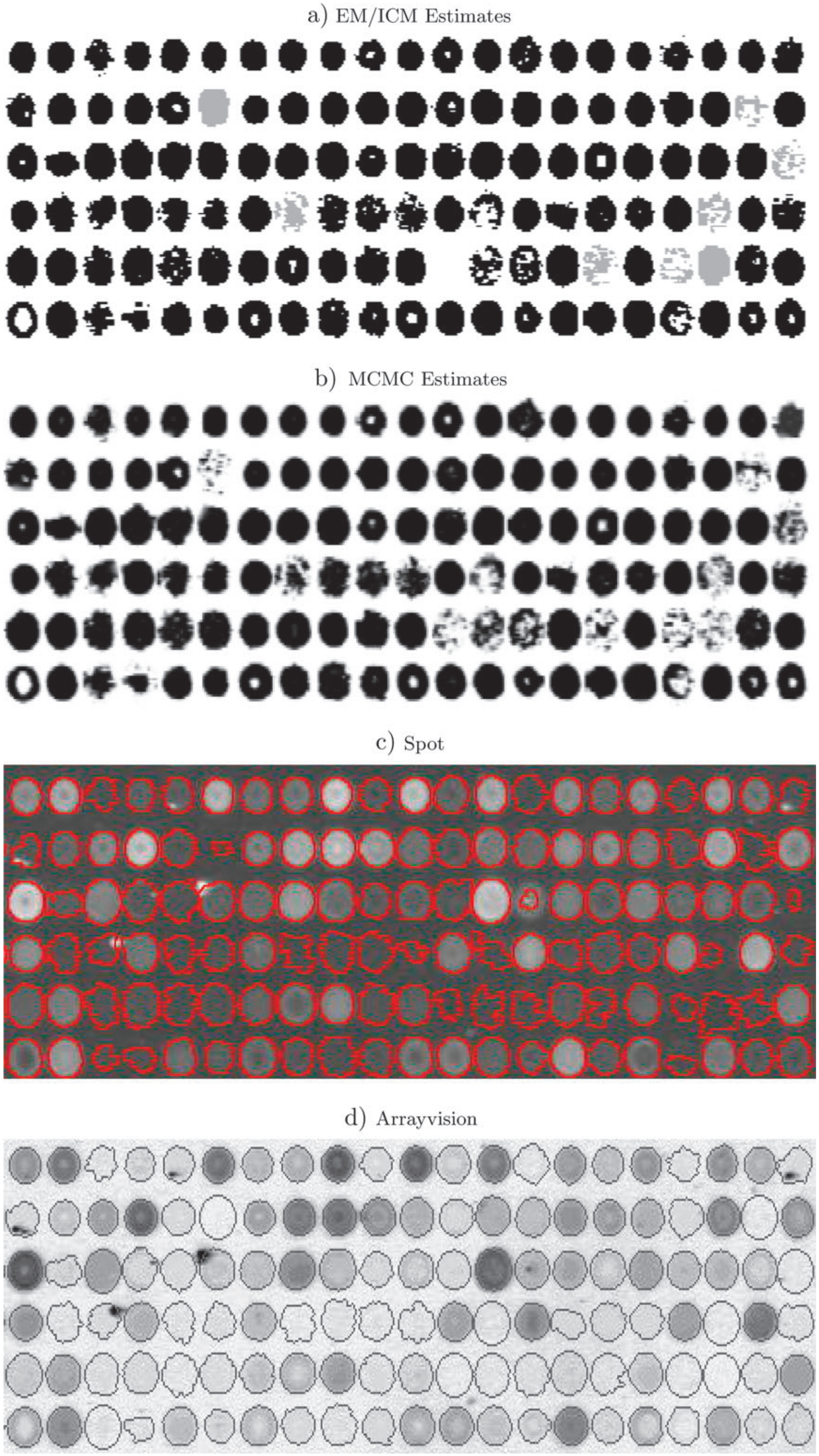
Figure 5 shows the log-ratio estimates as a function of the overall intensity. For the HIV genes, the estimated ratios should be infinite since the true intensity for one of the channels is exactly zero but not the other. For EM/ICM, the two values are ∞ (not displayed) and 9.93; for the fully Bayesian approach, 17.95 and 9.88; and for Spot, 10.48 and 9.26, suggesting a downward bias. The values are on the log scale and the difference between estimates would be even larger on the natural scale. Negative estimates occurred for several genes using Arrayvision, including one of the HIV genes.

Log-ratio estimates tend to be more variable at low intensity for all the methods except Spot (Figure 5). In Spot, the background estimation is based on morphological opening and usually leads to smaller estimates than competing methods. Because the size of the running rectangle is chosen to be quite large, the estimates also tend to be constant over larger regions (Figure 6). The associated background-subtracted intensity estimates are larger, perhaps overestimated, diminishing the number of estimated intensities close to zero. The background estimates from our model and Arrayvision are more comparable.

Figure 5(a) also shows the spots filtered out by our quality measures. Most of these have low overall intensity. The low-quality spot with largest overall intensity corresponds to a bright artifact.

5. Discussion

We have introduced a model for the analysis of microarray images, combining both segmentation and intensity estimation. Our model is robust, with t-distributed errors, and flexible, allowing the segmentation of all sorts of spot shapes. We claim that the segmentation results from our model are superior to two



a) EM/ICM Estimates

b) MCMC Estimates

c) Spot

d) Arrayvision

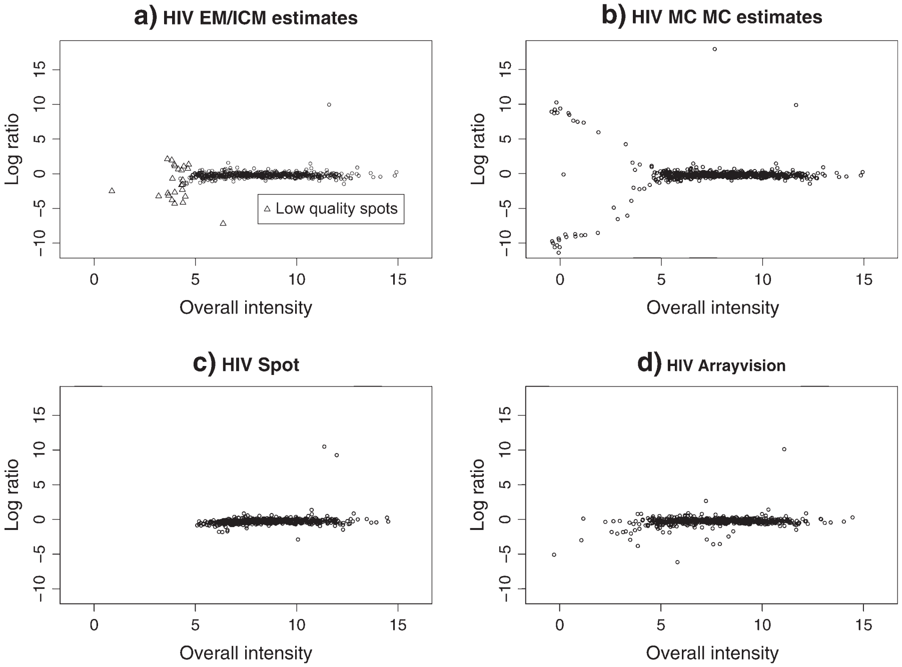


Fig. 5. Log-ratio estimates as a function of the overall intensity. All ratio estimates but the ones from Spot tend to be more variable at low intensity. The two genes with the largest ratios correspond to two HIV control genes. Their true ratio should be arbitrarily large. Some of the intensity estimates from Arrayvision, e.g. one of the HIV genes, were negative and cannot be displayed on a log scale. Some of the estimates from the EM/ICM fitting method are equal to 0 and cannot be displayed. This is a case for one of the HIV genes for which the log-ratio is equal to infinity.

competing methods. In addition to the segmented regions and point estimates, we provide spot quality measures that can be used for filtering unreliable spots.

The difference between log-ratio estimates obtained from Spot, Arrayvision, and our model is quite large, and suggests that background estimation is a significant factor. This is consistent with the results in Yang *et al.* (2002). To accurately estimate the hybridization intensity in each channel, an estimate of the background is usually subtracted from the foreground intensity. Using our model, negative differences cannot occur as we model the hybridization effects directly. Even though negative estimates are possible with Spot, they rarely occur in practice. This is mainly due to the nature of the morphological opening, which provides smaller background estimates. The variability of the estimated log ratios from Spot is also reduced, but this does not necessarily mean that it performs better. To quote the authors (Yang *et al.*, 2002): "… the variability of replicate log ratios is not in itself a useful measure of performance, as smaller variability can be achieved simply by using lower, or darker background estimates." There is a need for test data sets where the true log ratios of many of the genes are known. This would allow us to evaluate the different methods in terms of accuracy in addition to variability. Of course, such data sets would require

Fig. 4. Comparison of the different segmentation methods on a piece of the HIV images. For Spot and Arrayvision, the segmented region contours are drawn over the combined raw images. For the EM/ICM results, spots with quality measures greater than 0.1 are colored in gray. Both the EM/ICM and fully Bayesian implementations of our model properly segment doughnut-shaped spots.

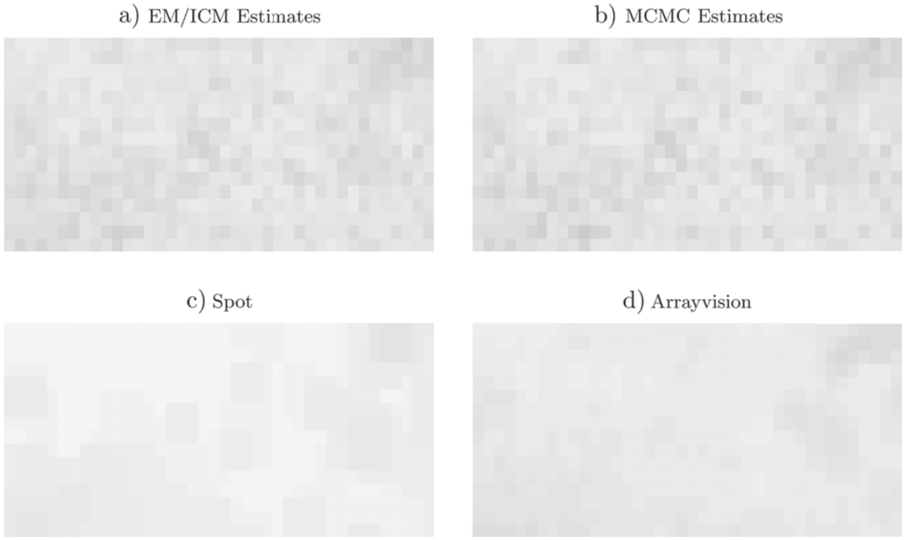


Fig. 6. Image plot of the background estimates for the four methods compared on the HIV data. All the estimates show spatial variation of the background intensities. The estimates from Spot are the lowest and are constant over larger regions. The EM/ICM and Bayesian estimates are almost identical.

many laboratory experiments and so far as we are aware none are currently in the public domain. Using two HIV control genes, for which the log ratios should be arbitrarily large, we showed that Spot had the largest bias.

One possible cause of the increased variability at low intensity is that the detection range of the scanner is set too high. As a consequence, the low-intensity spots are not visible in the raw images, leading to problematic segmentation. On the other hand, if the detection range is set too low, the high-intensity spots become saturated. Although one can model the truncation (Tadesse *et al.*, 2003), this severely increases the complexity of the model and does not seem worthwhile. In practice, the detection range is high enough so that high intensities are not affected. Therefore, we prefer to flag and/or discard unreliable spots.

According to the Bayesian paradigm, it might seem natural to include other levels of the analysis in our hierarchical model. In fact, this was a goal in Gottardo (2005). For example, we have tried modifying our model to incorporate replicates but we found no improvement in performance versus combining the replicates after the image analysis step. It appears helpful to flag and perhaps filter out low-quality spots before going to the next stage, which may become more difficult in a more complex model. In particular, if a spot replicate is deficient, it often degrades the quality of the estimates for the associated gene, and potentially other genes. If instead, one analyzes each image separately, the deficient spots can be removed before combining the estimates.

Finally, we have compared our EM/ICM implementation to a fully Bayesian one and found little improvement in terms of intensity estimates and segmentation. One advantage of the fully Bayesian approach is that it permits more realistic measures of uncertainty by combining information from both segmentation and estimation (Gottardo, 2005). The additional computational price does not seem worthwhile at the present time.

ACKNOWLEDGMENTS

The authors thank Roger Bumgarner, Quhua Li, Chris Fraley, and Adrian Raftery for helpful discussions. Gottardo's research was supported by the National Institutes of Health Grant 8 R01 EB002137-02. The authors also thank two anonymous referees and the associate editor for suggestions that clearly improved an earlier draft of the paper.

APPENDIX A

Quality measures

We first assume a simplified model, in which there is no spatial effect for the background (i.e. $\lambda_s^\beta = 0$). Conditional on $\mathbf{w}$ and $\mathbf{x}$, the Fisher information matrix for each parameter vector $\boldsymbol{\theta}_{rs} = (\beta_{rs}, \phi_{rs}, \lambda_{rs}^\epsilon)$ is given by

$$I_{rs}^{\boldsymbol{\theta}} = \begin{pmatrix} \lambda_{rs}^\epsilon \sum_p w_{rp} & \lambda_{rs}^\epsilon \sum_p w_{rp} x_{rp} & 0 \\ \lambda_{rs}^\epsilon \sum_p w_{rp} x_{rp} & \lambda_{rs}^\epsilon \sum_p w_{rp} x_{rp} & 0 \\ 0 & 0 & 0.5 P_r (\lambda_{rs}^\epsilon)^{-2} \end{pmatrix},$$

and therefore an estimate of the asymptotic variance for $\tilde{\phi}_{rs}$, the maximum likelihood estimate in this simplified model, is $\tau(w_{rp}, x_{rp}, \tilde{\lambda}_{rs}^\epsilon) \equiv [(\sum_p w_{rp}(1 - x_{rp}))^{-1} + (\sum_p w_{rp} x_{rp})^{-1}]/\tilde{\lambda}_{rs}^\epsilon$. Researchers usually focus on the log (base 2) transformation, and using the delta method we obtain $\tau'(w_{rp}, x_{rp}, \tilde{\lambda}_{rs}^\epsilon, \tilde{\phi}_{rs}) = \tau/(\tilde{\phi}_{rs} \log 2)^2$ as an estimate of the asymptotic variance of $\log_2(\tilde{\phi}_{rs})$. Although this is not an estimate of the asymptotic variance for $\log_2(\hat{\phi}_{rs})$ computed with Algorithm 1, we use it to derive a quality measure. For each spot, we define the quality measure Q_r by

$$Q_r^2 = \frac{\sum_s \tau'(\hat{w}_{rp}, \hat{x}_{rp}, \hat{\lambda}_{rs}^\epsilon, \hat{\phi}_{rs}) \mathbf{1}(\hat{\phi}_{rs} > 0)}{\sum_s \mathbf{1}(\hat{\phi}_{rs} > 0)}, \tag{A.1}$$

which can be seen as an average of the asymptotic variance estimates τ' using the parameter estimates obtained with Algorithm 1.

APPENDIX B

A fully Bayesian approach

The model introduced in Section 3.1 can be extended to a fully Bayesian approach. All equations introduced in Section 3.1 remain the same but we add priors for some of the unknown parameters.

The hybridization effect of the spot included in rectangle r of sample s, denoted by ϕ_{rs}, is modeled as a random effect with log Normal distribution $(\phi_{rs} | \xi_s^\phi, \lambda_s^\phi) \sim \log\text{Normal}(\xi_s^\phi, 1/\lambda_s^\phi)$, where $\xi_\phi \sim N(0, 100)$ and $\lambda_\phi \sim \mathcal{G}a(1, 0.005)$. We assume that the error precisions λ_{rs}^ϵ arise from a common gamma distribution, $\mathcal{G}a((a_s^\epsilon)^2/b_s^\epsilon, a_s^\epsilon/b_s^\epsilon)$, where $a_s^\epsilon \sim U_{[0,10]}$ and $b_s^\epsilon \sim U_{[0,50]}$. The prior for the centers of the circles c_r used in the Ising prior is taken to be uniform over all possible values such that the entire disk of center c_r is contained in the rth rectangle of the grid.

Finally, the prior for the degrees of freedom ν is uniform on the set $\{1, 2, \ldots, 10, 20, \ldots, 100\}$. These priors are vague but proper and have little influence on the posterior distribution because the parameters are shared across pixels, and so there is ample information in the data. Realizations are generated from the posterior distribution via MCMC algorithms (Smith and Roberts, 1993; Besag *et al.*, 1995); further details can be found in Gottardo (2005). The posterior mode of the degrees of freedom for the t-distribution ν is 2, indicating that the sampling errors are heavier tailed than the Gaussian distribution. The estimates for λ_1^β and λ_2^β are of the order 10^{-3}, which is close to the values used in Section 3.2.

References

Adams, R. and Bischof, L. (1994). Seeded region growing. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **16**, 641–647.

Angulo, J. and Serra, J. (2003). Automatic analysis of DNA microarray images using mathematical morphology. *Bioinformatics* **19**, 553–562.

Axon Instruments Inc. (2003). *Genepix 5.0, User's Guide*. Axon Instruments, Inc. (http://www.axon.com).

Besag, J. (1986). On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society, Series B, Methodological* **48**, 259–279.

Besag, J. E., Green, P., Higdon, D. and Mengersen, K. (1995). Bayesian computation and stochastic systems. *Statistical Science* **10**, 3–66.

Besag, J. and Kooperberg, C. (1995). On conditional and intrinsic autoregressions. *Biometrika* **82**, 733–746.

Brändle, N., Bischof, H. and Lapp, H. (2003). Robust DNA microarray image analysis. *Machine Vision and Applications* **15**, 11–28.

Buhler, J., Ideker, T. and Haynor, D. (2000). Dapple: improved techniques for finding spots on DNA microarrays. *Technical Report UWTR 2000-08-05*. Computer Science Department, University of Washington, Seattle, WA.

Chen, Y., Dougherty, E. R. and Bittner, M. L. (1997). Ratio-based decisions and the quantitative analysis of cDNA microarray images. *Journal of Biomedical Optics* **2**, 364–374.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B: Methodological* **39**, 1–22.

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**, 111–139.

Efron, B., Tibshirani, R., Storey, J. D. and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *Journal of the American Statistical Association* **96**, 1151–1160.

Eisen, M. (1999). *Scanalyze, User Manual*. Stanford, CA: Stanford University.

Eisen, M., Spellman, P., Brown, P. and Botsein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America* **95**, 14863–14868.

Golub, T., Slonim, D., Tomayo, P., Huard, C., Gaasenbeek, M., Merisov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M. A., *et al.* (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* **286**, 531–537.

Gottardo, R. (2005). Bayesian robust analysis of gene expression data. *Ph.D. Thesis*, University of Washington, Seattle, WA.

Gottardo, R., Pannucci, J. A., Kuske, C. R. and Brettin, T. (2003). Statistical analysis of microarray data: a Bayesian approach. *Biostatistics* **4**, 597–620.

GSI Lumonics (1999). *Quantarray Analysis Software, Operator's Manual*. GSI Lumonics (http://www.gsilumonics.com).

Hedge, P., Qi, R., Abernathy, K., Gay, C., Dharap, S., Gaspard, R., Earle-Hugues, J., Snesrud, E., Lee, N. and Quackenbush, J. (2000). A concise guide to cDNA microarray analysis. *Biotechniques* **29**, 548–562.

Imaging Research Inc. (2001). *Arrayvision Application Note: Spot Segmentation*. Imaging Research Inc. (http://www.imagingresearch.com).

Katzer, M., Kummert, F. and Sagerer, G. (2003). A Markov random field model of microarray gridding. In *Proceedings of the 2003 ACM Symposium on Applied Computing*. New York: ACM Press, pp. 72–77.

LI, Q., FRALEY, C., BUMGARNER, R. E., YEUNG, K. Y. AND RAFTERY, A. E. (2005). Donuts, scratches and blanks: robust model-based segmentation of microarray images. *Bioinformatics* **21**, 2875–2882.

MENG, X.-L. AND VAN DYK, D. (1997). The EM algorithm—an old folk-song sung to a fast new tune. *Journal of the Royal Statistical Society, Series B: Methodological* **59**, 511–540. (Discussion: pp. 541–567.)

NEWTON, M. A., KENDZIORSKI, C. M., RICHMOND, C. S., BLATTNER, F. R. AND TSUI, K. W. (2001). On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology* **8**, 37–52.

SCHENA, M., SHALON, D., DAVIS, R. W. AND BROWN, P. (1995). Quantitative monitoring of gene-expression patterns with a complementary-DNA microarray. *Science* **270**, 467–470.

SERRA, J. (1982). *Image Analysis and Mathematical Morphology*. London: Academic Press.

SMITH, A. F. M. AND ROBERTS, G. O. (1993). Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society, Series B, Methodological* **55**, 3–23.

SOILLE, P. (1999). *Morphological Image Analysis: Principles and Applications*. New York: Springer.

TADESSE, M., IBRAHIM, J. AND MUTTER, G. (2003). Identification of differentially expressed genes in high-density oligonucleotide arrays accounting for the quantification limits of the technology. *Biometrics* **59**, 542–554.

TAMAYO, P., SLONIM, D., MESIROV, J., ZHU, Q., KITAREEWAN, S., DMITROVSKY, E., LANDER, E. S. AND GOLUB, T. R. (1999). Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proceedings of the National Academy of Sciences of the United States of America* **96**, 2907–2912.

TIBSHIRANI, R., HASTIE, T., NARASIMHAN, B. AND CHU, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 6567–6572.

TUSHER, V., TIBSHIRANI, R. AND CHU, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America* **98**, 5116–5121.

VAN'T WOUT, A. B., LEHRMA, G. K., MIKHEEVA, S. A., O'KEEFFE, G. C., KATZE, M. G., BUMGARNER, R. E., GEISS, G. K. AND MULLINS, J. I. (2003). Cellular gene expression upon human immunodeficiency virus type 1 infection of CD4^{+}-T-cell lines. *Journal of Virology* **77**, 1392–1402.

YANG, Y. H., BUCKLEY, M. J., DUDOIT, S. AND SPEED, T. P. (2002). Comparison of methods for image analysis on cDNA microarray data. *Journal of Computational and Graphical Statistics* **11**, 108–136.

YEUNG, K. Y., FRALEY, C., MURUA, A., RAFTERY, A. E. AND RUZZO, W. L. (2001). Model-based clustering and data transformations for gene expression data. *Bioinformatics* **17**, 977–987.

[*Received November 12, 2004; revised July 6, 2005; accepted for publication July 11, 2005*]