

Methodology article

Open Access

## An algorithm for automatic evaluation of the spot quality in two-color DNA microarray experiments

Eugene Novikov\* and Emmanuel Barillot

Address: Service Bioinformatique, Institut Curie, 26 Rue d'Ulm, 75248 Paris Cedex 05, France.

Email: Eugene Novikov\* - Eugene.Novikov@curie.fr; Emmanuel Barillot - Emmanuel.Barillot@curie.fr

\* Corresponding author

Published: 09 December 2005

Received: 17 June 2005

BMC Bioinformatics 2005, 6:293 doi:10.1186/1471-2105-6-293

Accepted: 09 December 2005

This article is available from: <http://www.biomedcentral.com/1471-2105/6/293>

© 2005 Novikov and Barillot; licensee BioMed Central Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

### Abstract

**Background:** Although DNA microarray technologies are very powerful for the simultaneous quantitative characterization of thousands of genes, the quality of the obtained experimental data is often far from ideal. The measured microarrays images represent a regular collection of spots, and the intensity of light at each spot is proportional to the DNA copy number or to the expression level of the gene whose DNA clone is spotted. Spot quality control is an essential part of microarray image analysis, which must be carried out at the level of individual spot identification. The problem is difficult to formalize due to the diversity of instrumental and biological factors that can influence the result.

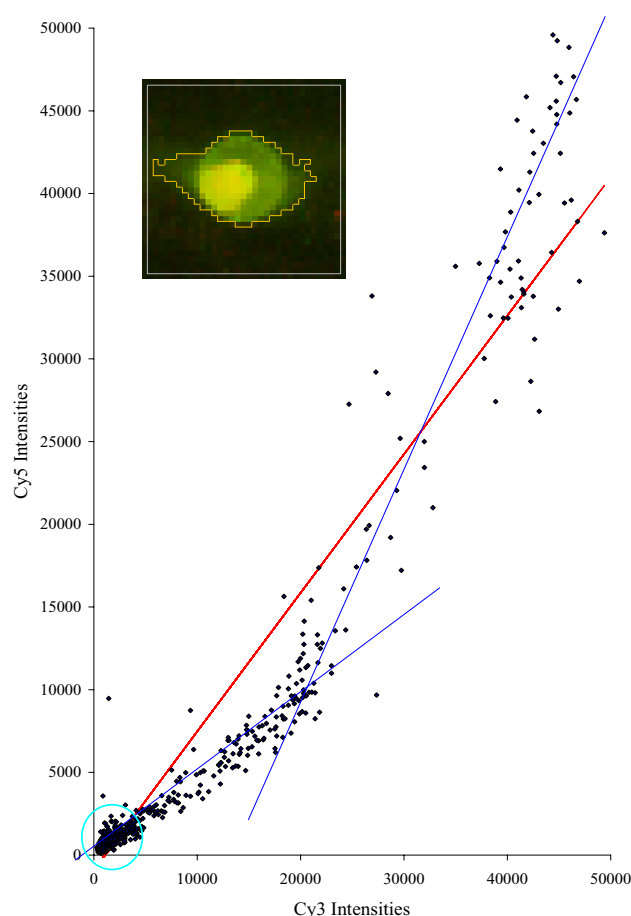
**Results:** For each spot we estimate the ratio of measured fluorescence intensities revealing differential gene expression or change in DNA copy numbers between the test and control samples. We also define a set of quality characteristics and a model for combining these characteristics into an overall spot quality value. We have developed a training procedure to evaluate the contribution of each individual characteristic in the overall quality. This procedure uses information available from replicated spots, located in the same array or over a set of replicated arrays. It is assumed that unspoiled replicated spots must have very close ratios, whereas poor spots yield greater diversity in the obtained ratio estimates.

**Conclusion:** The developed procedure provides an automatic tool to quantify spot quality and to identify different types of spot deficiency occurring in DNA microarray technology. Quality values assigned to each spot can be used either to eliminate spots or to weight contribution of each ratio estimate in follow-up analysis procedures.

### Background

In comparative DNA microarray experiments compared test and control samples are labeled with different fluorescent dyes (typically the red-fluorescent Cy5 and the green-fluorescent Cy3), mixed up and co-hybridized with the DNA clones regularly spotted on the microarray. The array is scanned at a high spatial resolution at the corresponding fluorescent wavelengths, and the fluorescence intensi-

ties are recorded in two color channels (Cy5 and Cy3) for each pixel. The ratio of the measured intensities (Cy5/Cy3) for each microarray spot reveals either differential gene expression (cDNA technology [1]) or change in DNA copy numbers (comparative genome hybridization (CGH) technology [2]) between the test and control samples for the corresponding gene. Each ratio estimate



**Figure 1**  
**Example of a spot with a low DW parameter (0.45).**  
 Although the coefficient of determination of the linear regression plot (red line) is relatively high (0.92), it is obvious that the linear regression model is not appropriate in this case. It is possible that there are contributions (blue lines) from two different species occurring within the given spot, leading to two different Cy5/Cy3 ratios. The background pixels are grouped near the origin of the linear regression plot (cyan circle).

should be accompanied by some measure of quality demonstrating the level confidence in the obtained ratios.

The main components of the microarray image analysis pipeline for spots include localization, quantification and quality control. Among these, quality control is the least formalized and least developed. To determine spot quality we need to have a clear definition of a good spot, or a list of all possible distortions that may spoil the spot. The diversity of instrumental platforms and instrumental and biological factors that may influence the result makes formalization difficult and unlikely to be universal.

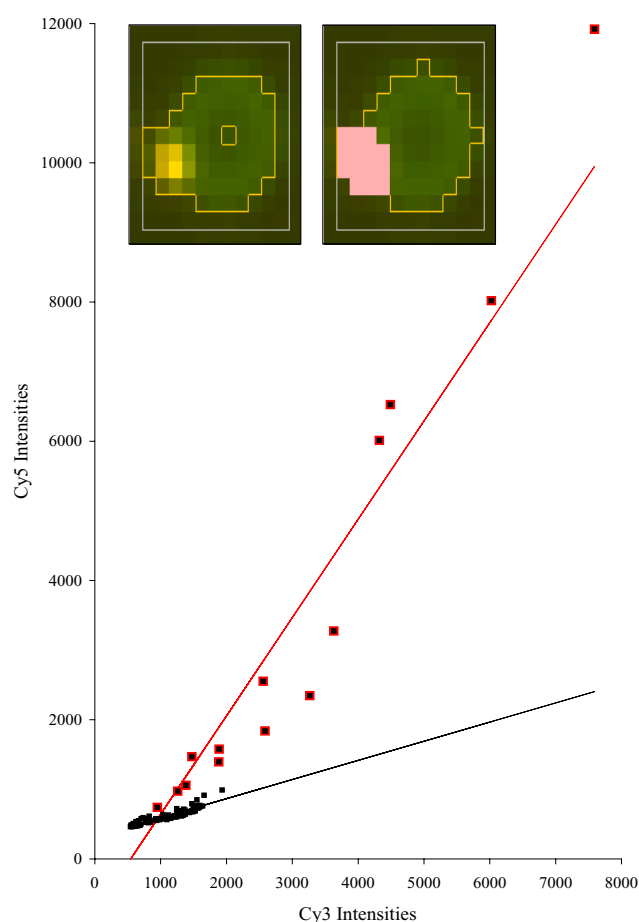
In this paper, we consider the problem of quantifying spot quality in comparative DNA microarray experiments. Several attempts have been made to approach the problem [3-7]. Generally a number of parameters characterizing the spot, such as signal-to-noise ratio, size, circularity, etc., are introduced. These parameters have to be combined into an overall quality value to be used as a confidence level in the follow-up analysis. There are different methods for deriving such a parameter. For example, in two studies [5,6], it was assumed that individual quality scores contribute equivalently to the composite quality score. This may not be true, depending on the instrumental setup and experimental design. Therefore we need an approach that allows us to evaluate the weights that control the input of each of the marginal quality characteristics into the overall score. For that, different training procedures, in which the user classified a set of representative spots into three (accepted, rejected or intermediate spots) [3] or four (bad, close to bad, close to good or good spots) [7] groups, were proposed. This requires an expert to evaluate at least a couple of hundred spots to achieve a good approximation, which is a difficult and time-consuming task.

Here, we develop an automatic training procedure to evaluate the contribution (or weight) of each marginal quality characteristic into the overall quality score, together with an original set of quality characteristics and a model that maps this set into an overall quality value. This procedure is based on information from replicated spots, located on the same array or over a set of replicated arrays, and assumes that unspoiled replicated spots must have very close intensity ratios, whereas poor spots yield greater diversity in the obtained ratio estimates. The obtained weights can then be used to establish a critical limit for each quality characteristic, such that if a spot's characteristic exceeds its critical limit, the spot can be declared a "bad" spot.

We demonstrate the applicability of the developed algorithms using simulated artificial images and experimental images of different array designs used within our Institute and CGH images obtained from the UCSF Cancer Center.

## Results

We assume that the spots are identified and well localized [8], (Novikov E, Barillot E: unpublished data)] in squares (called spot cells). This involves: (i) identifying the position of each spot on the array to associate it with the spotted clone; and (ii) establishing the borders between the neighboring spots to allow further independent data processing (extracting quantitative information) for each spot. Visually this results in the generation of a grid covering the microarray image.



**Figure 2**  
**Example of a spot with contamination.** The filtering procedure removes aberrant pixels (dots with the red contours), improving the Cy5/Cy3 ratio estimation and increasing the coefficient of determination of the linear regression plot. Before filtering (red line):  $CD = 0.86$ ; linear regression ratio is 1.41; segmentation ratio is 0.712. After filtering (black line):  $CD = 0.91$ ; linear regression ratio is 0.275; segmentation ratio is 0.273. However, larger amounts of aberrant pixels may result in a less reliable estimation.

### Ratio estimation

We have implemented [8,9] two approaches for calculating the Cy5/Cy3 ratio for the spot: (i) based on spot segmentation and (ii) based on linear regression.

### Spot segmentation

In the spot segmentation approach a direct arithmetic ratio of the background-corrected fluorescence intensity estimates in the two color channels is evaluated. This approach requires the identification of both the foreground – the measured spot – and the background – typically the level of non-specific hybridization. The segmentation procedure, in our implementation, is based

on the *k*-means adaptive pixel-clustering algorithm [10] modified to explicitly take into consideration the geometrical constraints on the spots and to improve identification of the background areas for the spots with smooth edges.

### Linear regression estimation

In the linear regression approach, the ratio is represented as the slope of the linear regression fit of the pixel intensities in two color channels (Figs. 1 and 2). As measured fluorescence intensities are statistically distorted in both color channels, orthogonal regression [11,12] is used to estimate the slope. In this method spot segmentation is unnecessary, as background pixels are concentrated at the origin of the linear regression plot and do not influence the slope of the regression line (Fig. 1). However, outlier or aberrant pixels within the spot cells, even in small numbers, can strongly influence the regression line, thus biasing the ratio. We have improved [8,9] the linear regression approach by developing a statistical filtering procedure that systematically searches and removes aberrant or outlier pixels.

Briefly, this procedure can be outlined as follows. Suspicious pixels are examined by evaluating the quality of the linear regression fit with and without the suspicious pixel. We quantify the fit quality by the residual variance,  $s^2$  [13]. The smaller  $s^2$  is, the closer the linear regression line is to the experimental data. The ratio of the  $s^2$  values is calculated for the fit with the tested pixel and for the fit without. If this ratio is larger than a critical value of the *F*-distribution at a user-defined confidence level, the pixel will be marked as aberrant. We select pixels with the highest intensity in either of two channels first and then select pixels having the largest deviation from the fitted regression line. To take into account the fact that the distortions caused by pixels from the top of the intensity scale and by pixels lying off of the linear regression line, may be different, we apply different confidence levels for the *F*-statistics for these pixels.

For the high-intensity pixels we also perform another test to determine how far their intensities are from the averaged intensity of the other pixels within the spot cell. This detects pixels, far away from the other pixels, that do not distort the linear regression line. Although these pixels may not change the Cy5/Cy3 ratio, they could be considered as aberrant pixels, as we expect to see an almost continuous distribution of pixels intensity.

The procedure performs iteratively until no more aberrant pixels are detected. An example of the outlier detection is presented in Fig. 2. Note that the regression approach is capable of detecting contamination pixels that are geometrically inseparable from the spot. Therefore, the devel-

oped procedure can be considered not only as a procedure for correcting ratio recovery, but also as a procedure to repair the spot and to improve the quality of experimental material. It requires, however, that the contamination clearly deviates from the straight regression line, which is defined by the majority of "good" pixels from the spot.

Although this procedure gives a high level of confidence in the linear regression ratio estimates, we still apply spot segmentation, because linear regression estimates may be biased when there is a high level of statistical noise (low correlation between the Cy3 and Cy5 color channels). However, after removing aberrant pixels, the segmentation algorithm also gives more robust estimates, and there is a greater agreement in the ratio values obtained with both methods.

As well as the Cy5/Cy3 ratio estimates, each of these approaches generates a series of parameters that can be used to evaluate the spot quality.

#### Quality characteristics

We define a set of quality parameters, characterizing different features of the spot. These parameters are scaled between 0 (bad spot) and 1 (good spot) to facilitate further quality analysis.

First, we use the characteristics from the linear regression approach. The **coefficient of determination** ( $CD = r^2$ , where  $r$  is the correlation coefficient) of linear regression signifies the degree of linear relationship between the intensities in the Cy3 and Cy5 channels. High values of  $CD$  (approaching 1) are expected for good spots. Low values suggest either relatively bright but non-correlated contamination, or strong statistical noise normally characterizing low-level (or missing) spots. Although the removal of aberrant pixels increases the  $CD$  of linear regression, it can still be low for noisy spots. These spots must be either flagged out or assigned a lower quality value. This parameter takes the range [0;1]:  $q_1(CD) = CD$ .

The **Durbin-Watson statistic** ( $DWS$ ) [14] evaluates the presence of the first-order autocorrelation in the residuals of the linear regression fit. It ranges from 0 to 4, 0 being a positive correlation and 4 being a negative correlation. A  $DWS$  value close to two indicates that the residuals are uncorrelated and the model is appropriate. Large deviations from two, resulting from systematic patterns in the residuals plot suggest that the spot cannot be modeled in terms of a simple linear regression. Low  $DWS$  values typically imply strong contamination that was not removed by filtering (Fig. 1). The Durbin-Watson quality parameter ( $q_2 \in [0;1]$ ) is obtained from the  $DWS$  value by the following transformation:  $q_2(DWS) = 1 - |DWS - 2|/2$ .

One more indicator of the quality directly available from the linear regression is the number ( $N$ ) of the aberrant pixels flagged out by the filtering procedure. Although aberrant pixels can be found everywhere within the spot cell, we count only those pixels within the spot contours. This value can be used as a quantitative measure of **spot contamination**. Small numbers of aberrant pixels do not influence intensity ratio estimates, whereas the removal of large numbers of pixels from the spot may indicate inconsistency with the linear regression model (Fig. 2). We scale  $N$  to fit the range [0;1] as:  $q_3(N) = 1 - N/S$ , where  $S$  is the number of pixels within the spot contour.

We now define three quality parameters from the analysis of the contoured spot. The **diameter** of the spot is calculated as  $D = 2(S/\pi)^{1/2}$ . As the true value for the spot diameter may be difficult to establish, we use a typical value taken as the median diameter over all spots on the array. Spots with exceptionally small diameters should normally be penalized. We define the diameter quality parameter as  $q_4(D) = \exp\{D-T\}$ , if  $D < T$  and  $q_4(D) = 1$ , if  $D > T$ , where  $T$  is the typical diameter.

The **geometrical symmetry** parameter measures deviation of the contoured spot from the ideal circle. The center and the diameter of the ideal circle correspond to the center and the diameter of the real spot. We divide both the real spot and the ideal circle into eight segments (pie slices defined as  $[k\pi/4; (k+1)\pi/4]$ ,  $k = 0, \dots, 7$ ) and we count the number of pixels belonging to the spot ( $N_{si}$ ,  $i = 1, \dots, 8$ ) and to the circle ( $N_{ci}$ ,  $i = 1, \dots, 8$ ) for each segment. The sum of

the absolute relative differences  $GS = \sum_{i=1}^8 |N_{si} - N_{ci}| / N_{ci}$

is then taken as an indicator of quality. For ideal circular spots  $GS$  should approach 0, whereas highly deformed (un-circular) spots can be recognized by high  $GS$  values (Fig. 3A). We transform the  $GS$  values as  $q_5(GS) = \exp(-GS)$  to fit the range [0;1].

Using the same partition of the spot into eight segments, we can also calculate the mean intensities for each of the segments ( $I_i$ ,  $i = 1, \dots, 8$ ). The **intensity symmetry** of the

spot is defined as  $IS = \sum_{i=1}^8 |I_i - I| / I$ , where  $I$  is the mean

intensity within the spot. We scale the  $IS$  values in the range [0;1] as  $q_6(IS) = \exp(-IS)$ . Although a spot may have perfect circular shape, it may contain very bright (or dark) and highly concentrated groups of pixels originating from pieces of dust or other contamination (Fig. 3B).  $IS$  is cal-

culated for each of two channels (Cy3 and Cy5) and the maximal value is taken as a final estimate.

We estimate two Cy5/Cy3 ratios; one by the linear regression approach (*RR*), and the other by the segmentation algorithm (*RS*). Despite the different methods of estimation, the variation between the two obtained ratios should be as small as possible. Consistent results should be expected, as most of the contaminating pixels have been removed by the filtering procedure. Large variations between the two estimates may indicate a problematic spot. Therefore, we use the **coefficient of variation of two ratio estimates**  $CVR = 2^{1/2}|RR-RS|/(RR+RS)$  as a characteristic of quality. The *CVR* quality parameter is defined as  $q_7(CVR) = \exp(-CVR)$  fitting the range [0;1]. This measure of quality is the least explicit of all the quality characteristics: we can only determine that there is big discrepancy between the estimates but not why there is a big discrepancy, unless it is accompanied by lower values in any of the other quality parameters.

Finally, we introduce two parameters evaluating the quality of the background estimates. The first is the **uniformity of the background** (*UB*) around the spot, more precisely, along the grid lines separating neighborhood spots. We divide the grid line surrounding the spot into eight segments and calculate the mean intensity in each segment ( $B_i$ ,  $i = 1, \dots, 8$ ). The *UB* parameter is defined as:

$$UB = \sum_{i=1}^8 |B_i - B| / B, \text{ where } B \text{ is the mean intensity for the}$$

whole grid line around the spot. Normally we would not expect to observe a big variability in the fluorescence intensity in the background areas. Large *UB* values may discover presence of relatively bright contamination around the spot, large variability in the background or merged neighboring spots (Fig. 3C). *UB* is rescaled to the range [0;1] by the exponential transformation:  $q_8(UB) = \exp(-UB)$ .

The second background quality characteristic is the **absolute level of the background** (*AB*). We calculated this from the local area around the spot. As for the spot diameter *D*, there is no predefined ideal value for the absolute background. Therefore, as a benchmark for comparisons, we take the typical value as the median background level over all spots on the array. Spots with exceptionally high *AB* values should be treated with care or discarded. Of two background estimates obtained in two-color microarray experiments, we use one that gives the highest value as the most indicative of possible problems. We define the *AB* quality parameter as  $q_9(AB) = \exp\{-(AB-B)/B\}$ , if  $AB > B$

and  $q_9(AB) = 1$ , if  $AB \leq B$ , where *B* is the typical background level.

We cannot claim that the developed quality parameters are the optimal. However, they have led to reasonable results for most of the experimental and simulated situations we tested. Of course, there may be a possibility to formalize some of these parameters more precisely and/or to develop new parameters accounting for other types of distortions.

### Spot quality analysis

We consider two aims of spot quality analysis. The first is to combine the nine marginal quality parameters into an overall quality value. This value can be used either to flag out directly spots with a quality lower than a user-defined threshold, or, in the follow-up image analysis procedures (normalization, classification, clustering, etc.) as a parameter characterizing the level of confidence in the obtained Cy5/Cy3 ratios. The second aim is to identify a critical range (a sort of a confidence interval) for each quality characteristic. If a certain quality characteristic of the spot falls in this range, the corresponding spot could be classified as a "bad" spot.

### Overall quality

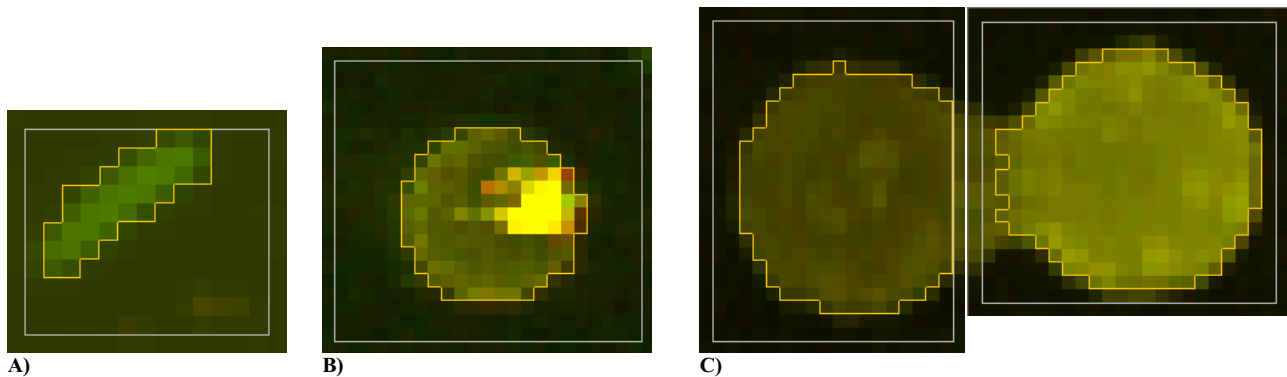
We used the following definition for the overall quality value:

$$Q = \min_{i=1 \dots 9} \{q_i^{w_i}\}, \quad (1)$$

where  $q_i = q_i(x_i) \in [0;1]$ ,  $i = 1, \dots, 9$  are the marginal quality parameters for  $x_1 = CD$ ,  $x_2 = DWS$ ,  $x_3 = N$ ,  $x_4 = D$ ,  $x_5 = GS$ ,  $x_6 = IS$ ,  $x_7 = CVR$ ,  $x_8 = UB$ ,  $x_9 = AB$ , and  $w_i$  are the weights that control the input of the corresponding quality components into the overall quality value. A link between the weight  $w_i$  and the critical value  $x_i^{\lim}$  can be established for each quality characteristic  $i = 1, \dots, 9$ :

$$w_i = \frac{\log\{Q^{\lim}\}}{\log\{q_i(x_i^{\lim})\}}, \text{ or } x_i^{\lim} = q_i^{-1}\left(\{Q^{\lim}\}^{1/w_i}\right), \quad (2)$$

where  $Q^{\lim} \in [0;1]$  is the user-defined overall quality threshold, and  $q_i(x_i^{\lim})$  is the quality parameter calculated for  $x_i^{\lim}$ . The critical value  $x_i^{\lim}$  sets up the limit such that if a certain characteristic *i* exceeds this limit, the corresponding quality parameter  $q_i(x_i^{\lim})$  will become lower

**Figure 3**

**Examples of spots with different types of distortions.** A) Large deviation from the circular shape ( $GS = 1.6$ ); B) Bright piece of contamination within a larger circular spot, resulting in a low value of the  $IS$  parameter ( $IS = 1.71$ ); C) Merged spots ( $UB = 2.98$ ).

than  $Q^{\lim}$ . The correspondence between  $x_i$ ,  $x_i^{\lim}$ ,  $q_i(x_i)$ ,  $q_i(x_i^{\lim})$ ,  $w_i$ ,  $Q$  and  $Q^{\lim}$  is demonstrated in Fig. 4.

#### Quality weights $w_i$

The experimental quality parameters  $q_i$ ,  $i = 1, \dots, 9$  are directly available from the quantification procedure, whereas the weights  $w_i$  (or the critical values  $x_i^{\lim}$ ) are unknown and are not easily guessed or derived from theory. Therefore, the problem of spot quality analysis becomes a problem of weights ( $w_i$ ) estimation. This can only be solved if additional information is available. Here we consider three possibilities:

1. The additional information may come, for example, from the user expertise. The user has to classify the spots manually [3,7] and assign a quality value to each spot from a representative subset. These values are then used for training the model (1) leading to a combination of the weights ( $w_i$ ) such that the overall quality values reproduce the user classification reasonably well.
2. We can manually apply different combinations of the weights  $w_i$  and visually appreciate, which spots have been flagged out. The trials must be continued until most of the user classified "bad" spots are eliminated by the chosen combinations of the weights.
3. The weights can be estimated automatically using information available from replicated spots on the same array or over a set of replicated arrays. Unspoiled replicate spots should have very similar estimated Cy5/Cy3 ratio values. Large differences between the observed ratios in the repli-

cate spots would signal that some spots from this replicate were irregular. We formalize this approach by first defining the quality value for the replicate:

$$Q_k = \min_{j=1 \dots n} \{Q_{kj}\}, \quad (3)$$

where  $k$  indicates the replicates,  $n$  is the number of spots in a replicate and  $Q_{kj}$  is a spot quality value given by Eq. (1). Substituting Eq. (1) into (3) yields

$$Q_k = \min_{j=1 \dots n} \left\{ \min_{i=1 \dots 9} \{q_{kji}^{w_i}\} \right\}, \quad (4)$$

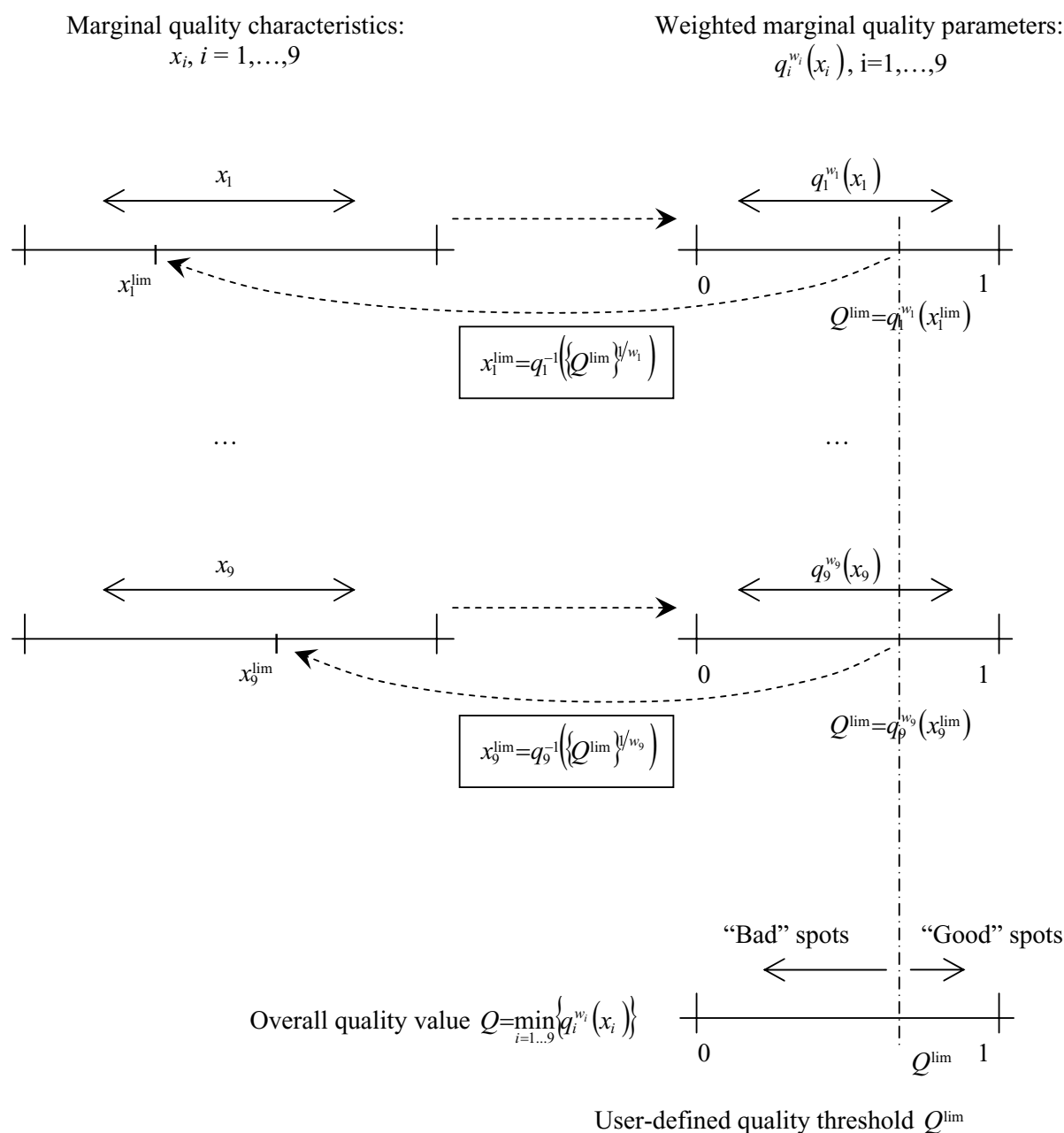
where  $q_{kji}$  is the  $i$ -th quality parameter of the  $j$ -th replicated spot in the  $k$ -th replicate. The weights  $w_i$ ,  $i = 1, \dots, 9$  are the parameters that ensure the best fit of the experimental quality values ( $Q_k$  versus  $V_k$ ) to a user-defined ideal quality curve  $f(V_k)$ , where  $V_k$  is the intensity ratio variation coefficient in the  $k$ -th replicate. The best fit of  $Q_k(V_k)$  to  $f(V_k)$  can be achieved by minimizing the sum of squared differences between  $Q_k(V_k)$  and  $f(V_k)$  (least-squares fit):

$$\sum_k \psi_k \left[ f(V_k) - \min_{j=1 \dots n} \left\{ \min_{i=1 \dots 9} \{q_{kji}^{w_i}\} \right\} \right]^2, \quad (5)$$

where  $\psi_k$  are the weights of the contribution of each replicate into the sum (5). The quality weights  $w_i$  can be estimated by minimization of Eq (5) using one of the algorithms for non-linear fitting [15].

#### Ideal quality curve $f(V_k)$

$f(V_k)$  defines how fast the overall quality of the replicates should decrease with increasing ratio variation. The shape



**Figure 4**  
**The correspondence between the quality characteristics, quality parameters and overall quality value.**

of the ideal quality curve  $f(V_k)$  is somewhat arbitrary; the only requirement is that it should demonstrate monotonic decay. Further formalization for  $f(V_k)$  is hardly possi-

ble until more information regarding the ideal quality behavior becomes available. Therefore we have to look for empirical approaches to define  $f(V_k)$ . For example,  $f(V_k)$

can be implemented as a non-parametric function, which can be constructed by the user through the properly designed user interface.  $f(V_k)$  can also be represented as a function (such as  $f(V_k) = V_k^\alpha \exp\{-\beta V_k\}$  with two adjustable parameters  $\alpha$  and  $\beta$ ), which can yield different shapes depending on the parameters values. Finally, a library of different functional shapes for  $f(V_k)$  can be created. In our work we follow the third approach: three shapes were implemented for testing:

$$f(V_k) = \begin{cases} \exp\{-V_k / \bar{V}\}, & (a) \\ \exp\{-V_k^2 / \bar{V}^2\}, & (b) \\ 1/\{1 + V_k / \bar{V}\}, & (c) \end{cases} \quad (6)$$

and for each shape only the characteristic ratio variation coefficient  $\bar{V}$  must be predefined. A typical example of the quality plot with the exponential  $f(V_k)$  (Eq. (6),a) is shown in Fig. 5.

In our quality analysis algorithm, user participation is limited to the definition of the ideal quality curve shape  $f(V_k)$ . This is somewhat simpler than deciding on the quality of several hundred spots, which is used to teach the algorithm in the manual approach. However, as with other solutions, this algorithm requires representative images to train the model. It is impossible to evaluate confidently the weight of the contribution of the diameter quality parameter, for example, if all spots in the array have the same diameter. Therefore, a careful selection of training images containing a realistic diversity of all possible distortions and artifacts is needed.

#### Fitting weights $\psi_k$

An important issue of the minimization of Eq (5) is reasonable selection of the fitting weights  $\psi_k$ . The easiest way is to set all  $\psi_k$  equal to one. However, most of the replicates are often observed in the initial part of the quality plot (high-quality spots), whereas there may be only a few replicates in the tail of the plot (poor-quality spots). In this case, equalized weights would give a very accurate fit for the initial part of the quality curve while ignoring the tail. However, mainly the tail contains the relevant information (spots with different distortions and artifacts) for identifying the quality weights. Therefore, we try to increase the input from regions with a smaller number of replicates by defining the weights  $\psi_k$  in the following way:

1. All replicates are sorted according to their ratio variation  $V_k$ .

2. The set of sorted replicates is divided into bins with ten replicates in each bin. The difference  $\lambda_j$  between the smallest and the largest ratio variation coefficients in each bin  $j$  is calculated. The larger this difference, the smaller the concentration of the replicates in the given region of the quality plot.

3. The weights  $\psi_k$  for Eq. (5) are then calculated as  $\psi_k = \psi_{j+10+i} = \lambda_j^{0.5}$ ,  $i = 0, \dots, 9$ . That is, they are equalized for ten replicates from the same bin  $j$ .

#### Follow-up image analysis

As it was mentioned earlier, the overall quality value,  $Q$  (Eq. (1)), can be used as a parameter characterizing the level of confidence in the obtained Cy5/Cy3 ratios. If, for example,  $n$  ratios should be averaged, the weighted mean would ensure a more robust estimate for the average:

$$\bar{R} = \frac{\sum_{l=1}^n Q_l R_l}{\sum_{l=1}^n Q_l}, \quad (7)$$

where  $R_l$  is the Cy5/Cy3 ratio and  $Q_l$  is the corresponding overall quality value ( $l = 1, \dots, n$ ). The weighted coefficient of variation is defined as

$$V = \frac{\sqrt{\sum_{l=1}^n Q_l (R_l - \bar{R})^2 / \sum_{l=1}^n Q_l}}{\bar{R}}. \quad (8)$$

Note that the ratio variation coefficient  $V_k$  from Eq. (5) can be determined from Eq. (8), if we set  $Q_l = 1$ ,  $l = 1, \dots, n$ , with  $n$  being the number of spots in the  $k$ -th replicate.

#### Image simulation

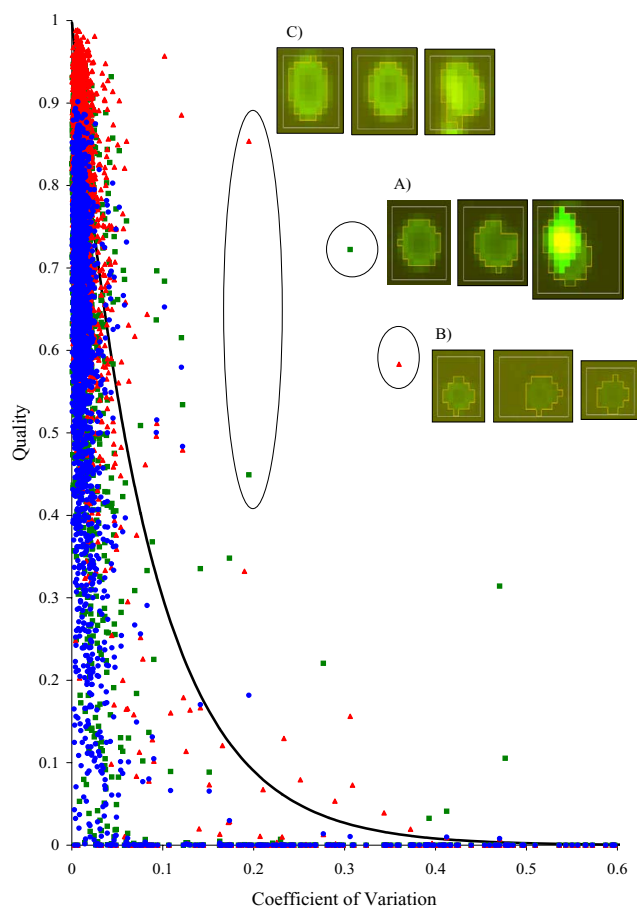
In [9] we have described the Monte-Carlo simulation model for generating artificial microarray images. The advantage of using artificial images is that we always know the exact Cy5/Cy3 ratio values. This allows us to test and compare objectively different image analysis algorithms. The general model for the two-color (Cy3, Cy5) microarray image is given by [9]:

$$F_{Cy3}(x, y) = \sum_{k=1}^N g\left(x, y, c_k^{sx}, c_k^{sy}, r^s, I^s\right) + \sum_{k=1}^M g\left(x, y, c_k^{dx}, c_k^{dy}, r^d, I^d\right), \quad (9)$$

$$F_{Cy5}(x, y) = R \sum_{k=1}^N g\left(x, y, c_k^{sx}, c_k^{sy}, r^s, I^s\right) + \sum_{k=1}^M g\left(x, y, c_k^{dx}, c_k^{dy}, r^d, I^d\right), \quad (10)$$

where  $N$  is the number of spots and  $M$  is the number of dust clusters,  $c_k^{sx}$  and  $c_k^{sy}$  are the coordinates of the center





**Figure 5**  
**Quality plots ( $Q_k$  versus  $V_k$ ) for image 7A.** Green dots – using only the CD quality parameter; red dots – using only the CRV quality parameter; blue dots – using overall quality parameter  $Q$ . The black solid line is the exponential ideal quality curve  $f(V_k)$  (Eq. (6),a). Three triplicates showing poor quality (outlined by circles) are given in the insets. The main characteristics of the spots from the selected triplicates are given in Table 3.

of a spot,  $c_k^{dx}$  and  $c_k^{dy}$  are the coordinates of the center of a dust cluster,  $r^s$  and  $r^d$  are the approximate radii of the spot and dust cluster, respectively,  $I^s$  and  $I^d$  are the fluorescence intensity in the center of the spot in the Cy3 color channel and in the center of the dust cluster, respectively, and  $R$  is the ratio of the test and control samples for each spot. Dust is represented by the random distribution over the array of clusters of pixels of varying brightness. We consider that these pixel clusters have an identical shape to the spots and therefore the same analytical representation is used for an ideal spot shape and dust cluster:

$$g(x, y, c^x, c^y, r, I) = I \exp \left( - \left\{ \left[ \frac{x - c^x}{r} \right]^4 + \left[ \frac{y - c^y}{r} \right]^4 + \left[ \frac{x - c^x}{r} \right]^2 \left[ \frac{y - c^y}{r} \right]^2 \right\} / 2 \right) \quad (11)$$

The parameters characterizing the spots ( $c_k^{sx}$ ,  $c_k^{sy}$ ,  $r^s$ ,  $I^s$  and  $R$ ) are user-defined. For example, the coordinates  $c_k^{sx}$  and  $c_k^{sy}$ , the radius  $r^s$  and the ranges for  $x$  and  $y$  for each spot are defined from a user-defined array design. The user should also specify the number of dust clusters  $M$  on the array. The other parameters characterizing the dust are random variables, and the probability laws for their generation is a matter of choice. We use uniform distributions for  $r^d$  (in the interval 0 to  $r_m$ ) and  $I^d$  (in the interval 0 to  $I_m$ ), where  $r_m$  and  $I_m$  are a user-defined maximal dust cluster radius and maximal dust intensity, respectively. We also assume that  $c_k^{dx}$  and  $c_k^{dy}$  are uniformly distributed over the array. Statistical laws of the dust characteristics can generally be different in the two (Cy3, Cy5) channels.

In the developed simulation model we also account for the nonspecific hybridization and statistical noise:

$$\tilde{F}_i(x, y) = F_i(x, y) + \bar{B}_i + \eta_{Bi} \bar{B}_i G_B + \sigma(x, y) G_S, \quad (12)$$

where  $i$  represents either Cy3 or Cy5,  $\bar{B}_i$  and  $\eta_{Bi}$  are the user-defined average and noise-to-signal ratio of nonspecific fluorescence intensity in the color channel  $i$ ,  $\sigma(x, y)$  is the standard deviation of the pixel statistical noise, and  $G_B$  and  $G_S$  are independent Gaussian random variables with zero mean and unit standard deviation. The exact representation for  $\sigma(x, y)$  is defined by the experimental set-up. There are currently three possibilities:  $\sigma(x, y)$  can be (i) constant, (ii) proportional to the signal, or (iii) proportional to the square root of signal. The type and quantitative characteristics of the statistical noise are defined by the user.

## Discussion Software

The developed algorithm for quality control is included in the software package MAIA, which offers a complete solution for DNA microarray image analysis, including automatic spot localization and spot quantification procedures. A demo version of the software can be downloaded from [8].

### Artificial images

All artificial images were generated using the same array design:  $4 \times 4$  blocks and  $21 \times 21$  spots within each block with the inter-spot distance of 15 pixels. For all spots in the generated arrays the spot radius,  $r$ , was about 4 pixels, the intensity,  $I$ , in the Cy3 color channel was 5000 and the ratio,  $R$ , of the Cy5 and Cy3 channels was 3. Non-specific hybridization was generated using  $\bar{B}_i = 1000$  and  $\eta_{Bi} = 0.5$ . The standard deviation of the statistical noise,  $\sigma(x, y)$ , at each pixel was proportional to the signal at the corresponding pixel with the noise-to-signal ratio of 0.1. The three generated images (Fig. 6 (insets)) differed in the number of dust clusters. The percentages of dust clusters with respect to the number of good spots were: 0% (image 6A), 5% (image 6B) and 25% (image 6C). In all cases the maximal dust cluster radius,  $r_m$ , was set to 8, and maximal dust intensity,  $I_m$ , to 64000.

As all spots have the same theoretical ratio ( $R = 3$ ), they can be considered as replicates and therefore we can use the quantitative characteristics defined in Eqs. (7) and (8) to characterize the performance of the algorithm. We expect the best estimators to provide the averaged Cy5/Cy3 ratio closer to the true ratio ( $R = 3$ ) with the least spread around this value.

For each artificial array we compared the weighted statistical characteristics for the average  $\bar{R}$  (Eq. (7)) and for the ratio variation coefficient  $V$  (Eq. (8)) with the un-weighted ones. The un-weighted characteristics were obtained from Eqs. (7) and (8) by setting all  $Q_l$ ,  $l = 1, \dots, n$  to 1. The weighted characteristics were calculated with the overall quality values  $Q_l$  available from the quality analysis algorithm. As all spots from the simulated image can be considered as replicates, we artificially split up the total number of spots into the groups of three closely placed spots. These groups, regarded as independent triplicates, can be used to calculate the experimental quality values  $Q_k$  (Eq. (4)) and to build up the corresponding quality plot ( $Q_k$  versus  $V_k$ ). Three functional shapes (Eq. (6)) for the ideal quality curve  $f(V_k)$  were tested.

The results are collected in Table 1. The weighted statistical characteristics with the developed quality control eliminated bias in the average Cy5/Cy3 ratio estimates and reduced the ratio variation coefficient. Note that for "ideal" image 6A (no dust clusters) the applied quality measures retained the estimates unchanged.

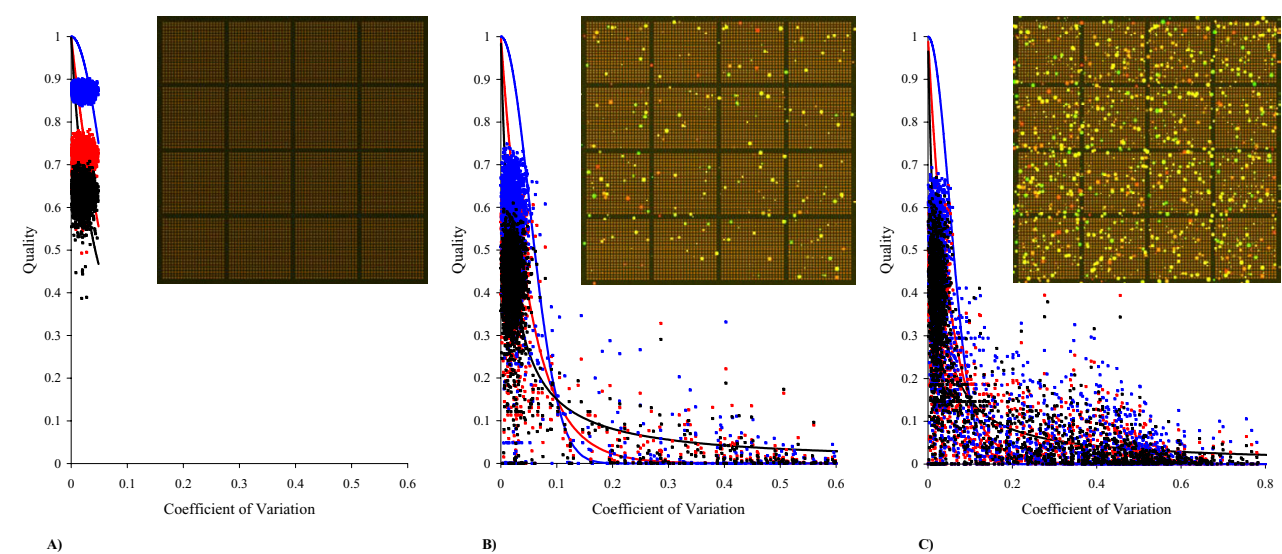
The three compared ideal quality shapes  $f(V_k)$  (Eq. (6)) demonstrated similar performance. A small advantage (slightly smaller  $V$ ) of the Gaussian-like function (Eq. (6),b) for image 6B (5% of dust clusters) was compensated by the lowest performance for image 6C (25% of dust clusters). The difference between the three shapes can be seen from Fig 6, where the corresponding quality plots are drawn for the three generated images. The inverse shape (Eq. (6),c) is the least stringent whereas the Gaussian-like shape is the most stringent with respect to the replicates variability. This means that the inverse shape does not require the replicates with higher variability to have lower quality values (and correspondingly the replicates with smaller variability to have higher quality values) as strictly as the Gaussian-like one. Although it seems that the Gaussian-like shape is the best choice, there is still a drawback. Due to its relatively abrupt decrease it is difficult to keep the balance in fitting weights  $\psi_k$  between the head and the tail of the shape. Despite the adaptive selection of the fitting weights  $\psi_k$  (see section Spot quality analysis), the fitting procedure, trying to ensure the highest quality for the replicates with the lowest variability, may still overlook the replicates with the higher variability. This depends on the image quality (replicate variability) and can explain why the Gaussian-like shape yielded the greatest ratio variation for image 6C. This also indicates that more work is needed to find out an optimal combination of the ideal quality curve and the fitting weights  $\psi_k$  to be used in the training procedure.

The results of simulation study and our experience with the experimental images suggest that the exponential shape (Eq. (6),a) offers a reasonable compromise between the fitting weights used and stringency with respect to the replicates variability. Therefore in the further quality analysis we will apply exponential  $f(V_k)$ .

### Experimental images

We performed quality analysis of two experimental images with different array designs and signal-to-noise levels.

One image (Fig. 7A) (image 7A) was provided as a demonstration example for UCSF Spot 2.0 (downloadable from [16]). It contains  $4 \times 4$  blocks with  $21 \times 21$  spots in each block, with a spot cell size of about 10 pixels. Cy3 and Cy5 color channels are strongly correlated, with the average correlation coefficient for the spots being about 0.97. Bright contamination spots can be seen irregularly scattered over the array. The magnified image of one such spot is shown in Fig. 2 (inset). Each clone was spotted in triplicate. The replicated spots are placed as neighbors in a row (see Fig. 7A).



**Figure 6**  
**Quality plots ( $Q_k$  versus  $V_k$ ) for artificial images.** Three generated images differed in the percentage of dust clusters with respect to the number of good spots: A) 0%, B) 5% and C) 25%. For the further details see the text. The solid lines represent the user-defined ideal quality curves  $f(V_k)$ : Red – exponential (Eq. (6),a); Blue – Gaussian-like (Eq. (6),b); Black – inverse (Eq. (6),c).

The second image (from the Institute Curie) (image 7B) contains  $12 \times 4$  blocks with  $12 \times 12$  spots in each block (Fig. 7B), with a spot cell size of about 30 pixels. The average correlation between the channels for the spots is about 0.85. This is lower than for the first image, although this image has no obvious contamination spots. Each clone was prepared in triplicate, with the replicated spots in three vertically distributed sub-arrays (see Fig. 7B).

We expect the Cy5/Cy3 ratios from the replicates to be similar. Therefore it is reasonable to take the coefficient of

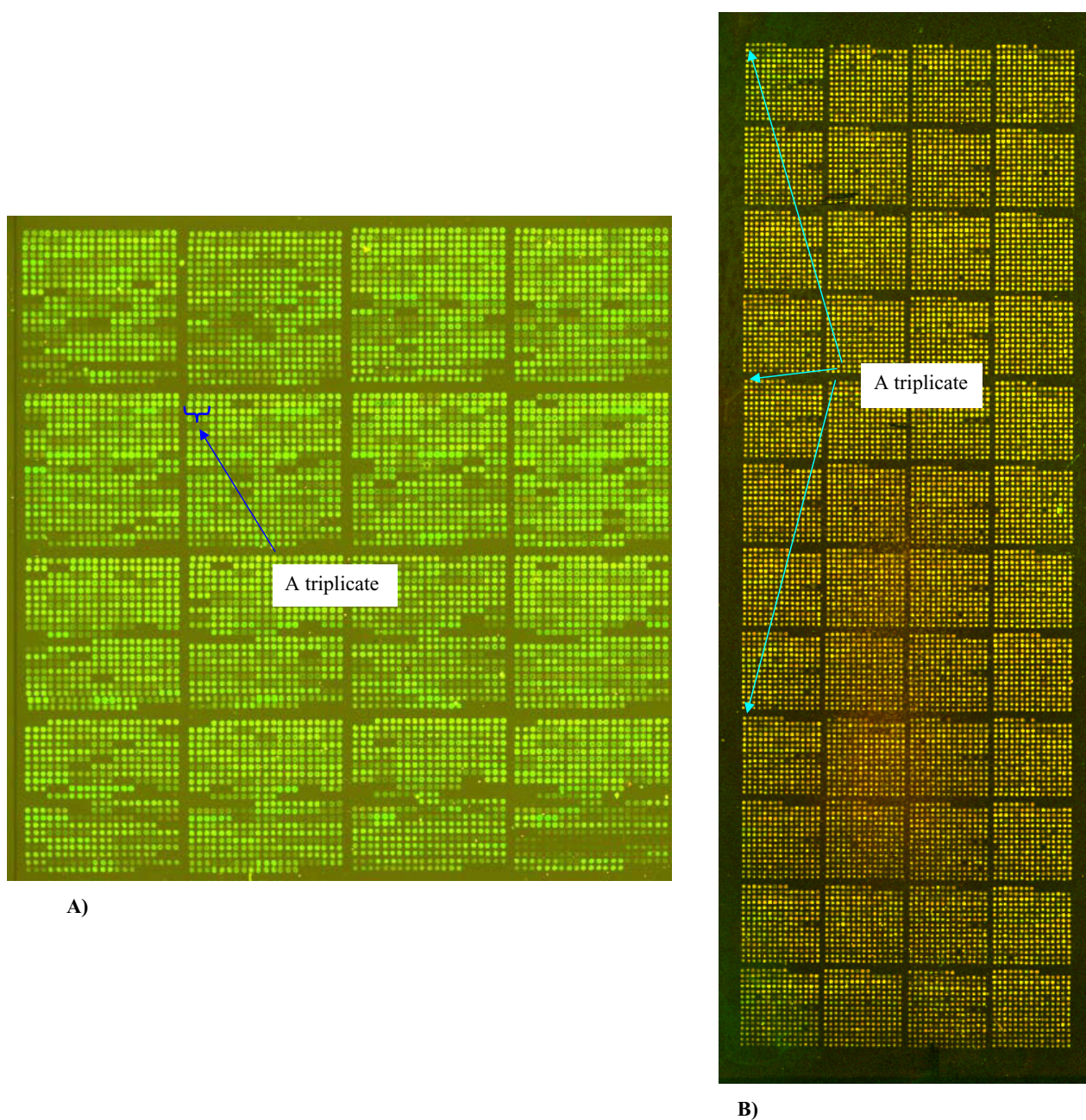
variation (Eq. (8)) of the replicates as a quantitative measure of the ratio estimation consistency. However, this measure may not be totally objective: (i) the estimates may be consistent, but systematically biased (the true values of the ratios are unknown); (ii) three replicated spots of very poor quality may give very similar ratio values just by chance (the number of replicates is low).

The average over all replicates at the given array coefficient of variation is taken as a global indicator of the Cy5/Cy3 ratio consistency of the array.

**Table 1: The average  $\bar{R}$  (Eq. (7)) and the coefficient of variation  $V$  (Eq. (8)) of the Cy5/Cy3 ratios over all spots for three artificial images A, B, and C from Fig 6**

| Image                                 | A         |       | B         |       | C         |       |
|---------------------------------------|-----------|-------|-----------|-------|-----------|-------|
|                                       | $\bar{R}$ | $V$   | $\bar{R}$ | $V$   | $\bar{R}$ | $V$   |
| No quality control                    | 3.00      | 0.028 | 2.96      | 0.20  | 2.84      | 0.35  |
| $f(V_k) = \exp\{-V_k / \bar{V}\}$     | 3.00      | 0.028 | 2.99      | 0.041 | 2.99      | 0.089 |
| $f(V_k) = \exp\{-V_k^2 / \bar{V}^2\}$ | 3.00      | 0.028 | 2.99      | 0.038 | 2.99      | 0.097 |
| $f(V_k) = 1 / (1 + V_k / \bar{V})$    | 3.00      | 0.028 | 2.99      | 0.042 | 2.99      | 0.094 |



**Figure 7**

**Experimental images.** A)  $4 \times 4$  blocks with  $21 \times 21$  spots per block, spot cell size is about 10 pixels; B)  $12 \times 4$  blocks with  $15 \times 15$  spots per block, spot cell size is about 30 pixels. The locations of triplicates are indicated.

We compared two quality characteristics (the coefficient of determination  $CD$  ( $q_1$ ) and the coefficient of variation of two ratio estimates  $CVR$  ( $q_7$ ), applied separately) and the overall quality parameter  $Q$  (Eq. (1)) to the case when no quality control was applied. The weights of the mar-

ginal quality parameters for  $Q$  were identified using the exponential ideal quality curve (Eq. (6),a) with  $\bar{V} \approx 0.08$  for image 7A, and with  $\bar{V} \approx 0.2$  for image 7B. The results are summarized in Table 2.

**Table 2: The average over all replicates coefficients of variation  $V$  (Eq. (8)) for two experimental images A and B from Fig. 7**

| Quality Control | A     | B     |
|-----------------|-------|-------|
| No control      | 0.066 | 0.120 |
| CD              | 0.019 | 0.105 |
| CVR             | 0.028 | 0.102 |
| Overall Q       | 0.017 | 0.101 |

We found a greater improvement for image 7A than for image 7B after applying the quality measures. This was not a surprise, as image 7B is characterized by a reasonably high signal-to-noise level, and it does not contain any obvious contaminated spots. However, even in this case the quality measures cannot be ignored, as there are still a few low-intensity spots that need to be specially treated (probably rejected). By contrast, image 7A has obvious randomly distributed pieces of dust, and the developed quality measures proved to be powerful enough to disregard the contaminated spots, thus increasing the consistency of the Cy5/Cy3 ratio estimates.

The comparison of the different quality measures has shown that they are array-specific. For image 7B, CD, CVR and Q gave almost equivalent performance, whereas for image 7A, CD was much better than CVR, and Q gave a slightly better ratio consistency than both of these. However, the gain from Q was not much greater, so that it would be reasonable to assume that only a single quality measure (such as CD or CVR) could always give good results. Unfortunately, this is not the case. Although we have found that CD and CVR are indeed the most powerful quality measures they cannot cover every type of distortion.

We demonstrate this with image 7A. There remained, after applying the individual quality measures, several triplicates with high coefficients of variation that were not correctly suppressed. Three such replicates are shown in Fig. 5, and the main quantitative characteristics of the corresponding spots are listed in Table 3. Note that Table 3 contains the quality characteristics from section Quality characteristics, which do not need to reside in the interval  $[0;1]$ .

In triplicate A, one spot is clearly contaminated; however, this contamination is highly correlated in the two channels ( $CD = 0.97$ ) and the spots cannot be filtered out using CD. A more powerful parameter for this type of problem is CVR (0.11), perhaps in combination with IS (1.02). Triplicate B includes low-intensity spots. Although CVR quality characteristic does not detect any deficiency in this case, the difference between the estimated intensity ratios (within the triplicate) is large, meaning that CVR

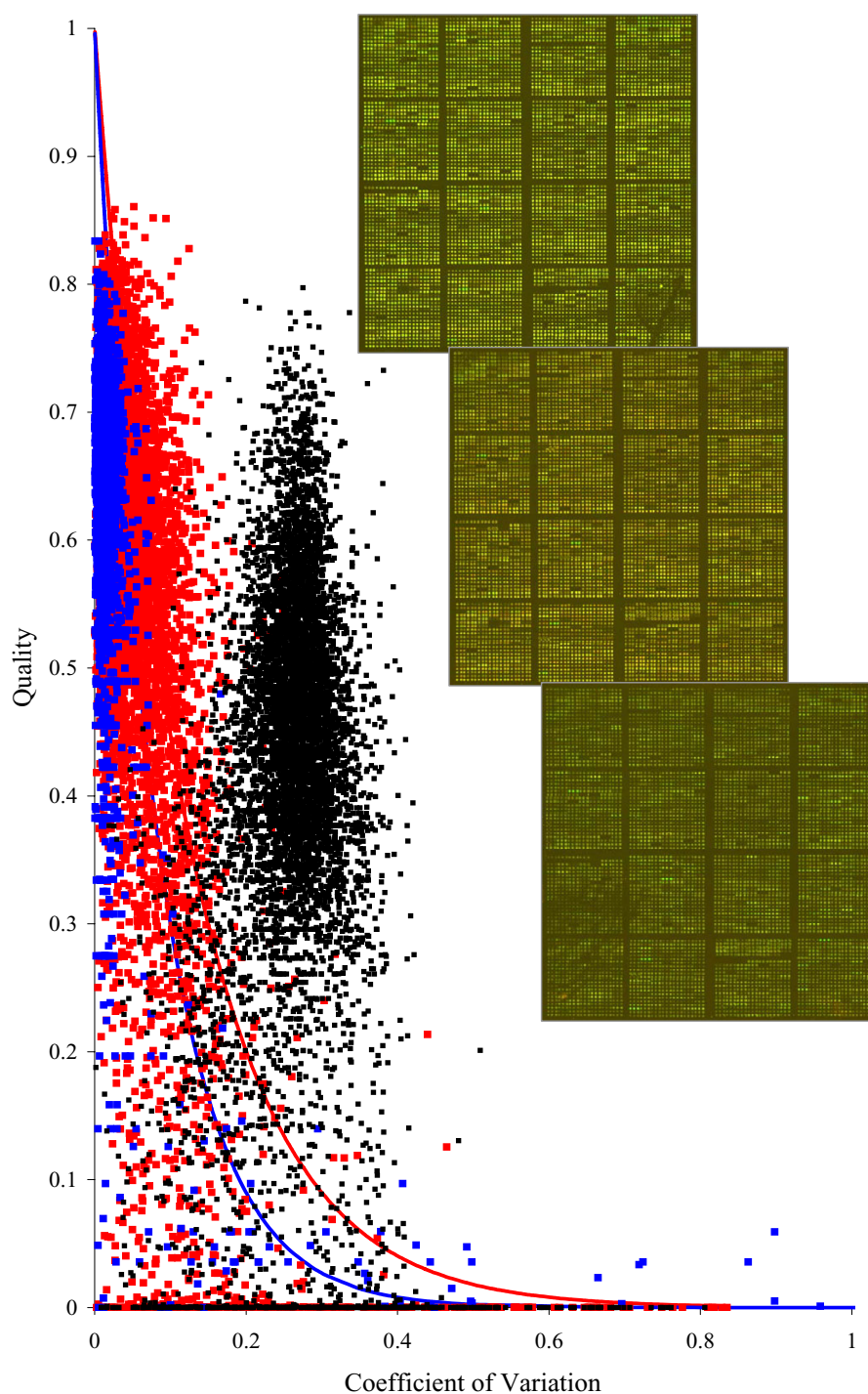
**Table 3: The main characteristics of the spots from three triplicates (see Fig. 5 (insets)), demonstrating excessive variability in the Cy5/Cy3 ratio estimates**

| Triplicate | Quality Parameter | Spot 1 | Spot 2 | Spot 3 |
|------------|-------------------|--------|--------|--------|
| A          | Ratio             | 0.27   | 0.26   | 0.12   |
|            | CD                | 0.98   | 0.98   | 0.97   |
|            | CVR               | 0.016  | 0.014  | 0.11   |
|            | IS                | 0.17   | 0.21   | 1.02   |
|            | UB                | 0.060  | 0.10   | 0.31   |
| B          | Ratio             | 0.19   | 0.13   | 0.072  |
|            | CD                | 0.85   | 0.60   | 0.33   |
|            | CVR               | 0.010  | 0.014  | 0.012  |
|            | IS                | 0.15   | 0.072  | 0.041  |
|            | UB                | 0.041  | 0.030  | 0.042  |
| C          | Ratio             | 0.29   | 0.30   | 0.44   |
|            | CD                | 0.97   | 0.97   | 0.93   |
|            | CVR               | 0.010  | 0.011  | 0.0052 |
|            | IS                | 0.12   | 0.15   | 0.36   |
|            | UB                | 0.24   | 0.11   | 0.99   |

alone is not sufficient to eliminate the outlier spots. CD seems to be more indicative in this case. Triplicate C cannot be confidently recognized either by CD or CVR quality measures. This problem (contamination penetrating into the spot from the outside area) can be figured out by applying either the uniformity of the background quality parameter ( $UB = 0.99$ ) or the Durbin-Watson quality parameter ( $DWS = 0.58$ ).

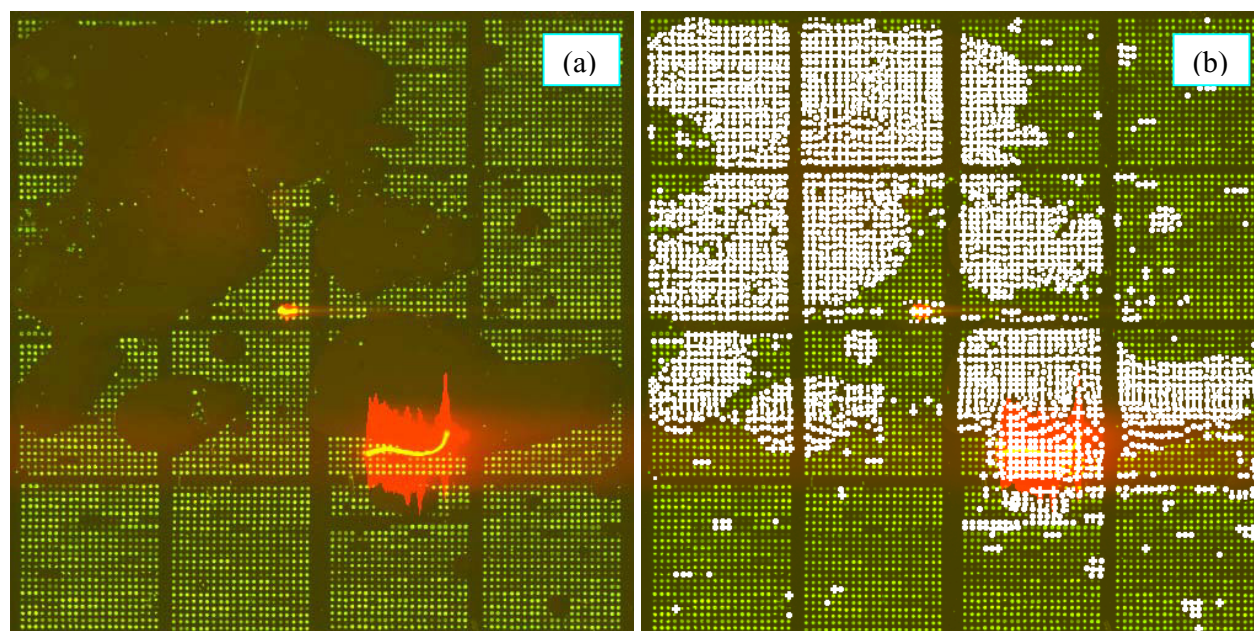
This type of analysis leads to the refined critical values  $x_i^{\lim}$  for each of the quality characteristics. These  $x_i^{\lim}$  values are then recalculated into the corresponding weights  $w_i$  for the user-selected overall quality threshold  $Q^{\lim}$ , using Eq. (2) (see Fig. 4). In general, however, the weights are derived automatically from Eq. (5) by the non-linear fitting procedure. If certain quality factors do not influence the shape of the experimental quality curve  $Q_E$  (Eq. (4)), the corresponding weights will be set close to 0. If a certain effect shows up in only a small number of spots, it may be neglected by the optimization procedure, and the corresponding weight will be erroneously small. In this case, manual correction of the weights would be necessary.

The quality value of three bad replicates from Fig. 5 (insets), as well as of many others demonstrating larger deviations in the obtained Cy5/Cy3 ratio estimates (Fig. 5), were decreased using the combined criteria Q, whereas the quality of replicates with smaller variations remained almost unaffected.

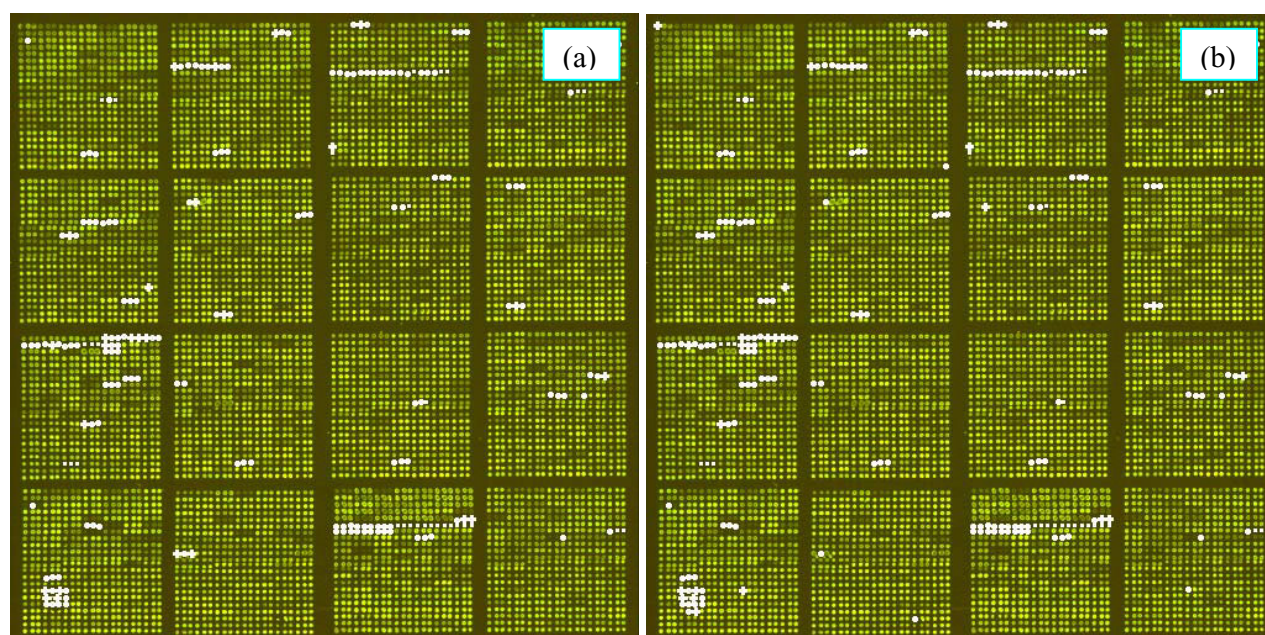
**Figure 8**

**Quality plots ( $Q_k$  versus  $V_k$ ) using replicated arrays.** *Black dots* – three replicated images (insets) are combined without normalization and the default quality weights are used; *red dots* – three replicated images (insets) are combined after the global normalization; *blue dots* – triplicate spots from the first array are used. The solid lines represent the exponential ideal quality curves  $f(V_k)$  (Eq. (6),a): *Red line* (for the red dots) –  $\bar{V} \approx 0.125$ ; *Blue line* (for the blue dots) –  $\bar{V} \approx 0.08$ .





A)



B)

### Figure 9

**Quality analysis of poor-quality (A) and good-quality (B) images.** A: (a) Poor-quality image; (b) Image A with the "bad" spots identified by the quality analysis based on its own triplicates. B: (a) Image B with the bad spots identified by the quality analysis based on the triplicates from image A; (b) Image B with the bad spots identified by the quality analysis based on its own triplicates. "Bad" spots are the spots with the overall quality below 0.3. "Bad" spots are indicated by the white crosses.

Quality measures can signal some of the replicates to be bad despite there being no big difference in the ratios. With a small number of replicated spots all spots from a replicate may indeed have very close ratios, but several of them may be really deficient. Therefore, it is more important to consider those replicates demonstrating unusually high diversity in the obtained ratio estimates. This clearly suggests problems in the spots. Algorithmically, this can be achieved by assigning larger weights  $\psi_k$  to these replicates in the non-linear fitting procedure (Eq. (5)). The fitted parameters  $w_i$  are then used to flag out all deficient spots, even those that belong to the replicates with the consistent ratio estimates. However, if *all* spots from a consistent replicate demonstrate poor quality, this can also be an indication that this replicate is actually good, but it was erroneously flagged out by the quality analysis procedure. This would require user intervention to correct for the selected set of quality characteristics and/or for the corresponding quality weight  $w_i$ . Additional procedures for automatic analysis of the replicates with consistent ratios but low quality values can be envisaged.

The fact that overall  $Q$  does not show up much better performance (Table 2) is due to rather good general quality of the images, and a few problematic triplicates cannot influence very much the averaged coefficients of variation. For example, in image 7A, we have less than 9% of triplicates with the ratio variation coefficients larger than the selected  $\bar{V}$  ( $\sim 0.08$ ), and 7% for image 7B ( $\bar{V} \approx 0.2$ ).

### Replicated arrays

Quality analysis can also be performed with replicated spots from different arrays. Three replicated images (Fig. 8 (insets)) were used for quality analysis. Although each image contains three replicated spots placed as neighbors in a row (similar to image A from Fig. 7), we pretended that there were no replicates within each array. Therefore we made available only replicated spots from different arrays for quality analysis. In Fig. 8 we present three quality plots. Black dots correspond to the case when the three replicated images were combined without normalization and the default quality weights were used. Relatively large coefficients of variation are due to the bias in the obtained Cy5/Cy3 ratio estimates between the three images. To apply our algorithm of quality analysis in this case the ideal quality curve should account for this bias, i.e. it should implicitly incorporate image normalization model. A simpler way would be to separate the normalization and quality analysis procedures: the image normalization should precede the quality analysis, or we can also envisage an iterative procedure where the steps of normalization and quality analysis are performed in turn. Taking into account that a variety of normalization methods [17]

are currently available, we leave the detailed development of the quality analysis strategy in this case for the future.

As an example, we applied a global normalization algorithm [18]. It is assumed that most genes are not differentially expressed and therefore the expected averaged over all spots of the array Cy5/Cy3 ratio should be close to one. The normalization constant  $a$  for each array can be calculated as

$$a = \exp \left\{ \frac{\sum_{i=1}^n \ln R_i}{n} \right\},$$

where  $n$  is the number of

spots and  $R_i$  is the ratio estimate for the  $i$ -th spot. Red dots in Fig 8 represent the quality curve using the replicated arrays after the global normalization. For comparison, blue dots show the quality curve using the triplicates from the first array. The red-dot cloud spreads wider, because, even after normalization, the variability of replicates from different arrays is higher than from the same array (where replicates were closely placed). Despite this replicated arrays can be used in the quality analysis since the required tendency – the decrease of the overall quality with the increase of the replicate variability – can be ensured. To account for higher levels of inter-array statistical variability in replicates, we have to apply less stringent ideal quality curve. This can be achieved by selecting larger values for the parameter  $\bar{V}$  in Eq. (6).

### Quality weight extrapolation

We demonstrate here a possibility to apply quality weights obtained from the analysis of one training image, which should contain replicated spots, to other arrays, which may not contain replicates. We used two images (Figs. 9A and 9B) of the same design as before: each image contains three replicated spots placed as neighbors in a row. Image 9A has obvious deficiencies (large amount of absent spots, scratches, contamination), whereas image 9B is less problematic.

The quality weights were estimated from the triplicates of image 9A, as it contains a wider diversity of possible artifacts and distortions. Using the obtained weights the "bad" spots were identified in image 9A. A spot was classified as a bad spot if its overall quality was below 0.3. Fig. 9A(b) shows the same image with the "bad" spots indicated by the white crosses.

To eliminate the "bad" spots ( $Q < 0.3$ ) from image 9B, we first applied the quality weights obtained for image 9A. The result is shown in Fig. 9B(a). Then we performed quality analysis for image 9B using its own triplicates. The spots flagged out by this approach are indicated in Fig. 9B(b). Comparing Figs. 9B(a) and 9B(b), one can con-



clude that the results are very similar, although not identical, of course.

This example attempts to reproduce an important possibility of designing microarray experiments. A small number of training arrays with replicated spots and representative diversity of possible artifacts can be measured and analyzed. The obtained results can then be used to evaluate the quality of other arrays of similar design, which may not contain replicated spots.

## Conclusion

We have described an algorithm for quantitative spot quality evaluation in DNA microarray image analysis that allows the automatic identification of the weights for the marginal quality parameters within the model combining these parameters into an overall spot quality value. The algorithm relies on the assumption that unspoiled replicated spots should have higher levels of consistency in the obtained Cy5/Cy3 ratio estimates than "bad" spots. The user is only required to define an ideal quality curve  $f(V_k)$  establishing how fast the overall quality of the replicates must decrease with increasing ratio variation in the corresponding replicates. For simple models from Eq. (6) only one parameter – the characteristic ratio variation coefficient  $\bar{V}$  – must be specified. In this paper the functional shape for  $f(V_k)$  was empirically selected through numerous experiments with artificial and experimental images. This, however, may not be an optimal choice and further improvements can be expected, preferably using more theoretical approaches. A complementary perspective is to further elaborate the algorithm for estimating the fitting weights  $\psi_{k'}$ , which may be implicitly dependent on the ideal quality curve.

We use nine marginal quality characteristics, which cover a broad range of different deviations from a normal (good) spot. Therefore, it is possible that some of these parameters will not be relevant for a certain image. For example, if there are no clearly un-circular spots, the corresponding parameter (GS) can be omitted. The optimization procedure will however report this, assigning a very low weight to the corresponding quality characteristic. These weights are then converted into the critical levels  $x_i^{\text{lim}}$  using Eq. (2), which will tolerate a big diversity in the possible values of the corresponding parameter. Thus, the developed procedure both searches for the weights and implicitly reports on the relevance of the quality parameters to each particular image. However, if the images used to train the model (that is to identify the weights) are not

representative enough and do not contain enough spots with certain types of artifacts, some important sources of spot deficiency may be overlooked. In this case, manual adjustment of the weights may be necessary.

Another problem is that generally a very small number of replicated spots (rarely more than 3) are available for the analysis. Therefore it is possible that all spots from a replicate, being defective, demonstrate consistent Cy5/Cy3 ratio values. However, we expect that is less probable than to observe unusually high diversity in the ratio estimates for these replicates. If all spots from a consistent replicate are actually good, but they are erroneously assigned low quality values, this is an indication of some problems in the quality analysis itself. This situation would possibly require user participation to correct for the selected set of quality characteristics and/or for the corresponding quality weights  $w_i$ .

We have demonstrated a possibility to carry out the quality analysis using replicated spots from different arrays. An additional procedure of the image normalization should precede the quality analysis in this case.

## Authors' contributions

EN conceived the algorithm, performed software implementation and drafted the manuscript. EB conceived of the study and participated in coordination. All authors read and approved the final manuscript.

## Acknowledgements

We would like to thank our colleagues from the different laboratories of the Institute Curie: (F. Radvanyi, UMR 144 CNRS/IC: Compartimentation et dynamique cellulaires; O. Delattre, INSERM 509: Pathologie moléculaire des cancers; M. Dutreix, UMR 2027 CNRS/IC: Génotoxicologie et cycle cellulaire) and Prof. D. Pinkel (UCSF Comprehensive Cancer Center) who have provided numerous microarray images allowing considerable improvement of the algorithms.

## References

1. Hegde P, Qi R, Abernathy K, Gay C, Dharap S, Gaspard R, Hughes JE, Snesrud E, Lee N, Quackenbush J: **A concise guide to cDNA microarray analysis.** *Biotechniques* 2000, **29**:548-562.
2. Pinkel D, Seagraves R, Sudar D, Clark S, Poole I, Kowbel D, Collins C, Kuo WL, Chen C, Zhai Y, Dairkee SH, Ljung BM, Gray JW, Albertson DG: **High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays.** *Nature Genetics* 1998, **20**:207-211.
3. Buhler J, Ideker T, Haynor D: **Dapple: improved techniques for finding spots on DNA microarrays.** *UW CSE Technical Report UWTP 2000-08-05* 2000.
4. Brown CS, Goodwin PC, Sorger PK: **Image metrics in the statistical analysis of DNA microarray data.** *PNAS* 2001, **98**:8944-8949.
5. Wang X, Ghosh S, Guo SW: **Quantitative quality control in microarray image processing and data acquisition.** *Nucleic Acids Res* 2001, **29**:e75.
6. Chen Y, Kamat V, Dougherty ER, Bittner ML, Meltzer PS, Trent JM: **Ratio statistics of gene expression levels and applications to microarray data analysis.** *Bioinformatics* 2002, **18**:1207-1215.
7. Hautaniemi S, Edgren H, Vesanen P, Wolf M, Järvinen AK, Yli-Harja O, Astola J, Kallioniemi O, Monni O: **A novel strategy for micro-**

- array quality control using Bayesian networks. *Bioinformatics* 2003, **19**:2031-2038.
8. **Software package for automatic microarray image analysis (MAIA)** [<http://bioinfo.curie.fr/projects/maia>]
  9. Novikov E, Barillot E: **A robust algorithm for ratio estimation in two-color microarray experiments.** *Journal of Bioinformatics and Computational Biology* 2005, **3**:1411-1428.
  10. Bozinov D, Rahnenführer J: **Unsupervised technique for robust target separation and analysis of DNA microarray spots through adaptive pixel clustering.** *Bioinformatics* 2002, **18**:747-756.
  11. Kendall MG, Stuart A: *The Advanced Theory Of Statistics Volume 2.* McMillan: New York; 1979.
  12. Dissanaikie G, Wang S: **A critical examination of orthogonal regression.** 2003 [<http://ssrn.com/abstract=407560>].
  13. Fisher LD, van Belle G: *Biostatistics. A Methodology for the Health Sciences* John Wiley & Sons: New York; 1993.
  14. Draper NR, Smith H: *Applied regression analysis* John Wiley & Sons: New York; 1998.
  15. Bevington PR, Robinson DK: *Data reduction and error analysis for the physical sciences* McGraw-Hill: New York; 1991.
  16. **Jain Lab Downloads** [<http://jainlab.ucsf.edu/Downloads.html>]
  17. Wu W, Xing EP, Myers C, Mian IS, Bissell MJ: **Evaluation of normalization methods for cDNA microarray data by k-NN classification.** *BMC Bioinformatics* 2005, **6**:191.
  18. Axon Instruments Inc: *GenePix Pro 5.0: User's Guide and Tutorial* 2003 [<http://www.moleculardevices.com>].

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:  
[http://www.biomedcentral.com/info/publishing\\_adv.asp](http://www.biomedcentral.com/info/publishing_adv.asp)

