BIOINFORMATICS ORIGINAL PAPER Vol. 23 no. 23 2007, pages 3162–3169
doi:10.1093/bioinformatics/btm487

Gene expression

MDQC: a new quality assessment method for microarrays based on quality control reports

Gabriela V. Cohen Freue[1,2,*], Zsuzsanna Hollander[6], Enqing Shen[6], Ruben H. Zamar[2], Robert Balshaw[2], Andreas Scherer[7], Bruce McManus[3,6], Paul Keown[4,8], W. Robert McMaster[5,8] and Raymond T. Ng[1]

[1]Department of Computer Science, [2]Department of Statistics, [3]Department of Pathology and Laboratory Medicine, [4]Department of Medicine, [5]Department of Medical Genetics, University of British Columbia, [6]The James Hogg iCAPTURE Centre, Providence Health Care-University of British Columbia, [7]Novartis Pharma AG, Basel and [8]Vancouver Coastal Health Research Institute, Vancouver, British Columbia, Canada

Received on June 6, 2007; revised on September 6, 2007; accepted on September 25, 2007

Advance Access publication October 12, 2007

Associate Editor: Joaquin Dopazo

ABSTRACT

Motivation: The process of producing microarray data involves multiple steps, some of which may suffer from technical problems and seriously damage the quality of the data. Thus, it is essential to identify those arrays with low quality. This article addresses two questions: (1) how to assess the quality of a microarray dataset using the measures provided in quality control (QC) reports; (2) how to identify possible sources of the quality problems.

Results: We propose a novel multivariate approach to evaluate the quality of an array that examines the 'Mahalanobis distance' of its quality attributes from those of other arrays. Thus, we call it Mahalanobis Distance Quality Control (MDQC) and examine different approaches of this method. MDQC flags problematic arrays based on the idea of outlier detection, i.e. it flags those arrays whose quality attributes jointly depart from those of the bulk of the data. Using two case studies, we show that a multivariate analysis gives substantially richer information than analyzing each parameter of the QC report in isolation. Moreover, once the QC report is produced, our quality assessment method is computationally inexpensive and the results can be easily visualized and interpreted. Finally, we show that computing these distances on subsets of the quality measures in the report may increase the method's ability to detect unusual arrays and helps to identify possible reasons of the quality problems.

Availability: The library to implement MDQC will soon be available from Bioconductor

Contact: gcohen@mrl.ubc.ca

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Microarray technologies have enabled researchers to monitor the expression levels of tens of thousands of genes simultaneously. However, the process of producing microarray data, from sample preparation to the final step of harvesting the data, involves multiple steps, some of which can be error-prone. Possible quality problems include poor RNA extraction, problems arising from the hybridization process, physical defects of the chips, and artifacts such as batching effects (see Zhang *et al*., 2004 and Brettschneider *et al*., 2007 for a more detailed discussion). As poor quality arrays may seriously distort the preprocessing as well as the data analysis procedures, examining the quality of the arrays is a critical step before any subsequent analysis can be performed. In this article, the first question we consider is how to assess the quality of a large microarray dataset. That is, after we receive n microarray chips from a facility that produces microarray data, we need to assess their quality, and if necessary, to identify those m chips that need to be rerun. The second question we consider is why each of the m chips is of unacceptable quality. Because of the resources involved (e.g. biological material, human time and production cost), this is an important step to reduce the effort required for the rerun.

Despite the extensive research on microarray data, the development of microarray quality assessment methods is still in its early stages. The standard practice of inspecting each image files to detect quality problems of each array is time consuming and difficult to apply in large studies. As a result, alternative automated quality assessment methods have been proposed. We briefly discuss some of these studies and refer the reader to Brettschneider *et al*. (2007) for a comprehensive review of the literature. For spotted arrays, several studies provide useful spot quality measures to examine features of each spot on the slides (Bylesjö *et al*., 2005; Hautaniemi *et al*., 2003; Sauer *et al*., 2005; Wang *et al*., 2001). In addition, Model *et al*. (2002) propose using multivariate statistical process control techniques based on the measurement values of cDNA arrays to detect problematic slides. For oligonucleotide arrays, quality control (QC) reports can be used to assess the quality of the arrays (Affymetrix 2004; Wilson and Miller 2005). Instead of relying on QC reports, Brettschneider *et al*. (2007) introduce

*To whom correspondence should be addressed.

several new quality measures based on probe level and probeset level information to assess the quality of Affymetrix GeneChips (Bolstad *et al*., 2005).

In this article, we develop a tool to identify quality problems based on the quality measures provided in any QC report. Although there is an open debate on whether the measures contained in the QC reports can identify quality problems, QC reports are commonly used to assess the quality of microarrays (Finkelstein, 2005; Landea *et al*., 2005). One common practice is to compare the values of each measure against *ad hoc thresholds*. Another practice is to account for the *similarity* of the measures across arrays and flag those arrays with one or more measures that substantially differ from those of the majority of the arrays. These methods can be implemented using softwares such as `simpleaffy` (Wilson and Miller, 2005) or GeneData Expressionist Refiner 5.0 (GeneData, Basel, Switzerland). The main drawback of these methods is that they are univariate, i.e. they ignore the correlation structure of the QC measures. As a result, these methods can only detect *univariate* outliers, i.e. observations that clearly depart from the bulk of the data in at least one dimension. However, they cannot detect *structural* outliers, i.e. observations that are not outliers in any single dimension, but are nonetheless outliers when multiple dimensions are considered (see Rousseeuw and Leroy 1987 or Model *et al*., 2002 for a discussion of this issue). In our context, we are interested in flagging arrays that violate any of the univariate checks, but also those that are of poor quality only when multiple QC parameters are simultaneously taken into account.

We propose a multivariate quality assessment method for microarrays that is based on the *similarity* of quality measures across arrays, i.e. on the idea of outlier detection. Intuitively, the 'distance' of an array's quality attributes measures the similarity of the quality of that array against the quality of the other arrays. Then, arrays with unusually high distances can be flagged as potentially low quality. Thus, our method computes a single distance measure, the Mahalanobis distance (MD), to summarize the quality of each array. The use of this distance allows us to perform a multivariate analysis of the information in QC reports taking the correlation structure of the quality measures into account. In addition, by using robust estimators to identify the typical quality measures of good-quality arrays, the evaluation is not affected by the measures of outlying arrays. This method can be based on all the quality measures simultaneously, or on subsets of them, which gives one distance value for each subset of parameters in the QC report. We show that the latter approach can be exploited to provide possible explanations of the source of the quality problems. In sum, we bring outlier detection methods widely used in statistics into the quality assessment of microarrays based on QC measures.

The method is specifically designed to identify a small fraction of potentially flawed arrays within a large set of arrays. Thus, it is useful to deal with the common problem that arises in microarray experiments when a small number of the arrays in the batch may have low quality and need to be identified. The method is not appropriate when a large fraction of the arrays or even the entire batch may be flawed due to incorrect laboratory procedures, contaminated samples or other reasons. However, such events can be easily detected using the univariate methods discussed above.

In addition to having a clear statistical foundation, our method has several salient features. First, it takes into account the correlation structure of the quality parameters in the QC report. We show that a multivariate analysis gives substantially richer information than the analysis of each parameter in isolation. Second, it is flexible and useful for any platform as it can be based on any QC report. We illustrate our method using two datasets of Affymetrix GeneChips and the QC reports generated by `simpleaffy` (Wilson and Miller, 2005 and GeneChip Operating Software (GCOS) (Affymetrix, 2005) respectively. However, all the ideas can be applied to other QC reports. Moreover, the user can choose how to group the different quality measures as well as the cutoff lines. Third, since our method is scale-invariant, the analysis does not change if different scales are used for the quality parameters. Last, once the QC reports are produced, our method is computationally light-weight and it summarizes the large number of quality parameters in a way that can be easily visualized and interpreted, which is especially valuable in large microarrays studies.

2 METHODS

2.1 Mahalanobis distances and quality control

Our method to assess the quality of an array is based on the MD (Mahalanobis, 1936) of its quality measures from those of the majority of the arrays. Thus, we refer to it as Mahalanobis Distance Quality Control (MDQC) and it can be described as follows. Let $\boldsymbol{X}_i=(X_{i1},\ldots,X_{ip})$, for $i=1,\ \ldots\ ,\ n$ and $p<n$, be p numeric[1] measures of the QC report for the ith array or p linear combinations of these measures obtained from a Principal Component Analysis (PCA). Let $\boldsymbol{X}$ be n times p matrix containing $\boldsymbol{X}_i$ in its rows, one for each array in the study. Also let $\boldsymbol{M}$ be a p-dimensional row vector and $\boldsymbol{S}$ a p times p positive definite matrix containing estimates of the center and covariance matrix of $\boldsymbol{X}$, respectively. Then, the MD of array i's quality measures from those corresponding to all the arrays characterized by $\boldsymbol{M}$ and $\boldsymbol{S}$, $d(\boldsymbol{X}_i;\boldsymbol{M},\boldsymbol{S})$, is defined by

$$d(\boldsymbol{X}_i;\boldsymbol{M},\boldsymbol{S})=\left[(\boldsymbol{X}_i-\boldsymbol{M})\boldsymbol{S}^{-1}(\boldsymbol{X}_i-\boldsymbol{M})^T\right]^{1/2} \quad (1)$$

Since we want to accurately compute the MD of one array's quality measures to those of other arrays, it is extremely important that outlying arrays do not contaminate our estimates of the center and correlation structure of all the arrays. If they did, then such distances would diverge from their true values simply because the reference point is imprecisely estimated and would not be useful in flagging problematic arrays. Thus, our method relies on *robust* $\boldsymbol{M}$ (location) and $\boldsymbol{S}$ (scatter) estimators to compute the MD defined in Equation (1). In Supplementary Material, we illustrate the relevance of using robust estimators to compute the MDs with a real data example. In this article, we use the *S*-estimator (Lopuhaä, 1989), however, any other robust multivariate location and scatter estimator can be used (e.g. minimum volume ellipsoid or minimum covariance determinant estimators). To increase the finite sample efficiency of these estimators and thus improve the approximation of the MDs distribution, we suggest using an estimator with 25% breakdown point. This is particularly important in studies containing a

[1]Note that the QC report may also contain some non-numeric measure (e.g. 'Spikes Decr' in GCOS QC reports) that should be inspected separately.

small number of arrays where the robust estimators are more unstable and the distributional approximations are less accurate.

The resulting MDs can be used to flag poor-quality arrays as their MDs will be large relative to those of undamaged arrays, i.e. they will be far from the center of the normal arrays. Assuming that $X_1, \ldots, X_p$ are multivariate normal random variables, the squared MDs have an approximate chi-squared distribution with p degrees of freedom. Thus, using the chi-squared distribution we can set a cutoff point to decide if the array is likely to be defective. For example, let $\boldsymbol{X}$ be a 20 times 14 matrix containing the 14 numeric quality measures of the GCOS QC report for 20 arrays in a study. Then, $\boldsymbol{M}$ is a 14-dimensional row vector that estimates the center of $\boldsymbol{X}$ and $\boldsymbol{S}$ is a 14 times 14 matrix that estimates its covariance matrix. Finally, the MDs would be distributed as a chi-squared with 14 degrees of freedom.

2.2 MDQC: different approaches

The most intuitive approach towards MDQC is to compute a single MD for each array based on *all* the quality measures in the QC report. However, this approach suffers from two drawbacks, one statistical and one conceptual. First, the low quality of an array may be reflected by extreme values in only a few of the measures in the report, while other measures may not significantly differ from those corresponding to the bulk of the arrays. Thus, it is possible that the combination of all the quality measures into a single MD 'masks' these outlying observations. Second, even when a single MD can be accurate in identifying poor quality arrays, it provides no information about the potential source of the quality problem. Thus, we recommend alternative approaches to address these issues.

Computing multiple MDs based on different groups with a reduced number of quality measures instead of a single MD based on all of them can help to 'unmask' outlying observations. As a result, the quality attributes of an array would be summarized by as many MDs as groups formed. In addition, QC reports usually contain more than one measure related to the same quality aspect of the array. Thus, grouping complementary measures according to the quality attribute they represent helps to identify possible reasons of the quality problems. We recommend to form these groups using the a priori grouping method, in which the groups are formed on the basis of an a priori interpretation of the quality measures in the report and according to the quality aspect they represent. To illustrate the use of this method, we now use the GCOS QC report for Affymetrix GeneChip arrays (Affymetrix, 2005). The QC measures in this report can be classified into four groups, according to whether they provide information on the quality of the chip and/or the sample, the chip, the sample and the RNA, respectively:

(1) RawQ, Noise, Background, Scale Factor and PercPresent
(2) Corner+, Corner−, Central−
(3) BioB, BioC, BioDN and CreX
(4) GapDH and B-Actin.

Then, we compute four MDs for an array, one for each group of quality measures. The MDQC method based on these a priori groups flags an array as potentially low quality if one or more of its MDs are abnormally high.

These groups contain valuable information about the possible sources of corruption. For example, if only the MD for Group 4 were abnormally high, then this would suggest that the array is defective due to poor RNA quality. However, the groups may sometimes provide less conclusive evidence about the source of the problem as a high MD in one group may manifest itself together with an abnormal MD in other groups. For example, a defective chip that should give large MDs in Group 2 may distort the expression of the housekeeping genes and thus also give large MDs in Group 3. Nevertheless, even in these cases the a priori approach usually allows the researcher to rule out at least some possible sources of corruption.

The MDQC method based on groups is versatile. The a priori approach described above can also be used on QC reports other than that provided by GCOS. This would result in different a priori groups based on the description of each measure in those reports. In addition, in the Supplementary Material we provide two data-driven methods to form the groups that serve as an alternative to the a priori approach. These are the clustering grouping method, which groups the quality measures using clustering analysis, and the loading PCA grouping method, which uses the loadings of a PCA to identify the quality measures that contain similar information. It is important to note that the groups formed using these approaches will vary from one dataset to another, and one may lose the interpretability of the groups provided by the a priori method.

We also propose an alternative approach to unmask low-quality arrays, which we refer to as the global PCA method. It uses PCA to create linear combinations of the original QC parameters, referred to as principal components (PCs), where the PCs retain most of the original variability in the data (Johnson and Wichern, 1999). Thus, the MD can be computed on a single group based on the *reduced* space of the first k PCs ($k<p$), which can help to 'unmask' outlying observations. It is important to note that, in this approach, the formed group does not contain a subset of the original quality measures sharing a common purpose as in the a priori groups. Thus, while this method can also flag low-quality arrays, it gives no indication of the source of the quality problem.

In a PCA, it is usually recommended to standardize the data by the mean and SD of each variable so that variables with a large variance will not dominate the first PCs (Johnson and Wichern, 1999). In addition, as the QC report may contain outlying measures associated with low-quality arrays, it is important to use a *robust* multivariate location and scatter estimator to standardize the variables. Thus, let $\boldsymbol{X}$ be a n times p matrix containing the quality parameters of each array in each row, and let $(\boldsymbol{M},\boldsymbol{S})$ be the robust location and scatter estimator of $\boldsymbol{X}$. Then, the standardized variables are given by $\boldsymbol{Z}_i^T = \boldsymbol{V}^{-1/2}(\boldsymbol{X}_i - \boldsymbol{M})^T$, for $i=1,\ldots,n$, where $\boldsymbol{V}$ is a p times p diagonal matrix containing the robust variance estimates. If $n>p$, a robust PCA can be performed deriving the PCs from a robust location and covariance matrix estimators of $\boldsymbol{Z}$, where $\boldsymbol{Z}$ is the n times p matrix with $\boldsymbol{Z}_i$ in its rows (Croux and Heasbroeck, 2000)[2]. That is, the j th PC is given by $\boldsymbol{Y}_j = \boldsymbol{Z}\boldsymbol{e}_j$, where $\boldsymbol{e}_j$ is the eigenvector corresponding to the jth largest eigenvalue λ_j of the covariance matrix of $\boldsymbol{Z}$, for $j=i,\ldots,p$.

If we use the same robust multivariate estimator to standardize the data, to derive the PCs and to estimate the PC's location and covariance matrix, then the estimated PC's location and covariance matrix become a zero vector, $\boldsymbol{0}_p$, and a diagonal matrix with the eigenvalues λ_j in its diagonal, $\boldsymbol{D}_p = \mathrm{diag}\{\lambda_1,\ldots,\lambda_p\}$, respectively. As a result, the MDs defined in Equation (1) reduces to an Euclidean distance weighted by the eigenvalues of the covariance matrix of Z (Johnson and Wichern, 1999). i.e.

$$d(\tilde{\boldsymbol{X}}_i, \boldsymbol{0}_k, \boldsymbol{D}_k) = \left[\sum_{j=1}^{k} \frac{Y_{ij}^2}{\lambda_j}\right]^{1/2} \quad (2)$$

where $\tilde{\boldsymbol{X}}_i = (Y_{i1}, \ldots, Y_{ik})$ is the set of the first k PCs for the ith array.

In this article, we use the S-estimator with 25% breakdown point in all the steps of the analysis, however, other robust estimators can be used. The scree plot (i.e., the plot of λ_j in decreasing order versus j, for $j=1,\ldots,p$) is used to determine the number k of principal components preserved in the analysis, looking for the 'elbow' or first important bend

[2]Robust PCA methods when $p>n$ can be found in Huber *et al.* (2002).

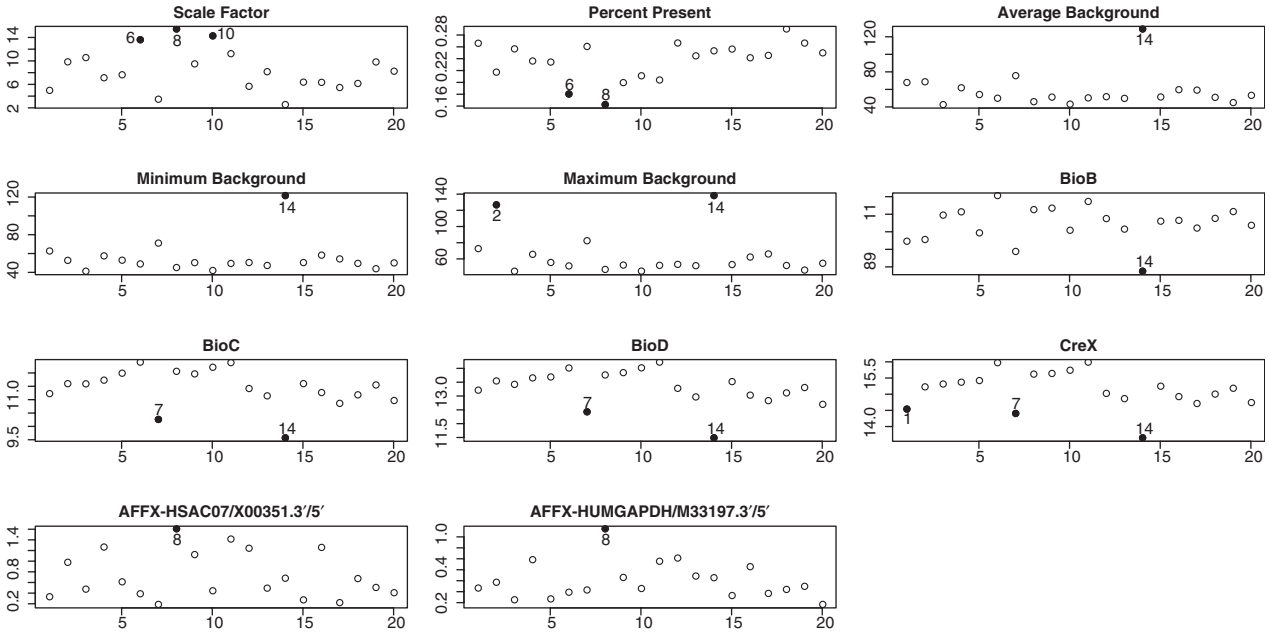

Fig. 1. Quality measures for the acute lymphoblastic leukemia study: univariate analysis. Arrays that do not pass recommended thresholds are identified using solid points. The *x* axis contains the index of each array and the *y* axis shows the different quality measures contained in the report in their original scale.

in the line (Johnson and Wichern, 1999). As before, we flag the array as potentially low quality if its distance defined in Equation (2) is unusually high.

3 RESULTS AND DISCUSSION

We use two case studies to evaluate the performance of the MDQC method. The first dataset is part of an acute lymphoblastic leukemia study described by Ross *et al*. (2003) and contains 20 Affymetrix HG-U133B microarrays. Bolstad *et al*. (2005) and Brettschneider *et al*. (2007) examined the quality of these arrays using histograms of probe-level data, MA-plots and probe-level model (PLM) methods (PLM).[3] According to their quality assessment, array 2 has a strong spatial artifact on the chip and array 14 presents other evidence of poor quality. Because of the existence of such 'ground truths', we first present the comparative analysis based on this small dataset. Our second dataset consists on 201 Affymetrix GeneChip Human Genome U133 Plus 2.0 on RNA isolated from whole blood of patients who have undergone kidney, liver or heart transplants. This dataset is owned by us, allowing us the opportunity to follow up on various aspects and to perform reruns. The analysis of both datasets was performed in R. The corresponding codes and QC reports are available upon request. In addition, the library to implement the MDQC method will soon be publicly available from Bioconductor.

[3]Note that the ID numbers used here are not the same as the ones used in previous references. However, array 2 in this study corresponds to array a in Bolstad *et al*. (2005) and 15 in Brettschneider *et al*. (2007).

3.1 Acute lymphoblastic leukemia study

Figure 1 illustrates the measures of the QC report generated by the R-package `simpleaffy` for 20 microarrays of the acute lymphoblastic leukemia study. A *univariate* analysis of these measures based on Affymetrix recommended thresholds (Affymetrix, 2004) flags some of the arrays as having potentially low quality and are identified using solid points. In particular, arrays 6, 8 and 10 have a 'scale factor' value above usual threshold of 10 and arrays 6 and 8 have also a 'percent present' value below 20. Array 14 has all background measures (average, minimum and maximum background) above usual threshold of 100. In addition, array 2 has a 'maximum background' value above this threshold. The values of the spiked hybridization controls (bioB, bioC, bioD and cre) are low for array 14, though they are always present with increasing signal values as recommended by Affymetrix. Note that arrays 1 and 7 has problems similar to those of array 14, though to a lesser extent. Finally, array 8 shows high values for the last two quality measures corresponding to RNA housekeeping genes. However, both values are below the recommended threshold of 2.

We now compare the previous univariate analysis with a *multivariate* one using MDQC based on *all* the quality measures in the report. Figure 2 shows that using this MDQC approach array 14 is flagged as having potential quality problems and array 2 appears only as a borderline case. Thus, collapsing *all* the quality measures into a single MD downweights array 2's quality problems and masks other outlying observations in the QC report, such as those of arrays 1, 7 or 8. Thus, we study the MDs on groups with a reduced number of variables such as those created by the a priori grouping method and the group of the first principal components. As it was previously discussed,

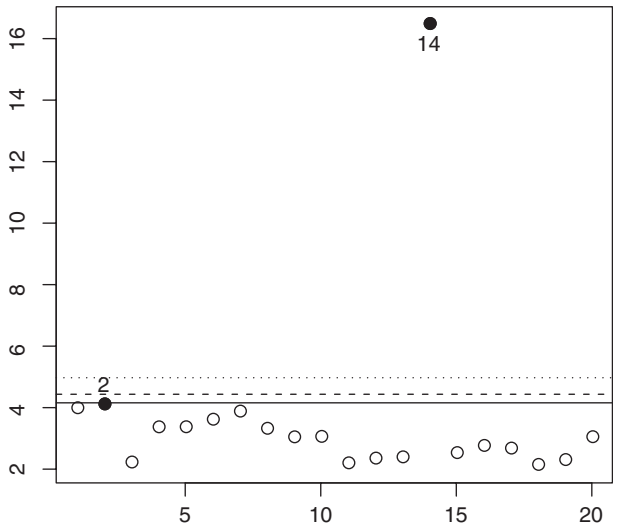

Fig. 2. Results of MDQC based on all measures of the QC report. The MDs (y axis) are computed using the robust S-estimator for each array (x axis). The solid, dashed and dotted lines correspond to the square root of the 90th, 95th and 99th percentile of the chi-squared distribution, respectively. Outlying arrays are identified using solid points.

these alternative methods reduce the possibility of masking outliers and may give information about the potential source of the quality problem.

Using the a priori grouping method, MDQC examines three MDs, one for each of the following three groups:

(1) Scale Factor, % Present, Avg BG, Min BG, Max BG
(2) BioB, BioC, BioD, CreX
(3) AFFX-HSAC07/X00351.3′/5′, GapDH.

Note that the quality measures of Group 2 in Section 2.2 are not available in the `simpleaffy` QC report, thus, there are only three groups to examine. These groups can be used to assess the quality of the chip/sample, the sample and the RNA, respectively. Each plot in Figure 3 shows the MD (y = axis) of each array (x = axis) within each a priori group. The solid, dashed and dotted lines correspond to the square root of the 90th, 95th and 99th percentile of the chi-squared distribution, respectively.

In Group 1, arrays 2 and 14 are *both* flagged as potentially defective, and array 17 as a borderline case. In Group 2, array 1 has an MD exceeding the 99% cutoff and arrays 7 and 14 have MDs exceeding the 95% cutoff line. Finally, array 8 is the only one flagged in Group 3. Thus, the MDs based on groups of lower dimension flag both arrays 2 and 14, which is consistent with the results in Bolstad *et al*. (2005) and Brettschneider *et al*. (2007). In addition, arrays 1 and 8 are flagged as potentially low quality and arrays 7 and 17 as borderline quality. Moreover, based on the interpretability of the groups, the problems in array 2 are most likely due to defects in the chip as this array is only identified in Group 1. Similarly, since arrays 1 and 14 are flagged in Group 2, their low quality is most likely due to low quality of the sample. Note that although array 14 is also flagged in Group 1, this can still be due to quality problems in the sample. Finally, array 8 is flagged only in Group 3, suggesting potential problems in the RNA quality. In the Supplementary Material, we also analyze this dataset using MDQC based on the clustering grouping method and the loading PCA grouping method. The groups formed by these two data-driven methods almost validate the a priori grouping described above, and thus the results are similar to those reported in Figure 3.

Comparing the previous multivariate analysis with the univariate one, it is important to note that besides the 'maximum background', all other quality measures for array 2 are similar to those of the other arrays (see all plots in Fig. 1). Thus, without 'maximum background', the univariate analysis does not identify this array as having quality problems. However, the top-right plot in Figure 3 shows that MDQC using the a priori grouping method still flags this array even when the 'maximum background' is not included in the analysis. This example illustrates that a *multivariate* analysis can flag a problematic array that a univariate analysis cannot detect. In addition, note that array 14 is flagged by both the univariate and the multivariate analyses. However, although the univariate analysis suggests that array 14 is more problematic than array 2, i.e. many of its quality measures are outlying, the MDQC using the a priori grouping method ranks array 2 as having lower quality (the MD of array 2 is larger than that of array 14 in Group 1). Thus, our MDQC method not only flags unusual arrays but also ranks them in a way that is not evident from the univariate analysis.

Finally, we examine the performance of MDQC using the global PCA method to reduce the dimensionality of the data. Using the scree plot, we retain $k = 4$ principal components in this analysis (see Supplementary Material). Figure 4 shows the results of the MDQC when a single MD is calculated based on the first four PCs derived from a robust PCA based on robustly standardized data (see Section 2.2). We note that this approach still flags arrays 2, 8 and 14 as having potential quality problems. However, the first two appear only as borderline cases. In addition, arrays 1, 7 and 17 are still masked using this method.

In sum, all three grouping approaches of MDQC (i.e. all variables, the a priori grouping method and the global PCA method) identify the problematic arrays 2 and 14 that were previously detected by Bolstad *et al*. (2005) and Brettschneider *et al*. (2007). However, the a priori grouping method outstands the problem of array 2, unmasks other potentially low-quality arrays and provides possible explanations of the quality problems.

3.2 Transplantation study

We use this dataset to illustrate the performance of our method in a large study with the ability to re-run potentially low-quality arrays. The analysis is based on the 14 numerical quality measures contained in the GCOS QC report for each array (see Section 2.2). Based on the MDQC analysis of the 201 original arrays, budget and sample material limitations, 22 arrays flagged with potentially low quality have been re-run. While the diagnostic of the original set is based on an analysis that does

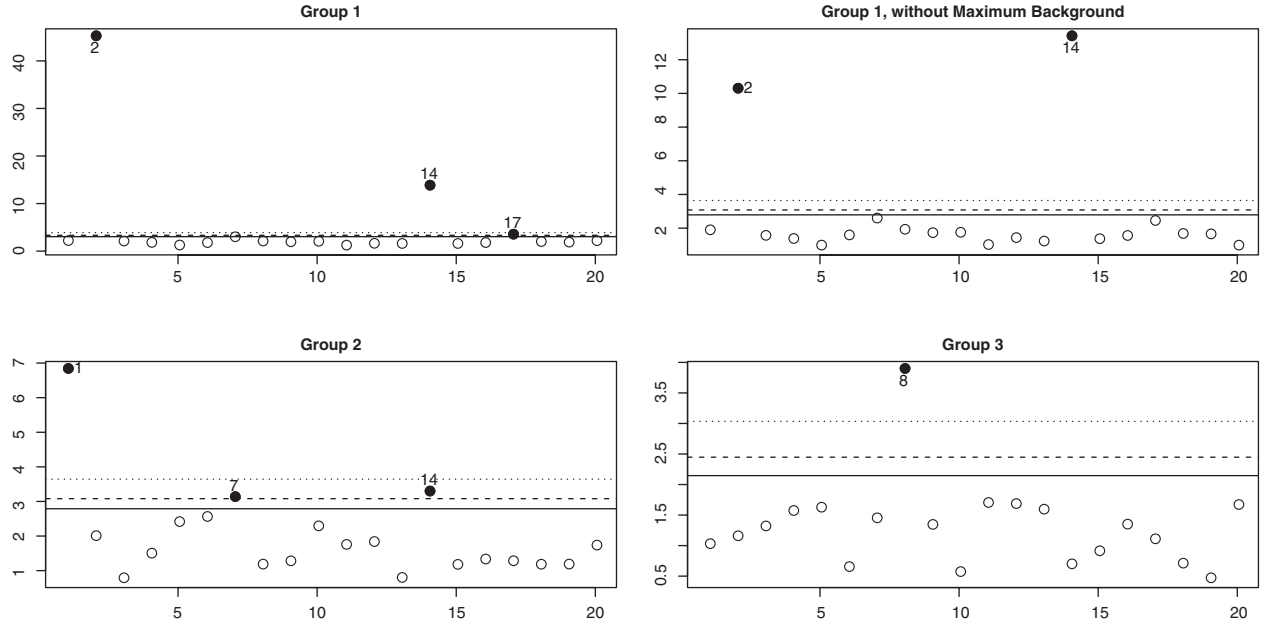

Fig. 3. Results of MDQC using the a priori grouping method. The MDs (*y* axis) within each group are computed using the robust *S*-estimator for each array (*x* axis). The scale of the *y* axis varies from one case to another. The solid, dashed and dotted lines correspond to the square root of the 90th, 95th and 99th percentile of the chi-squared distribution, respectively. Outlying arrays are identified using solid points.

not include the re-runs, to simplify the exposition, we include the results of all the arrays in the same plot. Thus, Figure 5 shows the MDs (*y* axis) of the 223 arrays within each of the four a priori groups defined in Section 2.2. We recall that Groups 1–4 provide information on the quality of the chip and/or sample, the chip, the sample and the RNA, respectively. Solid points are used to identify the 22 arrays that were re-run, solid triangles for the re-runs and open triangles for those that could not be re-run due to the lack of additional sample material. In addition, the array's IDs contain two numbers: the first one corresponds to the patient ID and the second one to the number of months after transplant. The re-run arrays are labeled with an R after this numeric ID. For example, 21-4 is the ID for the array corresponding to patient 21 at 4 months after transplant, and 21-4R is its re-run. The solid, dashed and dotted lines correspond to the square root of the 90th, 95th and 99th percentile of the chi-squared distribution, respectively.

Figure 5 shows the plots of the MDs in each of the four groups. Although our method identifies several outlying arrays, we focus our discussion on the subset of those arrays with the highest MDs that we were able to re-run: 21-4, 17-6, 25-5, 302-7, 36-6, 21-2, 21-3, 5 arrays of patient 13 and 10 arrays of patient 317.

Arrays 21-4, 17-6, 25-5 and 302-7 have outlying MDs in Groups 1–3, but not in Group 4. Thus, the quality problem may come from the chip, the sample, or both, but not from the RNA quality. To identify the source of the quality problem, we re-run them using the same sample material but a new chip. As the MDs of the re-run for array 21-4 (21-4R) are below the thresholds in all four groups, we conclude that the original chip was damaged. In contrast, the MDs of the re-runs of arrays 17-6, 25-5 and 302-7 continue to be flagged as outliers (data not shown),

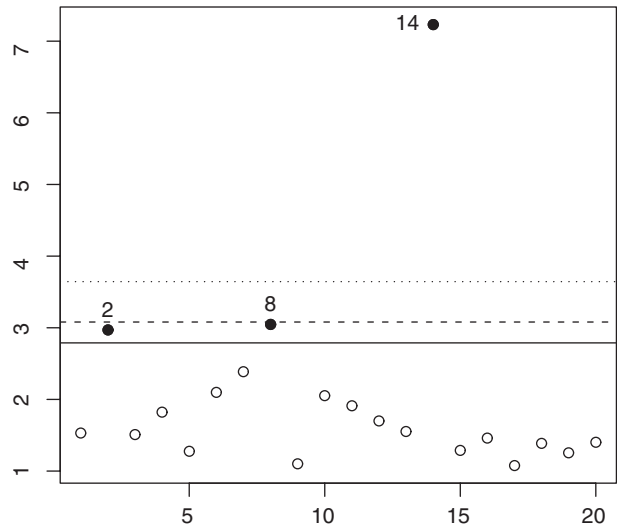

Fig. 4. Results of MDQC using the global PCA method. The MDs (*y* axis) are computed on the first four principal components for each array (*x* axis). The solid, dashed and dotted lines indicate the square root of the 90th, 95th and 99th percentile of the chi-squared distribution, respectively. Outlying arrays are identified using solid points.

suggesting that the original chips were not defective. The re-runs of these arrays using new sample material give MDs that are below the thresholds in all four groups. Thus, we conclude that these arrays suffered from low-quality sample material. Further, the array 36-6 has outlying MDs in Groups 1 and 2, while arrays 21-2 and 21-3 have outlying MDs in Groups 1 and 3. We re-run

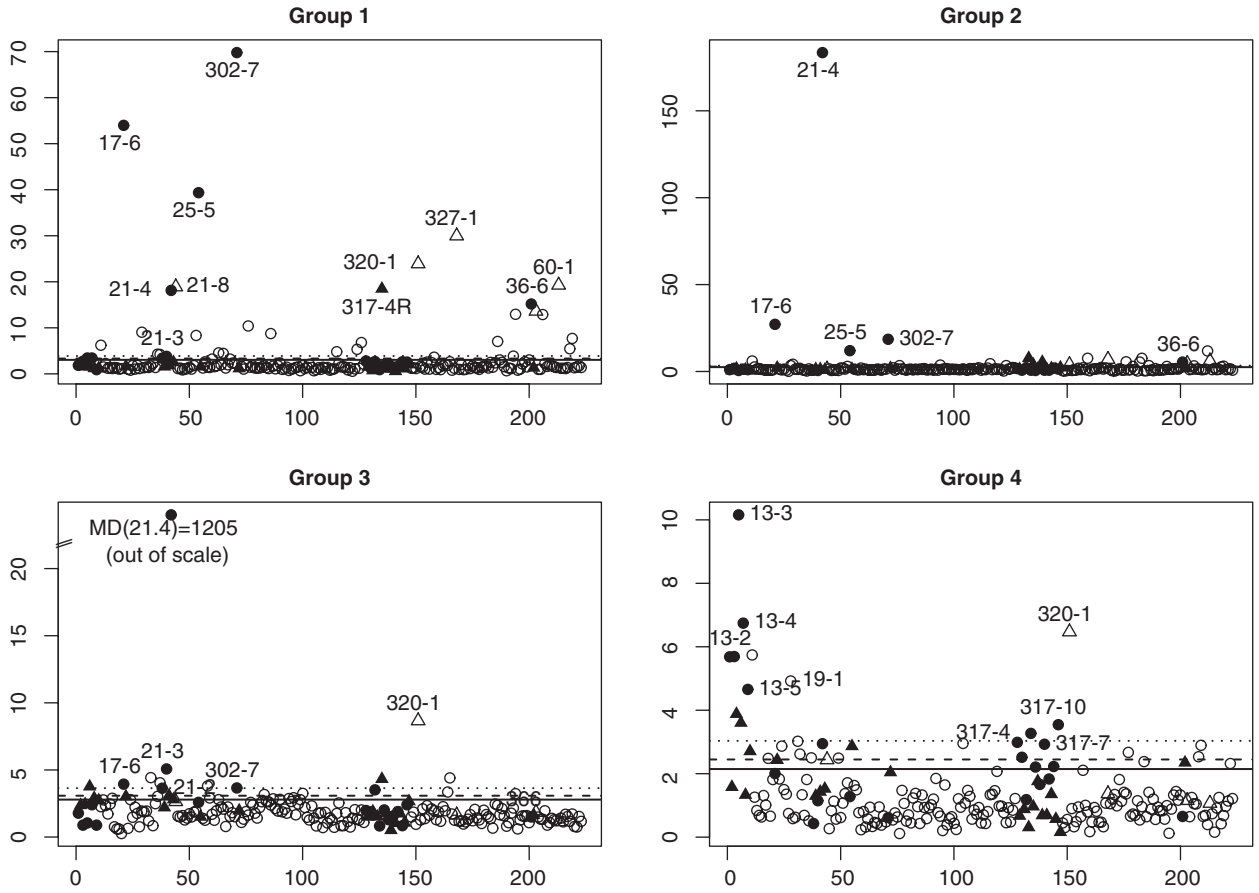

Fig. 5. Results of MDQC using a priori groups for the transplantation study. The scale of the *y* axis varies from one case to another. The solid, dashed and dotted lines correspond to the square root of the 90th, 95th and 99th percentile of the chi-squared distribution, respectively. Arrays that were re-run are identified with solid points, the results of the re-runs with solid triangles, and those arrays that could not be re-run (not enough RNA) with open triangles.

these arrays using both new chips and new sample material and our method ceased to flag them as low-quality arrays.

We additionally identify a set of arrays with unusual indicators of RNA quality measures (see Group 4). These arrays correspond to patients 13 and 317, although those for the latter patient are borderline cases. As the arrays of both patients were originally run in the same batch, these unusual values can correspond to either a batch effect or a quality problem in the RNA. We re-run each of these arrays in different batches when RNA was still available. The re-run arrays have MDs of the quality measures in Group 4 that are similar to those of the rest of the arrays.

We further use other quality assessment methods to assess the performance of MDQC. As some of these methods are computationally intensive or difficult to visualize in large studies, we select 12 potentially bad arrays and 10 good arrays based on the MDQC diagnostic and the inspection of the image files. We examine the histogram of probe-level data, the MA-plots and perform a PLM QC assessment (Bolstad *et al.*, 2005), including the inspection of array pseudo-images, RNA-degradation plots, relative log expressions (RLE) and normalized unscaled standard errors (NUSE). Here, we briefly describe the last quality measure and present its results for our data. The conclusions are similar for the other measures (see Supplementary Material).

The box plots in Figure 6 show the NUSE for the selected arrays. These errors are the standard error between probe intensities within a probe set for each array, normalized by dividing all values of a particular probe set by the median standard error for that probe set across arrays (Bolstad *et al.*, 2005). Their box plots are expected to be small and centered at one reflecting a small variability within the probe sets of an array. It is noticeable that those arrays identified by MDQC as having potential quality problems are also flagged by this quality measure (similar results are found using other diagnostic plots of PLM available in Supplementary Material). Their boxes are larger and in most cases not centered at one, indicating the existence of more outlying probes in those arrays with a larger variability within probe sets than in other arrays. In sum, the MDQC method is comparable in its effectiveness as the PLM method. Its main advantage over the PLM method is that it is not computer memory intensive, and is much more suitable for assessing the quality of a large number of arrays.

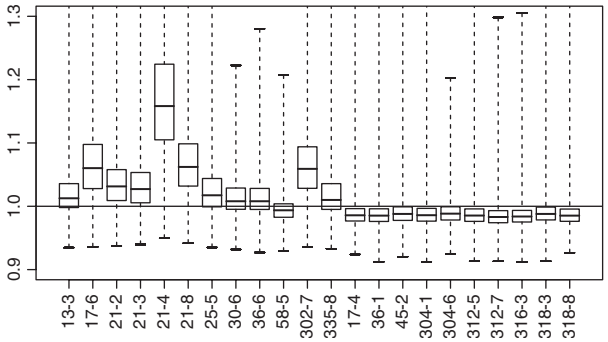

Fig. 6. NUSE plots for 12 low-quality and 10 good-quality arrays.

4 CONCLUSION

We propose a new method (MDQC) to identify potentially low-quality arrays. Its advantage is that it has a clear statistical foundation, it uses the correlation structure of the various QC measures, it is easy to apply, and it is computationally light-weight. These properties make MDQC a useful diagnostic technique suitable for large datasets. MDQC performs a robust multivariate analysis of the quality measures provided in the QC report while taking into account their correlation structure. More precisely, the method first identifies the typical quality measures of valid arrays using robust estimators of the center and correlation structure. It then uses the MD based on these estimators to flag arrays with quality measures that are far from those of valid ones. We show that a multivariate analysis gives substantially richer information than the inspection of individual measures in isolation. Moreover, the method gives a simple way to compare the quality across arrays that is useful to rank them according to their quality and to flag those likely to be defective. Finally, we show that computing these distances on subsets of the quality measures in the report, instead of on all of them, may increase the method's ability to detect unusual arrays. In our case studies, we find that the a priori grouping method and the global PCA identify almost the same set of multivariate outliers. However, using the a priori method, the interpretability of the groups may be used to provide useful information about the likely source of potential quality problem.

ACKNOWLEDGEMENTS

Special thanks go to John Quackenbush and Hernan Ortiz Molina for helpful comments and to Biomarkers in Transplantation's members, Alice Mui, Janet McManus, Pooran Quasimi, David Lin, Axel Bergman and Martha Casey-Knight, for useful discussions. We would also like to thank Tim Triche, Jonathan Buckley and Betty Schaub from MAC lab for processing the Transplantation samples and discussing its quality measures with us. G. C. F. is supported by an IBM Institutes of Innovation Fellowship Award.

Conflict of Interest: none declared.

REFERENCES

Affymetrix (2004) GeneChip Expression Analysis. Data Analysis Fundamentals. Affymetrix, Inc., Stanta Clara, CA.

Affymetrix (2005) GeneChip Operating Software. Technical Manual. Rev. 5.

Bolstad,B.M. *et al*. (2005) Quality assessment of Affymetrix GeneChip data. In Gentleman,R. *et al*. (ed.) *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Springer, New York.

Brettschneider,J. *et al*. (2007) Quality assessment for short oligonucleotide arrays. Forthcoming in *Technometrics*.

Bylesjö,M. *et al*. (2005) MASQOT: a method for cDNA microarray spot quality control. *BMC Bioinformatics*, **6**, 250.

Croux,C. and Haesbroeck,G. (2000) Principal component analysis based on robust estimators of the covariance or correlation matrix: influence functions and efficiencies. *Biometrika*, **87**, 603–618.

Finkelstein,D. (2005) Trends in the quality of data from 5168 oligonucleotide microarrays from a single facility. *J. Biomol. Tech.*, **16**, 143–153.

Hautaniemi,S. *et al*. (2003) A novel strategy for microarray quality control using Bayesian networks. *Bioinformatics*, **19**, 2031–2038.

Huber,M. *et al*. (2002) A fast method for robust principal components with applications to chemometrics. *Chemometrics and Intell. Lab. Syst.*, **60**, 101–111.

Johnson,R.A. and Wichern,D.W. (1999) *Applied Multivariate Statistical Analysis*. 4 edn. Prentice Hall, Upper Saddle River, New Jersey.

Landea,J.D. *et al*. (2005) Gene expression profiling in murine obliterative airway disease. *Am. J. Transplan.*, **5**, 2170–2184.

Lopuhaä,H.P. (1989) On the relation between S-estimators and M-estimators of multivariate location and covariance. *Ann. Stat.*, **17**, 1662–1683.

Mahalanobis,P. C. (1936) On the generalized distance in statistics. *Proc. Natl. Inst. Sci. India*, **2**, 49–55.

Model,F. *et al*. (2002) Statistical process control for large scale microarray experiments. *Bioinformatics*, **18**, 155–163.

Ross,M.E. *et al*. (2003) Classification of pediatric acute lymphoblastic leukemia by gene expression profiling. *Blood*, **102**, 2951–2959.

Rousseeuw,P.J. and Leroy,A. (1987) *Robust Regression and Outliers Detection*. Wiley, New York.

Sauer,U. *et al*. (2005) Quick and simple: quality control of microarray data. *Bioinformatics*, **21**, 1572–1578.

Wang,X. *et al*. (2001) Quantitative quality control in microarray image processing and data acquisition. *Nucleic Acids Res.*, **29**, e75.

Wilson,C. and Miller,C. (2005) Simpleaffy: a BioConductor package for Affymetrix quality control and data analysis. *Bioinformatics*, **21**, 3683–3685.

Zhang,W. *et al*. (2004) *Microarray Quality Control. Hoboken,* John Wiley and Sons, Inc., New Jersey.