Microarray data quality control improves the detection of differentially expressed genes

Audrey Kauffmann *, Wolfgang Huber

EMBL-European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, CB10 1SD, UK

Abstract

Microarrays have become a routine tool for biomedical research. Data quality assessment is an essential part of the analysis, but it is still not easy to perform objectively or in an automated manner, and as a result it is often neglected. Here, we compared two strategies of array-level quality control using five publicly available microarray experiments: outlier removal and array weights. We also compared them against no outlier removal and random array removal. We find that removing outlier arrays can improve the signal-to-noise ratio and thus strengthen the power of detecting differentially expressed genes. Using array weights is similarly effective, but its applicability is more limited. The quality metrics presented here are implemented in the Bioconductor package arrayQualityMetrics.

© 2010 Elsevier Inc. All rights reserved.

Introduction

Microarrays are widely used for applications such as molecular profiling [1], identification of new drug targets [2], discovery of biomarkers [3] or genome annotation [4]. The two major analysis approaches for microarray data are testing for differentially expressed genes between two or more conditions, and examining gene sets enriched in regulated genes. Numerous statistical methods are available for these tasks (e.g. [5–7]), and methods for controlling type I errors—for example, false discovery rates—are well established. However, there has been less attention on type II errors: how many discoveries are missed? In particular, the importance of data quality assessment and control for getting good sensitivity is sometimes underappreciated.

The term quality is not easily defined; in the context of microarray data, we use it to describe two potentially independent concepts: (i) the suitability of the data to answer the experimenter’s original question and (ii) the usefulness of the data for subsequent, integrative analyses. Great efforts have been made to provide tools to the community [8–10] to assess the quality of a microarray experiment and in particular, to identify outlier samples [11,12]. We recently presented a software to calculate a comprehensive set of widely used metrics that assesses data quality and helps identify apparent outlier arrays [13]. We anticipate multiple use cases for this software: an experimentalist can use it to assess and improve the experimental protocols, choose a platform or decide when to repeat certain parts of the experiment; a bioinformatician or statistical collaborator can use it to decide whether and how to proceed with an analysis; a microarray core facility can use it to decide whether their product is fit for delivery to the customer. It may also be useful to integrative biologists analysing public data, in helping them choose which experiments and which arrays of an experiment to consider.

Quality problems can stem from many different sources [14]. For instance, the hybridisation step can introduce uneven fluorescence on the chip [15], whereas differences in RNA quality can cause variable intensity distributions [16]. Quality problems can affect the data at different levels. For example, inappropriate experimental design may affect the dataset as a whole; poor probe design or probe mis-annotation will affect all data from a particular probe; sample mislabeling or inappropriate sample treatment may result in individual outlier arrays. Often, diagnostic tools can help track down these effects in the data.

Quality assessment can be performed before or after data preprocessing to answer different questions. Before normalisation, it helps the user choose the most appropriate preprocessing steps in order to correct the most dominant effects. It can also serve to assess alternative experimental protocols. After preprocessing, quality assessment can help determine the effectiveness of the chosen preprocessing steps and more importantly the suitability and usability of the data for the biological analysis.

Here, we focus on quality metrics aimed at identifying outlier arrays. They can be divided into two categories: relative and absolute. Relative quality metrics compare each array’s intensities against those of other arrays within the dataset. Absolute quality metrics make use of internal controls, spike-in, or variability among replicate probes and are in principle independent of the behaviour of other arrays.

In this paper, we review approaches to handling outlier arrays. We analysed five publicly available datasets using the same workflow. The
results demonstrate that carefully controlling the quality of a dataset and removing or weighting outlier arrays improves the sensitivity as measured by the statistical power to detect differential expression at a given level of type I error—and the biological sensitivity of the analysis.

Example datasets

We generated reports from the arrayQualityMetrics package on five datasets from the ArrayExpress database [17]. These datasets met two criteria: (i) a simple experimental design, with only two groups of samples and (ii) the array platform used was Affymetrix GeneChip. Table 1 lists these datasets.

Analysis

We compared four strategies of data analysis (details of Strategies 2 and 4 are explained in Section 4):

- **Strategy 1** No outlier removal
- **Strategy 2** Outlier removal guided by arrayQualityMetrics
- **Strategy 3** Removing random arrays (same number of arrays as in Strategy 2)
- **Strategy 4** Array weights using the function arrayWeights [23] from the limma Bioconductor package

For each of these strategies, the analysis pipeline was the following: (i) the datasets were imported into Bioconductor using the package ArrayExpress [24], (ii) the function rma (robust multi-array average) from the package affy [25] was used for background correction, quantile normalisation and probe set summarization and (iii) differentially expressed genes were identified using the moderated t-test from the limma package [5]. P-values were adjusted for multiple testing with the Benjamini and Hochberg method for control of the false discovery rate, and genes were called significant when the adjusted P-value was <0.01. Additionally, for two of the five datasets, in Strategies 1 and 2, we perform a KEGG pathway enrichment analysis for two datasets using the function gseattperm from the package GSEABase [26].

The full quality reports for each dataset before and after normalisation and the R code of the analysis of the first dataset is provided in the Supplementary Material.

Outlier detection

Three visualisations are particularly helpful for assessing whether an array differs substantially from other arrays in the same dataset. These are the MA-plot, the boxplot of the log-ratios and the heatmap plot of the distance between arrays, which are included in the arrayQualityMetrics reports. In order to formalise and automate the plot of the distance between arrays, which are included in the Supplementary Material, we discovered that there is a confounding effect of expression analysis on datasets where the same number of arrays was expressed between the two biological conditions. After further inspection, we discovered that there is a confounding effect of treatment or experiment date, which would need to be discounted in Table 1

<table>
<thead>
<tr>
<th>ArrayExpress ID</th>
<th>Ref.</th>
<th>Organism</th>
<th># Chips</th>
<th>Study</th>
</tr>
</thead>
<tbody>
<tr>
<td>E-GEOI-3419</td>
<td>[18]</td>
<td>H. sapiens</td>
<td>16</td>
<td>Hair follicle bulge cells enrichment in stem cells</td>
</tr>
<tr>
<td>E-GEOI-7258</td>
<td>[19]</td>
<td>M. musculus</td>
<td>17</td>
<td>Urethane treatment in bronchoalveolar lavage cells</td>
</tr>
<tr>
<td>E-GEOI-10211</td>
<td>[20]</td>
<td>M. musculus</td>
<td>10</td>
<td>Epithelial cell response to sendai virus infection</td>
</tr>
<tr>
<td>E-MEXP-170</td>
<td>[22]</td>
<td>H. sapiens</td>
<td>8</td>
<td>Sulforaphane treatment of colon cells</td>
</tr>
</tbody>
</table>

the others—arrayQualityMetrics computes these summary metrics for each array: the absolute value of M (log-ratio) to represent its MA-plot, the mean and interquartile range (IQR) of M to represent its boxplot, the mean of the L1-distances (mean absolute difference of M) to all other arrays to represent its position in the heatmap plot. Outlier arrays, with respect to these metrics, are then identified by an outlier detection algorithm borrowed from R’s boxplot function. We classify an array as an outlier if it is detected as such for two or more of the metrics above. This is “Strategy 2” in Analysis.

The array weighting strategy implemented in limma—“Strategy 4” in Analysis—is based on the empirical reproducibility of the gene expression measures from replicate arrays. A heteroscedastic linear model is fit to expression values, resulting in array level weights arrayweights.

The results of the outlier detection method are shown in Table 2. Briefly, we identified one outlier for E-MEXP-170, two for E-MEXP-774, E-GEOI-3419 and E-GEOI-10211, and three for E-GEOI-7258. The outlier arrays are indicated in the second column with the identification number given in the quality assessment report. The arrays detected as outliers are also the ones getting the lowest weights using the limma method as shown in the second and third columns of Table 2.

Differential expression results

The differential gene expression analysis showed that two of the methods—arrayWeights from limma and outlier detection from arrayQualityMetrics—resulted in a significant increase in the number of differentially expressed genes for the type I error. The results are summarized in Table 2 and Fig. 1. The effect was moderate in the case of E-GEOI-3419 and strong for E-GEOI-10211. E-MEXP-170 caught our attention as it is unlikely for over 4000 genes to be differentially expressed between the two biological conditions. After further inspection, we discovered that there is a confounding effect of treatment or experiment date, which would need to be discounted in order to make any biologically relevant interpretation of the data, both with or without outlier removal.

For datasets E-GEOI-10211 and E-MEXP-774, the outlier removal and weighting strategies output similar numbers of differentially expressed genes. The numbers are more different for the other three experiments; however with the exception of experiment E-GEOI-3419 the outlier removal strategy identifies almost all genes detected using the weighting method (Fig. 2).

Comparison against random array removal

We compared the outlier removal and array weighting strategies with the arbitrary removal of arrays. We performed the differential expression analysis on datasets where the same number of arrays was removed randomly. The results are presented in Fig. 3. Compared with using all arrays, removal of random arrays leads to a loss of power and
hence fewer genes are detected. In contrast, outlier removal and array weighting increased the numbers of differentially expressed genes.

Geneset enrichment analysis

Although the above strategies allowed us to identify more differentially expressed genes, these additional genes may not necessarily be biologically relevant. To check whether removing outliers results in better biological sensitivity, we performed an analysis of the enrichment of KEGG pathways for the two example datasets E-GEOD-3419 and E-GEOD-7258. For each experiment, five of the most highly enriched pathways are listed in Table 3. In both datasets, these pathways are relevant to the biological situation studied. Their enrichment p-values are smaller after removing outlier arrays, confirming that it improved the biological relevance of the analysis.

Other approaches for outlier detection

Finally, we compared our outlier-detection approach with three other methods: (i) generalized extreme studentized deviate (GESD) [27], (ii) the method of Hampel in which the sample median is used for location estimation, and the median absolute deviation is used for scale [28] and (iii) Rousseeuw’s rule, with the midpoint of the shortest half sample used as location estimator, and the length of this shortest half sample used as scale estimator [29].

For four out of the five datasets, we obtain similar results to the GSED method. GSED detected an additional outlier in dataset E-GEOD-3419, which was significant in only one of the metrics we use. The methods of Hampel and Rousseeuw gave the same outliers in 3 of 5 cases, whereas in two cases (E-GEOD-3419, E-GEOD-10211) only one sample was detected as an outlier instead of two (see Table 4).

Discussion

Correct use of microarray analysis can lead to good adjusted p-values, clustering, geneset enrichment results; however, many important genes can be missed if poor quality arrays are included in the dataset. Although using array weights might, in theory, be more...
efficient than simply removing outliers, we showed here—at least for
the datasets tested—that both methods perform equally well, and
provided an improvement compared to keeping all samples. The same
arrays spotted as outliers by arrayQualityMetrics were the ones
getting low weights with arrayWeights.

An outlier array can be interpreted as being of low quality, and this
is the reason why its presence would add noise and impair the
statistical and biological significance of the analysis. However, an
array can be detected as an outlier because of a real biological
property of the sample or an intentional protocol peculiarity. This
makes it difficult (and, in general, not advisable) to automate the
removal of outliers, as depending on the context, some “outliers”
might be of great interest for the analysis. The comprehensive report
offered by arrayQualityMetrics—thanks to the visualizations it pro-
vides—will help the user understand why a particular array is
identified as an outlier. We recommend manual inspection to decide
whether or not an array should then be removed. Such inspection can
also provide useful feedback to improve experimental protocols.

Acknowledgments

AK is funded by EU FP6 (EMERALD, Project no. LSHG-CT-2006-
037686). We thank Richard Bourgon for discussions and Nicholas
Luscombe for critical comments on the manuscript.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in

Table 3

<table>
<thead>
<tr>
<th>Pathway name</th>
<th>Genes</th>
<th>p-value when removing outliers</th>
<th>p-value when all arrays</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pyrimidine metabolism</td>
<td>37</td>
<td>$&lt;10^{-3}$</td>
<td>0.701</td>
</tr>
<tr>
<td>Base excision repair</td>
<td>17</td>
<td>0.001</td>
<td>0.542</td>
</tr>
<tr>
<td>DNA replication</td>
<td>19</td>
<td>0.003</td>
<td>0.451</td>
</tr>
<tr>
<td>Cell cycle</td>
<td>69</td>
<td>0.009</td>
<td>0.387</td>
</tr>
<tr>
<td>TGF-beta signaling pathway</td>
<td>48</td>
<td>0.009</td>
<td>0.558</td>
</tr>
<tr>
<td>Pentose phosphate pathway</td>
<td>13</td>
<td>0.003</td>
<td>0.588</td>
</tr>
<tr>
<td>Fructose and mannose metabolism</td>
<td>28</td>
<td>0.003</td>
<td>0.326</td>
</tr>
<tr>
<td>Biosynthesis of steroids</td>
<td>20</td>
<td>0.003</td>
<td>0.012</td>
</tr>
<tr>
<td>Oxidative phosphorylation</td>
<td>44</td>
<td>0.003</td>
<td>0.299</td>
</tr>
<tr>
<td>Starch and sucrose metabolism</td>
<td>16</td>
<td>0.003</td>
<td>0.317</td>
</tr>
</tbody>
</table>

Table 4

Comparison of four outlier detection methods. The one currently implemented in arrayQualityMetrics, based on the boxplot, the generalized extreme studentized deviate (GESD), the method of Hampel using the median absolute deviation and the one from Rousseeuw using the shorth. The results overlap mostly.

<table>
<thead>
<tr>
<th>ArrayExpress ID</th>
<th>arrayQuality Metrics</th>
<th>GESD</th>
<th>Hampel</th>
<th>Rousseeuw</th>
</tr>
</thead>
<tbody>
<tr>
<td>E-GEOD-3419</td>
<td>6, 12</td>
<td>3, 6</td>
<td>12</td>
<td>12</td>
</tr>
<tr>
<td>E-GEOD-7258</td>
<td>7, 15, 16</td>
<td>7, 15, 16</td>
<td>7, 15, 16</td>
<td>7, 15, 16</td>
</tr>
<tr>
<td>E-GEOD-10211</td>
<td>2, 7</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>E-MEXP-774</td>
<td>4, 17</td>
<td>4, 17</td>
<td>4, 17</td>
<td>4, 17</td>
</tr>
<tr>
<td>E-MEXP-170</td>
<td>6</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
</tbody>
</table>
References


