BIOINFORMATICS ORIGINAL PAPER Vol. 30 no. 10 2014, pages 1392–1399
doi:10.1093/bioinformatics/btu027

Gene expression

Advance Access publication January 21, 2014

IQRray, a new method for Affymetrix microarray quality control, and the homologous organ conservation score, a new benchmark method for quality control metrics

Marta Rosikiewicz[1,2,*] and Marc Robinson-Rechavi[1,2]

[1]Department of Ecology and Evolution, University of Lausanne and [2]Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland

Associate Editor: Janet Kelso

ABSTRACT

Motivation: Microarray results accumulated in public repositories are widely reused in meta-analytical studies and secondary databases. The quality of the data obtained with this technology varies from experiment to experiment, and an efficient method for quality assessment is necessary to ensure their reliability.

Results: The lack of a good benchmark has hampered evaluation of existing methods for quality control. In this study, we propose a new independent quality metric that is based on evolutionary conservation of expression profiles. We show, using 11 large organ-specific datasets, that IQRray, a new quality metrics developed by us, exhibits the highest correlation with this reference metric, among 14 metrics tested. IQRray outperforms other methods in identification of poor quality arrays in datasets composed of arrays from many independent experiments. In contrast, the performance of methods designed for detecting outliers in a single experiment like Normalized Unscaled Standard Error and Relative Log Expression was low because of the inability of these methods to detect datasets containing only low-quality arrays and because the scores cannot be directly compared between experiments.

Availability and implementation: The *R* implementation of IQRray is available at: ftp://lausanne.isb-sib.ch/pub/databases/Bgee/general/IQRray.R.

Contact: Marta.Rosikiewicz@unil.ch

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

Received on June 21, 2013; revised on December 9, 2013; accepted on January 14, 2014

1 INTRODUCTION

Thousands of microarray results are available in public repositories such as the Gene Expression Omnibus (Edgar *et al.*, 2002) and ArrayExpress (Brazma *et al.*, 2003). This wealth of expression data covering many organisms, tissues, developmental stages, diseases and treatments is now available for meta-analyses, system biology studies and use in secondary databases. Combining results from several independent studies allows improved detection of differentially expressed genes and analysis of biological pathways and co-expression networks (Tseng *et al.*, 2012). These vast transcriptomic resources have been also extensively used for functional gene annotation and reanalysis of lists of candidate genes obtained with high-throughput experiments. These tasks are facilitated by large secondary databases such as Genevestigator (Hruz *et al.*, 2008), BioGPS (Wu *et al.*, 2013), the Gene Expression Atlas (Kapushesky *et al.*, 2010) or Bgee (Bastian *et al.*, 2008) that allow mining of many microarray experiments at the same time. Additionally, there are many more specialized databases, which, for example, collect data only from a selected species (Dash *et al.*, 2012; Le Crom *et al.*, 2002) or for diseases (Hebestreit *et al.*, 2012; Rhodes *et al.*, 2007), or provide resources for more specific analyses, such as COXPRESdb for studying co-expressed genes (Obayashi *et al.*, 2013) or TiSGeD for the analysis of tissue-specific gene expression (Xiao *et al.*, 2010).

The quality of results that can be derived from meta-analysis and database searches of microarray data depends directly on the quality of the arrays themselves. Despite this, there are still no objective criteria that allow discrimination between low- and high-quality arrays (Brettschneider *et al.*, 2008; Wilkes *et al.*, 2007). By *quality* in this context, we understand the agreement between microarray gene expression-level estimations and true gene expression profile. A proper quality control procedure should report whether the amount of biological information captured by the microarray experiment is sufficient to answer a particular biological question.

The most common approach to quality assessment of microarray results makes the assumption that the majority of arrays in each dataset is of good quality, and various parameters serve to identify outlier arrays (Beisvag *et al.*, 2011; Bolstad *et al.*, 2005). This task is relatively easy when the dataset is large, but the analysis often lacks power when only a few arrays are analyzed together. Moreover, in the case of a dataset with a high proportion of low-quality arrays (e.g. microarray results derived from degraded RNA samples), the good-quality arrays might actually be removed to decrease variability of the dataset. This approach is also powerless to detect datasets composed only of low-quality samples.

Most of the existing methods of quality control of Affymetrix microarrays use specific features of these chips, such as control probes, pairs of mismatch (MM) and perfect match (PM) probes, or the fact that there are many probes gathered in probe sets targeting the same transcript. The degradation of transcripts starts from the 3′ end; thus, the ratio of abundance of 5′–3′ ends serves in molecular biology practice as an indicator of

*To whom correspondence should be addressed.

RNA quality. Affymetrix chips contain special probe sets designed to bind 5′ or 3′ ends of actin and GAPDH transcripts, and the ratio of their signal level is used as a quality parameter (Affymetrix, 2001). Another method that measures RNA degradation takes advantage of the fact that the expression level of each gene is estimated using multiple probes. When probes in every probe set are ordered according to the localization of their binding site in target transcript, the average signal of probes with binding sites closer to the 5′ end of transcripts is shifted toward lower values. The slope of the line from the graph illustrating this trend can be used as a quality measure (Gautier *et al.*, 2004). The most widely used Affymetrix arrays have, for each probe that perfectly matches the target transcript, also a 'MM probe' that has a single nucleotide mismatch in the middle of the sequence. The MM probes were intended to measure the level of unspecific background signal for corresponding PM probes. The difference in signal level between pairs of PM and MM probes from the same probe sets is used to generate present/absent calls with the MAS5 algorithm, and the percentage of present calls is one of the quality metrics proposed by the Affymetrix company (Affymetrix, 2001; Wilson and Miller 2005). Kauffmann *et al.* (2009) suggested the inspection of the difference between distributions of signal levels of PM and MM probes as part of the quality assessment of microarray experiments. In the current study, we quantify this tendency by computing the value of the paired *t*-test statistic from the signal levels of all PM/MM probe pairs. The Average Background, one of the measures of quality that can be obtained from the original Affymetrix GCOS software, is simply the average of the 2% lowest cell intensities on the chip, and higher values of this parameter suggest high levels of nonspecific binding. The scaling factor is a value that should be used to multiply all values of intensities on the chip to scale the 2% trimmed mean signal to a selected constant (Affymetrix, 2001). Finally, the most popular multi-array quality metrics are Relative Log Expression (RLE) and Normalized Unscaled Standard Error (NUSE), which are based on comparisons of the outcome of Robust Multichip Analysis/Probe Level Model fitting procedures between arrays from the same experiment (Bolstad *et al.*, 2004; Gautier *et al.*, 2004; Irizarry *et al.*, 2003). In the RLE method, medians of probe set expression values are subtracted from expression values of all arrays in the experimental series. If the quality of a given array does not differ greatly from the average quality in the dataset, then such subtracted expression values center around 0 and display interquantile ranges similar to other arrays. The NUSE values measure precision of estimation of expression values. Shifted or wider distribution of NUSE values indicates problems with the quality of a particular array. The more general version of this parameter, Global Normalized Uncalled Standard Error (GNUSE), compares the standard error of expression estimation with the values stored in precomputed frozen parameter vectors obtained on the basis of analysis of many arrays of the same type together (McCall *et al.*, 2011).

The beneficial influence of removing outlier arrays for microarray data analysis has been demonstrated in several studies (Asare *et al.*, 2009; Kauffmann and Huber, 2010; McCall *et al.*, 2011). Development and testing of methods for evaluating the quality of microarray data between experiments suffer from the lack of a good benchmark. In the current study, we propose to use the degree of conservation of expression profile between species as an independent indicator of quality and to assess performance of the most popular quality control methods along with a new method developed by us. We show that our new IQRray method is consistently the best in predicting the quality of microarrays.

2 MATERIALS AND METHODS

2.1 Distribution of probe set average ranks

The IQRray statistic is obtained by ranking all the probe intensities from a given array and by computing the average rank for each probe set. The interquartile range (IQR) of the probe sets average ranks serves then as quality score. The simulated distribution of average ranks from probe sets composed from random probes was obtained by random assignment of ranks to probes from arrays. The simulated distribution of average ranks from probe sets with consistent ranks within probe sets was obtained by assigning succeeding ranks to probes from the same probe sets. In both cases, the number of probe sets and the number of probes were identical to arrays of the HG-U133_Plus_2 type. Real examples of low- (GSM50702) and high (GSM371402) quality arrays of HG-U133_Plus_2 type were selected from the Bgee database.

2.2 Expression data

We selected from the Bgee database (Bastian *et al.*, 2008) seven homologous human and mouse organs represented by high numbers of microarray results, obtained using Affymetrix platforms HG-U133_Plus_2 and Mouse430_2, respectively. The number of samples and experiments for each organ are presented in Table 1. All results from prenatal development stages were excluded from the analysis. In the case of the testis dataset, only data from adult developmental stage were included. Five human and six mouse datasets composed of arrays from at least five independent experiments were used as a training set for benchmarking quality control metrics. The datasets that have not passed this criterion were used as reference only (see Table 1). The complete list of arrays and experiments used in the study is included in Supplementary Table S1. The initial source of all experiments whose names start with 'GSE' is the GEO database (Edgar *et al.*, 2002); all the other experiments were originally downloaded from ArrayExpress (Brazma *et al.*, 2003).

2.3 Computing quality control parameters

The raw data from CEL files were read into the *R* environment using the package affy (Gautier *et al.*, 2004) from Bioconductor (Gentleman *et al.*, 2004). The parameters of average background, percent present, scaling factor and ratios between probe sets for 3′ and 5′ end of actin and

Table 1. Number of samples and experiments in organ-specific datasets

Organ name	Mouse sample	Mouse exp	Human sample	Human exp
Blood	28*	4*	429	19
Liver	389	60	51	9
Kidney	95	15	41	5
Colon	47	6	103	7
Testis	47	11	12*	2*
Placenta	50	10	50	9
Cerebral cortex	95	11	19*	2*

*Used as reference only.

GAPDH transcripts were calculated using the *R* package simpleaffy (Wilson and Miller, 2005). The slope for RNA degradation was obtained using the package affy (Gautier *et al*., 2004). RLE and NUSE metrics were computed with the package affyPLM (Bolstad *et al*., 2005) using all arrays that belong to certain experiment. The GNUSE values for every array were computed with the package fRMA (McCall *et al*., 2010) and relevant packages with fRMA vectors available in Bioconductor. Quality parameters for all arrays used as a training set are included in Supplementary Table S1.

2.4 Preprocessing of raw data

The raw data from CEL files were read into *R*. The signal values for PM probes were averaged for every probe set. The mapping between probe set IDs and gene IDs were taken from Ensembl version 69 (Flicek *et al*., 2013). Probe sets that match more than one gene were excluded. The expression values of independent probe sets that match the same gene were averaged. Only results for genes with one to one orthology between mouse and human with matching probe sets on both microarray types (according to Ensembl version 69) were used in further analysis (13 136 genes in total).

2.5 Correlation of the expression profiles between homologous organs

The pairwise Spearman correlation coefficients between expression profiles of all arrays from organ-specific datasets and homologous organ reference datasets were computed. For every array from an organ-specific dataset, the highest correlation coefficient [the homology organ correlation (HOC)] obtained was selected and used in further analysis. We used the highest correlation to decrease the probability that the array would be classified as low quality because of natural biological variability related, for example, to age or sex. The HOC score for all arrays used as a training set are included in Supplementary Table S1.

2.6 Performance of quality metrics

The correlations between quality parameters and HOC scores were calculated using the Spearman correlation coefficient. For some parameters, larger values are better (e.g. percent present), whereas for others, smaller values are better (e.g. GNUSE); these are noted as 'ascending' or 'descending', respectively, in Supplementary Table S2a and b. Quality scores from all 'descending' methods were multiplied by −1 before computing correlations. Correlation for all organs together was obtained after transforming values of HOC score into quartiles for every organ separately. To evaluate the performance of quality metrics in detecting the lowest quality arrays, we selected the worst 5% samples according to each quality control method from all organ-specific datasets and analyzed the distribution of their quartile of HOC scores. The 5% cutoff value that was used for each quality control method can be found in Supplementary Table S2a and b.

2.7 Quality control of Bgee database

The distributions of IQRray and percent present quality score values were computed using large numbers of independent arrays from the Bgee (array type Mouse430_2 and HG-U133_Plus_2) and GEO (15 other types of arrays) databases. The arrays from GEO were selected using the GEOmetadb and GEOquery *R* packages (Zhu *et al*., 2008). The number of arrays taken from each experiment was limited to 10 to ensure that the arrays from large experiments do not bias the computation of the cutoff value. Quantiles for the percent present and IQRray scores, as well as the number of arrays used for computation, are reported in Supplementary Table S3A and S3B. In Bgee, we decided to use cutoff values that allow removing the worst 5% arrays according to either percent present or IQRray score. All arrays that did not satisfy one or both of the quality thresholds were removed from the database.

3 RESULTS

3.1 IQRray

Because of the limitations of available methods, we propose a new method for multi-experiment quality control. In Affymetrix technology, the final expression level is computed on the basis of intensity levels of several independent probes matching the same target messenger RNA. In our new IQRray method, we transform all probe signal values into ranks and subsequently compute the average rank of probes that belong to the same probe set. We expect that the higher the quality of a given array, the more consistent the levels of probe signal from the same probe set. The average rank of probes from a probe sets that match highly expressed genes should be high, whereas the average rank of probes sets that match lowly or not expressed genes should be low. All factors that increase signal noise, such as unspecific hybridization or spatial artifacts, are expected to lead to a more random distribution of probe signals among probe sets. Mixing of low and high ranks in the same probe set should shift the value of the average rank of a probe set toward the average rank of all probes on the array. Consequently, lower quality microarrays will have more narrow spreads of distribution of rank averages. As a measure of this tendency, we propose to use IQR of probe set average rank: the IQRray score. Figure 1 shows distributions of probe set average ranks from two idealized arrays: one where intensities of probes in probe sets had consistent values and a second where signal values were assigned randomly to the probe sets. We also selected from microarrays in the Bgee database examples of arrays with extreme IQRray scores. It can be seen that the IQR of probe set average ranks is much smaller when the signal values were distributed randomly among probe sets than when they have consistent signal values. The distribution of probe sets' average ranks of a presumptive low-quality array resembles the distribution of probes with randomly assigned values. The distribution of a presumptive high-quality array shows, in contrast, a bimodal shape due to probe sets targeting lowly or not expressed genes and highly expressed genes.

3.2 Benchmarking of QC metrics

To compare the performance of the newly proposed method and existing ones for quality control, we set up a test study using 11 large organ-specific datasets—5 for human and 6 for mice. We use the similarity of its expression profile to a reference profile of the homologous organ from the other species as an external independent quality indicator for each microarray. We determined for every array the Homologous Organ Correlation score (HOC score, see Section 2). We expect that the lower quality arrays will display less biological signal and consequently a lower correlation with the reference. All arrays analyzed displayed a positive correlation with the profile of the homologous organ (Fig. 2), which is consistent with the previously reported conservation of orthologous gene expression in homologous tissues (Brawand *et al*., 2011; Piasecka *et al*., 2012; Zheng-Bradley *et al*., 2010). However, in each dataset, outlier samples of suspicious quality with unusually low HOC scores can be found. These arrays should be preferentially removed through quality control procedure.

3.3 Correlation with external quality control method

Each array was evaluated separately by a set of microarray quality control methods. Then we checked how well the quality metrics of each method agreed with the HOC score. There was large variation in the correlation with this external quality indicator (Figs 3 and 4). For example, in the case of the human blood dataset, the largest dataset in the study (Table 1), the IQRray method displayed nearly perfect correlation with the HOC score (Spearman correlation of 0.97) (Figs 3a and 4a). In contrast, NUSE and RLE (McCall *et al.*, 2011), which are frequently used quality control methods, showed only a weak positive correlation (Figs 3b and 4b). For this human blood dataset, only a low fraction of samples came from experiments with fewer than six samples (Supplementary Table S1); thus, the low correlation with HOC cannot be explained simply by a lack of power, owing to an insufficient number of arrays in experiments. In general, across all analyzed datasets, NUSE and RLE performed poorly, which suggests that the scores returned by these methods are not directly comparable between independent experiments. All traditional single-array quality metrics, such as RNA degradation slope, average background (avbg), scaling factor and ratios between signal for the 3′ and 5′ ends of actin and GAPDH transcripts, show low performance, and the correlation was even negative for some of the methods for some datasets (Fig. 4a and b). The score from the GNUSE method, the only published method dedicated to absolute quantification of microarray quality (McCall *et al.*, 2011), correlates well with the external quality metric only for mouse datasets, whereas for humans, GNUSE obtained poor results for nearly all datasets (Fig. 4b). The IQRray performed the best in 8 of 11 datasets. The other methods that displayed high agreement with the HOC score for both mouse and human data were percent present and the PM/MM *t*-test.

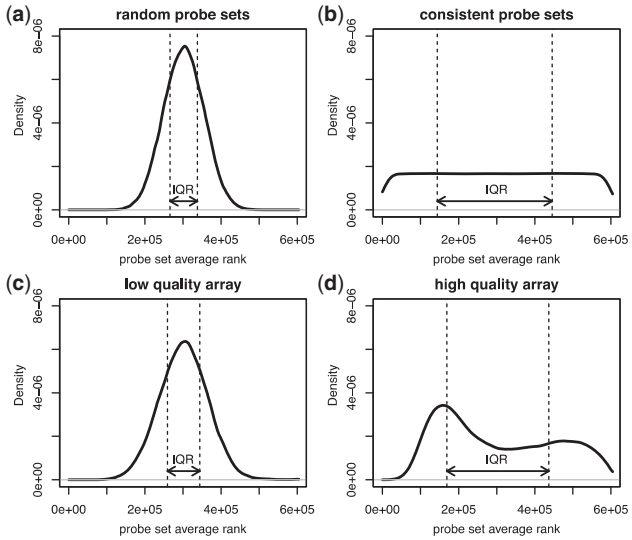

Fig. 1. Distribution of probe set average ranks of (**a**) simulated array with intensity values assigned randomly to probe sets, (**b**) simulated array with consistent intensity values in probe sets, (**c**) real array with a low IQRray score (GSM50702) and (**d**) real array with a high IQRray score (GSM371402)

3.4 Simulation of quality control assessment of large set of samples

The ultimate goal of quality assessment is to remove the worst quality samples from datasets. A good quality control method should both correctly identify the worst samples and avoid

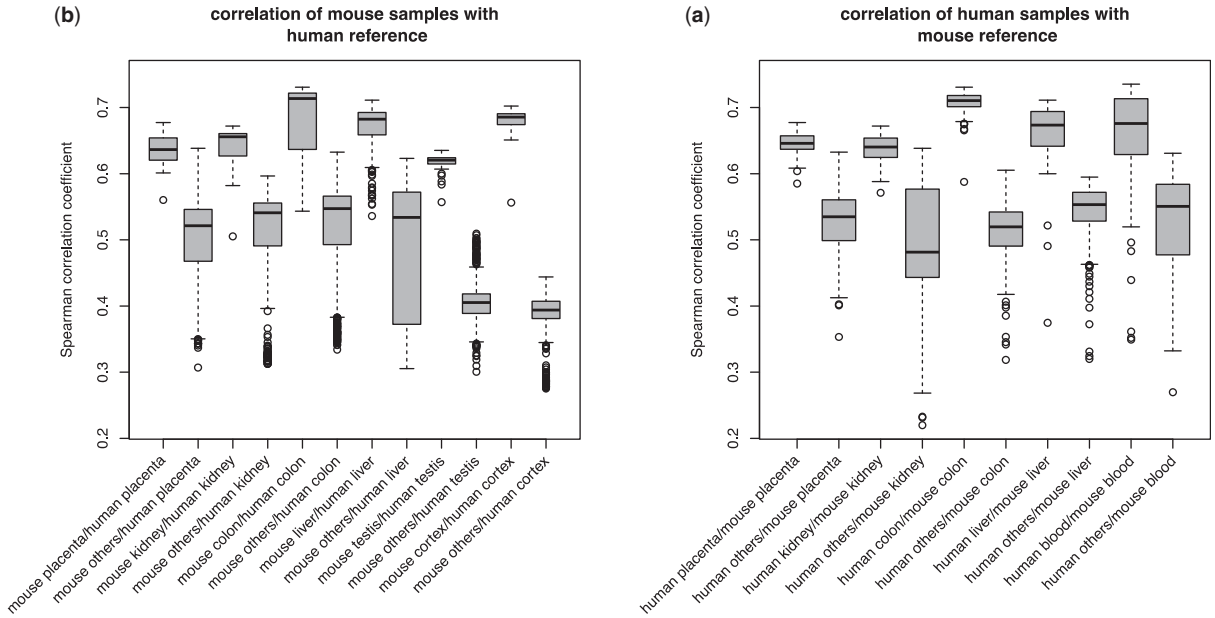

Fig. 2. Boxplots of Spearman ρ values from correlation test between organ-specific microarray results and reference results from another species (**a**) correlation between human samples and mouse reference or (**b**) correlation between mouse samples and human reference

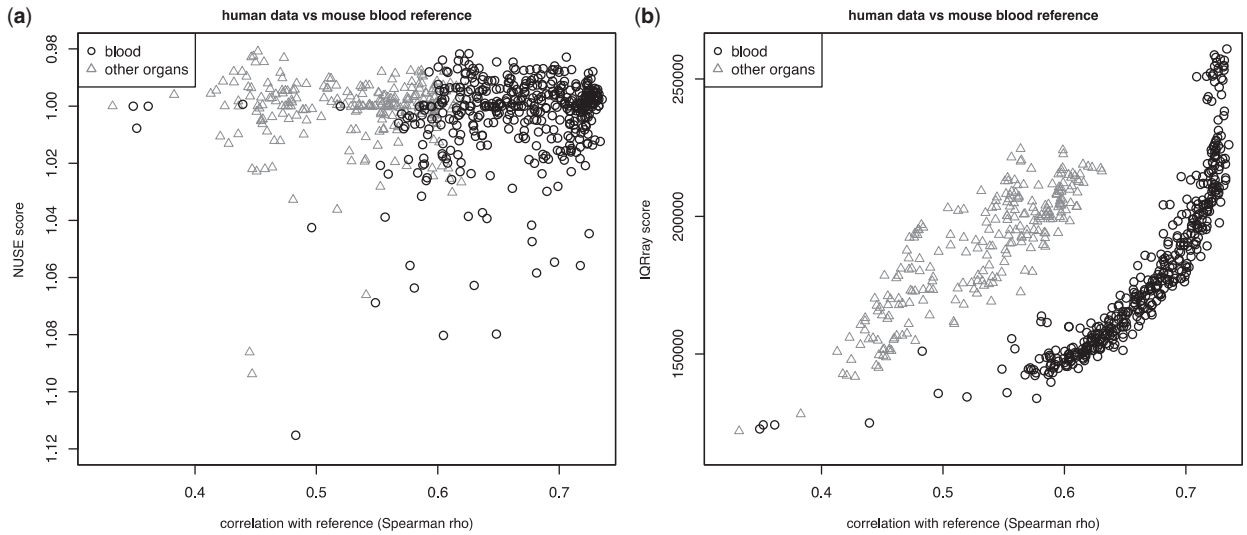

Fig. 3. Correlation between (**a**) NUSE and (**b**) IQRray scores and Spearman ρ values from correlation test between mouse blood reference and human blood samples (black circles) and other samples from other organs (gray triangles). Each point on the plot corresponds to a single array

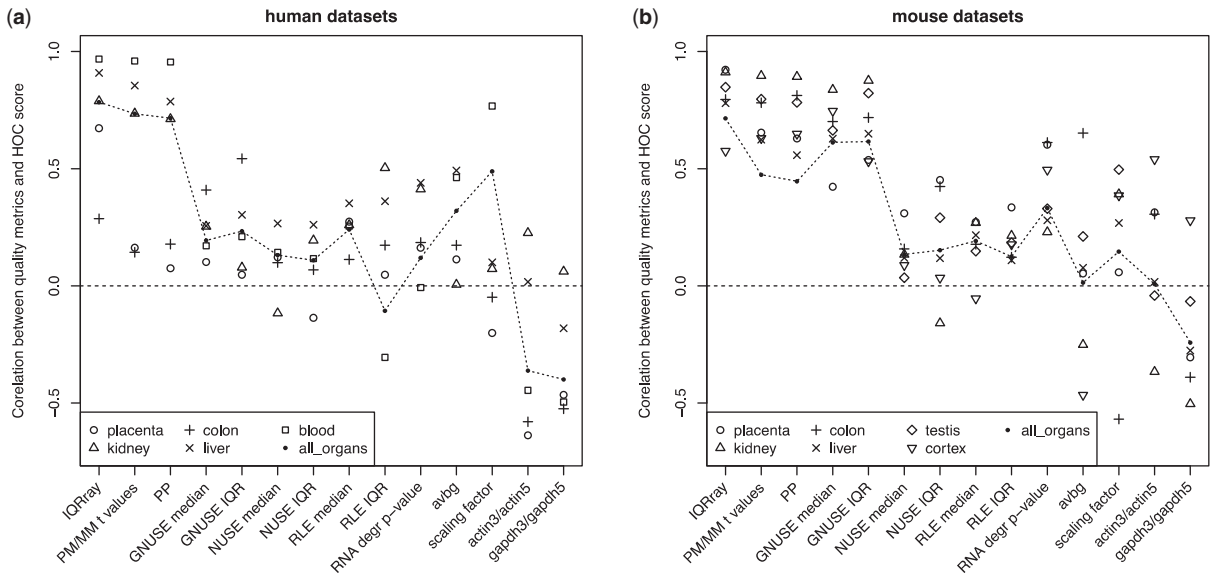

Fig. 4. Spearman ρ values from correlation test between quality metrics and HOC score for (**a**) human and (**b**) mouse organ-specific datasets

assigning falsely low scores to good-quality arrays. We simulated a quality control check for a collection of microarrays. We selected the 5 and 10% of samples with the lowest quality according to each quality control method from all samples used in the study (Supplementary Table S2). We show the distribution of HOC scores for arrays selected with a 5% cutoff (Fig. 5). We also measured the efficiency of methods in selecting the arrays with the lowest quality by computing the proportion of results below a chosen cutoff value (Supplementary Table S2) displaying a quantile of HOC values also below the same cutoff (e.g. quantiles <5% for 5% cutoff) (Supplementary Table S2). The method that consistently was the most efficient in identifying the worst quality arrays was IQRray.

For both mouse and human data, all samples selected by this method had quantiles <0.2, which means that they were of considerably lower quality than other samples derived from the same tissue. This method also identified the highest proportion (~60% in all cases) of low-quality arrays using both 5 and 10% cutoffs. Again, the other methods that performed relatively well were percent present and PM/MM *t*-test. Consistent with previous results, the GNUSE metric showed high performance for mouse data and confusing results for human data. Among the traditional methods used for quality assessment of microarrays, the scaling factor gave relatively good results for both human and mouse datasets.

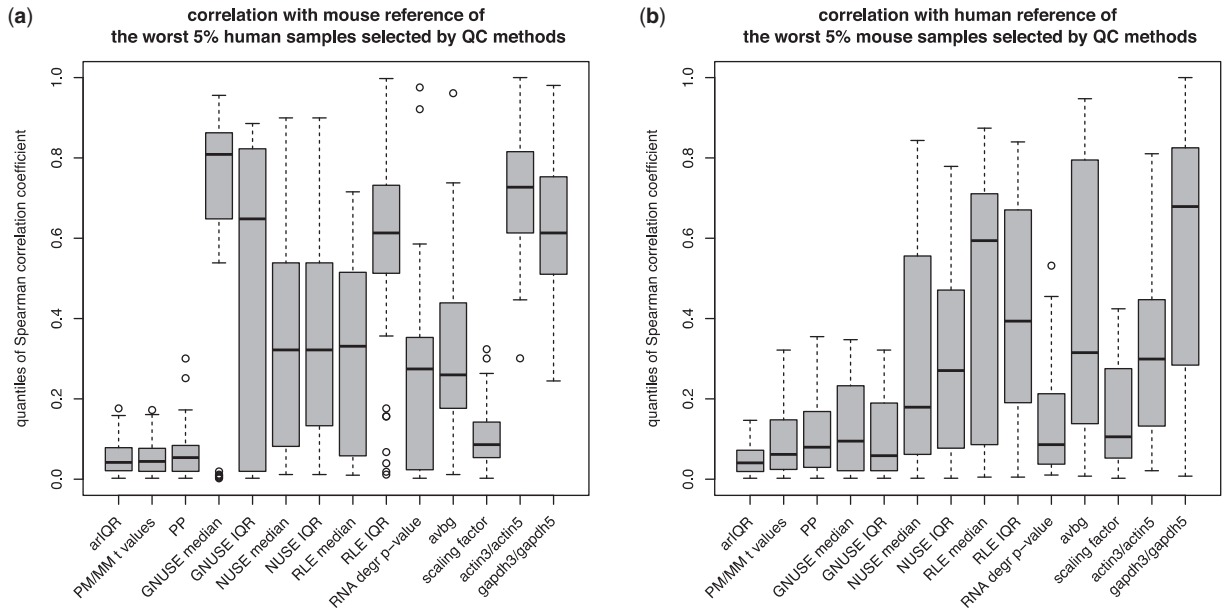

Fig. 5. Boxplot of quantiles of HOC scores of the worst 5% samples selected by different QC methods from (**a**) human and (**b**) mouse datasets

The Bgee database was created to facilitate comparative and evolutionary analyses of gene expression profiles. Ensuring that the microarrays used contain sufficient biological signal is of primary importance. Based on this benchmarking, in the Bgee processing pipeline, we decided to use two independent quality control methods: IQRray and percent present. The 5% quantiles used as quality thresholds were computed for each array type separately using a high number of representative arrays. The quantiles of distributions for both metrics for selected array types are available as supplementary material and can be used to evaluate the quality level of any new microarray experiment. Among 12 173 microarrays with raw CEL files from Bgee subjected to quality assessment, 887 (7.2%) did not pass at least one quality threshold, 441 (3.6%) did not pass either of them, 194 (1.6%) were eliminated because of a low IQRray score and 252 (2%) were eliminated because of a low percent present score. These 887 arrays come from 172 experiments (16.6% of 1033 experiments available in Bgee). Arrays from 22 experiments were excluded entirely, and at least 50% of samples were eliminated from another 37 experiments.

4 DISCUSSION

In this study, we introduce a new methodology for benchmarking the quality of results of high-throughput transcriptomic experiments. We decided to use evolutionary conservation of expression profile as an indicator of quality because it provides an independent assessment of biological relevance. Because of the use of correlation between species rather than of correlation between arrays obtained with the same platform, like in previous studies (Gagnon-Bartsch and Speed, 2012; McCall *et al.*, 2011), we can avoid giving high scores to low-quality results that cluster together. Such spurious results can easily be obtained with microarray technology because the GC content of probes and probe set design, which highly influence the background unspecific level of the measure, remain the same among samples. The HOC can be used not only for the assessment of performance of quality control methods, but also more generally for the benchmarking performance of any preprocessing step. Moreover, the method can be directly applied to results from different technologies such as RNA-seq and may be easily adapted for other types of experiments where an evolutionary conservation of the results is expected, such as ChIP-seq experiments analyzing transcription binding factors or chromatin methylation marks. A good quality control method should be sensitive to all factors that impede the specificity of binding of labeled targets, such as RNA degradation, insufficient labeling of target transcripts, suboptimal hybridization conditions or surface defects like bubbles or scratches (Novak *et al.*, 2002). The IQRray algorithm outperformed all the other tested methods in identifying low-quality arrays. The method owed its success to two features: (i) the fact that all factors decreasing the strength of specific signal or adding noise to the estimation of the final expression values of a substantial proportion of probe sets will change the distribution of average ranks computed for each probe set; and (ii) the possibility of direct comparison of IQRray score between arrays because of the transformation of original values to ranks, which can be considered as a between-arrays normalization step.

Our study showed that among the known methods for quality assessment of microarrays, the best results can be obtained using methodologies based on PM and MM probe pairs. Use of MM probes for estimation of unspecific background signal for PM probes has been highly criticized, and gives inferior results in comparison with other methods of background signal measurement, mostly because the MM probes also bind, to some extent, the specific target transcript (Irizarry, *et al.*, 2003; Lockhart, *et al.*, 1996; Shedden, *et al.*, 2005). However, when all PM/MM probe pairs on the array are analyzed simultaneously, the shift in signal distribution between PM and MM probes is clearly

visible, and its strength measured by either PM/MM paired *t*-test or by proportion of probe sets called present seems to be a good indicator of quality. This result might be explained by the fact that the occurrence of differences in hybridization strength between PM and MM probes is strictly dependent on the specificity of target binding, and all factors that negatively influence it also diminish this difference. These methods might be less suitable to detect loss of quality caused by spatial artifacts, although because of the fact that the pairs of MM/PM probes are located next to each other on the surface of the microarray, the large intensive surface artifacts also probably have an impact of the final quality score.

The performances of GNUSE scores, the only quality parameter intended for direct comparison of quality between different experiments, differ significantly between mouse and human datasets. Such a discrepancy may indicate that although the assumptions of the methodology are correct, the differences in experimental protocol strongly influence the final results, as was suggested by the GNUSE authors (McCall, *et al*., 2011). The other possible explanation of this phenomenon is simply lower quality of the fRMA vector specific for the human HG-U133_Plus_2 array. If a higher proportion of human microarray experiments is of lower quality, then the frozen parameter vector prepared on the basis of these data may underestimate the natural variance in probe signal from the same probe set.

Despite the fact that methods for the analysis of microarray results are already in their maturity, and a lot of effort and attention has been made toward improving methods of quality assessment of microarrays, including calling two international consortia EMERALD and MAQC (Beisvag *et al*., 2011; Canales *et al*., 2006), there are still no objective rules for defining absolute quality of microarrays. Our new IQRray algorithm for quality control of microarrays appears to be powerful in detecting of low-quality arrays, as are the PM/MM *t*-test and the percentage called present, as measured by our independent evolutionary conservation-based quality metric. We strongly recommend using one or several of these metrics before performing meta-analyses or integrating microarrays into databases, as we do in Bgee.

ACKNOWLEDGEMENT

The authors thank Frédéric Bastian, Aurélie Comte, Nadezda Kryuchkova and Anne Niknejad (UNIL and SIB) for helpful discussions.

Funding: Swiss Institute of Bioinformatics (SIB) support to Bgee; Swiss National Science Foundation (31003A 133011/1); Etat de Vaud.

Conflict of Interest: none declared.

REFERENCES

Affymetrix. (2001) *Guidelines for assessing data quality*. Affymetrix Inc, Santa Clara, CA.

Asare,A.L. *et al.* (2009) Power enhancement via multivariate outlier testing with gene expression arrays. *Bioinformatics*, **25**, 48–53.

Bastian,F. *et al.* (2008) Bgee: integrating and comparing heterogeneous transcriptome data among species. In: Bairoch,A., Cohen-Boulakia,S. and Froidevaux,C. (eds) *Data Integration in the Life Sciences*. Springer, Berlin, Heidelberg, pp. 124–131.

Beisvag,V. *et al.* (2011) Contributions of the EMERALD project to assessing and improving microarray data quality. *Biotechniques*, **50**, 27–31.

Bolstad,B.M. *et al.* (2005) Quality assessment of Affymetrix GeneChip data. In: Gentleman,R., *et al.* (eds) *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*. Springer, New York, pp. 33–47.

Bolstad,B.M. *et al.* (2004) Experimental design and low-level analysis of microarray data. *Int. Rev. Neurobiol.*, **60**, 25–58.

Brawand,D. *et al.* (2011) The evolution of gene expression levels in mammalian organs. *Nature*, **478**, 343–348.

Brazma,A. *et al.* (2003) ArrayExpress–a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res.*, **31**, 68–71.

Brettschneider,J. *et al.* (2008) Quality assessment for short oligonucleotide microarray data. *Technometrics*, **50**, 241–264.

Canales,R.D. *et al.* (2006) Evaluation of DNA microarray results with quantitative gene expression platforms. *Nat. Biotechnol.*, **24**, 1115–1122.

Dash,S. *et al.* (2012) PLEXdb: gene expression resources for plants and plant pathogens. *Nucleic Acids Res.*, **40**, D1194–D1201.

Edgar,R., Domrachev,M. and Lash,A.E. (2002) Gene expression omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, **30**, 207–210.

Flicek,P. *et al.* (2013) Ensembl 2013. *Nucleic Acids Res.*, **41**, D48–D55.

Gagnon-Bartsch,J.A. and Speed,T.P. (2012) Using control genes to correct for unwanted variation in microarray data. *Biostatistics*, **13**, 539–552.

Gautier,L. *et al.* (2004) affy–analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*, **20**, 307–315.

Gentleman,R.C. *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, **5**, R80.

Hebestreit,K. *et al.* (2012) Leukemia gene atlas–a public platform for integrative exploration of genome-wide molecular data. *PLoS One*, **7**, e39148.

Hruz,T. *et al.* (2008) Genevestigator v3: a reference expression database for the meta-analysis of transcriptomes. *Adv. Bioinform.*, **2008**, 420747.

Irizarry,R.A. *et al.* (2003) Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res.*, **31**, e15.

Irizarry,R.A. *et al.* (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, **4**, 249–264.

Kapushesky,M. *et al.* (2010) Gene expression atlas at the European bioinformatics institute. *Nucleic Acids Res.*, **38**, D690–D698.

Kauffmann,A., Gentleman,R. and Huber,W. (2009) arrayQualityMetrics–a bioconductor package for quality assessment of microarray data. *Bioinformatics*, **25**, 415–416.

Kauffmann,A. and Huber,W. (2010) Microarray data quality control improves the detection of differentially expressed genes. *Genomics*, **95**, 138–142.

Le Crom,S. *et al.* (2002) yMGV: helping biologists with yeast microarray data mining. *Nucleic Acids Res.*, **30**, 76–79.

Lockhart,D.J. *et al.* (1996) Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat. Biotechnol.*, **14**, 1675–1680.

McCall,M.N., Bolstad,B.M. and Irizarry,R.A. (2010) Frozen robust multiarray analysis (fRMA). *Biostatistics*, **11**, 242–253.

McCall,M.N. *et al.* (2011) Assessing affymetrix GeneChip microarray quality. *BMC Bioinformatics*, **12**, 137.

Novak,J.P., Sladek,R. and Hudson,T.J. (2002) Characterization of variability in large-scale gene expression data: implications for study design. *Genomics*, **79**, 104–113.

Obayashi,T. *et al.* (2013) COXPRESdb: a database of comparative gene coexpression networks of eleven species for mammals. *Nucleic Acids Res.*, **41**, D1014–D1020.

Piasecka,B. *et al.* (2012) Comparative modular analysis of gene expression in vertebrate organs. *BMC Genomics*, **13**, 124.

Rhodes,D.R. *et al.* (2007) Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles. *Neoplasia*, **9**, 166–180.

Shedden,K. *et al.* (2005) Comparison of seven methods for producing Affymetrix expression scores based on False Discovery Rates in disease profiling data. *BMC Bioinformatics*, **6**, 26.

Tseng,G.C., Ghosh,D. and Feingold,E. (2012) Comprehensive literature review and statistical considerations for microarray meta-analysis. *Nucleic Acids Res.*, **40**, 3785–3799.

Wilkes,T., Laux,H. and Foy,C.A. (2007) Microarray data quality - review of current developments. *OMICS*, **11**, 1–13.

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

Wu,C., Macleod,I. and Su,A.I. (2013) BioGPS and MyGene.info: organizing online, gene-centric information. *Nucleic Acids Res.*, **41**, D561–D565.

Xiao,S.J. *et al.* (2010) TiSGeD: a database for tissue-specific genes. *Bioinformatics*, **26**, 1273–1275.

Zheng-Bradley,X. *et al.* (2010) Large scale comparison of global gene expression patterns in human and mouse. *Genome Biol.*, **11**, R124.

Zhu,Y. *et al.* (2008) GEOmetadb: powerful alternative search engine for the Gene Expression Omnibus. *Bioinformatics*, **24**, 2798–2800.