Genomics 95 (2010) 138–142

Contents lists available at ScienceDirect

Genomics

journal homepage: www.elsevier.com/locate/ygeno

Minireview

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

ARTICLE INFO

Article history:
Received 12 November 2009
Accepted 8 January 2010
Available online 14 January 2010

Keywords:
Microarray
Quality
Outlier

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 controling 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

* Corresponding author.
E-mail address: audrey@ebi.ac.uk (A. Kauffmann).

0888-7543/$ – see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2010.01.003

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 probeset 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 are 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 detection of unusual arrays–i. e., arrays that are highly different from 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 L_1-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-GEOD-3419 and E-GEOD-10211, and three for E-GEOD-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.

Table 1
The five datasets used in this review were obtained from the ArrayExpress database ae. The array platform used was Affymetrix GeneChip, the experiments encompass two species, different cell types and a range of array types.

ArrayExpress ID	Ref	Organism	# Chips	Study
E-GEOD-3419	[18]	*H. sapiens*	16	Hair follicle bulge cells enrichment in stem cells
E-GEOD-7258	[19]	*M. musculus*	17	Urethane treatment in bronchoalveolar lavage cells
E-GEOD-10211	[20]	*M. musculus*	10	Epithelial cell response to sendai virus infection
E-MEXP-774	[21]	*M. musculus*	30	Dexamethasone treatment of preadipocytes
E-MEXP-170	[22]	*H. sapiens*	8	Sulforaphane treatment of colon cells

Table 2
Results from differential gene expression analysis on: all arrays, after removing outlier arrays, and using array weights (see Outlier detection for details). The second column lists the identified outlier arrays as numbered in the quality assessment reports provided in Supplementary material, as well as, in brackets, their weights. For comparison, the third column gives the range of the weights for each dataset. The last three columns give the number of differentially expressed genes identified by each method.

ArrayExpress ID	Outlier arrays (weight)	Weights range	All arrays	Number of genes Outlier(s) removed	Weights
E-GEOD-3419	6 (0.5), 12 (0.3)	0.3–1.8	11	41	16
E-GEOD-7258	7 (0.1), 15(0.3), 16 (0.2)	0.1–3.3	6	547	33
E-GEOD-10211	2 (0.1), 7 (0.4)	0.1–2.1	93	396	356
E-MEXP-774	4 (0.1), 17 (0.2)	0.1–2.2	1125	1694	1587
E-MEXP-170	6 (0.01)	0.01–4.9	262	4221	3268

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-GEOD-3419 and strong for E-GEOD-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-GEOD-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-GEOD-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

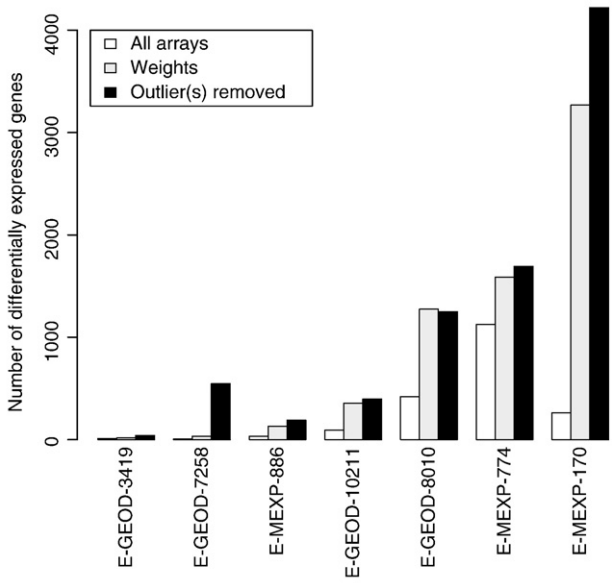

Fig. 1. Number of differentially expressed genes identified on the whole dataset (white bars), after removing outliers identified by arrayQualityMetrics (black bars) and using weights obtained by arrayWeights from *limma* (grey bars).

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

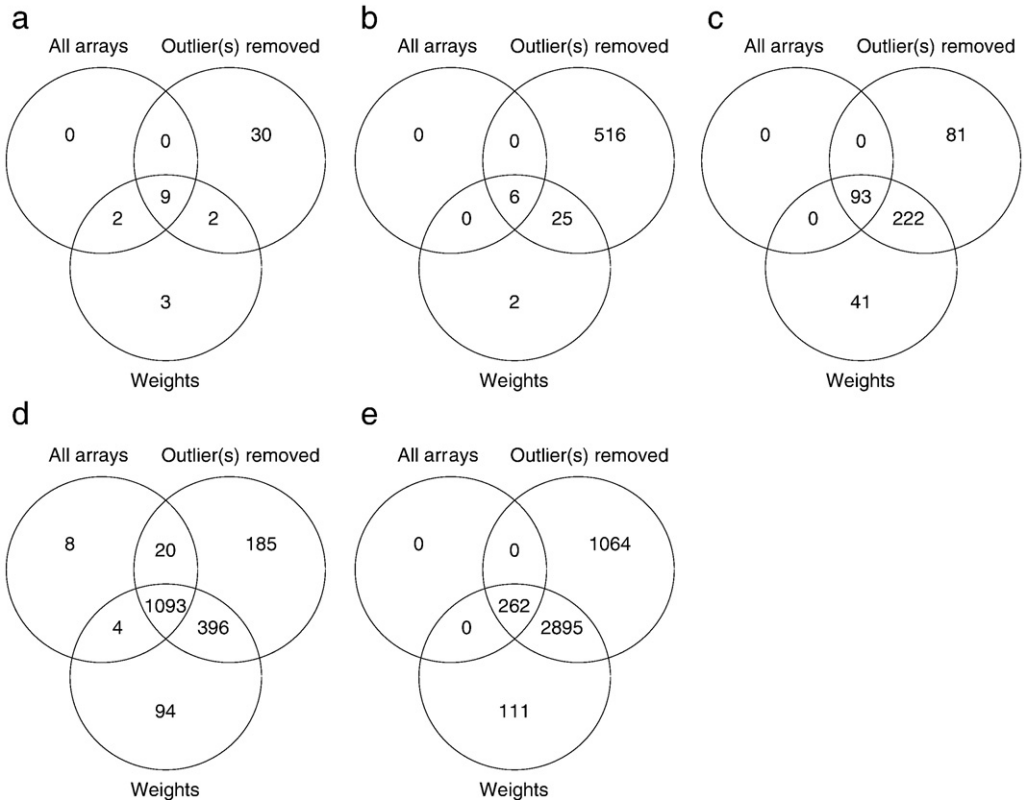

Fig. 2. Venn diagrams representing the number of differentially expressed genes identified by each method: all arrays, after removing outlier arrays, using array weights. (a) E-GEOD-3419, (b) E-GEOD-7258, (c) E-GEOD-10211, (d) E-MEXP-774, (e) E-MEXP-170.

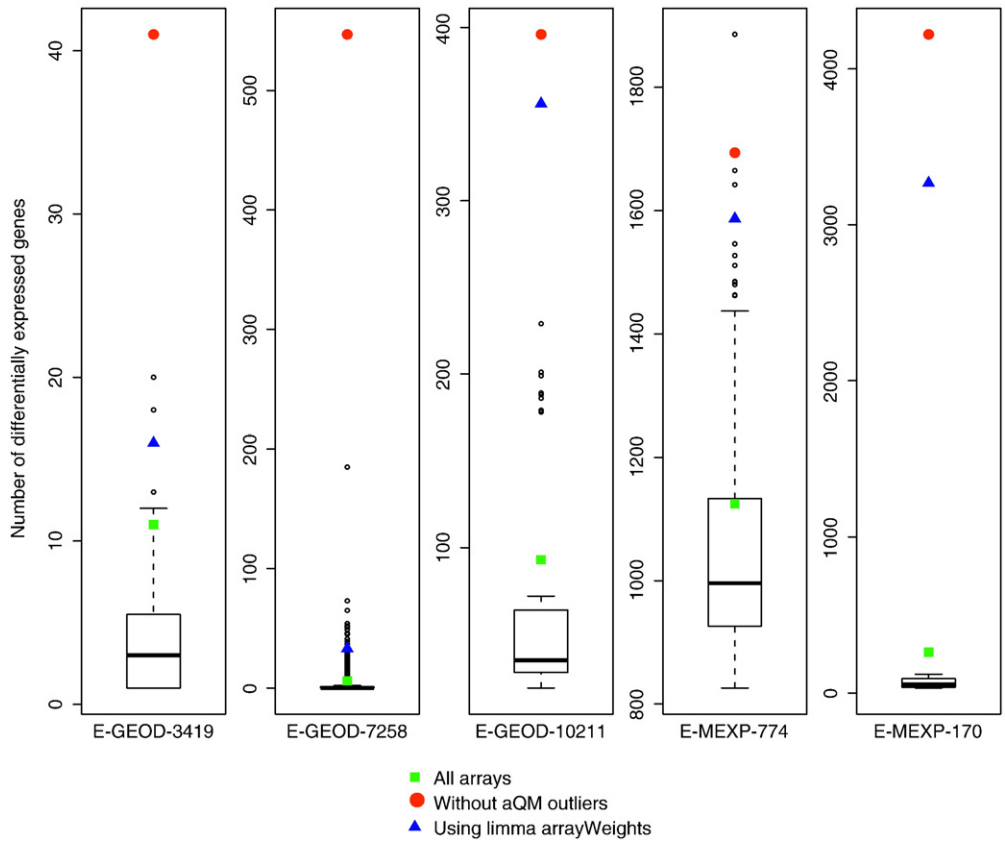

Fig. 3. Boxplots representing the number of genes called differentially expressed in each experiment when removing arbitrary subsets of size *K*, the number of outlier arrays identified by arrayQualityMetrics , from the *N* samples. When *N* over *K* ($N!/(K!(N\text{-}K)!)$) was less than 1000, all possible subsets were considered, otherwise 1000 subsets were sampled randomly.

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 provides–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.

Table 3
KEGG pathway enrichment analysis. Five of the most enriched pathways according to gene set enrichment analysis are listed for experiments E-GEOD-3419 and E-GEOD-7258, with and without outlier removal. The pathways are related to the biology studied in the experiments, and their enrichment is more significant after outlier removal.

Pathway name	Genes	p-value when removing outliers	p-value when all arrays
E-GEOD-3419			
Pyrimidine metabolism	37	$<10^{-3}$	0.701
Base excision repair	17	0.001	0.542
DNA replication	19	0.003	0.451
Cell cycle	69	0.009	0.387
TGF-beta signaling pathway	48	0.009	0.558
E-GEOD-7258			
Pentose phosphate pathway	13	0.003	0.588
Fructose and mannose metabolism	28	0.003	0.326
Biosynthesis of steroids	20	0.003	0.012
Oxidative phosphorylation	44	0.003	0.299
Starch and sucrose metabolism	16	0.003	0.317

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 the online version, at doi:10.1016/j.ygeno.2010.01.003.

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.

ArrayExpress ID	arrayQuality Metrics	GESD	Hampel	Rousseeuw
E-GEOD-3419	6, 12	3, 6, 12	12	12
E-GEOD-7258	7, 15, 16	7, 15, 16	7, 15, 16	7, 15, 16
E-GEOD-10211	2, 7	2, 7	2	2
E-MEXP-774	4, 17	4, 17	4, 17	4, 17
E-MEXP-170	6	6	6	6

References

[1] J.W. Bauer, H. Bilgic, E.C. Baechler, Gene-expression profiling in rheumatic disease: tools and therapeutic potential, Nat. Rev. Rheumatol. 5 (2005) 257–265.
[2] M. Jayapal, A.J. Melendez, DNA microarray technology for target identification and validation, Clin. Exp. Pharmacol. Physiol. 33 (2006) 496–503.
[3] C.S. Cooper, C. Campbell, S. Jhavar, Mechanisms of disease: biomarkers and molecular targets from microarray gene expression studies in prostate cancer, Nat. Clin. Pract. Urol. 4 (2007) 677–687.
[4] G. Sherlock, C.A. Ball, Storage and retrieval of microarray data and open source microarray database software, Mol. Biotechnol. 30 (2005) 239–251.
[5] G.K. Smyth, Linear models and empirical Bayes for assessing differential expression in microarray experiments, Stat. Appl. Genet. Mol. Biol. 3 (2004) Article 1.
[6] A.P. Oron, Z. Jiang, R. Gentleman, Gene set enrichment analysis using linear models and diagnostics, Bioinformatics 24 (2008) 2586–2591.
[7] F. Hahne, W. Huber, R. Gentleman, S. Falcon, Bioconductor case studies, UseR! book (2008).
[8] A. Petri, J. Fleckner, M.W. Matthiessen, Array-A-Lizer: a serial DNA microarray quality analyzer, BMC Bioinformatics 5 (2004) 12.
[9] A. Buness, W. Huber, K. Steiner, et al., arrayMagic: two-colour cDNA microarray quality control and preprocessing, Bioinformatics 21 (2005) 554–556.
[10] C. Parman, C. Halling, R. Gentleman (2005) affyQCReport: QC Report Generation for affyBatch objects. R package version 1.24.0.
[11] G.V. Cohen Freue, Z. Hollander, E. Shen, et al., MDQC: a new quality assessment method for microarrays based on quality control reports, Bioinformatics 23 (2007) 3162–3169.
[12] A.D. Shieh, Y.S. Hung, Detecting outlier samples in microarray data, Stat. Appl. Genet. Mol. Biol. 8 (2009) Article 13.
[13] A. Kauffmann, R. Gentleman, W. Huber, arrayQualityMetrics—a bioconductor package for quality assessment of microarray data, Bioinformatics 25 (2009) 415–416.
[14] J. Schuchhardt, D. Beule, A. Malik, et al., Normalization strategies for cDNA microarrays, Nucleic Acids Res. 28 (2000) E47.
[15] T.L. Fare, E.M. Coffey, H. Dai, et al., Effects of atmospheric ozone on microarray data quality, Anal. Chem. 75 (2003) 4672–4675.
[16] H. Laux, T. Wilkes, A. Kauffmann et al. (2009) Impact of RNA integrity on microarray based measurements. In preparation.
[17] H. Parkinson, M. Kapusheksy, N. Kolesnikov, et al., ArrayExpress update—from an archive of functional genomics experiments to the atlas of gene expression, Nucleic Acids Res. 37 (2009) D868–D872.
[18] M. Ohyama, A. Terunuma, C.L. Tock, et al., Characterization and isolation of stem cell-enriched human hair follicle bulge cells, J. Clin. Invest. 116 (2006) 249–260.
[19] R.S. Stearman, L. Dwyer-Nield, L. Zerbe, et al., Analysis of orthologous gene expression between human pulmonary adenocarcinoma and a carcinogen-induced murine model, Am. J. Path. 167 (2005) 1763–1775.
[20] L.P. Shornick, A.G. Wells, Y. Zhang, et al., Airway epithelial versus immune cell Stat1 function for innate defense against respiratory viral infection, J. Immunol. 180 (2008) 3319–3328.
[21] M. Miron, O.Z. Woody, A. Marcil, et al., A methodology for global validation of microarray experiments, BMC Bioinformatics 7 (2006) 333.
[22] M. Traka, A.V. Gasper, J.A. Smith, et al., Transcriptome analysis of human colon Caco-2 cells exposed to sulforaphane, J. Nutr. 135 (2005) 1865–1872.
[23] M.E. Ritchie, D. Diyagama, J. Neilson, et al., Empirical array quality weights in the analysis of microarray data, BMC Bioinformatics 7 (2006) 261.
[24] A. Kauffmann, T. Rayner, H. Parkinson, et al., Importing ArrayExpress datasets into R/Bioconductor, Bioinformatics 25 (2009) 2092–2094.
[25] L. Gautier, L. Cope, B.M. Bolstad, R.A. Irizarry, affy—analysis of Affymetrix GeneChip data at the probe level, Bioinformatics 20 (2004) 307–315.
[26] M. Morgan, S. Falcon, R. Gentleman (2009) GSEABase: Gene set enrichment data structures and methods. R package version 1.8.0.
[27] B. Rosner, (1983) Percentage points for a generalized ESD many-outlier procedure Technometrics, 25, Article 2.
[28] P.L. Davies, U. Gather, The identication of multiple outliers, J. Am. Stat. Asso. 88 (1993) 782–801.
[29] P.J. Rousseeuw, A.M. Leroy, A robust scale estimator based on the shortest half, Stat. Neer. 42 (1988) 103–116.