Bioinformatics Advance Access published December 23, 2008

arrayQualityMetrics - a Bioconductor package for quality assessment of microarray data.

Audrey Kauffmann [1,*], Robert Gentleman [2], Wolfgang Huber [1]

[1]EMBL European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK.
[2]Computational Biology - FHCRC, 1100 Fairview Avenue North, Seattle, WA, USA 98109

Associate Editor: Prof. David Rocke

ABSTRACT

Summary: The assessment of data quality is a major concern in microarray analysis. arrayQualityMetrics is a Bioconductor package that provides a report with diagnostic plots for one or two colour microarray data. The quality metrics assess reproducibility, identify apparent outlier arrays and compute measures of signal-to-noise ratio. The tool handles most current microarray technologies and is amenable to use in automated analysis pipelines or for automatic report generation, as well as for use by individuals. The diagnosis of quality remains, in principle, a context-dependent judgement, but our tool provides powerful, automated, objective and comprehensive instruments on which to base a decision.

Availability: arrayQualityMetrics is a free and open source package, under LGPL license, available from the Bioconductor project at www.bioconductor.org. A users guide and examples are provided with the package.

Supplementary Information: Some examples of HTML reports generated by arrayQualityMetrics can be found at http://www.microarray-quality.org

Contact: audrey@ebi.ac.uk

INTRODUCTION

As microarray data quality can be affected at each step of the microarray experiment processing (Schuchhardt *et al.*, 2000), quality assessment is an integral part of the analysis. There are freely available tools allowing quality assessment for a specific microarray type such as Affymetrix (Craig Parman and Conrad Halling, 2005), Illumina (Dunning *et al.*, 2007), two-colour cDNA arrays (Buness *et al.*, 2005). Other free tools are designed to identify a particular problem among which, spot quality (Li *et al.*, 2005) or hybridization quality (Petri *et al.*, 2004). Some tools perform outlier detection from quality metrics done before (Freue *et al.*, 2007), or propose interactive quality plots (Lee *et al.*, 2006). We developed a Bioconductor (Gentleman *et al.*, 2004) package, arrayQualityMetrics, with the aim to provide a comprehensive tool that works on all expression arrays and platforms and produces a self-contained report which can be web-delivered. The Supplementary Table shows a comparison with the functionality and scope of other Bioconductor packages concerned with quality assessment or outlier detection.

DESCRIPTION

Input: To perform an analysis using the arrayQualityMetrics package, one needs to provide the matrix of microarray intensities and optionally, information about the samples and the probes in a Bioconductor object of class *AffyBatch*, *ExpressionSet*, *NChannelSet* or *BeadLevelList*. These classes are widely used and well documented. The manner of calling the arrayQualityMetrics function to create a report is the same for all of these classes, and it can be applied to raw array intensities as well as to normalised data. Applied to raw intensities, the quality metrics can help with monitoring experimental procedures and with the choice of normalisation procedure; application to the normalised data is more relevant for assessing the utility of the data in downstream analyses.

Individual array quality: The MA-plot allows the evaluation of the dependence between the intensity levels and the distribution of the ratios (Figure 1a) (Dudoit *et al.*, 2002). For two-colour arrays, a probe's M-value is the log-ratio of the two intensities and the A-value is the mean of their logarithms. In the case of one colour arrays, the M-value is computed by dividing the intensity by the median intensity of the same probe across all arrays. A false colour representation of each array's spatial distribution of feature intensities (Figure 1b) helps in identifying spatial effects that may be caused by, for example, gradients in the hybridization chamber, air bubbles or printing problems.

Homogeneity between arrays: To assess the homogeneity between the arrays, boxplots of the $\log_2$ intensities and density estimate plots (Figure 1c) are presented.

Between array comparison: Figure 1d shows a heatmap of between array distances, computed as the mean absolute difference of the M-value for each pair of arrays,

$$d_{xy} = \underset{i}{\text{mean}}|M_{xi} - M_{yi}|, \tag{1}$$

where M_{xi} is the M-value of the i-th probe on the x-th array. Consider the decomposition of M_{xi}:

$$M_{xi} = z_i + \beta_{xi} + \varepsilon_{xi} \tag{2}$$

where z_i is the probe effect for probe i (the same across all arrays), ε_{xi} are i.i.d random variables with mean zero and β_{xi} is a sparse matrix representing differential expression effects. Under these assumptions, all values d_{xy} are approximately the same and deviations from this can be used to identify outlier arrays. The dendrogram can serve to check if the experiments cluster in accordance with the sample classes.

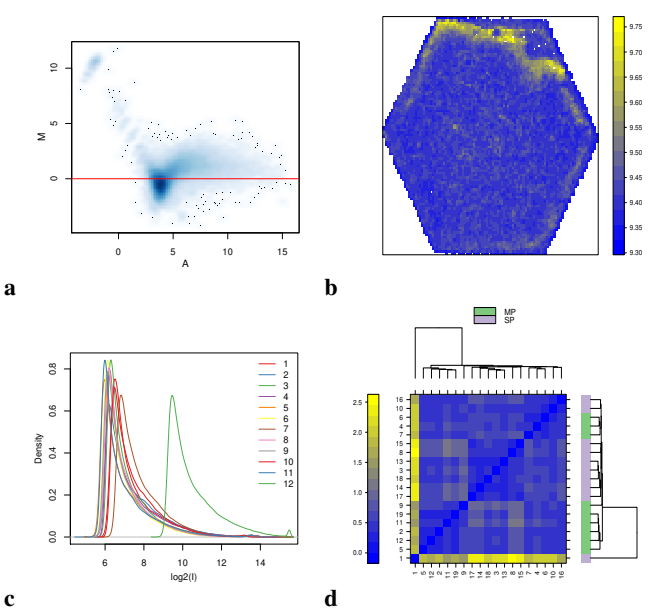

Fig. 1. a) MA-plot for an Agilent microarray. The M-values are not centered on zero meaning that there is a dependency between the intensities and the log-ratio. b) Spatial distribution of the background of the green channel for an Illumina chip. There is an abnormal distribution of high intensities at the top border of the array. c) Density plot of the log-intensities of an Affymetrix set of arrays (E-GEOD-349 ArrayExpress set). The density of one of the arrays is shifted on the x-axis. d) Heatmap of the ArrayExpress Affymetrix data set E-GEOD-1571. Array 18 is an outlier.

Affymetrix specific plots: Four Affymetrix specific metrics are evaluated if the input object is an *AffyBatch*. The RNA degradation plot from the affy package (Gautier *et al.*, 2004), the Relative Log Expression boxplots (RLE) and the Normalized Unscaled Standard Error (NUSE) boxplots from the affyPLM package (Brettschneider *et al.*, 2007) and the QC stat plot from the simpleaffy package (Wilson and Miller, 2005) are represented.

Scores: To guide the interpretation of the report, we have included the computation of numeric scores associated with the plots. Outliers are detected on the MA-plot, spatial distributions of the features' intensities, boxplot, heatmap, RLE and NUSE. The mean of the absolute value of M is computed for each array and those that lie beyond the extremes of the boxplot's whiskers are considered as possible outliers arrays. The same approach, i.e. using the whiskers of the boxplot, is applied to the following: the mean and interquartile range (IQR) from the boxplots and NUSE, the sums of the rows of the distance matrix, and the relative amplitude of low versus high frequence components of the Fourier transformation. In the case of the RLE plot, any array with a median RLE higher than 0.1 is considered an outlier.

Report: The metrics are rendered as figures with legends in a detailed report and the scores are used to provide a summary table. Examples of reports are provided at http://www.microarray-quality.org/quality_metrics.html.

CONCLUSION

arrayQualityMetrics supports the quality assessment of many types of microarrays in R. After preparation of the data, a single command line is used to create the report. The main benefits of arrayQualityMetrics are its simplicity of use, the ability to have the same report for different types of platforms, and the opportunity for users or developers to extend it for their needs. This tool can be used for individual data analyses or in routine data production pipelines, to provide fast uniform reporting.

ACKNOWLEDGEMENT

We would like to thank the developers of the R and Bioconductor packages that we are using, especially Ben Bolstad, Mark Dunning, Crispin Miller, Gregoire Pau and Deepayan Sarkar. AK is funded by EU FP6 (EMERALD, Project no. LSHG-CT-2006-037686). RG is funded by NIH P41HG004059.

REFERENCES

Brettschneider, J., Collin, F., Bolstad, B. M., and Speed, T. P. (2007). Quality assessment for short oligonucleotide arrays. *arXiv:0710.0178v2*.

Buness, A., Huber, W., Steiner, K., Sltmann, H., and Poustka, A. (2005). arrayMagic: two-colour cDNA microarray quality control and preprocessing. *Bioinformatics*, **21**, 554–6.

Craig Parman and Conrad Halling (2005). *affyQCReport: QC Report Generation for affyBatch objects*. R package version 1.17.0.

Dudoit, S., Yang, Y. H., Callow, M., and Speed, T. (2002). Statistical methods for identifying differentially expressed genes in replicated cdna microarray experiments. *Statist. Sinica*, **12**, 111–39.

Dunning, M. J., Smith, M. L., Ritchie, M. E., and Tavare, S. (2007). beadarray: R classes and methods for Illumina bead-based data. *Bioinformatics*, **23**, 2183–4.

Freue, G. V. C., Hollander, Z., Shen, E., Zamar, R. H., Balshaw, R., Scherer, A., ans Paul Keown, B. M., McMaster, W. R., and Ng, R. T. (2007). MDQC: a new quality assessment method for microarrays based on quality control reports. *Bioinformatics*, **23**, 3162–9.

Gautier, L., Cope, L., Bolstad, B. M., and Irizarry, R. A. (2004). affy – analysis of affymetrix genechip data at the probe level. *Bioinformatics*, **20**, 307–15.

Gentleman, R. C., Carey, V. J., Bates, D. J., Bolstad, B. M., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A. J., Sawitzki, G., Smith, C., Smyth, G. K., Tierney, L., Yang, Y. H., and Zhang, J. (2004). Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology*, **5**, R80.

Lee, E.-K., Yi, S.-G., and Park, T. (2006). arrayQCplot: software for checking the quality of microarray data. *Bioinformatics*, **22**, 2305–7.

Li, Q., Fraley, C., Bumgarner, R. E., Yeung, K. Y., and Raftery, A. E. (2005). Donuts, scratches and blanks: robust model-based segmentation of microarray images. *Bioinformatics*, **21**, 2875–82.

Petri, A., Fleckner, J., and Matthiessen, M. W. (2004). Array-a-lizer: a serial DNA microarray quality analyzer. *BMC Bioinformatics*, **5**, 12.

Schuchhardt, J., Beule, D., Malik, A., Wolski, E., Eickhoff, H., Lehrach, H., and Herzel, H. (2000). Normalization strategies for cDNA microarrays. *Nucleic Acids Res*, **28**, E47.

Wilson, C. L. and Miller, C. J. (2005). Simpleaffy: a Bioconductor package for Affymetrix quality control and data analysis. *Bioinformatics*, **21**, 3683–5.