Bioinformatics Advance Access published September 28, 2004

BIOINFORMATICS

arrayMagic: two-colour cDNA microarray quality control and preprocessing

Andreas Buneß, Wolfgang Huber, Klaus Steiner, Holger Sültmann and Annemarie Poustka

Department of Molecular Genome Analysis, German Cancer Research Center, INF 580, Heidelberg, 69120, Germany

ABSTRACT

Summary: *arrayMagic* is a software package for quality control and preprocessing of two-colour cDNA microarray data. The automated analysis pipeline comprises data import, normalisation, replica merging, quality diagnostics and data export. The script based processing combines reproducibility and flexibility at high throughput and provides quality assured and preprocessed microarray data to high-level follow-up analysis.

Availability: The R package *arrayMagic* is available with BSD license at http://www.bioconductor.org.

Supplementary Information: The package contains documentation in the form of manual pages and a vignette with a guided tour of a typical workflow.

Contact: a.buness@dkfz.de

Keywords: microarray; quality control; Bioconductor.

INTRODUCTION

Two-colour cDNA microarray technology has evolved to a routine laboratory procedure. Our motivation in implementing *arrayMagic* was to deal with the large amount of data generated by microarray projects in an efficient, reliable and reproducible way. We focused on preprocessing and quality assurance, leaving out high-level analysis which has to be adressed specifically.

The main design goal was to allow for the rapid construction of customised quality assessment and control (QA/QC) and preprocessing pipelines for such projects from a small set of building blocks. *arrayMagic* bridges the gap between the image quantification software and subsequent statistical and explorative analyses like testing for differential expression or classification. It simplifies the task of building processing pipelines that are *reproducible*, which means that even for idiosyncratic experimental designs and non-trivial combinations and selections of the data the whole procedure from raw data to normalised, quality-controlled, annotated, and summarised data is documented in a not too verbose script that can at any time be re-run or extended. The compendium technology (Gentleman, 2004) can be used to produce distributable objects containing the data as well as revivable documents reporting the processing.

We aimed to integrate normalisation methods, quality scores and visualisations that had been reported before. In addition, we provide tools for dealing with different microarray layouts within one experiment and for merging data from replicate probes or hybridisations. The researcher obtains an instant overview on the quality of the experiment.

NORMALISATION

Normalisation strategies for two-colour microarrays can be divided into two groups: adjustment of the colour channels or of the log-ratios. Moreover, depending on the experimental design and the objectives either a single channel intensity or a log-ratio based analysis might be more appropriate.

The tool offers log-ratio based normalisation by means of the *loess* method (Yang et al., 2002) and and direct intensity-based normalisation by means of *vsn* (Huber et al., 2002) and *quantile normalisation* (Bolstad et al., 2003). We will use the terms “log-ratios” and “log-transformed intensities” also for the data resulting from the *vsn* method. Groups of hybridisations, subsets of spots, e.g. by grid, print-tip or PCR plate, as well as colour channels can be normalised separately. Plots characterizing the distributions of the log-ratios and colour channels before and after normalisation are generated, for an example see Figure 1 (middle).

QUALITY CONTROL AND ASSESSMENT

Quality assured data are prerequisite for any reliable high-level analysis. In addition, quality control allows to monitor and improve the laboratory procedures.

The quality of hybridisations is best assessed in the context of normalisation. In a model based approach like *vsn*, the model is a summary of past experience and our expectations on the data. Thus, it can be used to identify hybridisations or groups of measurements that do not fit. Other methods like *loess* or *quantile* normalisation put more emphasis on making the data conform in any

situation. In these cases, statistics of the data distribution can be calculated (for example, location and scale of the distribution of normalised log-ratios) and compared against expectations. Moreover, as long as the majority of the data is thought to be acceptable, outlier detection methods can be used for quality control.

Visual inspection of the data is supported by spatial false–colour representations of fore- and background intensities and the log-ratios. This allows to detect scratches and artifacts (Figure 1). Most notably, the spatial plots of the normalised data are useful for assessing the necessity of background correction and for assuring spatial homogeneity of the data.

Several quality scores are calculated, stored in a report file, and are visualised in part. These scores include spot replicate concordance, the correlation of the two colour channels and a robust measure of noise W for each hybridisation. W is defined as the median absolute deviation of the normalised log-ratios q_i, i. e. $W = \text{mad}_i(q_i) = \text{median}_i(|q_i - \text{median}_j(q_j)|)$. A minority of differentially expressed genes should not disturb W.

We do not find it practical to define universally applicable thresholds on quality scores. They should be evaluated not on the level of a single hybridisation, but in the context of all data in the experiment. In our experience this has been very useful in detecting outliers in large scale experiments. In particular, a global view on all pairwise similarities between all hybridisations as shown in Figure 1 has proved to be useful.

For two arrays a and b, we define a similarity score $S_{ab} = \text{mad}_i(x_{ia} - x_{ib})$, where x_{ia} can be the log-ratio of the i-th probe on the a-th array, or the log-transformed normalised intensity of an individual colour channel. Especially in the case of biologically related samples, this is an informative measure of similarity.

IMPLEMENTATION

The software is implemented in the R language (Ihaka and Gentleman, 1996) and integrates into the Bioconductor project (Gentleman et al., 2004), an open source software project for bioinformatics. It uses building blocks from the packages *Biobase*, *vsn*, and *limma*. The software works on Linux, Windows and MacOS.

CONCLUSION

The open source software tool *arrayMagic* facilitates the analysis of two colour cDNA microarray data. It aims to provide quality assured and normalised data. The script based pipeline supports reproducible batch-like processing. The workflow starts with quantified image scan result files. Several quality scores and diagnostics are calculated and visualised, offering a broad view. The processed data can be exported as HTML-file or as tab-delimited file with spot and sample annotation and may serve as input for follow-up analysis in commonly used tools of choice. Naturally, high level follow-up analysis in the framework of R and Bioconductor is supported by adequate representation of the data. Documentation of all functionality and a step-by-step example following a typical workflow is part of the package.

REFERENCES

Huber W., von Heydebreck A., Sültmann H., Poustka A., and Vingron M. (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics* **18** Suppl. 1, S96–S104.

Gentleman R. (2004) Reproducible Research: A Bioinformatics Case Study. *Statistical Applications in Genetics and Molecular Biology* **3**.

Bolstad, B. M., Irizarry R. A., Astrand, M., and Speed, T. P. (2003) A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Bias and Variance. *Bioinformatics* **19**, 185–193.

Yang, Y. H., Dudoit S., Luu P., Lin D. M., Peng V., Ngai J., Speed T. P. (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.* **30**(4):e15.

Gentleman, R., V. J. Carey, D. J. Bates, B. M. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. A. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. K. Smyth, L. Tierney, Y. H. Yang, and J. Zhang (2004) Bioconductor: Open software development for computational biology and bioinformatics. *Bioconductor Project Working Papers.* Working Paper 1. http://www.bepress.com/bioconductor/paper1

Ihaka, R. and R. Gentleman (1996) R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics* **5**, 299–314.

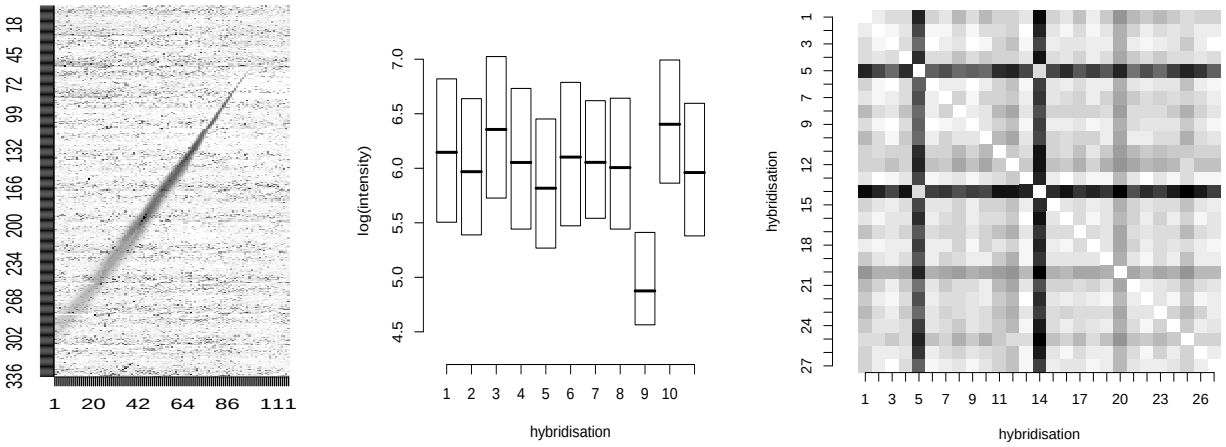

Fig. 1. *Left:* A comet tail like artifact that was discovered in the green intensity channel of a microarray. *Middle:* Simplified box plots of the distributions of the red intensities for 11 hybridisations. Low intensity levels are found for the ninth hybridisation. *Right:* (Dis)similarities between all pairs of a set of 27 hybridisations. White indicates high and black low similarity. The plot allows to unveil outliers and potential technical artifacts like batch effects. The two crosses identify outliers. Visualizations like in the middle and the right are useful for data sets of up to about 200 microarrays.