Analysis of Microarray Data

Matthias Kohl*

Department of Mechanical and Process Engineering, Furtwangen University, Jakob-Kienzle-Str. 17, D-78054 VS-Schwenningen, Germany

Abstract: We give a brief overview over necessary steps in the analysis of microarray data. We cover quality control, preprocessing, statistical as well as enrichment analysis.

Keywords: Microarray, quality control, preprocessing, normalization, statistical analysis, enrichment analysis.

INTRODUCTION

We consider microarrays more precisely oligonucleotide arrays which are employed for various applications today and have been validated by the microarray quality control (MAQC) projects I and II [1, 2]. Due to the complexity of the experiments and the high dimensionality of the data the analysis consists of several steps to obtain reliable and reproducible results.

QUALITY CONTROL

A detailed quality control (QC) of the data is an important first step in every data analysis as good data quality is the key to reliable and reproducible results. The goal of QC besides assessing the quality of the data is the verification of necessary assumptions for the statistical analysis. There are even two QCs in case of microarray data. First, a QC of the raw data and secondly, the QC of the so-called preprocessed data which usually also includes the selection of the most appropriate preprocessing method.

The quality control of microarray data for the very most part consists of diagnostic plots. Typically, one considers images or image plots of the processed arrays to identify spatial artifacts, box plots and density plots to find arrays with unusual signal distributions, and plots of the negative and positive control probes incorporated into the arrays for checking the experimental procedures. Furthermore, one applies dimension reduction techniques for outlier identification such as hierarchical clustering, principal component analysis, or similarity plots. For the preprocessed data it is also important to check the dependence of the signal variability on the signal intensity as homogenous variances (homoscedasticity) is a basic assumption for many statistical models.

PREPROCESSING

The preprocessing of microarray data usually begins with a background correction aiming at a reduction of the background noise. In the simplest case it consists of a subtraction of the background signal intensities from the foreground signal intensities. The next step for many microarray platforms is the aggregation of technical replicates; e.g., in case of Affymetrix GeneChip Arrays the perfect match (PM) and mismatch (MM) probe pairs of a probe set are aggregated to a single intensity value, in case of Illumina BeadArrays the so-called bead level data are summarized to so-called bead summary data.

These two preprocessing steps are often integrated in the next and most important step, the so-called normalization. The normalization typically includes a variance stabilizing transformation of the data which in most cases is a log-transformation. The goal of normalizing the data is the reduction or ideally the compensation of the technical bias unavoidable in complex experiments. This bias may for instance be caused by different amounts of starting RNA, different labeling efficiencies, different dye detection efficiencies if two or more dyes are used or other variations in experimental conditions such as minimal fluctuations in temperature, minimal differences of the pipetted amounts, a change of the experimenter, time lags between experiments. From a statistical point of view normalizing the data leads to a harmonization of the signal intensities of the processed microarrays thus increases the reliability and reproducibility of the results [3,4].

There are quite a lot of preprocessing methods e.g. more than 30 in case of Affymetrix GeneChip Arrays [5] and there is no gold standard. Hence, several preprocessing methods are usually applied and it is an

*Address corresponding to this author at the Department of Mechanical and Process Engineering, Furtwangen University, Jakob-Kienzle-Str. 17, D-78054 VS-Schwenningen, Germany; Tel: +49 (0) 7720 307-4746; Fax: +49 (0) 7720 307-4727; E-mail: Matthias.Kohl@hs-furtwangen.de

E-ISSN: 1929-6029/12 © 2012 Lifescience Global

important part of the QC of the preprocessed data to identify the most appropriate preprocessing method for the data at hand.

STATISTICAL ANALYSIS

For the statistical analysis of microarray data all kind of uni- and multivariate statistical procedures are employed depending on the experimental design of the study. In particular, the widespread use of microarrays has forced the development of new methods for data with a large number of variables at small or moderate sample sizes, e.g. empirical Bayes methods or various multiple-testing procedures [6, 7]. As has been shown by the MAQC-I project the reproducibility is likely to benefit from a combination of adjusted p values with criteria such as fold changes [1].

ENRICHMENT ANALYSIS

The statistical analysis of microarray data usually ends with a large list of hundreds or even thousands of interesting genes. The biological interpretation of such a large list of genes is very challenging and cannot be achieved without modern bioinformatics methods accessing the biological knowledge collected in public databases. These so-called enrichment methods split large gene lists into smaller subsets and attempt to identify the most enriched and pertinent biology. As conclusions are based on a group of genes instead of an individual gene the likelihood for identifying the correct biological processes is increased. By comparing the gene list subsets with the gene background (gene universe) of the selected population the enrichment can be quantitatively measured by classical statistical methods such as Fisher's exact test. Today there are numerous tools for enrichment analysis [8].

ACKNOWLEDGEMENT

We would like to thank two anonymous referees for valuable comments on the manuscript.

APPENDIX OF SYMBOLS

MAQC = microarray quality control

QC = quality control

PM = perfect match

MM = mismatch

RNA = ribonucleic acid

REFERENCES

[1] MAQC-Consortium. The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements. Nat Biotechnol 2006; 24: 1151-61. http://dx.doi.org/10.1038/nbt1239

[2] MAQC-Consortium. The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. Nat Biotechnol 2010; 28: 827-38. http://dx.doi.org/10.1038/nbt.1665

[3] Quackenbush J. Microarray data normalization and transformation. Nat Genet 2002; 32(Suppl): 496-501. http://dx.doi.org/10.1038/ng1032

[4] De Bruyne V, Al-Mulla F, Pot B. Methods for microarray data analysis. Methods Mol Biol 2007; 382: 373-91. http://dx.doi.org/10.1007/978-1-59745-304-2_23

[5] Irizarry RA, Wu Z, Jaffee HA. Comparison of Affymetrix GeneChip expression measures. Bioinformatics 2006; 22(7): 789-94. http://dx.doi.org/10.1093/bioinformatics/btk046

[6] Smyth GK. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol 2004; 3(1): Article 3.

[7] Dudoit S, van der Laan MJ. Multiple Testing Procedures and Applications to Genomics. Springer, New York 2008.

[8] Huang da W, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res 2009; 37(1): 1-13. http://dx.doi.org/10.1093/nar/gkn923

Received on 18-09-2012 Accepted on 03-10-2012 Published on 08-10-2012

http://dx.doi.org/10.6000/1929-6029.2012.01.01.09