Detection and correction of probe-level measurement artefacts on microarrays

Tobias Petri [1], Evi Berchtold [1], Ralf Zimmer [1] and Caroline C. Friedel [1,2]

[1]Institute for Informatics, Ludwig-Maximilians-Universität München, Munich 80333, Germany
[2]Institute of Pharmacy and Molecular Biotechnology, Heidelberg University, Heidelberg 69120, Germany

Quality control for microarrays remains a major issue in gene expression analysis as reliability of measurements may be affected by many factors in each step of the experimental process . Accordingly, many methods and software tools have been developed for quality assessment of microarrays and identification of microarrays of poor quality [1–4].

While most of these methods are appropriate for detecting experimental artefacts in individual samples, it is usually not clear how to proceed once such artefacts have been detected. So far, the two alternatives are (1) to either completely exclude or (2) include the corresponding arrays for any subsequent analysis. Both of these approaches have disadvantages. In the first case, the corresponding measurements are not available for gene expression profiling and may even have to be repeated if they are crucial to the analysis. This can be cost-intensitve given appropriate samples are no longer available. In the second case, one has to assume that normalization and summarization methods can correct for the measurement errors.

This assumption is based on the specific construction of microarrays where probes of the same probeset are not contiguous on the array. Thus, smaller artefacts due to uneven hybridization or other experimental problems may only affect a subset of probes for a probeset. In this case, it is usually presumed that summarization methods, such as RMA [5], which combine the values for individual probes to a probeset value, can estimate the probeset value correctly despite measurement errors for some probes. In this study, we show that this assumption is risky and in some cases invalid. We illustrate this by showing that even small artefacts on the array can have a significant effect on the overall expression levels even for probesets not affected by the artefacts. Using simulations, we find a clear dependency between the fraction of probes for a probeset affected by the artefacts and the performance of summarization methods in correctly estimating the probeset value. However, while one or two affected probes per probeset still allow a reasonable estimation of the probeset value, the reproducibility for these probesets is still considerably lower than for probesets which are not affected at all.

These results clearly show that summarization applied to all probes on the array is not an appropriate approach to address measurement artefacts. Instead, we propose an alternative strategy which is based on identifying corrupted probes first and replacing them by the mean of unaffected probes belonging to the same probeset before summarization. Different criteria can be used for evaluating the noise level for individual probes based either on technical replicates, simultaneous measurements of RNA synthesis and decay or residuals determined by the summarization method itself. We describe two strategies for identifing the corrupted probes based on the probe noise levels: one simple threshold-based approach and one approach which incorporates the neighbourhood information on the array. We show that by including this neighborhood information in the analysis, defective probes can be identified with higher accuracy and the reproducibility of the estimated probeset levels can be increased. In this way, measurement errors can be corrected for and the corresponding microarray experiments can be salvaged.

References

1. Bolstad, B., Collin, F., Brettschneider, J., Simpson, K., Cope, L., Irizarry, R., and Speed, T. (2005) Quality Assessment of Affymetrix GeneChip Data. In Gail, M., Krickeberg, K., Samet, J., Tsiatis, A., Wong, W., Gentleman, R., Carey, V. J., Huber, W., Irizarry, R. A., and Dudoit, S., editors, *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, Statistics for Biology and Health, 33–47. Springer New York.
2. Wilson, C. L. and Miller, C. J. (2005) Simpleaffy: a BioConductor package for Affymetrix Quality Control and data analysis. *Bioinformatics*, **21**, 3683–3685.
3. Freue, G. V. C., Hollander, Z., Shen, E., Zamar, R. H., Balshaw, R., Scherer, A., McManus, B., Keown, P., McMaster, W. R., and Ng, R. T. (2007) MDQC: a new quality assessment method for microarrays based on quality control reports. *Bioinformatics*, **23**, 3162–3169.
4. Kauffmann, A., Gentleman, R., and Huber, W. (2009) arrayQualityMetrics–a bioconductor package for quality assessment of microarray data. *Bioinformatics*, **25**, 415–416.
5. Irizarry, R. A., Bolstad, B. M., Collin, F., Cope, L. M., Hobbs, B., and Speed, T. P. (2003) Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res*, **31**, e15.