Data quality checking in highly replicated 2-dye microarray data sets

Stuart McHattie[1], Andrew Mead[2], Linda Hughes[2] and Vicky Buchanan-Wollaston[2]

[1]Systems Biology Centre, University of Warwick, UK
[2]Warwick HRI, University of Warwick, UK

2-dye microarrays are an increasingly common, but typically noisy, method of identifying gene expression changes between biological samples. As with all biological experiments, microarray data sets may be affected by various extraneous sources of variation, including dye biased binding affinity, slide printing errors and physical damage to slides caused by uneven drying or poor handling before scanning. These sources of variation will inevitably introduce more noise into any data analysis, potentially masking important treatment effects. But often the vast quantity of data generated in a microarray experiment appears to be a barrier to the data checking that is common in almost all other areas of biological data analysis, with standard "normalisation" approaches often employed automatically without looking at the data. Systematic, but semi-automated, data quality checking should lead to the identification of erroneous data points, and a consequent reduction in the background variability for microarray data, thus allowing a more powerful assessment of treatment effects.

Working with a data set obtained from 176 2-dye customised CATMAv3 arrays (32448 spots), to study patterns of gene expression in the leaves of *Arabidopsis thaliana* during natural senescence, we are further developing a suite of data quality checking functions within the R [1] microarray analysis package MAANOVA [2]. In contrast to many microarray data analysis approaches, MAANOVA uses a mixed model approach to analyse the observed data from each of the two dyes on each array directly, rather than first converting the data to ratios. This method was chosen as it fitted more closely with the classical experimental design approach (treating both arrays and dyes as blocking factors) used to construct the experiment.

The data quality checking tools originally included within the MAANOVA package provided a good graphical summary of the observed data, perfectly adequate for a small experiment (relatively few spots per array, relatively few arrays). But the size of our experiment meant that it would be almost impossible to visually check every graphical summary to identify potentially erroneous observations. Hence some quantification of the graphical summaries was needed allowing a more automated approach to data quality checking. Comparisons of the responses for each channel (two different treatments) on each array (GRIDCHECK) and of the responses for every pair of technical replicates (same treatment on different arrays) (TECHREPCHECK) were quantified using major axis regression (via principal component analysis). Identification of large residuals in these analyses provides an approach to eliminating anomalous technical replicates on a gene-by-gene basis. Quantification within another data quality checking approach (ARRAYVIEW) provides an assessment of spatial trends in expression intensities that can then be eliminated via various normalisation procedures. Further work is needed to establish the effect of the omission of data points, using these tools, on the formal analysis of the complete data set.

References

[1] R Development Core Team (2007). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.
[2] Hao Wu, modified by Hyuna Yang with ideas from Gary Churchill, Katie Kerr and Xiangqin Cui (2008). maanova: Tools for analyzing Micro Array experiments. R package version 1.8.1. http://www.jax.org/staff/churchill/labsite/software/Rmaanova/