SCHNABLE LAB
Plant Genomics

2035C Roy J. Carver Co-Laboratory
Plant Sciences Institute
Iowa State University
Ames, Iowa 50011-3650
515 294-7209
FAX 515 294-5256

STATISTICAL DESIGN AND ANALYSIS PROTOCOLS FOR THE MAIZE SHOOT APICAL MERISTEM (SAM) PROJECT

funded by
National Science Foundation Award DBI-0321595

Experimental Design

Each microarray experiment in the SAM project involves a comparison of two cell types using two-color microarrays. The work of Dobbin, Shih, and Simon (2003) shows that the optimal design for identifying cell-type differences for a fixed number of microarray slides involves

1. pairing samples of different cell types on each slide,

2. assuring that the two possible assignments of dyes to cell types are each used on an equal number of slides, and

3. measuring each independent cell-type sample on exactly one slide.

Such a design is depicted in Figure 1 along with some competing designs that could be used to compare two cell types with two color microarrays. The superiority of Design A to Design B in Figure 1 illustrates that adding new pairs of samples measured on only one slide each is better than measuring existing samples with multiple slides. The superiority of Design B to C in Figure 1 illustrates that it is better to measure existing samples twice rather than only once if additional biological samples are unavailable. Finally the superiority of Design A to Design C indicates that increasing the number of independent sample pairs increases the quality of the design. We typically use at least 6 independent pairs of samples. It is quite difficult to separate genes differentially expressed between cell types from other genes using fewer than 6 independent sample pairs.

Another design (not depicted in Figure 1) that is often used for the two cell-type comparison is the so called reference design in which samples of each cell type are each compared on slides against a common reference sample. Dobbin, Shih, and Simon (2003) show that this design is inferior to our proposed design when the goal is to identify genes that are differentially expressed between cell types using a fixed number of slides.

SCHNABLE LAB
Plant Genomics

2035C Roy J. Carver Co-Laboratory
Plant Sciences Institute
Iowa State University
Ames, Iowa 50011-3650
515 294-7209
FAX 515 294-5256

Data Analysis

We typically use separate mixed linear model analysis for each gene (Wolfinger et al., 2001) coupled with False Discovery Rate considerations (Storey and Tibshirani, 2003) to identify genes whose mean expression differs between cell types. Prior to mixed linear model analyses, we normalize data within each slide using the LOWESS normalization approach (Dudoit et al., 2002; Yang et al., 2002) to account for the possibility of intensity-dependent dye bias. We normalize data across slides by adding a constant to the log-scale data for each combination of slide and dye so that the median of the log-scale data will be the same for all combinations of slide and dye (Yang et al., 2002).

The mixed linear model that we consider for each gene includes fixed effects for cell types and dyes along with random effects for slides. (Though we apply normalization strategies intended to remove dye effects, gene-specific dye effects often persist following normalization; thus we find it necessary to include dye effects in our gene-specific mixed models.) A simple and effective way to conduct these mixed linear model analyses is to perform a regression analysis on the difference in normalized log-scale signals between the two dye channels for each gene and slide. An R script for normalizing, analyzing, and plotting data is provided here.

Other Resouces

Co-PI Dan Nettleton teaches a course on the statistical design and analysis of microarray experiments each spring. Additional information addressing many aspects of statistical design and analysis of microarray experiments is available on his course Web page.

Many R functions for microarray data analysis are freely available at Bioconductor, an open source and open development software project for the analysis and comprehension of genomic data.

SCHNABLE LAB
Plant Genomics

2035C Roy J. Carver Co-Laboratory
Plant Sciences Institute
Iowa State University
Ames, Iowa 50011-3650
515 294-7209
FAX 515 294-5256

References

Dobbin, K., Shih, J. H., and Simon, R. (2003). Statistical design of reverse dye microarrays. *Bioinformatics*, **19**(7), 803-810.

Dudoit, S., Yang, Y. H., Callow, M. J., and Speed, T. P. (2002). Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. *Statistica Sinica*, **12**, 111-140.

Kerr, M. K., and G. A. Churchill. (2001). Experimental design for gene expression microarrays. *Biostatistics*, **2**, 183-201.
Storey, J. D., and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, **100**, 9440-9445.

Wolfinger, R. D., Gibson, G., Wolfinger, E. D., Bennett, L., Hamadeh, H., Bushel, P., Afshari, C., and Paules, R. S. (2001). Assessing gene significance from cDNA microarray expression data via mixed models. *Journal of Computational Biology*, **8**, 625-637.

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 Research*, **30**, No. 4, e15.