

## STATISTICAL ISSUES IN MICROARRAY EXPERIMENTS INCLUDE:

- Experimental design
- Preprocessing and data cleansing
- Normalization
- Differential expression testing
- Clustering and prediction
- Annotation

### Experimental Design

A sound experimental design is crucial for an informative microarray experiment. Issues are described by Pan et al. (2002), Simon et al. (2002) and Kerr and Churchill (2001); and S-PLUS code for sample size calculations and power curve construction from the Pan et al. (2002) paper are available from the first author's [website](#).

### Data Preparation and Normalization

Exploratory data analysis (EDA), quality control and normalization are achieved quite simply in S-PLUS. There have been many normalization methods proposed and many in current use. Yang et al. (2002) and Bolstad et al. (2003) provide good reviews and methods. Simple normalization of replicate chips to the same interquartile range and median is achieved as follows for chip data in an S-PLUS dataframe `x` with rows as genes and columns as expression intensities:

```
iqrfn <- function(xx) quantile(xx,0.75,na.rm=T)-
quantile(xx,0.25,na.rm=T)
btwn.norm <- function(tmp)
{
#Adjust IQ ranges to be the same as max of IQRs
divisor <- matrix(rep(apply(tmp,2,iqrfn)/max(apply(tmp,2,iqrfn)),
dim(tmp)[1]), nrow=dim(tmp)[1],byrow=T)
tmp.adj <- tmp/divisor
#Adjust medians to be the same as max of medians
adjustment <- matrix(rep(max(apply(tmp.adj,2,median,na.rm=T))
-apply(tmp.adj,2,median,na.rm=T),
dim(tmp.adj)[1]),nrow=dim(tmp.adj)[1],byrow=T)
tmp.adj2 <- tmp.adj+adjustment
return(tmp.adj2)
}
x.norm <- btwn.norm(x)
```

## Differential Expression Analysis using Hierarchical Bayes approach

Direct calculation of posterior probabilities of differential expression using a hierarchical Bayes approach is described by Newton et al. (2001) and S-PLUS code for this approach is available from the first author's [website](#).

## Clustering and Prediction

There are many clustering and prediction methods available in S-PLUS. The hierarchical clustering methods include agglomerative: `hclust()`, `agnes()`; and divisive: `diana()`. Partitioning clustering methods include `pam()` and `kmeans()`. The dendrograms produced by the hierarchical methods may be visualized using `plclust()` and layered over heatmaps of the expression intensity data produced using `image()` to produce the now familiar visualization of microarray experimental data (Eisen et al., 1998). Model based clustering methods, using mixture models, are available using the function `mclust()` (Fraley and Raftery, 2002), and a repository of S-PLUS code is available at [this link](#).

There are many other supervised and unsupervised learning approaches to microarray data analysis. S-PLUS code for some of these methods is available from the [GeneClust site](#). This includes S-PLUS code for gene shaving (Hastie and Tibshirani, 2000; Do et al., in press).

## Acknowledgement

Many of the statistical methods included in Insightful ArrayAnalyzer are courtesy of the open-source [Bioconductor Project](#), which is included in part with ArrayAnalyzer. We recognize particularly the assistance provided by Robert Gentleman (Harvard Medical School), Sandrine Dudoit (UC Berkeley) and Rafael Irizarry (Johns Hopkins University).

## References

Benjamini Y, Hochberg Y (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B, Methodological* 57:289-300.

Bolstad B.M., Irizarry RA, Astrand M, Speed TP (2003). A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Bias and Variance. To appear in *Bioinformatics*.

Chambers JM (1998). *Programming with Data: A guide to the S language*. Springer.

Do K, Broom, Wen (2003). GeneClust. To appear in *The Analysis of Gene Expression Data: Methods and Software*. Edited by G Parmigiani, ES Garrett, RA Irizarry and SL Zeger. Published by Springer, New York.

Dudoit S, Yang YH, Speed TP, Callow MJ (2002). Statistical methods for identifying genes with differential expression in replicated cDNA microarray experiments. *Statistica Sinica* Vol. 12, No. 1, p. 111-139.

Eisen MB, Spellman PT, Brownand PO, Botstein D (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences USA* 95(25):14863-14868.

- Fraley C. and Raftery A. E. (2002). MCLUST: Software for Model-Based Clustering, Discriminant Analysis and Density Estimation. Technical Report no. 415, Department of Statistics, University of Washington.
- Hastie T, Tibshirani R, Eisen MB, Alizadeh A, Levy R, Staudt L, Chan WC, Botstein D, Brown P (2000). "Gene Shaving" as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biology* 1:research0003.1-research0003.21
- Irizarry, RA, Hobbs, B, Collin, F, Beazer-Barclay, YD, Antonellis, KJ, Scherf, U, Speed, TP (2003) Exploration, Normalization, and Summaries of High Density Oligonucleotide Array Probe Level Data. Accepted for publication in *Biostatistics*.
- Kerr MK, Churchill GA (2001). Statistical design and the analysis of gene expression microarray data. *Genetic Research* 77:123-128.
- Lee JK and O'Connell M (2003). An S-PLUS library for the analysis of differential expression. To appear in *The Analysis of Gene Expression Data: Methods and Software*. Edited by G Parmigiani, ES Garrett, RA Irizarry and SL Zeger. Published by Springer, New York.
- Li C, Wong W (2001). Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *Proceedings of the National Academy of Science U S A* 98:31-36.
- Newton MA, Kendzioriski CM, Richmond CS, Blattner FR, Tsui KW (2001). On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *Journal of Computational Biology* 8:37-52.
- Pan W, Lin J and Le C (2002) "How Many Replicates of Arrays are Required to Detect Gene Expression Changes in Microarray Experiments? A Mixture Model Approach". *Genome Biology*, 3:5: research0022.1- research0022.10.
- Simon R, Radmacher MD, Dobbin K (2002). Design of studies using dna microarrays. *Genetic Epidemiology* 23:21-36.
- Storey JD. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society, Series B*, 64: 479-498.
- Tibshirani R, Hastie T, Narashiman and Chu (2002): [Diagnosis of multiple cancer types by shrunken centroids of gene expression](#). PNAS 2002 99:6567-6572.
- Wolfinger RD, Gibson G, Wolfinger E, Bennett L, Hamadeh H, Bushel P, Afshari C, Paules RS (2001). Assessing gene significance from cDNA microarray expression data via mixed models. *Journal of Computational Biology* 8:625-637.
- Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed T (2002). Normalization for cdna microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Research* 30(4):e15.