Genomics in Breast Cancer
Ready for Prime-Time?

Richard Simon, D.Sc.
Chief, Biometric Research Branch
National Cancer Institute
http://linus.nci.nih.gov/brb

http://linus.nci.nih.gov/brb

- Powerpoint presentation
- Reprints & Technical Reports
  - Myths & Truths
    About Microarray Expression Profiling
- BRB-ArrayTools software

- *Design and analysis of DNA Microarray Investigations*
  - R Simon, EL Korn, MD Radmacher, L McShane, G Wright, Y Zhao
  - Springer, 2003

Myth

- That the greatest challenge is managing the mass of microarray data
- Greater challenges are:
  - Effectively designing and analyzing experiments that utilize microarray technology
  - Organizing and facilitating effective interdisciplinary collaboration with statisticians, clinicians & biologists
  - Designing and conducting proper validation studies

- Myth that microarray studies should be based on data mining of archives to find interesting patterns that give clear answers to questions that were never asked
- Good microarray studies have clear objectives, but not generally gene specific mechanistic hypotheses

Design and Analysis Methods Should Be Tailored to Study Objectives

- Class Comparison
  - Find which genes are differentially expressed among pre-defined classes of samples
    - Responders vs non-responders
    - Patients who develop mets vs those who don't
- Class Prediction
  - Prediction of class (phenotype) using gene expression profile
- Class Discovery
  - Discover clusters of specimens with similar expression profiles

- Cluster analysis is appropriate only for class discovery
  - Cluster analysis is subjective and always produces clusters
  - Cluster analysis if frequently used in a misleading way
- *Supervised methods* are more appropriate for class comparison and class prediction

Many Microarray Studies Do Not Address A Medically Relevant Question

- Comparing expression in AML vs ALL
- Finding genes whose expression correlates with RFS in a heterogeneous group of primary breast cancer patients is usually not therapeutically meaningful
  - N+, N-, ER+, ER-, Systemic rx

Fit of a Model to the Same Data Used to Develop it is No Evidence of Predictive Accuracy

- When the number of candidate predictors (p) exceeds the number of cases (n), perfect prediction on the same data used to create the predictor is always possible

Validation of a Predictor

- In-study validation
  - Re-substitution estimate
    - Horribly biased
  - Split-sample validation
    - Develop one fully specified model
  - Cross-validation
    - Often used incorrectly
- Independent data validation

Leave-one-out Cross Validation

- Omit sample 1
  - Develop multivariate classifier *from scratch* on training set with sample 1 omitted
  - Predict class for sample 1 and record whether prediction is correct

Leave-one-out Cross Validation

- Repeat analysis for training sets with each single sample omitted one at a time
- e = number of misclassifications determined by cross-validation
- Subdivide e for estimation of sensitivity and specificity

Independent Data Validation

- From different clinical centers
- Specimens assayed at different time from training data
- Positive and negative samples collected in the same way
- Study sufficiently large to give precise estimate of sensitivity and specificity of the multivariate classifier
- The validation study is prospectively planned
  - patient selection pre-specified to address a therapeutically relevant question
  - endpoints and hypotheses pre-specified
  - predictor fully pre-specified
  - Study addresses assay reproducibility
  - Specimens may be either prospective or archived

Development of Whole Genome Technologies Has Created Serious Challenges for Biomedicine

- The journal review process is disfunctional for properly screening microarray and proteomic studies
  - Major publications with poor design, erroneous statistical analysis and misleading conclusions
- Interdisciplinary collaboration between biologists, clinicical investigators and biostatisticians is inadequate for publishing valid results and for effective utilization of new technology
- The nature of biological investigation is changing
  - biomedical scientists need training in design and analysis of experiments with high-dimensional data
  - Statistical bioinformatics needs to be recognized and supported as an essential research area, not as a support service

Genomics in Breast Cancer Ready for Prime Time?

- Some important applications of genomics to breast cancer are ready for the prime time
- There are too few adequate validation studies of therapeutically important hypotheses and the nature of proper validation is not sufficiently understood
- Biomedical science cannot afford to have major genomic projects with the capacity to alter medical practice conducted without the collaboration of statisticians with sufficient expertise and independence to ensure that statistically valid methods of design and analysis are employed.