BRB-ArrayTools

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

- Reprints, Presentations, Technical reports
- BRB-ArrayTools
  - Registration & Download
    - 5000+ registered users in 60 countries
  - Message board
  - Archive of human tumor gene expression data with clinical/pathological accompanying data
- Microarray Myths


Challenges in Effective Use of DNA Microarray Technology

• Design & Analysis are bigger challenges than data management.
  – Much greater opportunity for misleading yourselves and others than traditional single gene/protein studies

• Limited availability of experienced statistical collaborators

• Predominance of hype, mis-information, and dangerous methods promulgated by biomedical scientists as well as professional statistical/computational scientists

• Predominance of flashy software that encourages misleading analyses
Objectives of BRB-ArrayTools

• Provide biomedical scientists access to statistical expertise for the analysis of expression data

• Provide biomedical scientists and statistical/computational fellows
  – training in analysis of high dimensional data
  – access to critical assessment of methods published in a rapidly expanding literature
BRB-ArrayTools

• Integrated package
• Excel-based user interface
  – Doesn’t use Excel analyses
  – state-of-the art analysis methods programmed in R, Java & Fortran
  – Data not stored as worksheets
    • >1000 arrays and 65000 genes per project
• Based on continuing evaluation of validity and usefulness of published methods
  – Methods carefully selected by R Simon
  – Not a repository like Bioconductor
• Publicly available for non-commercial uses from BRB website:
BRB-ArrayTools

• Not tied to any database
  – Importer for common databases and platforms
    • MadB, GenePix, MAS5/GCOS
    • Imports .cel files
    • Import wizzard for any files output by image analysis program
  – Import (collate)
    • Expression data (eg separate file for each array)
    • Spot (probeset) identifiers
    • Experiment descriptor worksheet
      – Rows correspond to arrays
      – Columns are user defined phenotypes to drive the analyses
        » Can be updated during analysis
  – Imported data saved as project folder containing project workbook and binary files
    • Project workbook can be re-opened in Excel at any time
    • Output saved in html files in output folder
BRB-ArrayTools

• Highly computationally efficient
  – Non-intensive analyses in R
  – Intensive analyses in FORTRAN
    • eg BRB-AT version of SAM is 9x + more efficient than Bioconductor or web based versions
      – And more accurate

• Extensive gene and pathway annotation features
BRB-ArrayTools

• Plug-in facility for user written R functions
• Message board and listserve
• Extensive built-in help facilities, tutorials, datasets, usersguide, data import and analysis wizzards, sample statistical analysis sections, …
BRB-ArrayTools Archive of Human Tumor Expression Data

- Archive of BRB-ArrayTools zipped project folders of expression profiles of human tumors and associated clinical/pathological descriptors
  - Published data
- Easy way to archive your data and to analyze someone else’s data
  - Download, unzip, open in Excel
• Effective microarray research requires clear objectives, careful planning and appropriate statistical analysis

• Clear objectives, but not gene specific mechanistic hypotheses
Design and Analysis Methods Should Be Tailored to Study Objectives

- **Class Comparison**
  - Find genes that are differentially expressed among conditions or tissues

- **Class Prediction**
  - Prediction of response to treatment using gene expression profile

- **Class Discovery**
  - Discover clusters of specimens or genes whose expression profiles are similar
BRB-ArrayTools
Unsupervised Analysis Tools

• Scatterplot
  – One array vs another array
  – Phenotype averages
    • eg arrays for males vs females

• Cluster Analysis
  – Includes Cluster & Treeview internally
  – Native hierarchical cluster analyses
  – Cluster stability and reproducibility for clustering arrays
  – Multicolored dendrograms
BRB-ArrayTools
Unsupervised Analysis Tools

• Rotating 3-D principal component plots
  – Controls for direction of spin
  – Brushing of points for identification
  – Color coding of points
  – Saves plot as Powerpoint presentation with active controls
• 10,000 non-differentially expressed genes x 5% false positivity rate equals 500 false positives

• 10,000 x 0.1% = 10 false positives
BRB-ArrayTools
Class Comparison

- Distinguish biological from technical variability
- Univariate significance (p<0.001)
  - Based on normality
    - Hierarchical (random) variance model
  - Based on permutation
- Multivariate permutation test controlling false discovery rate with specified confidence
- Multivariate permutation test controlling number of false discoveries with specified confidence
- SAM
BRB-ArrayTools
Class Comparison

• Easy to adjust for pairing or blocking variable
  – eg genes whose expression is related to patient outcome after adjustment for tumor grade
• Identifies GO categories with exceptional number of genes in resulting gene list
• Provides chromosome analysis of resulting gene list
• Provides hyperlinks to multiple genomic databases for resulting gene list
• Gene list saved for subsequent analysis and annotation
Gene Set Class Comparison

• Uses built in pre-defined gene sets
  – Gene Ontology sets
  – Biocarta pathways
  – Kegg Pathways
  – BROAD/Whitehead Signatures
  – Adding TF target gene sets
  – User defined gene sets

• Computes summary of differential expression for each gene set

• Evaluates statistical significance of summary
  – Permutation analysis
  – Resampling random gene sets of same number of genes
Gene Set Class Comparison

- More powerful than post-hoc annotation
- Valid measures of statistical significance available
BRB-ArrayTools
Analysis of Variance Tools

• Fixed effects ANOVA to find genes associated with quantitative variable
• Mixed model for repeated measures on the same experimental subjects
• Model for analysis of single channel intensities in dual label arrays to analyze non-reference designs
• Regression model for time-series analysis of laboratory data
Class Prediction Examples

• Predict from expression profiles which patients are likely to experience severe toxicity from a new drug versus who will tolerate it well

• Predict which breast cancer patients will relapse within two years of diagnosis versus who will remain disease free
Class Prediction

• Cluster analysis is frequently used in publications for class prediction in a misleading way
Fallacy of Clustering Classes Based on Selected Genes

- Even for arrays randomly distributed between classes, genes will be found that are “significantly” differentially expressed.
- With 10,000 genes measured, about 500 false positives will be differentially expressed with $p < 0.05$.
- Arrays in the two classes will necessarily cluster separately when using a distance measure based on genes selected to distinguish the classes.
Class Prediction Paradigm

• Select genes \( (G) \) to be included in predictive model using training data in which class membership of the samples is known
• Fit predictive model containing features \( (G) \) using training data
  – e.g. linear discriminant analysis
• Evaluate predictive accuracy of model on completely independent data not used in any way for development of the model
Leave-One-Out Cross-validation Paradigm for Evaluating Classification Error Rate

• Leave-out one specimen
  – Perform gene selection and model fitting on the training set consisting of the remaining specimens
  – Evaluate whether the model predicts correctly for the left-out specimen

• Repeat the above procedure leaving-out all specimens, one at a time, re-doing feature selection and model fitting for each training set separately

• Total the number of classification errors
Misconceptions About Cross Validation

• Too numerous to mention here
• Often used improperly in biomedical and bioinformatic literature
BRB-ArrayTools
Class Prediction

• Classifiers
  – Compound covariate predictor
  – Diagonal LDA
  – K-Nearest Neighbor Classification
  – Nearest Centroid
  – Support Vector Machines
  – Random Forest Classifier
  – Shrunken Centroids (PAM)
  – Top Scoring Pairs
  – Binary Tree Classifier
BRB-ArrayTools
Class Prediction

• Validation
  – Split Sample
  – Leave one out cross validation
  – K-fold cross validation
  – Repeated K-fold cross validation
  – .632+ Bootstrap resampling
BRB-ArrayTools
Class Prediction

• Gene Selection
  – Re-done for each re-sampled training set
  – Univariate significance level less than specified threshold
    • Option for threshold for gene selection optimized by inner loop of cross-validation
  – Pairs of genes that work well together
  – Shrunken centroids
BRB-ArrayTools
Class Prediction

• Permutation test of significance of cross-validated misclassification rate
• Predictions for new patients
BRB-ArrayTools
Survival Risk Group Prediction

- No need to transform data to good vs bad outcome. Censored survival is directly analyzed
- Gene selection based on significance in univariate Cox Proportional Hazards regression
- Uses k principal components of selected genes
- Gene selection re-done for each resampled training set
- Develop k-variable Cox PH model for each leave-one-out training set
BRB-ArrayTools
Survival Risk Group Prediction

• Classify left out sample as above or below median risk based on model not involving that sample
• Repeat, leaving out 1 sample at a time to obtain cross-validated risk group predictions for all cases
• Compute Kaplan-Meier survival curves of the two predicted risk groups
• Permutation analysis to evaluate statistical significance of separation of K-M curves
BRB-ArrayTools
Survival Risk Group Prediction

• Compare Kaplan-Meier curves for gene expression based classifier to that for standard clinical classifier

• Develop classifier using standard clinical staging plus genes that add to standard staging
Acknowledgements

• Amy Lam and the BRB-ArrayTools Development Team
• Dr. Yingdong Zhao