Myths & Truths About Microarray Expression Profiling

Richard Simon, D.Sc.
Chief, Biometric Research Branch
NCI
rsimon@nih.gov
http://linus.nci.nih.gov/brb

Myth

- That the greatest challenge is managing the mass of microarray data

Truth

- Commercial and non-commercial software for managing large volumes of microarray data are available. If you have lots of money, you can develop your own.
- Greater challenges are:
  - Effectively designing and analyzing experiments that utilize microarray technology
  - Organizing and facilitating effective interdisciplinary collaboration with statistical scientists

BRB ArrayTools: An integrated package for the analysis of DNA microarray data

http://linus.nci.nih.gov/BRB-ArrayTools.html

BRB ArrayTools Design Objectives

- Encapsulates BRB experience in analysis of data and development of methods
- Educating biologists in microarray data analysis
- Easy user interface
  - Excel front-end
- Ease of data loading
  - integrated
- Drill-down linkage to genomic databases
- State-of-the-art analytic tools
  - Based on BRB critically evaluating literature
- Easily extensible
  - R add-ins
- Portable
  - Non-proprietary
  - Free for non-commercial use

Myth

- That data mining is an appropriate paradigm for analysis of microarray data
- That planning microarray investigations does not require “hypotheses” or clear objectives

Truth

- Effective microarray research requires clear objectives, but not gene specific mechanistic hypotheses
  - Class comparison
  - Class prediction
  - Class discovery

Design and Analysis Methods Should Be Tailored to Study Objectives

- Class Comparison (supervised)
  - For predetermined classes, establish whether gene expression profiles differ, and identify genes responsible for differences
- Class Prediction (supervised)
  - Prediction of phenotype using information from gene expression profile
- Class Discovery (unsupervised)
  - Discover clusters among specimens or among genes

Class Comparison Examples

- Establish that expression profiles differ between two histologic types of cancer
- Identify genes whose expression level is altered by exposure of cells to an experimental drug

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 Discovery Examples

- Discover previously unrecognized subtypes of lymphoma
- Identify co-regulated genes

Myth

- That cluster analysis is the generally appropriate method of data analysis

Truth

- Cluster analysis is only effective for class discovery and for identifying potentially co-regulated genes
- *Supervised* methods are more powerful for class comparison and class prediction
  - Clusters are not sensitive to the minority of genes that distinguish the classes
  - Multiple comparison issues not addressed by cluster methods

Do Expression Profiles Differ for Two Defined Classes of Arrays?

- Not a clustering problem
  - Global similarity measures generally used for clustering arrays may not distinguish classes
- Supervised vs unsupervised methods
- Requires multiple biological samples from each class

Do Expression Profiles Differ for Two Defined Classes of Arrays?

- Global test
  - Number of genes significantly differentially expressed among classes at specified nominal significance level
  - Cross-validated mis-classification rate
- Multiple comparison adjustment for finding differentially expressed genes
  - Experiment-wise error
  - Univariate screening with $p<0.001$ threshold
  - False discovery rate

Myth

- That comparing tissues or experimental conditions is based on looking for red or green spots on a single array
- That comparing tissues or experimental conditions is based on using Affymetrix MAS software to compare two arrays

Truth

- Comparing expression in two RNA samples tells you (at most) only about those two samples and may relate more to sample handling and assay artifacts than to biology. Robust knowledge requires multiple samples that reflect biological variability.

Sample Size Planning

GOAL: Identify genes differentially expressed in a comparison of pre-defined classes of specimens.

- Total sample size when comparing two equal sized, independent groups:

$$n = 4(z_{\alpha/2} + z_{\beta})^2/(\delta/\sigma)^2$$

where δ = mean difference between classes

σ = standard deviation

$z_{\alpha/2}$, z_{β} = standard normal percentiles

- Choose α small to limit expected number of false discoveries, e.g. $\alpha = .001$

Number of Biological Specimens for Class Comparison

δ	σ	α	β	n/2 Specimens per class
1	0.25	0.001	0.05	6
1	0.50	0.001	0.05	15

Myth

- The reference RNA for two-label arrays must be biologically relevant

Truth

- The reference generally serves only to control variation in the size of corresponding spots on different arrays and variation in sample distribution over the slide.
- The reference provides a relative measure of expression for a given gene in a given sample that is less variable than an absolute measure.
- The relative measure of expression will be compared among biologically independent samples from different classes.

Myth

- Multiple testing issues are not important with microarray data
- Multiple testing issues make microarray based conclusions inherently fallacious

Truth

- Comparing two classes of samples with regard to expression of 10,000 genes, one expects 500 erroneous findings of genes that appear significant at the 5% significance level. This is true regardless of the correlation pattern of the genes.
- Eyeball analysis of multi-colored image plots for genes that appear differentially expressed is similarly unreliable.

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 8000 genes measured, 400 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

Truth

- There are statistical methods for limiting the number of false discoveries in finding genes that are differentially expressed in comparing two or more classes of samples. These methods are based on more stringent levels of significance than 0.05
- Multivariate permutation methods are the most powerful and robust methods for class comparison problems in microarray studies.

Myth

- That complex classification algorithms such as neural networks perform better than simpler methods for class prediction.

Truth

- Artificial intelligence sells to journal reviewers and peers who cannot distinguish hype from substance when it comes to microarray data analysis.
- Comparative studies have shown that simpler methods work better for microarray problems because the number of candidate predictors exceeds the number of samples by orders of magnitude.

Myth

- A prediction model that fits the data used to develop it should predict well for future samples

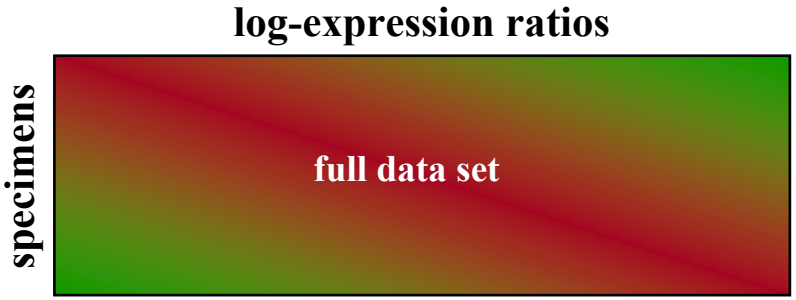
Truth

- “Prediction is difficult, particularly the future.”
  - Quail or Bohr?
- A straight line can fit 2 points perfectly.
- An n’th degree polynomial can fit n-1 points perfectly.
- A predictor based on 10,000 genes can be made to fit class labels for 100 samples perfectly.

Truth

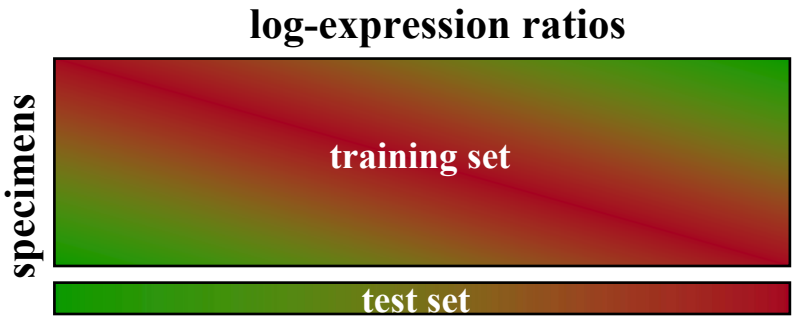
- Fit of a model to the same data used to develop it is no evidence of prediction accuracy for independent data.
- Leave-one-out cross-validation simulates the process of separately developing a model on one set of data and predicting for a test set of data not used in developing the model

Non-Cross-Validated Prediction



1. Prediction rule is built using full data set.
2. Rule is applied to each specimen for class prediction.

Cross-Validated Prediction (Leave-One-Out Method)



1. Full data set is divided into training and test sets (test set contains 1 specimen).
2. Prediction rule is built using the training set.
3. Rule is applied to the specimen in the test set for class prediction.
4. Process is repeated until each specimen has appeared once in the test set.

Prediction on Simulated Null Data

Generation of Gene Expression Profiles

- 14 specimens (P_i is the expression profile for specimen i)
- Log-ratio measurements on 6000 genes
- $P_i \sim \text{MVN}(\mathbf{0}, \mathbf{I}_{6000})$
- Can we distinguish between the first 7 specimens (Class 1) and the last 7 (Class 2)?

Prediction Method

- Compound covariate prediction
- Compound covariate built from the log-ratios of the 10 most differentially expressed genes.

Percentage of simulated data sets with *m* or fewer misclassifications

m	Non-cross-validated class prediction	Cross-validated class prediction
0	99.85	0.60
1	100.00	2.70
2	100.00	6.20
3	100.00	11.20
4	100.00	16.90
5	100.00	24.25
6	100.00	34.00
7	100.00	42.55
8	100.00	53.85
9	100.00	63.60
10	100.00	74.55
11	100.00	83.50
12	100.00	91.15
13	100.00	96.85
14	100.00	100.00

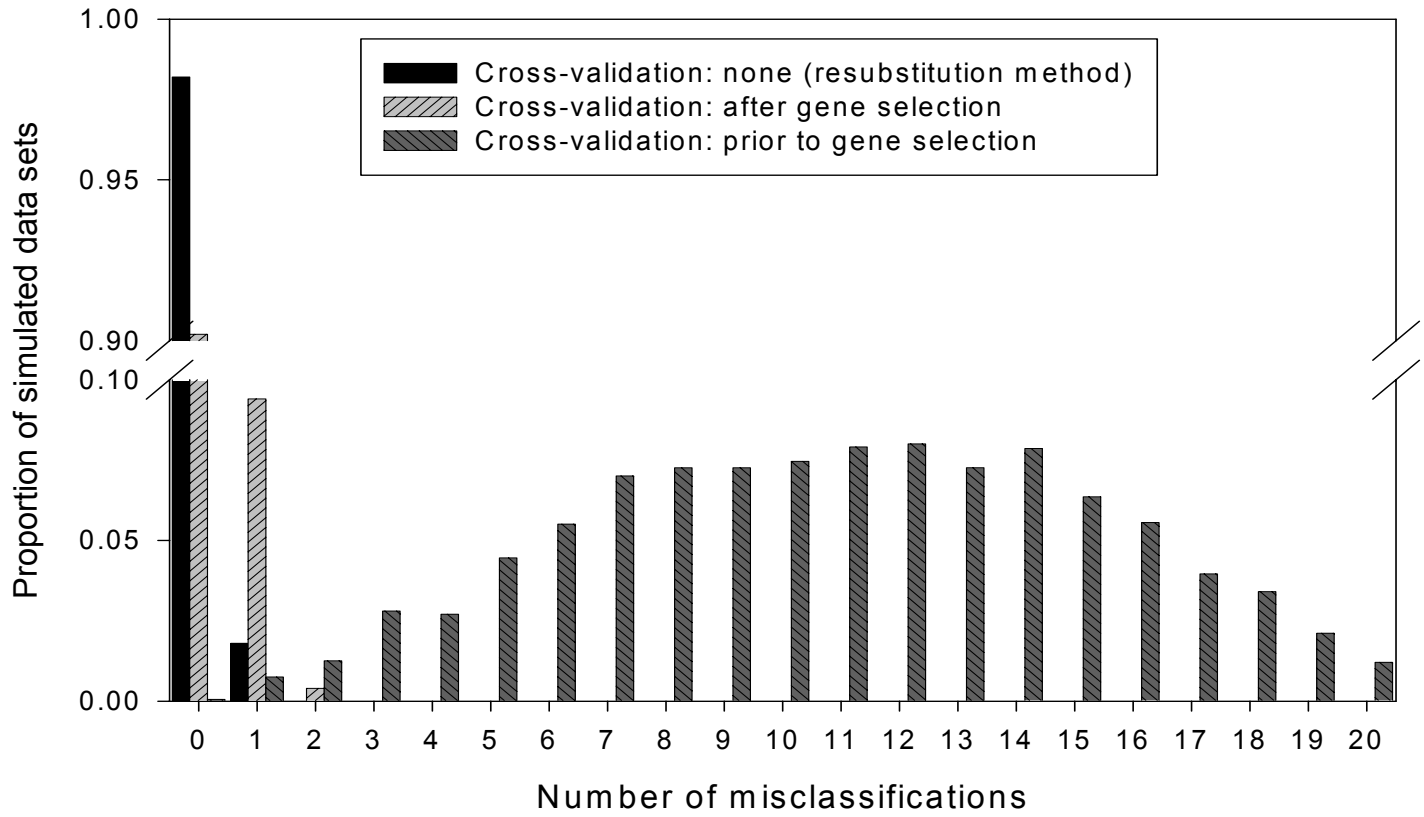
From Radmacher *et al.*, *Journal of Computational Biology* (in press)

Myth

- Cross-validation of a model can occur after selecting the genes to be used in the model

Truth

- Cross validation is only valid if the training set is not used in any way in the development of the model. Using the complete set of samples to select genes violates this assumption and invalidates cross-validation.
- With proper cross-validation, the model must be developed from scratch for each leave-one-out training set. This means that gene selection must be repeated for each leave-one-out training set.
- The cross-validated estimate of misclassification error applies to the model building process, not to the particular model or the particular set of genes used in the model.


Cross-validation: none (resubstitution method)
Cross-validation: after gene selection
Cross-validation: prior to gene selection
Proportion of simulated data sets
1.00
0.95
0.90
0.10
0.05
0.00
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Number of misclassifications

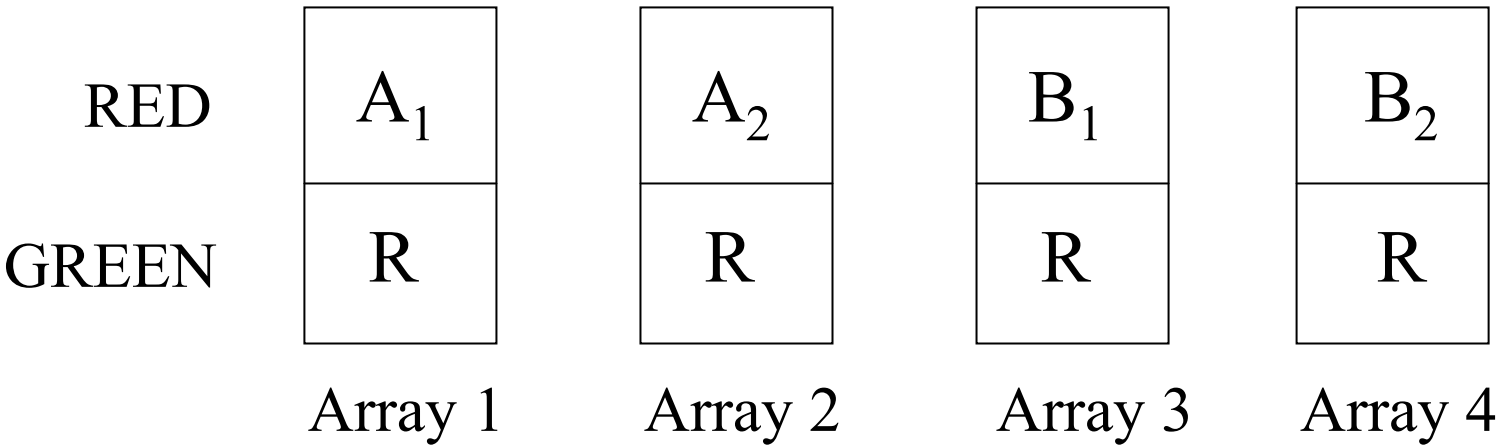
Classification of hereditary breast cancers with the compound covariate predictor

Class labels	Number of differentially expressed genes	m = number of misclassifications	Proportion of random permutations with m or fewer misclassifications
BRCA1$^+$ vs. *BRCA1*$^-$	9	1 (0 *BRCA1*$^+$, 1 *BRCA1*$^-$)	0.004
BRCA2$^+$ vs. *BRCA2*$^-$	11	4 (3 *BRCA2*$^+$, 1 *BRCA2*$^-$)	0.043

Myth

- Common reference designs for two-color arrays are inferior to “loop” designs.

Reference Design

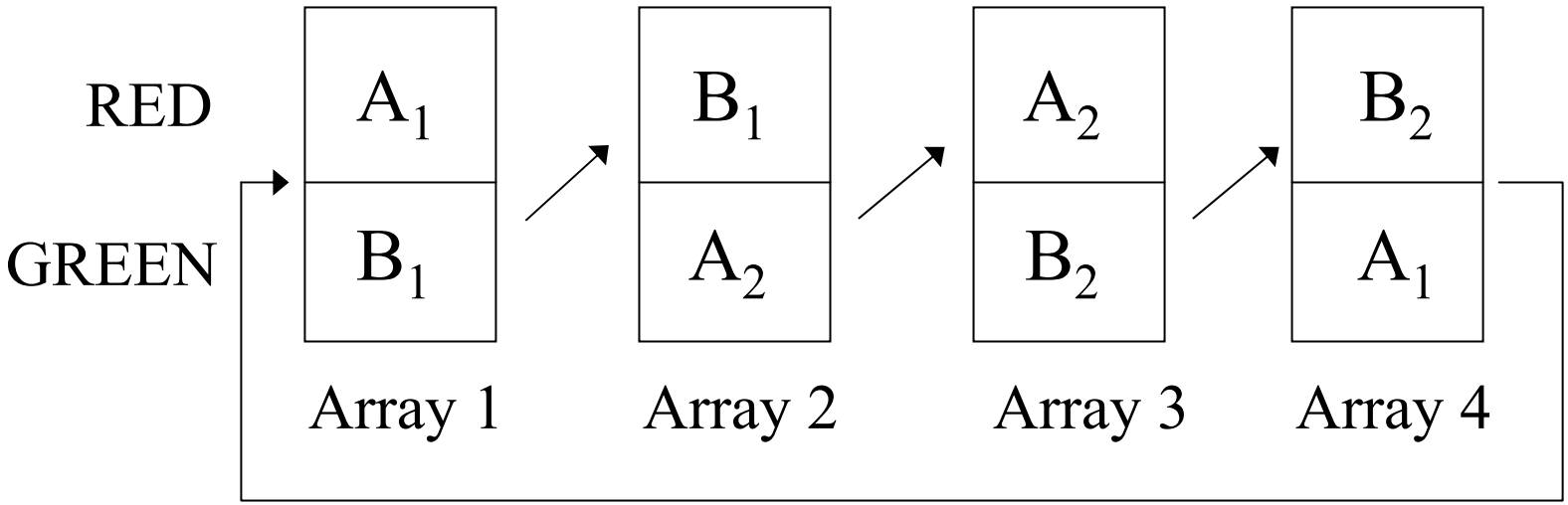


A_i = ith specimen from class A

B_i = ith specimen from class B

R = aliquot from reference pool

Loop Design

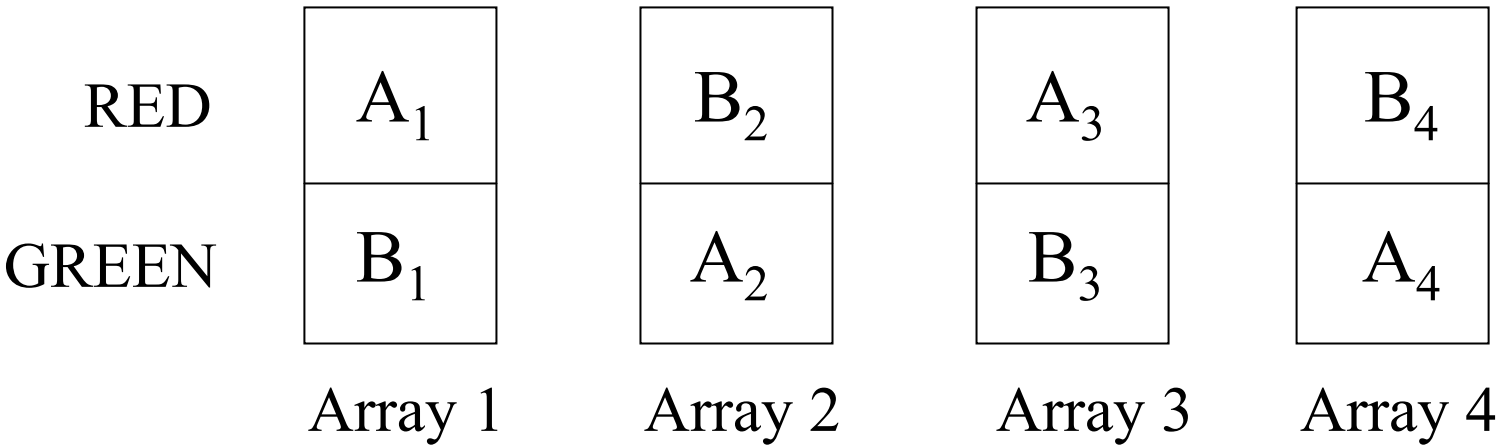


A_i = aliquot from *i*th specimen from class A

B_i = aliquot from *i*th specimen from class B

(Requires two aliquots per specimen)

Balanced Block Design



A_i = ith specimen from class A

B_i = ith specimen from class B

Truth

- Common reference designs are very effective for many microarray studies. They are robust, permit comparisons among separate experiments, and permit many types of comparisons and analyses to be performed.
- Loop designs are non-robust, are very inefficient for class discovery analyses, and are not widely applicable.
- For simple two class comparison problems, balanced block designs are very efficient and require half the number of arrays as common reference designs. They are not appropriate for class discovery, class prediction or more complicated class comparison problems.

Myth

- Three technical replicates of each array should be run

Truth

- Technical replicates do not hurt, but also do not help much.
- Biological conclusions require independent biological replicates. The power of statistical methods for microarray data depends on the number of biological replicates.
- Technical replicates are useful insurance to ensure that at least one good quality array of each specimen will be obtained.

Myth

- For two color microarrays, each sample of interest should be labeled once with Cy3 and once with Cy5 in dye-swap pairs of arrays.

Truth

- Dye swap technical replicates of the same two rna samples are rarely necessary.
- Using a common reference design, dye swap arrays are not necessary for valid comparisons of classes or for cluster analysis. The reference rna should be consistently labeled with the same dye.
- Statisticial concerns about gene specific labeling bias does not effect class comparisons in reference design experiments.
- With balanced block designs, each label should be used equally for biologically independent samples from both classes, but dye swaps of the same rna samples are not necessary or efficient.

Myth

- That software is a good substitute for collaboration with statistical scientists on complex problems

Truth

- Biologists need both good software analysis tools and good statistical collaborators.
- Both are in short supply.

References

- Dobbin K, and Simon R. (2002). Comparison of microarray designs for class comparison and class discovery. Bioinformatics 18:1438-1445.
- Dobbin K, Shih J, and Simon R. (2003) Statistical design of reverse dye microarrays. Bioinformatics (In Press).
- Dobbin K, Shih J, and Simon R. (2003) Questions and answers about the design of dual label microarrays for identifying differentially expressed genes. Journal of the National Cancer Institute (In Press)
- McShane LM, Radmacher MD, Freidlin B, Yu R., Li M., Simon R. (2002) Methods for assessing reproducibility of clustering patterns observed in analyses of microarray data. *Bioinformatics* 18:1462-1469.

References

- Simon R, Radmacher MD, and Dobbin K. (2002) Design of studies using DNA microarrays. Genetic Epidemiology 23:21-36
- Simon R, Radmacher MD, Dobbin K, and McShane LM. Pitfalls in the analysis of DNA microarray data for diagnostic and prognostic classification. (2003a) Journal of the National Cancer Institute 95:14-18
- Simon R, and Lam A. (2003b) BRB-ArrayTools 3.0 User's Guide, http://linus.nci.nih.gov/~brb
- Simon R, Dobbin K. Experimental design of DNA microarray experiments. Biotechniques 34:1-5
- Radmacher MD, McShane LM, and Simon R. (2002) A paradigm for class prediction using gene expression profiles. Journal of Computational Biology 9:505-511