Key Features in the Design and Analysis of DNA Microarray Studies

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, Presentatio
- BRB-ArrayTools software

Experimental Design

- Dobbin K, Simon R. Comparison of microarray designs for class comparison and class discovery. Bioinformatics 18:1462-9, 2002
- Dobbin K, Shih J, Simon R. Statistical design of reverse dye microarrays. Bioinformatics 19:803-10, 2003
- Dobbin K, Shih J, Simon R. Questions and answers on the design of dual-label microarrays for identifying differentially expressed genes, JNCI 95:1362-69, 2003
- Simon R, Korn E, McShane L, Radmacher M, Wright G, Zhao Y. *Design and analysis of DNA microarray investigations*, Springer Verlag (2003)
- Simon R, Dobbin K. Experimental design of DNA microarray experiments. Biotechniques 34:1-5, 2002
- Simon R, Radmacher MD, Dobbin K. Design of studies with DNA microarrays. Genetic Epidemiology 23:21-36, 2002
- Dobbin K, Simon R. Sample size determination in microarray experiments for class comparison and prognostic classification. Biostatistics (In Press)

Class Prediction

- Simon R, Radmacher MD, Dobbin K, McShane LM. Pitfalls in the analysis of DNA microarray data: Class prediction methods. Journal of the National Cancer Institute 95:14-18, 2003

- Radmacher MD, McShane LM and Simon R. A paradigm for class prediction using gene expression profiles. Journal of Computational Biology 9:505-511, 2002

- Simon R. Using DNA microarrays for diagnostic and prognostic prediction. Expert Review of Molecular Diagnostics 3:587-595, 2003

- Simon R. Diagnostic and prognostic prediction using gene expression profiles in high dimensional microarray data. British Journal of Cancer 89:1599-1604, 2003

Class Comparison

- Korn EL, McShane LM, Troendle JF, Rosenwald A and Simon R. Identifying pre-post chemotherapy differences in gene expression in breast tumors: a statistical method appropriate for this aim. British Journal of Cancer 86:1093-1096, 2002

- Korn EL, Troendle JF, McShane LM, and Simon R. Controlling the number of false discoveries: Application to high-dimensional genomic data. Journal of Statistical Planning and Inference 124:379-398, 2004

- Wright G.W. and Simon R. A random variance model for detection of differential gene expression in small microarray experiments. Bioinformatics 19:2448-55, 2003

Outline of Presentation

- Design
- Development and validation of predictive models
- Software for microarray data analysis

Myth

- That microarray investigations are unstructured data-mining adventures without clear objectives

- Good microarray studies have clear objectives, but not generally gene specific mechanistic hypotheses
- Design and Analysis Methods Should Be Tailored to Study Objectives

Common Types of Objectives

- Class Comparison
  - Identify genes differentially expressed among predefined classes.
- Class Prediction
  - Develop multi-gene predictor of class label for a sample using its gene expression profile
- Class Discovery
  - Discover clusters among specimens or among genes

Do Expression Profiles Differ for Defined Classes of Samples?

- Not a clustering problem
  - Global similarity measures generally used for clustering arrays may not distinguish classes
  - Selecting features and then clustering will give good separation even for classes which do not differ unless the false discovery rate is controlled in feature selection
- Supervised methods are better
- Requires multiple biological samples from each class

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
- Many published statistical methods are limited to comparing rna transcript profiles from two samples

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

Levels of Replication

- Technical replicates
  - RNA sample divided into multiple aliquots and re-arrayed
- Biological replicates
  - Multiple subjects
  - Re-growth of cell culture under fixed conditions

- For comparing two rna samples, technical replicates are important.
- For comparing average expression between two or more conditions, time points after an intervention or between kinds of tissues, technical replicates 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.

Allocation of Specimens to Dual Label Arrays for Simple Class Comparison Problems

- Reference Design
- Balanced Block Design
- Loop Design

Reference Design

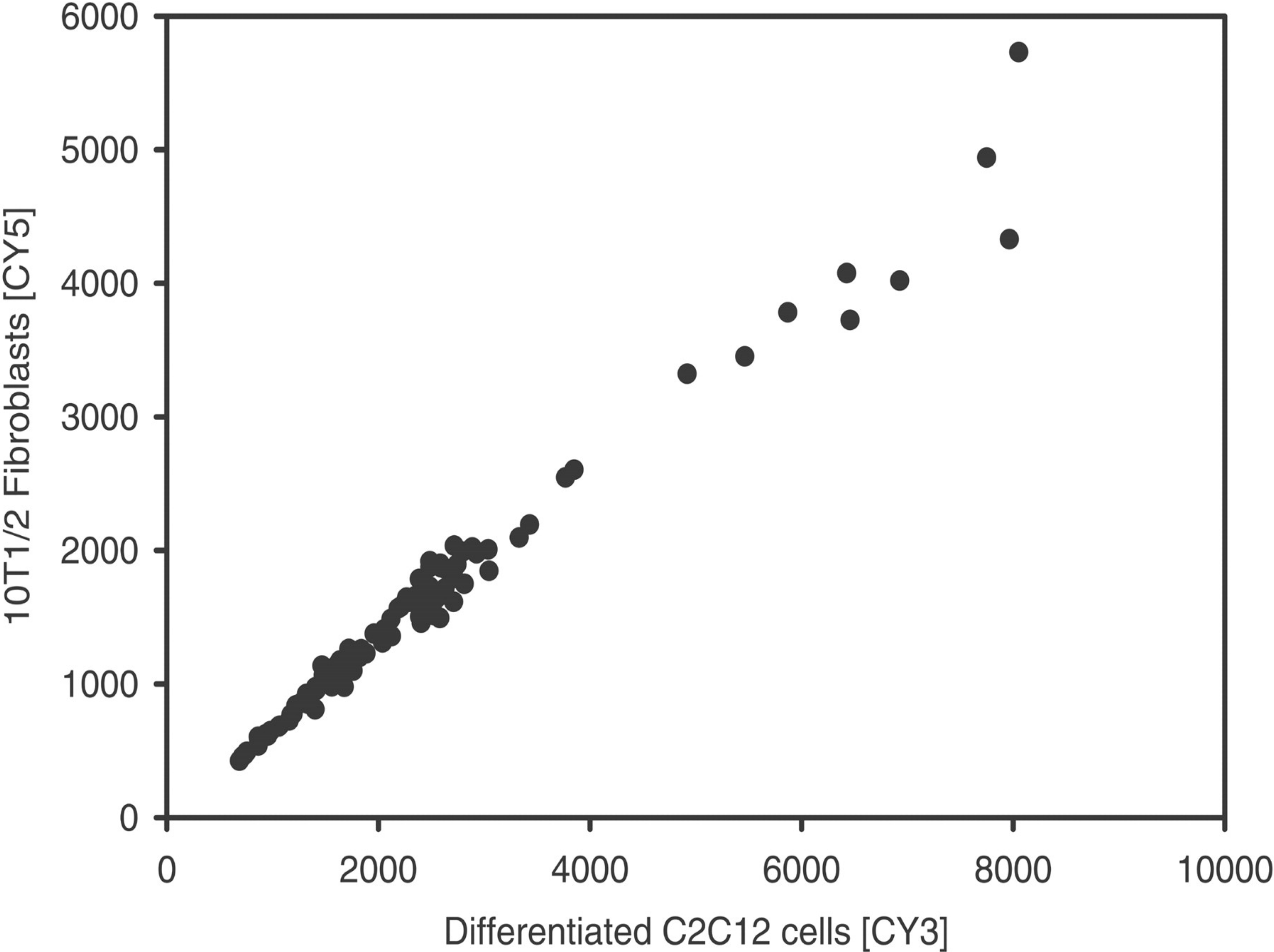
	Array 1	Array 2	Array 3	Array 4
RED	A_1	A_2	B_1	B_2
GREEN	R	R	R	R

A_i = *i*th specimen from class A

B_i = *i*th specimen from class B

R = aliquot from reference pool

- 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.


6000
5000
4000
3000
2000
1000
0
10T1/2 Fibroblasts [CY5]
0
2000
4000
6000
8000
10000
Differentiated C2C12 cells [CY3]

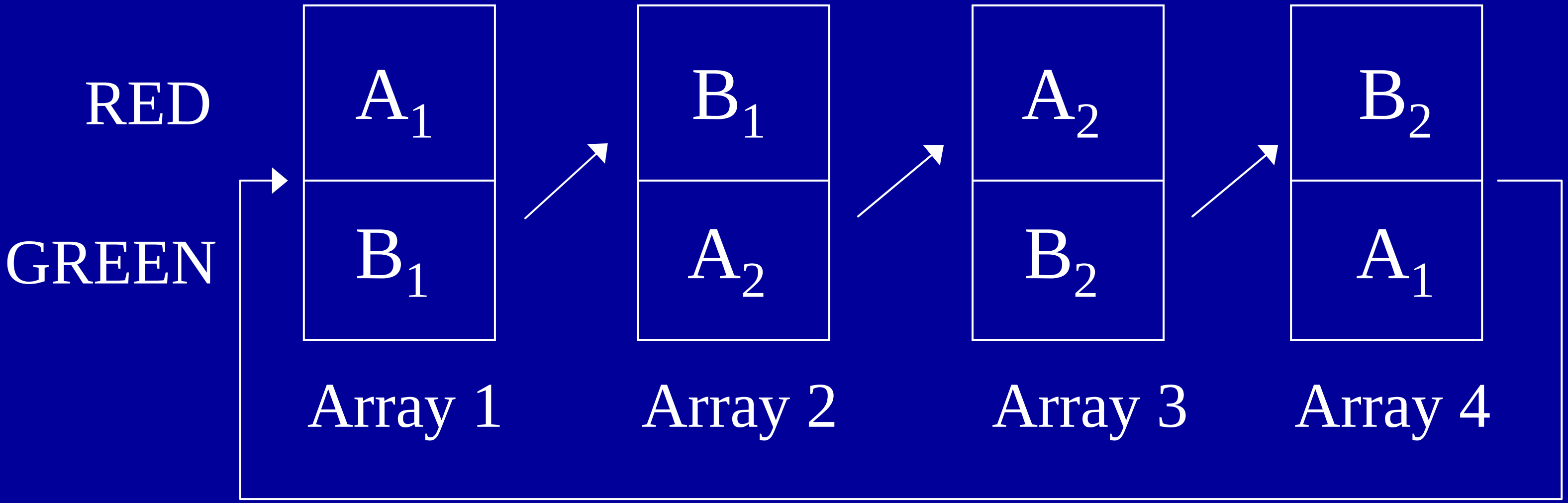
Balanced Block Design

	Array 1	Array 2	Array 3	Array 4
RED	A_1	B_2	A_3	B_4
GREEN	B_1	A_2	B_3	A_4

A_i = *i*th specimen from class A

B_i = *i*th specimen from class B

Loop Design



	Array 1	Array 2	Array 3	Array 4
RED	A_1	B_1	A_2	B_2
GREEN	B_1	A_2	B_2	A_1

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)

- Detailed comparisons of the effectiveness of designs:
  - Dobbin K, Simon R. Comparison of microarray designs for class comparison and class discovery. Bioinformatics 18:1462-9, 2002
  - Dobbin K, Shih J, Simon R. Statistical design of reverse dye microarrays. Bioinformatics 19:803-10, 2003
  - Dobbin K, Simon R. Questions and answers on the design of dual-label microarrays for identifying differentially expressed genes, JNCI 95:1362-1369, 2003

Common Reference Designs

- Very effective for many microarray studies.
- Robust to bad arrays
- Permit many class variables to be examined
- Efficient for clustering
- Permit class predictors to be develooped
- Permit comparisons among separate experiments utilizing the same common reference

Loop Designs

- Useful for studies using technical replicates with one rna sample from each class
- Useful for simple time series experiments
- Not robust to poor arrays
- Inefficient for class discovery (clustering) analyses
- Not applicable to class prediction analyses

Balanced Block Designs

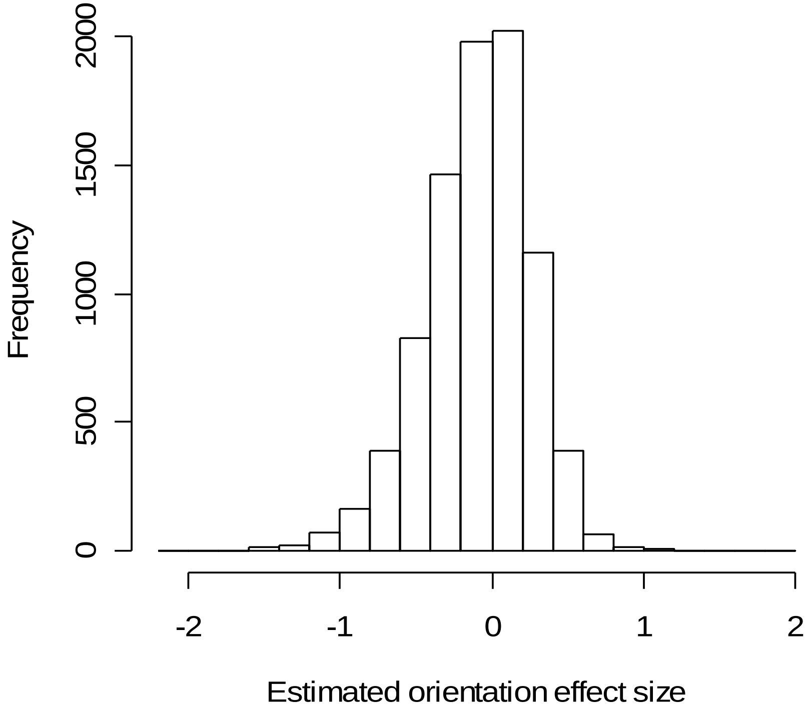
- Very efficient for simple two class comparison problems
  - require many fewer arrays for the same number of samples as the common reference design
- More difficult to apply to class comparison studies with more than two classes or multiple class variables
- Not effective for class discovery or class prediction

Dye Bias

- *Average* differences among dyes in label concentration, labeling efficiency and photon detection efficiency are corrected by normalization procedures
- Gene specific dye bias may not be corrected by normalization

Cell Line Name	Number of oligonucleotide arrays (Number with reference green/Cy3)	Number of cDNA Arrays (Number with reference green/Cy3)	Description
MCF10a	4 (2)	4 (2)	Human mammary epithelial cell line
LNCAP	4 (2)	4 (2)	Human prostate cancer cell line
L428	9 (4)	7 (4)	Hodgkins disease cell line
SUDHL	4 (2)	4 (2)	Human lymphoma cell line
OCILY3	5 (3)	5 (3)	Human lymphoma cell line
Jurkat	4 (2)	4 (2)	Human T lymphocyte acute T cell leukemia cell line
Total	30 (15)	28 (15)	

cDNA experiment estimated sizes of the gene-specific dye bias for each of the 8,604 genes. An effect of size 1 corresponds to a 2-fold change in expression



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.

Dye swap technical replicates of the same two rna samples are rarely necessary

Common Reference Design

- Dye swap arrays are not necessary for valid comparisons of classes since specimens labeled with different dyes are never compared.
- Dye bias is the same for all classes and cancels in comparing classes

Direct Comparison Designs

- Analysis of variance should be used to analyze the data in a manner that adjusts class comparisons for dye bias
- It is more efficient to balance the dye-class assignments for independent biological specimens (balanced block design) than to do dye swap technical replicates
  - Dye bias is estimatable in the balanced block design without using dye-swap technical replicates

Sample Size Planning

- GOAL: Identify genes differentially expressed in a comparison of two pre-defined classes of specimens using single label arrays

- Compare classes separately by gene with adjustment for multiple comparisons

- Approximate expression levels (log signal) as normally distributed

- Determine number of samples n/2 per class to give power 1-β for detecting mean difference δ at level α

Single Label Arrays

Comparing 2 equal size classes

$$n = 4m\left[\frac{z_{\alpha/2} + z_{\beta}}{\delta}\right]^2\left(\tau_g^2 + \gamma^2 / m\right)$$

- m = number of technical reps per sample
- n = total number of arrays
- δ = mean difference between classes in log signal
- τ^2 = biological variance within class of log signal
- γ^2 = variance among technical replicates
- α = significance level e.g. 0.001
- $1-\beta$ = power
- z = normal percentiles (use t percentiles for better accuracy)

Dual Label Arrays With Reference Design

Comparing 2 equal size classes

$$n = 4m\left[\frac{z_{\alpha/2} + z_{\beta}}{\delta}\right]^2 \left(\tau_g^2 + 2\gamma^2 / m\right)$$

- m = number of technical reps per sample
- n = total number of arrays
- δ = mean difference between classes in log ratio
- τ^2 = biological variance within class of log ratios
- $\gamma 2$ = technical variance of log ratios
- α = significance level e.g. 0.001
- 1-β = power
- z = normal percentiles (use t percentiles for better accuracy)

$\alpha=0.001$ $\beta=0.05$ $\delta=1$

$\tau^2+2\gamma^2=0.25$, $\tau^2/\gamma^2=4$

human tumors

m technical reps	n arrays required	samples required
1	25	25
2	42	21
3	60	20
4	76	19

Controlling Expected False Discovery Rate

π Proportion of differentially expressed genes	α Significance level per test	β Statistical power per test	FDR
0.01	0.001	0.10	9.9%
0.01	0.005	0.10	35.5%

Can I reduce the number of arrays by pooling specimens?

- Pooling all specimens is inadvisable because conclusions are limited to the specific RNA pool, not to the populations since there is no estimate of variation among pools
- With multiple biologically independent pools, some reduction in number of arrays may be possible at the cost of a large increase in number of samples required

Dual Label Arrays With Reference Design
Pools of k Biological Samples

$$n = 4m\left[\frac{z_{\alpha/2} + z_{\beta}}{\delta}\right]^2 \left(\tau_g^2 / k + 2\gamma^2 / m\right)$$

Number of arrays and samples required for various pooling levels. An independent pool is constructed for each array, so that no sample is represented on more than one array.

$\tau_g^2/\sigma_g^2 = 4$ $\tau_g^2 + 2\sigma_g^2 = .25$

α=0.001, β=0.05, δ=1, τ^2=0.384, γ^2=0.054, m=1

Number of samples pooled on each array	**Number of arrays required**	**Number of samples required**
1	48	48
2	30	60
3	23	69
4	20	80

Avoid Confounding

- Avoid confounding tissue handling and microarray assay procedures with the classes to be distinguished
  - Date assay performed
  - Print set

Components of Class Prediction

- Feature selection
  - Which genes or proteins will be included in the model
- Select model type
  - E.g. DLDA, Nearest-Neighbor, …
- Fitting parameters (regression coefficients) for model

Feature Selection

- Usually features are selected that are univariately differentially expressed among the classes at a specified significance level (e.g. 0.001)
- Complex methods attempt to identify features which together give accurate predictions.
- Very limited evidence that complex feature selection is useful in microarray problems
  - Failure to compare to simpler methods
  - Some published complex methods for selecting combinations of features do not appear to have been properly evaluated

Linear Classifiers for Two Classes

$$l(\underline{x}) = \sum_{i \varepsilon F} w_i x_i$$

$\underline{x} =$ vector of features

$F =$ features included in model

$w_i =$ weight for i'th feature

decision boundary $l(\underline{x})$ > or < d

Linear Classifiers for Two Classes

- Fisher linear discriminant analysis (weights based on assumed multivariate normal distribution of expression vector in each class with common covariance matrix)
- Diagonal linear discriminant analysis (DLDA) assumes features are uncorrelated
  - Naïve Bayes estimator
- Compound covariate predictor (Radmacher) and Golub’s method are similar to DLDA in that they can be viewed as weighted voting of univariate classifiers

Linear Classifiers for Two Classes

- Support vector machines with inner product kernel are linear classifiers with weights determined to minimize errors
- Perceptrons are linear classifiers

When p>>n

- For the linear model, many weight vectors w can always be found that give zero classification errors for the training data.
  - p>>n problems are almost always linearly separable
- Why consider more complex models?
- The number of parameters for this simple model is generally too large relative to the number of specimens to achieve accurate prediction for future samples if we select a single model by minimizing training errors

Myth

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

- Artificial intelligence sells to naïve journal reviewers and readers.
- Comparative studies indicate that simpler methods that avoid overfitting work better for p>>n problems.
  - DLDA, Compound covariate predictor, linear SVM, nearest neighbor and nearest (shrunken) centroid methods

Evaluating a Classifier

- Fit of a model to the same data used to develop it is no evidence of prediction accuracy for independent data.

Split-Sample Evaluation

- Training-set
  - Used to select features, select model type, determine parameters and cut-off thresholds
- Test-set
  - Withheld until a *single* model is *fully* specified using the training-set.
  - Fully specified model is applied to the expression profiles in the test-set to predict class labels.
  - Number of errors is counted
  - Ideally test set data is from different centers than the training data and assayed at a different time

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

- Cross validation is only valid if the test 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 feature selection must be repeated for each leave-one-out training set.
- The cross-validated estimate of misclassification error is an estimate of the prediction error for model fit using specified algorithm to full dataset
- If you use cross-validation estimates of prediction error for a set of algorithms and select the algorithm with the smallest cv error estimate, you do not have a valid estimate of the prediction error for the selected model

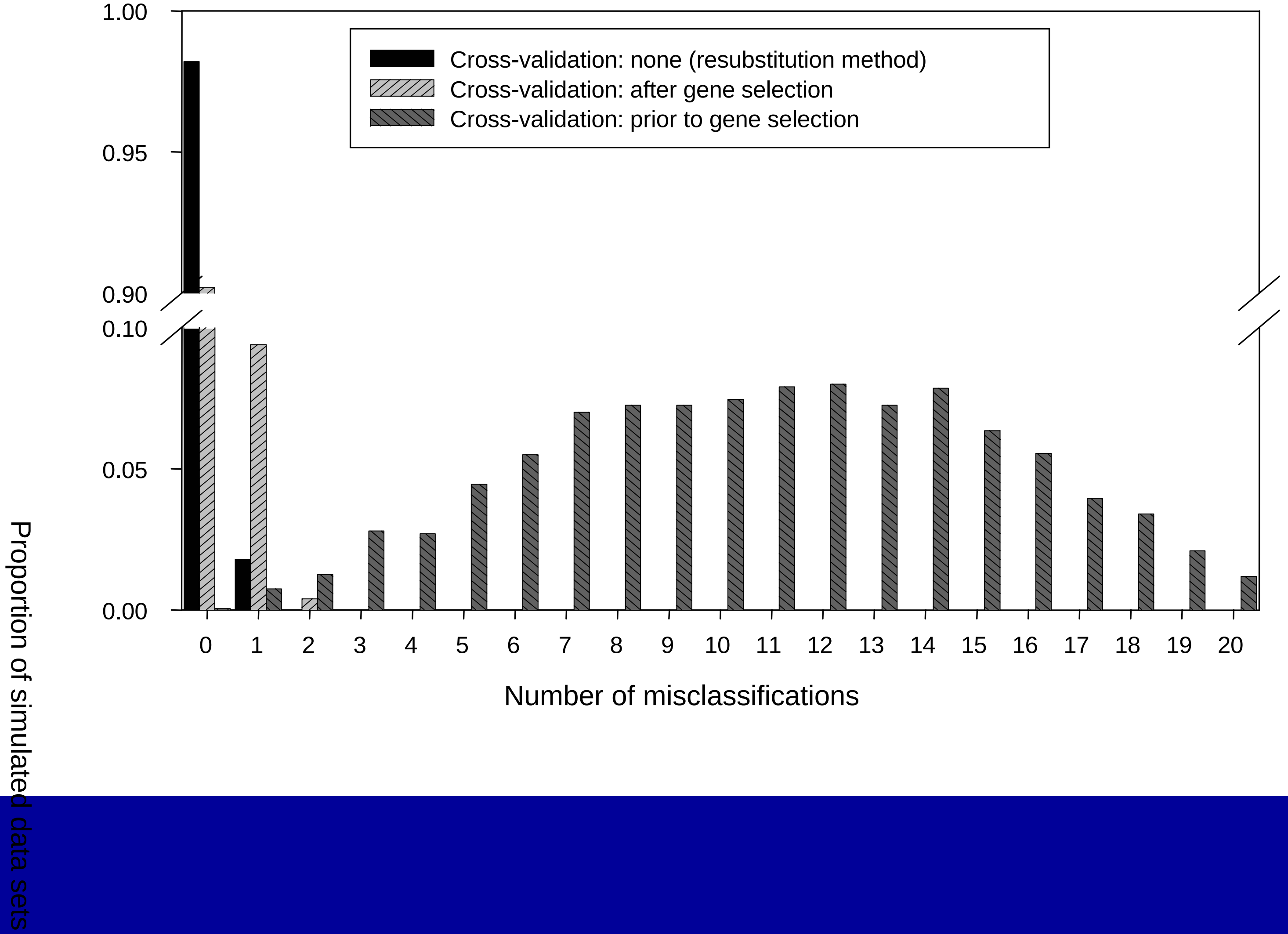
Prediction on Simulated Null Data

Generation of Gene Expression Profiles

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

Prediction Method

- Compound covariate prediction (*discussed later*)
- Compound covariate built from the log-ratios of the 10 most differentially expressed genes.



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

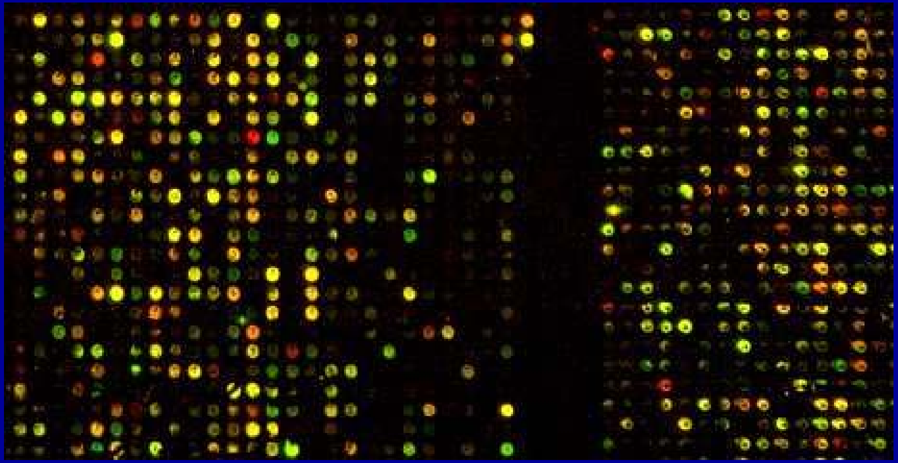
Permutation Distribution of Cross-validated Misclassification Rate of a Multivariate Classifier

- Randomly permute class labels and repeat the entire cross-validation
- Re-do for all (or 1000) random permutations of class labels
- Permutation p value is fraction of random permutations that gave as few misclassifications as e in the real data

Gene-Expression Profiles in Hereditary Breast Cancer

cDNA Microarrays

Parallel Gene Expression Analysis



- Breast tumors studied:
  - 7 *BRCA1*+ tumors
  - 8 *BRCA2*+ tumors
  - 7 sporadic tumors

- Log-ratios measurements of 3226 genes for each tumor after initial data filtering

Can we distinguish *BRCA1*+ from *BRCA1*– cancers and *BRCA2*+ from *BRCA2*– cancers based solely on their gene expression profiles?

BRCA1

α_g	# of significant genes	# of misclassified samples (m)	% of random permutations with *m* or fewer misclassifications
10^{-2}	182	3	0.4
10^{-3}	53	2	1.0
10^{-4}	9	1	0.2

BRCA2

α_g	# of significant genes	m = # of misclassified elements (misclassified samples)	% of random permutations with m or fewer misclassifications
10^{-2}	212	4 (s11900, s14486, s14572, s14324)	0.8
10^{-3}	49	3 (s11900, s14486, s14324)	2.2
10^{-4}	11	4 (s11900, s14486, s14616, s14324)	6.6

Classification of BRCA2 Germline Mutations

Classification Method	LOOCV Prediction Error
Compound Covariate Predictor	14%
Fisher LDA	36%
Diagonal LDA	14%
1-Nearest Neighbor	9%
3-Nearest Neighbor	23%
Support Vector Machine (linear kernel)	18%
Classification Tree	45%

Invalid Criticisms of Cross-Validation

- “You can always find a set of features that will provide perfect prediction for the training and test sets.”
  - For complex models, there may be many sets of features that provide zero training errors.
  - A modeling strategy that either selects among those sets or aggregates among those models, will have a generalization error which will be validly estimated by cross-validation.

BRB ArrayTools:
An integrated Package for the Analysis of DNA Microarray Data

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

BRB-ArrayTools

- Integrated software package using Excel-based user interface but state-of-the art analysis methods programmed in R, Java & Fortran
- Publicly available for non-commercial use

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

Selected Features of BRB-ArrayTools

- Multivariate permutation tests for class comparison to control false discovery proportion with any specified confidence level
- SAM
- Find Gene Ontology groups and signaling pathways that are differentially expressed
- Survival analysis
- Analysis of variance
- Class prediction models (7) with prediction error estimated by LOOCV, k-fold CV or .632 bootstrap, and permutation analysis of cross-validated error rate
  - DLDA, SVM, CCP, Nearest Neighbor, Nearest Centroid, Shrunken Centroids, Random Forests
- Clustering tools for class discovery with reproducibility statistics on clusters
  - Built in access to Eisen's Cluster and Treeview
- Visualization tools including rotating 3D principal components plot exportable to Powerpoint with rotation controls
- Import of Affy CEL files and apply RMA probe processing and quantile normalization
- Extensible via R plug-in feature
- Links genes to annotations in genomic databases
- Tutorials and datasets

Acknowledgements

- Design
  - Kevin Dobbin, Joanna Shih
- Analysis
  - Michael Radmacher, Ed Korn, Lisa McShane, George Wright, Yingdong Zhao
- BRB-ArrayTools
  - Amy Lam