Design of DNA Microarray Studies

Richard Simon, D.Sc.
Chief, Biometric Research Branch
Head, Computational & Systems Biology Group
National Cancer Institute
rsimon@nih.gov
http://linus.nci.nih.gov/brb

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

- Powerpoint presentation
- Bibliography
- Reprints, Technical Reports, Presentations
  - Microarray myths
- BRB-ArrayTools software

Acknowledgements

- Kevin Dobbin
- Sudhir Varma
- Michael Radmacher
- Joanna Shih

Myth

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

Truth

- 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

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 6:27-38, 2005.

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

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

Which Genes are Differentially Expressed In Two Conditions or Two Tissues?

- Not a clustering problem
  - Global similarity measures generally used for clustering arrays may not distinguish classes
  - Feature selection should be performed in a manner that controls the false discovery rate
- Supervised methods
- 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

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.

Experimental Design Issues for Single Label Arrays (Affymetrix GeneChips)

Avoid Bias

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

Avoid Bias

- If test set is used for evaluating classifier developed on the training set, the test samples should not be analyzed in any way until a single completely specified classifier is determined on the training set

Levels of Replication

- Technical replicates
  - RNA sample divided into multiple aliquots and re-arrayed
- Biological replicates
  - Multiple subjects
  - Re-growing the cells under the defined conditions

- 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.

Sample Size Planning for Class Comparison: Single Label Arrays

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

n arrays

k pools per class

m technical reps per pool

$\tau^2 =$ variance among specimens

$\gamma^2 =$ variance among technical reps

α=0.001 β=0.05 δ=1
$\tau^2+\gamma^2$=0.25, τ^2/γ^2=4, unpooled

m technical reps	n arrays required	samples required
1	30	30
2	56	28
3	78	26
4	104	26

Should I pool specimens?

- Pooling all specimens is inadvisable because there is no estimate of variation among pools of the same class
- With multiple biologically independent pools, some reduction in number of arrays may be possible, but at the expense of a greater required number of independent specimens

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.

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

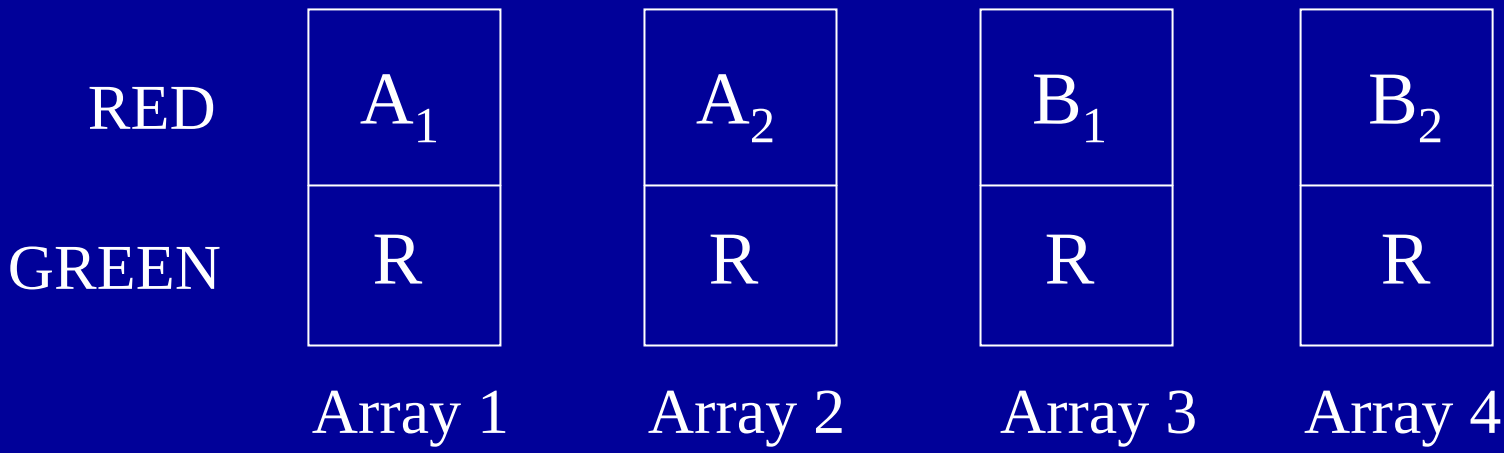
α=0.001, β=0.05, δ=1, τ^2=0.20, γ^2=0.05, m=1

Number of samples pooled on each array	**Number of arrays required**	**Number of samples required**
1	30	30
2	20	40
3	18	54
4	16	64

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

- Reference Design
- Balanced Block Design
- Loop Design

Reference Design

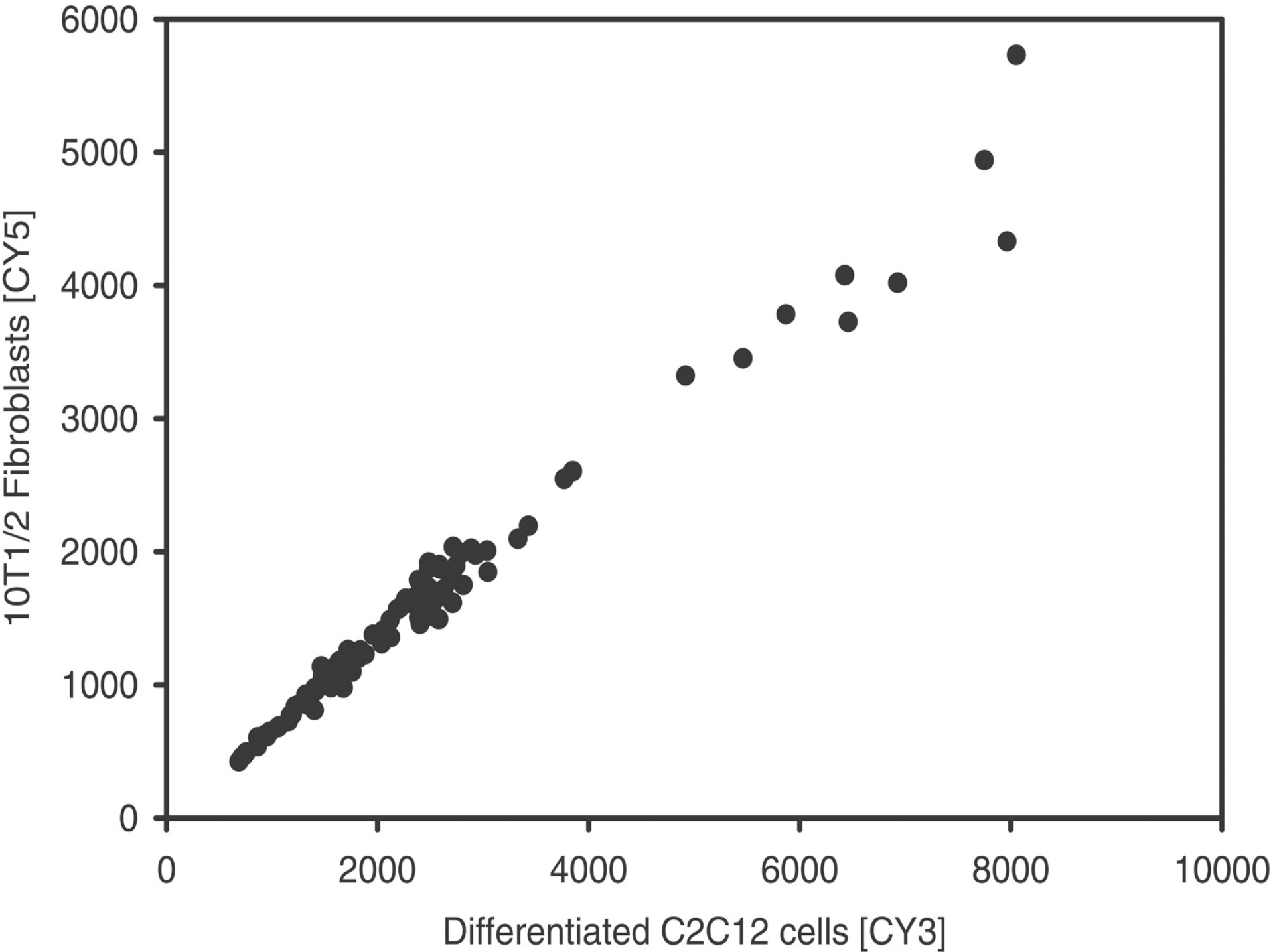


A_i = *i*th specimen from class A

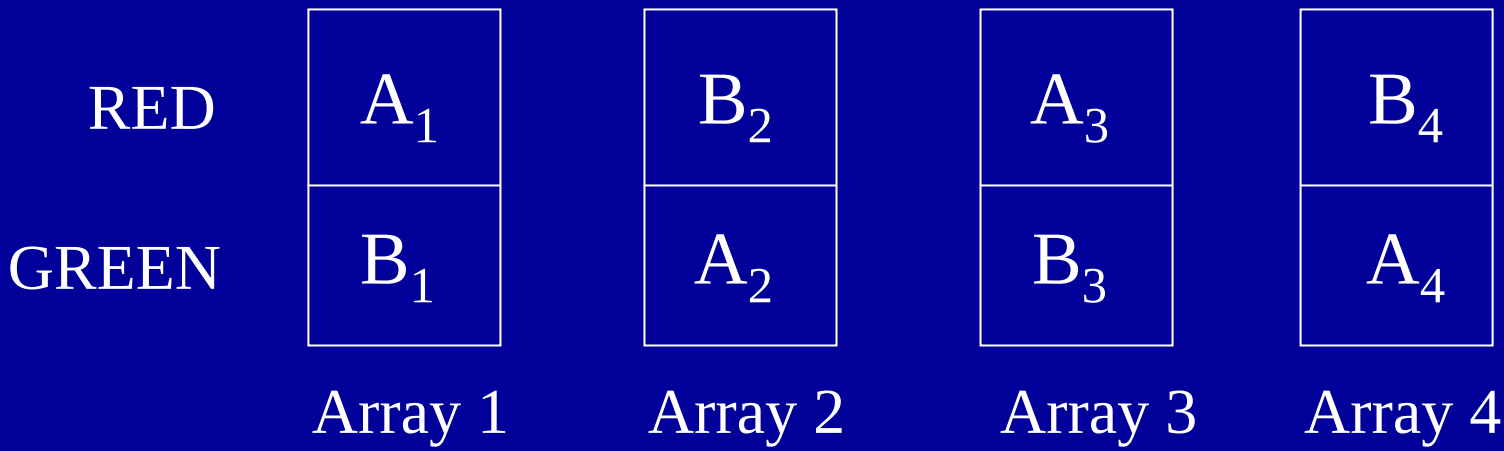
B_i = *i*th specimen from class B

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



A_i = *i*th specimen from class A

B_i = *i*th specimen from class B

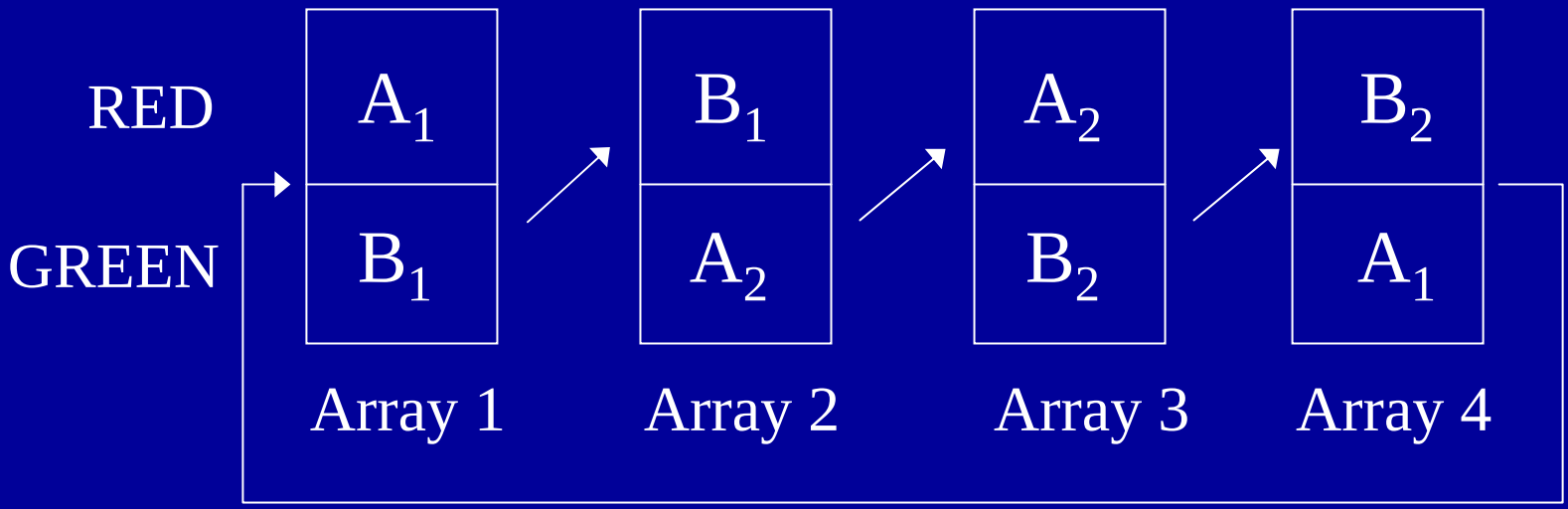
Balanced Block and Reference Designs With 5 Classes A,B,C,D,E.

Each non-reference sample hybridized to a single array.

Array	1	2	3	4	5	6	7	8	9	10
Cy3	A	C	A	E	B	D	B	C	E	D
Cy5	B	A	D	A	C	B	E	D	C	E

Array	1	2	3	4	5	6	7	8	9	10
Cy3	R	R	R	R	R	R	R	R	R	R
Cy5	A	B	C	D	E	A	B	C	D	E

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)

Relative Efficiency of Designs Evaluated Based on ANOVA for Logarithm of Background Adjusted Normalized Intensities

- Model Effects
  - Gene
  - Array by Gene (spot)
  - Variety by Gene
  - Sample within Variety by Gene

Gene-Variety Model

- $r = G_g + AG_{ag} + VG_{vg} + SG_{sg} + \varepsilon$
- $\varepsilon \sim N(0,\sigma_g^2)$
- Efficiency of design based on variance of estimators of VG_{ig}-VG_{jg}
- To study efficiency, assume $SG_{sg} \sim N(\mu_g,\tau_g^2)$

- 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

Myth

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

Truth

- Common reference designs are effective for many microarray studies. They are robust, permit comparisons among separate experiments, permit unplanned types of comparisons to be performed, permit cluster analysis and class prediction analysis.

- Loop designs are non-robust, are very inefficient for class discovery (clustering) analyses, are not applicable to class prediction analyses and do not easily permit inter-experiment comparisons.

- For simple two class comparison problems, balanced block designs are the most efficient and require many fewer arrays than reference designs. They are not appropriate for class discovery or class prediction and are more difficult to apply to more complicated class comparison problems.

Designs for Class Discovery

- Loop designs with 2 sub-samples per sample make clustering possible
- Reference designs do not require sub-sampling

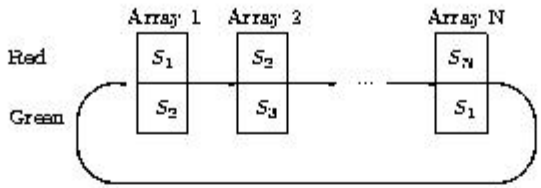
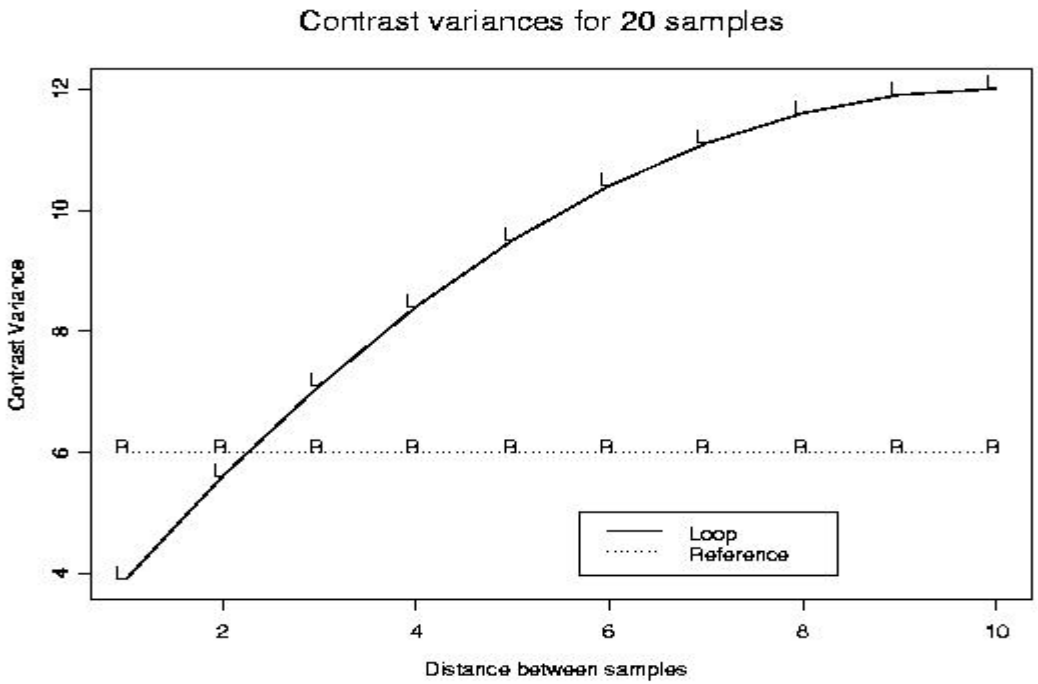


Figure 1: Loop Design for Cluster Analysis

Designs for Class Discovery

- For the loop design, variance of inter-sample contrasts depends on how close the samples appear in the loop


Contrast variances for 20 samples
Contrast Variance
Distance between samples
Loop
Reference
4
6
8
10
12
2
4
6
8
10

Evaluation of Designs for Class Discovery

- Generate data from two-varieties

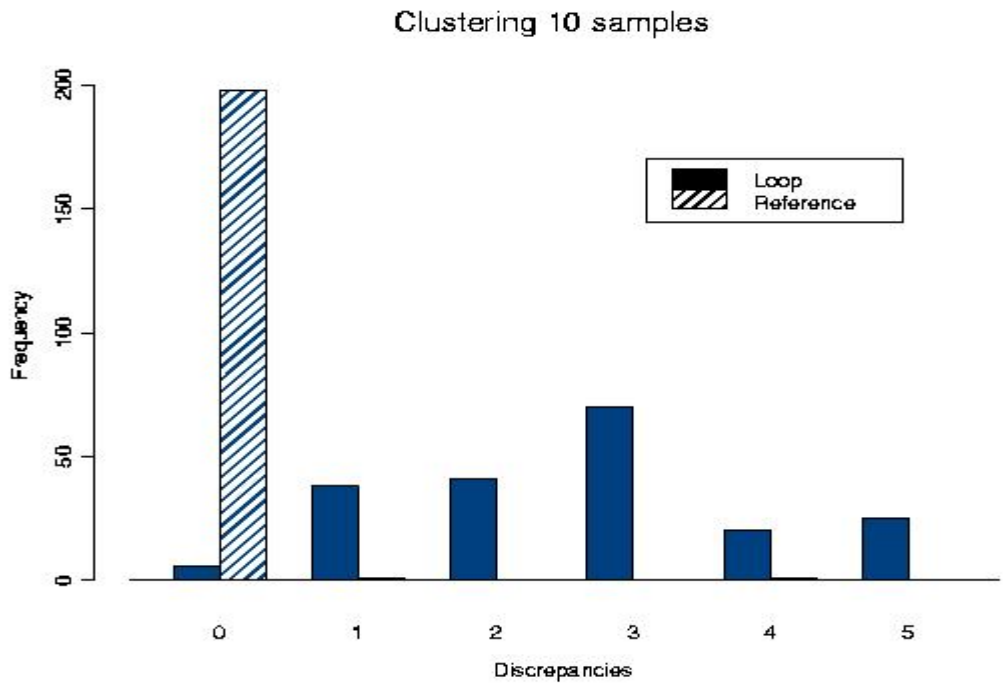
- $\tau_g^2 = \sigma_g^2$

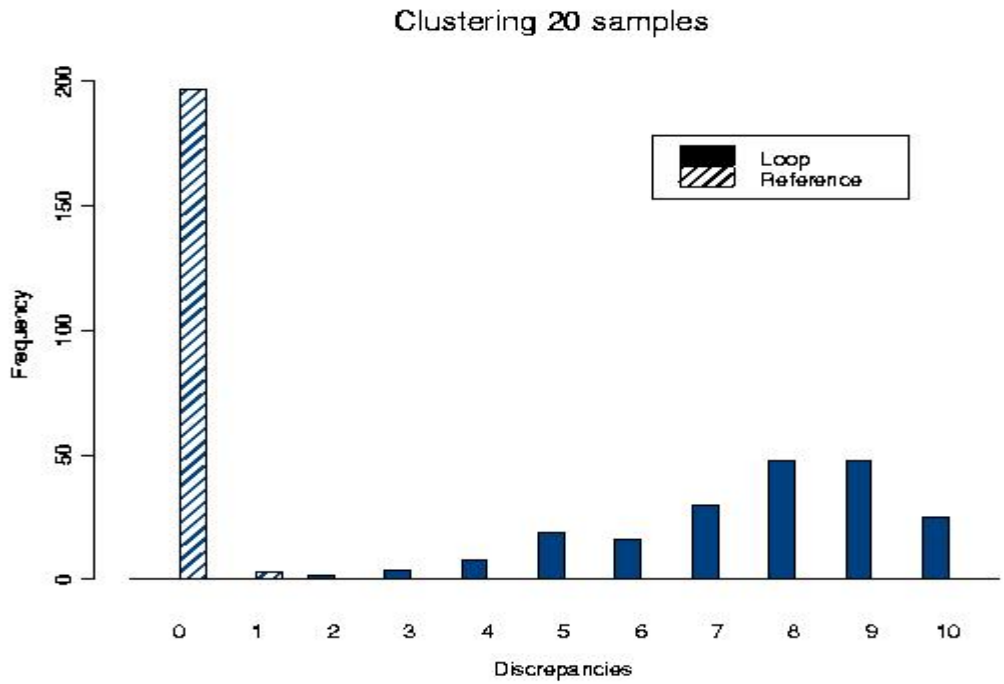
- Fit gene model without varieties
- Cluster data using hierarchical clustering
- Cut dendrogram at level giving 2 clusters

Evaluation of Designs for Class Discovery

- Associate clusters with varieties used to generate the data in manner that maximizes correspondence

- Count number of misclassifications

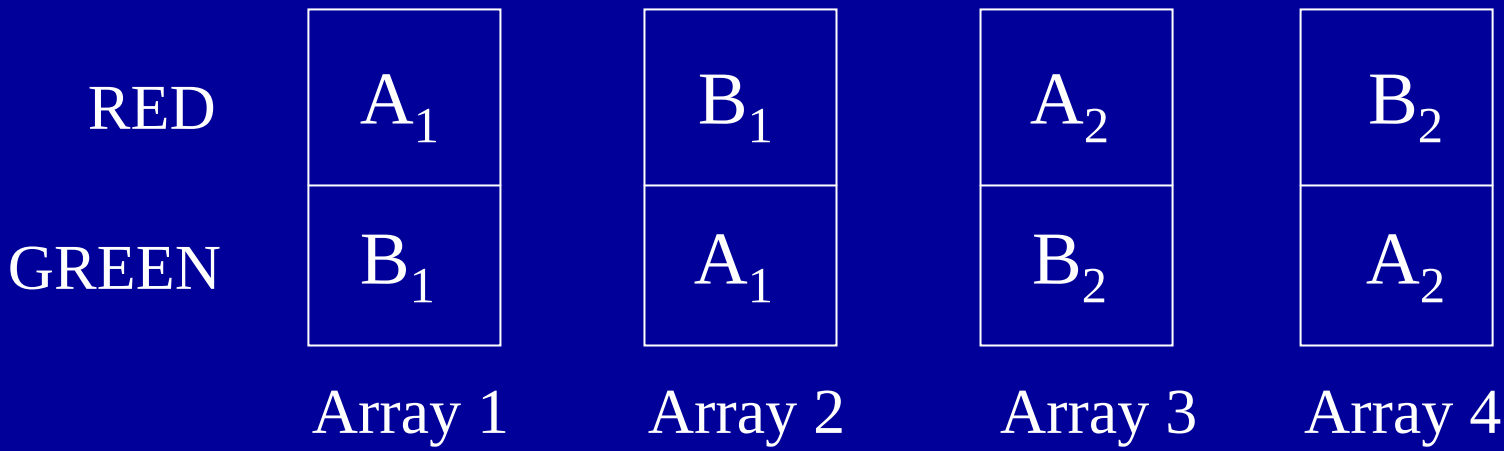

Clustering 10 samples
Loop
Reference
Frequency
0
50
100
150
200
0
1
2
3
4
5
Discrepancies


Clustering 20 samples
Loop
Reference
Frequency
0
50
100
150
200
0
1
2
3
4
5
6
7
8
9
10
Discrepancies

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 Design



A_i = *i*th specimen from class A

B_i = *i*th specimen from class B

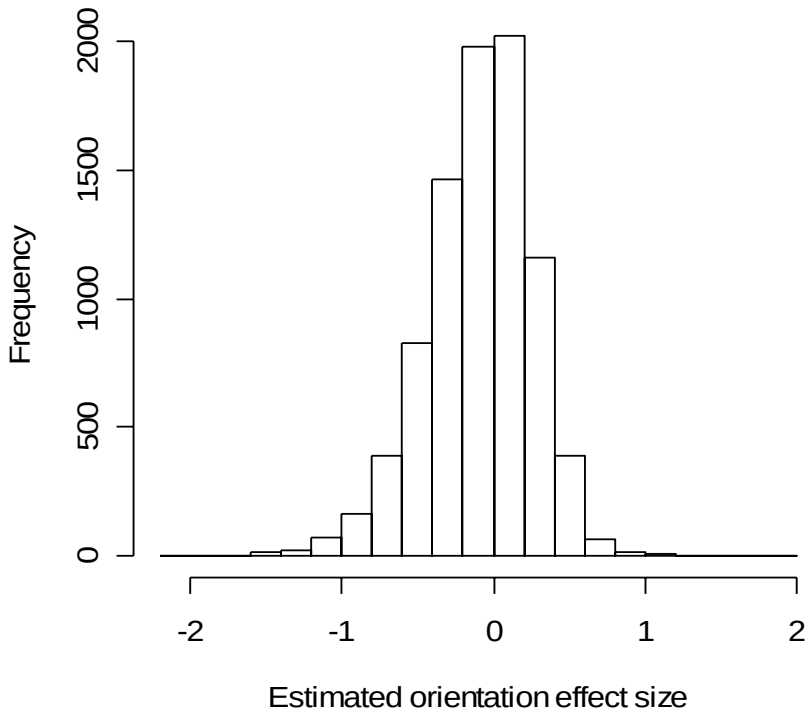
Dye Bias

- Average differences among dyes in label concentration, labeling efficiency, photon emission efficiency and photon detection 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)	

- Gene-specific dye bias
  - 3681 genes with $p<0.001$ of 8604 evaluable genes
- Gene and sample specific dye bias
  - 150 genes with $p<0.001$

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



cDNA agreement between models with and without gene-specific dye effects included

		All data: no dye effects	
		P-value < .001	P-value > .001
All data: dye effects	P-value < .001	4801 (56%)	559 (6%)
	P-value > .001	81 (1%)	3163 (37%)

(a) Reference design comparing tumor tissue to normal tissue. (b) A confounded design comparing tumor tissue to normal tissue. (c) Balanced block design comparing tumor tissue to normal tissue.

(a)

	Array 1	Array 2	Array 3	Array 4	Array 5	Array 6
Cy3	Tumor	Tumor	Tumor	Normal	Normal	Normal
Cy5	Reference	Reference	Reference	Reference	Reference	Reference

(b)_

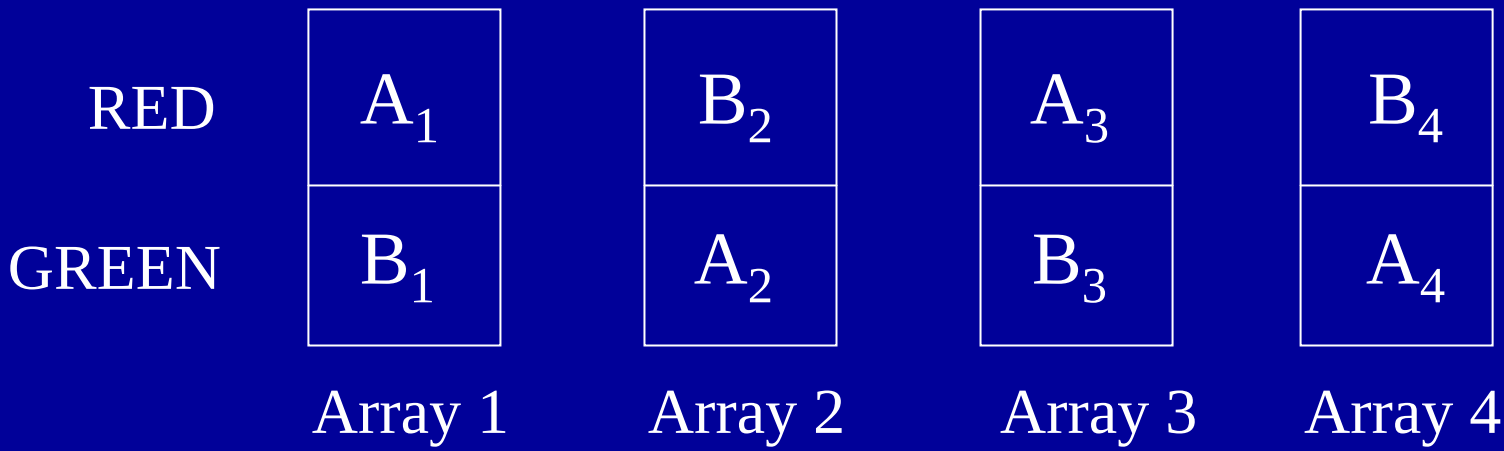
	Array 1	Array 2	Array 3	Array 4	Array 5	Array 6
Cy3	Tumor	Tumor	Tumor	Tumor	Tumor	Tumor
Cy5	Normal	Normal	Normal	Normal	Normal	Normal

(c)

	Array 1	Array 2	Array 3	Array 4	Array 5	Array 6
Cy3	Tumor	Normal	Tumor	Normal	Tumor	Normal
Cy5	Normal	Tumor	Normal	Tumor	Normal	Tumor

- 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 since specimens labeled with different dyes are never compared.
- For two-label direct comparison designs for comparing two classes, it is more efficient to balance the dye-class assignments for independent biological specimens than to do dye swap technical replicates

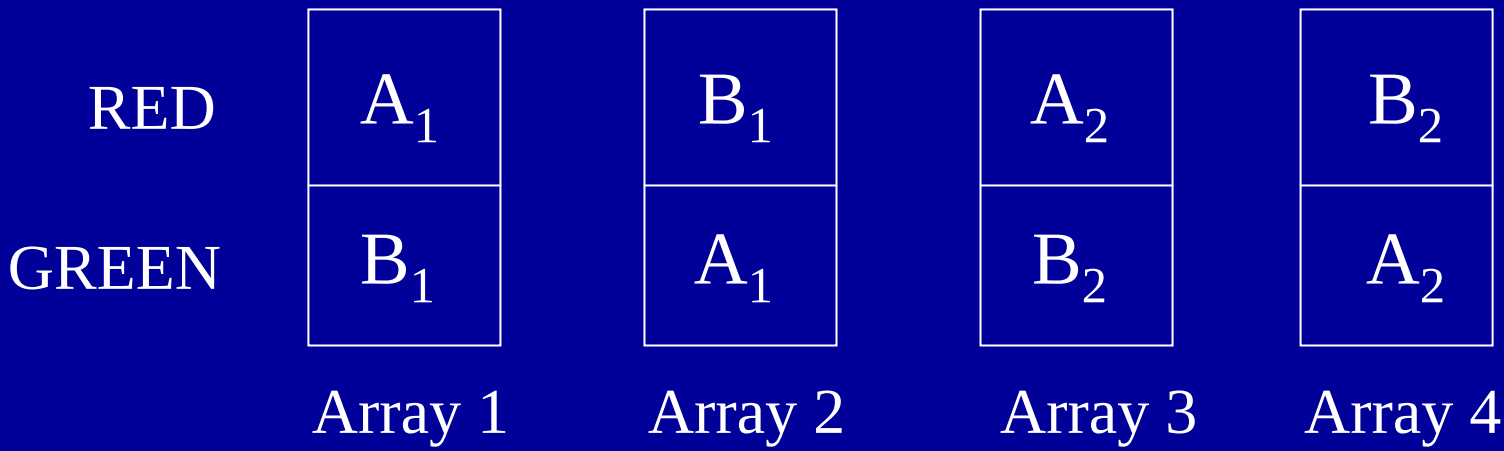
Balanced Block Design



A_i = *i*th specimen from class A

B_i = *i*th specimen from class B

Dye Swap Design



A_i = *i*th specimen from class A

B_i = *i*th specimen from class B

Balanced Block Designs for Two Classes

- Half the arrays have a sample from class 1 labeled with Cy5 and a sample from class 2 labeled with Cy3;
- The other half of the arrays have a sample from class 1 labeled with Cy3 and a sample from class 2 labeled with Cy5.
- Each sample appears on only one array. Dye swaps of the same rna samples are not necessary to remove dye bias and for a fixed number of arrays, dye swaps of the same rna samples are inefficient

Time Series Experiment

- Grow cell lines untreated and harvest at time points $t_1,\ldots t_n$
- Grow treated cell lines and harvest at same time points
- Find genes whose expression is effected by treatment

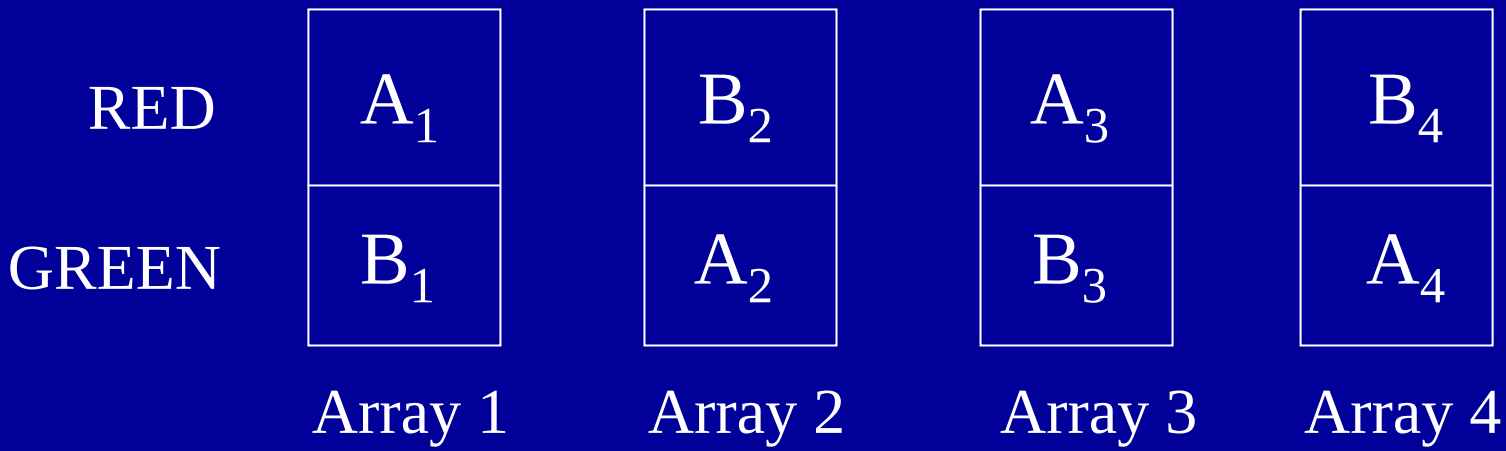
Analysis Strategy for Time Series Experiment

- Two way analysis of variance (ANOVA) of log intensities including main (marginal) effects of treatment and time points and interaction between treatment and effects of time points
- Identify genes for which interaction is statistically significant at $p<0.001$ level.

Design Options for Dual Label Expression Analysis of Time Series

- Use common reference for all arrays
  - Pooled baseline samples
- Balanced block design for comparing treated to untreated within each time point
- Designs which use multiple aliquots for each sample
- Designs when each experimental unit can be observed at multiple time points

Balanced Block Design at Time t



A_i = *i*th specimen from untreated cells at time t

B_i = *i*th specimen from treated cells at time t

Sample Size Planning

- GOAL: Identify genes differentially expressed in a comparison of two pre-defined classes of specimens on two-color arrays using reference design or single label arrays
- Compare classes separately by gene with adjustment for multiple comparisons
- Approximate expression levels (log ratio or log signal) as normally distributed
- Determine number of samples n/2 per class to give power $1-\beta$ for detecting mean difference δ at level α

- 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
- γ^2 = technical variance
- 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

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

- 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
- σ^2 = technical variance
- α = significance level e.g. 0.001
- $1-\beta$ = power
- z = normal percentiles (use t percentiles for better accuracy)

Comparing 2 equal size classes
No technical reps (m=1)

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

where δ = mean log-ratio difference between classes

γ = within class standard deviation of log-ratio

- Choose α small, e.g. $\alpha = .001$
- Use percentiles of t distribution for improved accuracy

Total Number of Samples for Two Class Comparison

α	β	δ	γ	Total Samples
0.001	0.05	1 (2-fold)	0.5 human tissue	26
			0.25 transgenic mice	12 (t approximation)

- π = proportion of genes on array that are differentially expressed between classes
- N = number of genes on the array
- FD = expected number of false discoveries
- TD = expected number of true discoveries
- FDR = FD/(FD+TD)

- $FD = \alpha(1-\pi)N$
- $TD = (1-\beta)\ \pi N$
- $FDR = \alpha(1-\pi)N/\{\alpha(1-\pi)N + (1-\beta)\ \pi N\}$
- $= 1/\{1 + (1-\beta)\pi/\alpha(1-\pi)\}$

Controlling Expected False Discovery Rate

π	α	β	FDR
0.01	0.001	0.10	9.9%
	0.005		35.5%
0.05	0.001		2.1%
	0.005		9.5%

Dual Label Arrays With Balanced Block Design For 2 classes

$$n = \left[\frac{z_{\alpha/2} + z_{\beta}}{\delta} \right]^2 \left(\eta_g^2 + 2\sigma^2 \right)$$

- n = total number of arrays
- δ = mean log ratio (class 1 / class 2)
- η^2 = biological variance of log-ratio
- σ^2 = technical variance of log intensity
- α = significance level e.g. 0.001
- $1-\beta$ = power
- z = normal percentiles (use t percentiles for better accuracy)

Number of Events Needed to Detect Gene Specific Effects on Survival

- σ = standard deviation in log2 ratios for each gene
- δ = hazard ratio (>1) corresponding to 2-fold change in gene expression

$$\left[\frac{z_{1-\alpha/2} + z_{1-\beta}}{\sigma \log_2 \delta} \right]^2$$

Number of Events Required to Detect Gene Specific Effects on Survival $\alpha=0.001, \beta=0.05$

Hazard Ratio δ	σ	Events Required
2	**0.5**	**26**
1.5	**0.5**	**76**

Some Other Design Issues

- Selection of common reference
- Assignment of specimens to arrays and dyes for dual label time series experiments
- Design of class prediction studies
  - Split sample or cross validation
  - Sample size

Class Prediction

Statistical Methods Appropriate for Class Comparison Differ from Those Appropriate for Class Prediction

- Demonstrating statistical significance of prognostic factors is not the same as demonstrating predictive accuracy.
- Demonstrating goodness of fit of a model to the data used to develop it is not a demonstration of predictive accuracy.
- Most statistical methods were developed for inference, not prediction.
- Most statistical methods for were not developed for p>>n settings

Components of Class Prediction

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

Univariate Feature Selection

- Genes that are univariately differentially expressed among the classes at a significance level α (e.g. 0.01)
  - The α level is selected to control the number of genes in the model, not to control the false discovery rate
  - The accuracy of the significance test used for feature selection is not of major importance as identifying differentially expressed genes is not the ultimate objective

Linear Classifiers for Two Classes

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

$\underline{x}$ = vector of log ratios or log signals

F = features (genes) 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
- Diagonal linear discriminant analysis (DLDA) assumes features are uncorrelated
  - Naïve Bayes classifier
- Compound covariate predictor (Radmacher et al.) 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 subject to regularization condition
  - Can be written as finding hyperplane with separates the classes with a specified margin and minimizes length of weight vector
- Perceptrons are linear classifiers

When p>n

- For the linear model, an infinite number of 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?

Myth

- That complex classification algorithms perform better than simpler methods for class prediction
  - Many comparative studies indicate that simpler methods work as well or better for microarray problems

- Fitting complex functions to training data results in unstable classifiers unless there is a huge training dataset
- For unstable classifiers, the test sample error rate is generally much less than the generalization error rate

Model Stability Can Be Improved By

- Restriction to models with fewer parameters
  - Complexity depends on number of parameters per candidate feature, not per selected feature
- Reducing number of candidate features
  - Principal components
  - Cluster averages
- Not minimizing training error
  - Equivalent to including penalty for complexity
- Aggregating models
- Use fitting criterion incorporating robustness to changes in data

- With unstable classifiers, we obtain both large bias and large variance in estimating the true classifier function
  - Large bias because there are many classifiers with zero training set errors that are far from the true classifier function
  - Large variance because the selected classifier varies substantially with small variations in the data

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

- 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 indexed by a tuning parameter 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

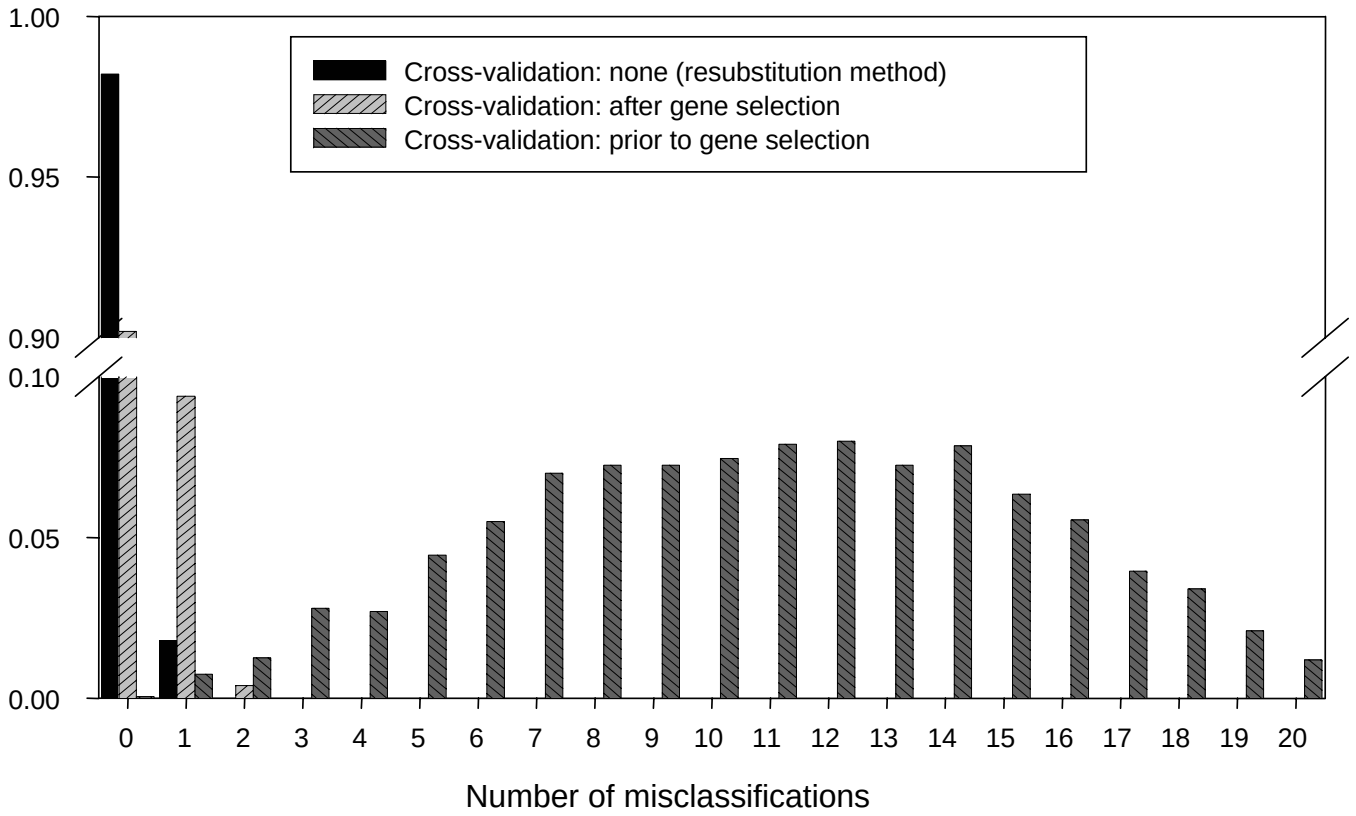
Prediction on Simulated Null Data

Generation of Gene Expression Profiles

- 14 specimens (P_i is the expression profile for specimen i)
- Log-ratio measurements on 6000 genes
- $P_i \sim \text{MVN}(\mathbf{0}, \mathbf{I}_{6000})$
- 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.


Cross-validation: none (resubstitution method)
Cross-validation: after gene selection
Cross-validation: prior to gene selection
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
Proportion of simulated data sets

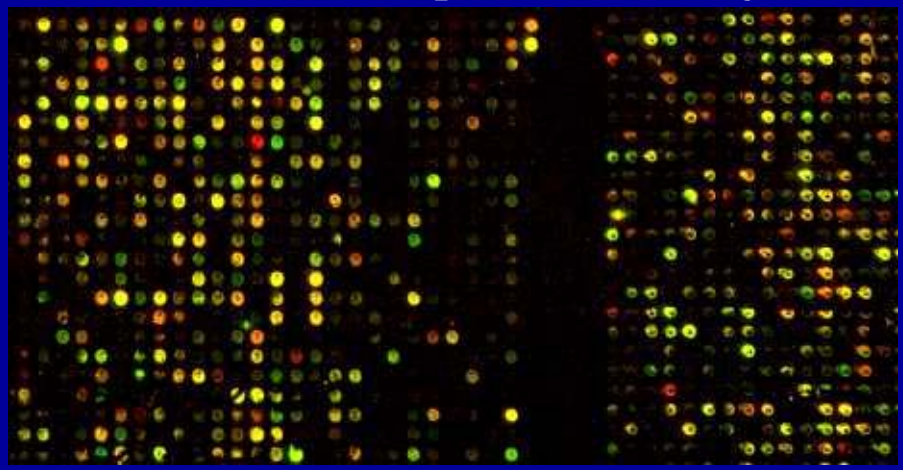
Permutation Distribution of Cross-validated Misclassification Rate of a Multivariate Classifier

- 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

RESEARCH QUESTION

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."
  - There is often many sets of features that provide zero training errors. Cross validation will provide an unbiased estimate of the generalization error for a specified algorithm that selects a specific model on a training set.

Cross-Validation

- Estimates prediction error for future data
  - For prediction using model developed using full current dataset
- Cross-validation is used to estimate prediction error of a defined algorithm, not as part of a model building algorithm
- If you use the results of cross-validation for model building, then a double nested cross-validation is needed to obtain a valid estimate of prediction error for the resulting model

Comparison of Internal Validation Methods

Molinaro, Pfiffer & Simon

- For small sample sizes, LOOCV is much more accurate than split-sample validation
  - Split sample validation is highly positively biased
- For small sample sizes, LOOCV is preferable to 10-fold, 5-fold cross-validation or repeated k-fold versions

- *Design & Analysis of DNA Microarray Investigations*, RM Simon, EL Korn, LM McShane, MD Radmacher, GW Wright, Y Zhao, Springer, 2003

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