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

Collaborators

- Kevin Dobbin
- 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

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
  - 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.

How many replicates are needed?

Levels of Replication

- Technical replicates
  - RNA sample divided into multiple aliquots and re-arrayed
- Biological replicates
  - Multiple subjects
  - Replication of the tissue culture experiment

- 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.

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

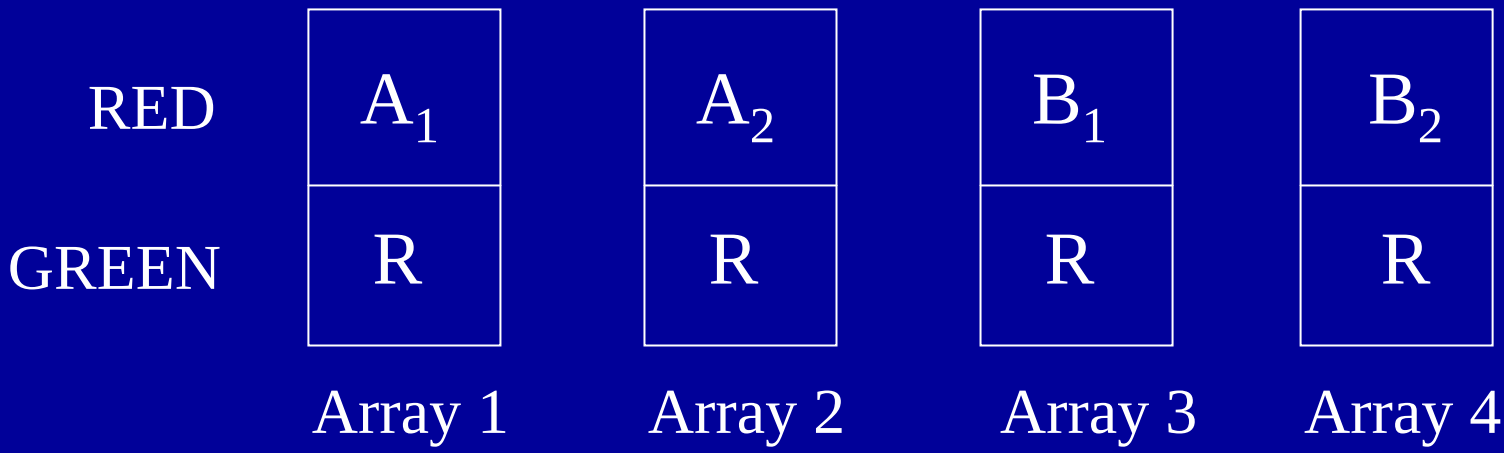
- Korn EL, Troendle JF, McShane LM, and Simon R. Controlling the number of false discoveries: Application to high-dimensional genomic data. Journal of Statistical Planning and Inference 124:379-398, 2004

- 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

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

- Reference Design
- Balanced Block Design
- Loop Design

Reference Design

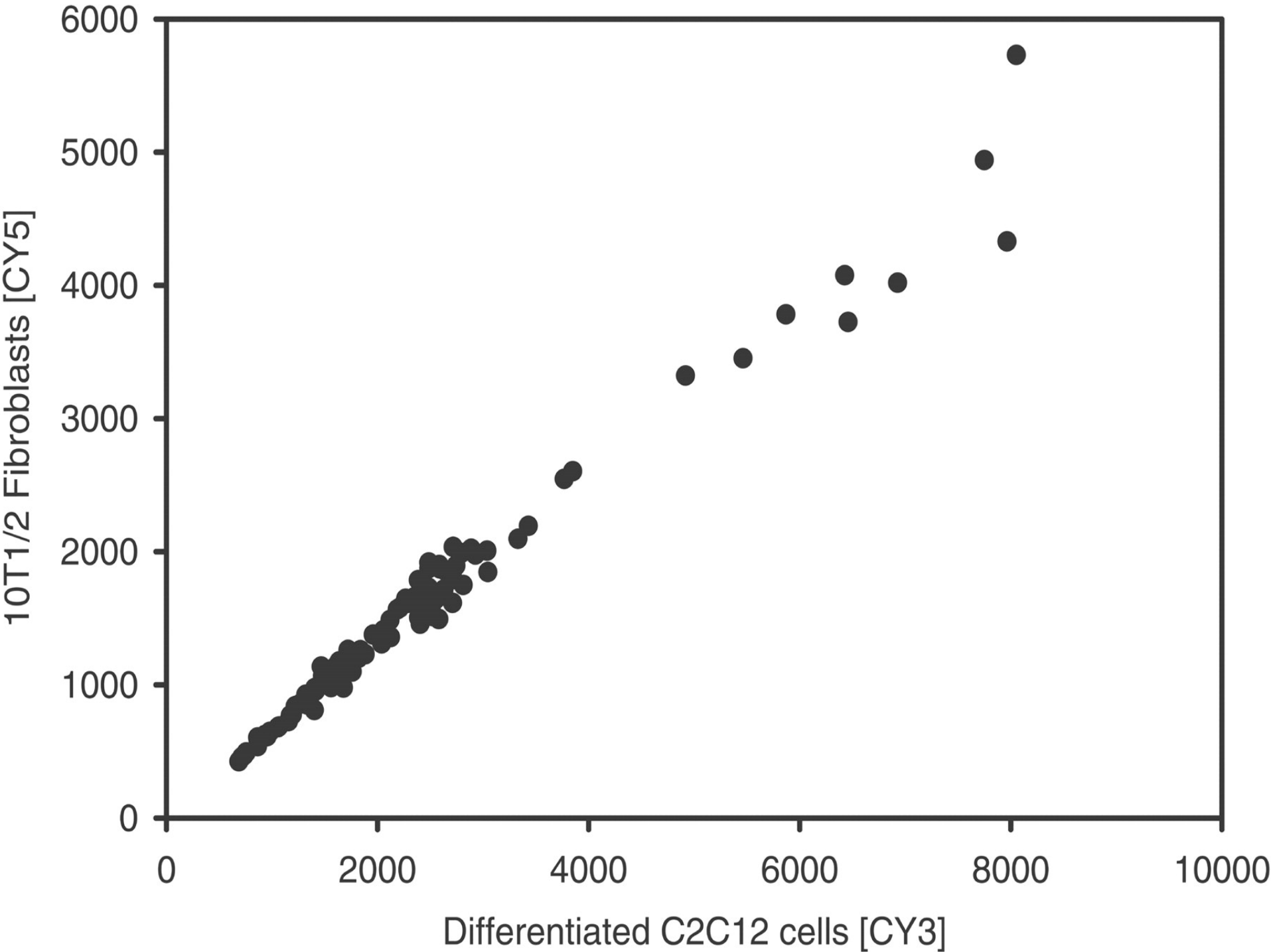


A_i = *i*th specimen from class A

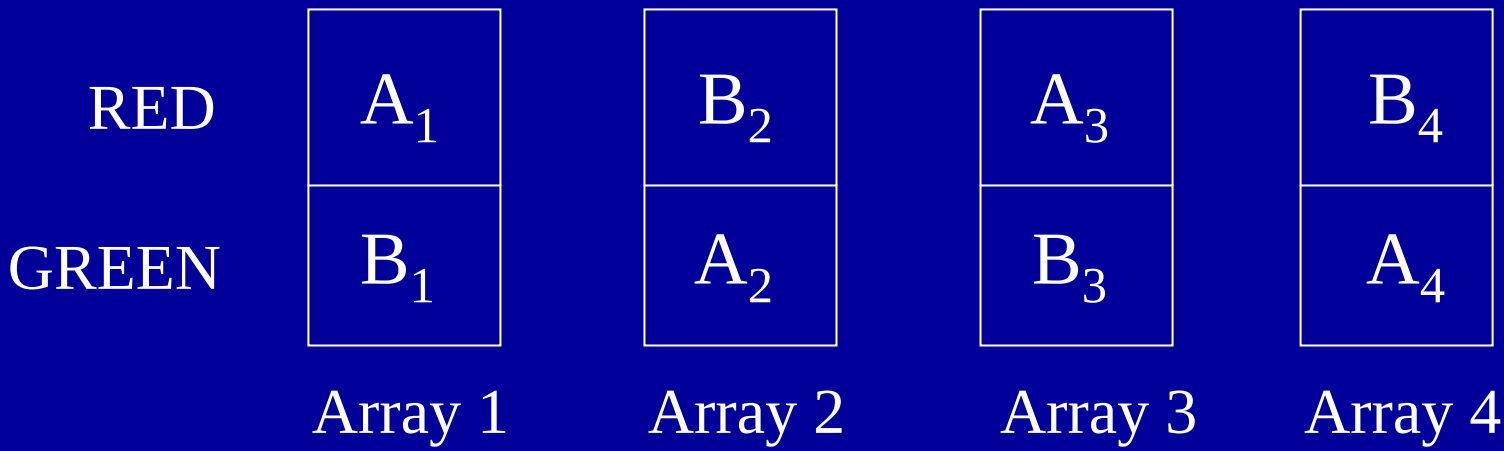
B_i = *i*th specimen from class B

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

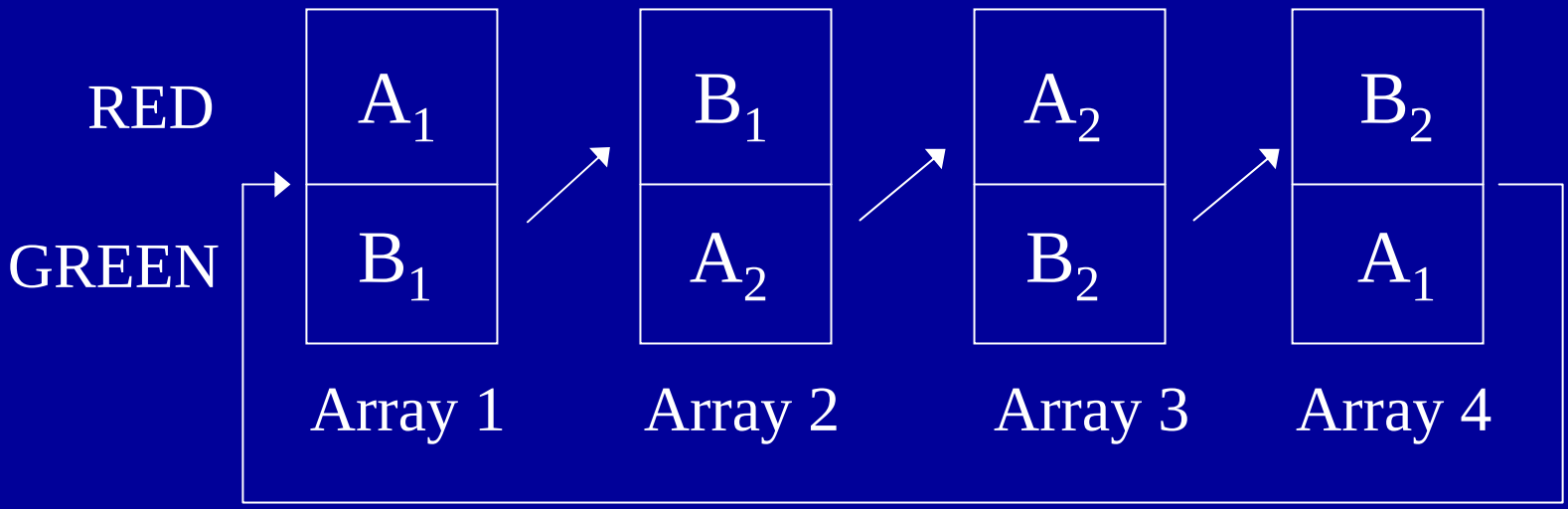
B_i = *i*th specimen from class B

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

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

- $r = G_g + AG_{ag} + VG_{vg} + SG_{sg} + \mathcal{M}$
- $\mathcal{M} \sim N(0, \blacklozenge_g^2)$
- Efficiency of design based on variance of estimators of VG_{ig}-VG_{jg}
- To study efficiency, assume $SG_{sg} \sim N(\mathbb{O}_g, \blacklozenge_g^2)$

- 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 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 (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 very efficient and require many fewer arrays than common reference designs. They are not appropriate for class discovery or class prediction and are more difficult to apply to more complicated class comparison problems.

Relative efficiencies of balanced block to reference designs.

Variance ratio of 4.

Relative Efficiencies		**Number of Classes**	
		2	**3**
Limiting Factor	**Same number of arrays used**	2.4	1.8
	Same non-reference samples used	1.2	0.9

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.

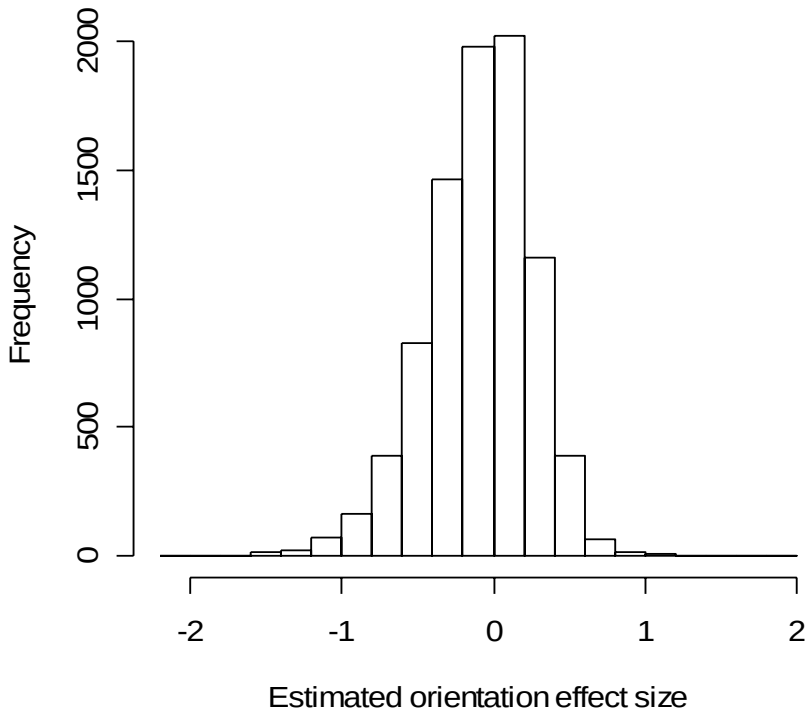
Dye Bias

- 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)_

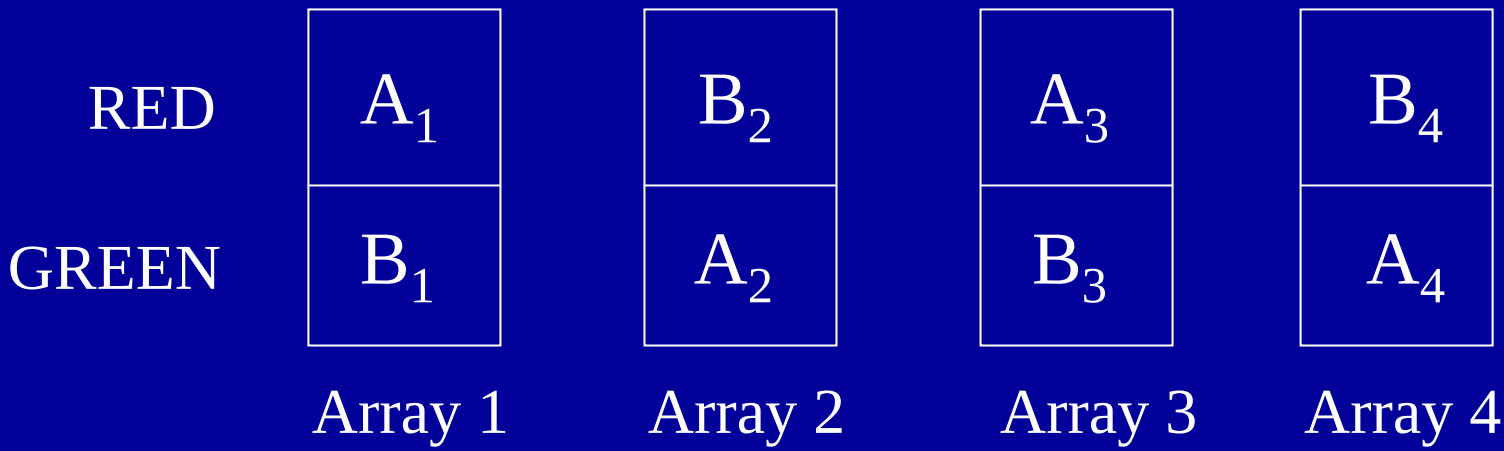
	Array 1	Array 2	Array 3	Array 4	Array 5	Array 6
Cy3	Tumor	Tumor	Tumor	Tumor	Tumor	Tumor
Cy5	Normal	Normal	Normal	Normal	Normal	Normal

(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

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

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

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
- 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

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

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

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

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

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

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

where δ = mean log-ratio or log signal difference between classes

σ = within class standard deviation of log-ratio or log signal

- 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	σ	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%

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

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

Dual Label Arrays With Balanced Block Design For 2 classes

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

- n = total number of arrays
- δ = mean log ratio (class 1 / class 2)
- η^2 = biological variance of log-ratio
- γ^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
- ♎ = 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

α=0.001, β=0.05

Hazard Ratio Ψ	Φ	Events Required
2	**0.5**	**26**
1.5	**0.5**	**76**

Avoid Confounding

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

Some Other Design Issues

- Selection of common reference
- Selection of sampling times for time series experiments
- Dye assignments for multi-factor experiments
- Design of class prediction studies
  - Split sample or cross validation
  - Sample size

- *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