

Design of DNA Microarray Studies

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

Collaborators

- Kevin Dobbin
- Joanna Shih

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 (supervised)
 - Identify genes differentially expressed among predefined classes.
- Class Prediction (supervised)
 - Develop multi-gene predictor of class label for a sample using its gene expression profile
- Class Discovery (unsupervised)
 - 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
- Supervised methods
- Requires multiple biological samples from each class

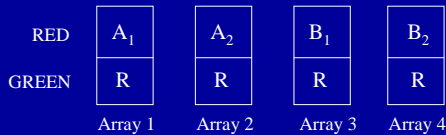
Allocation of Specimens to Dual Label Arrays for Simple Class Comparison Problems

- Reference Design
- Balanced Block Design
- Loop Design

References

- 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
- Simon R, Radmacher MD, Dobbin K. Design of studies with DNA microarrays. *Genetic Epidemiology* 23:21-36, 2002
- Simon R, Dobbin K. Experimental design of DNA microarray experiments. *Biotechniques* 34:1-5, 2002

Reference Design

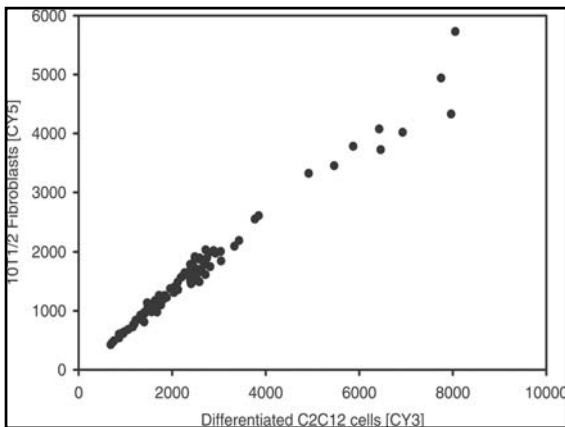


A_i = ith specimen from class A

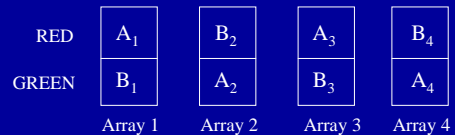
B_i = ith specimen from class B

R = aliquot from reference pool

- The common reference rna need not be biologically "relevant"
- The reference generally serves 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.



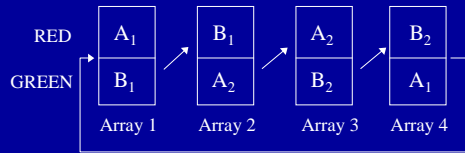
Balanced Block Design



A_i = ith specimen from class A

B_i = ith specimen from class B

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)

ANOVA for Logarithm of Background Adjusted Normalized Intensities

- Gene-Variety Models Fitted to Residuals after normalization separately by gene
 - Gene
 - Array by Gene (spot)
 - Variety by Gene
 - Sample within Variety by Gene

Gene-Variety Model

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

Myth

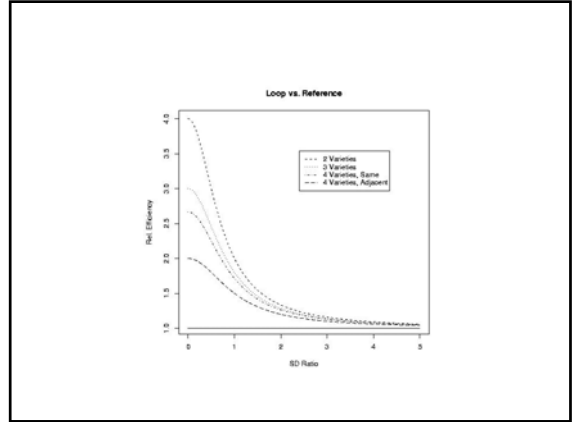
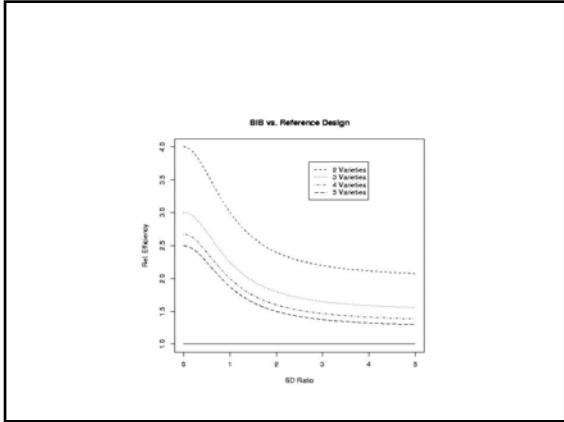
- Common reference designs for two-color arrays are inferior to “loop” designs.

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 inefficient for class discovery 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.

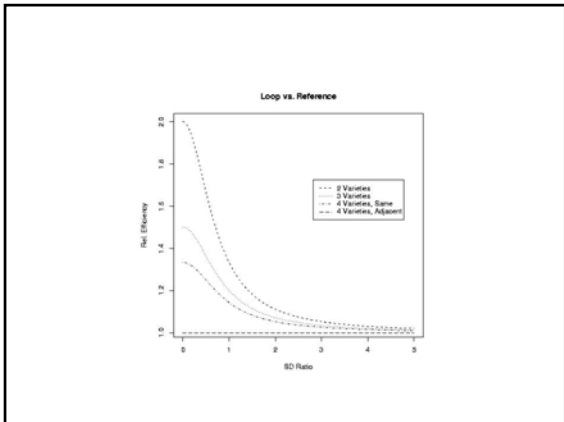
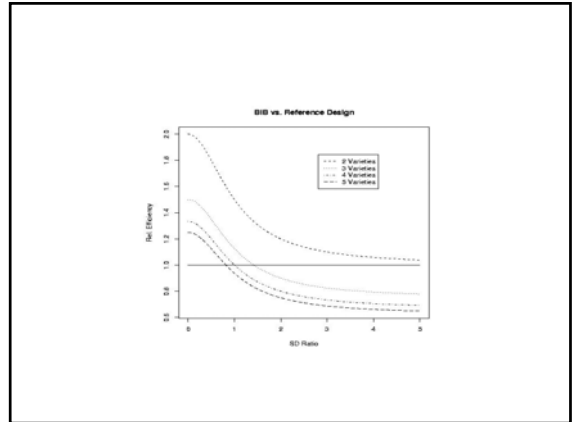
Comparison of Designs

- Equal number of arrays



Comparison of Designs

- Equal number of non-reference samples



Designs for Class Discovery

- Loop designs with 2 sub-samples per specimen make clustering possible without confounding array*gene effects with residual error
- Reference designs do not require sub-sampling
- Clustering is not possible with balanced block designs without confounding array*gene effects with residual error

Designs for Class Discovery

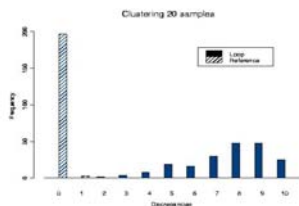
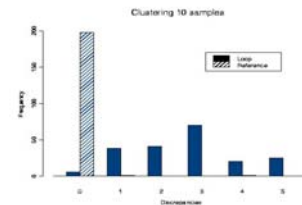
- For the loop design, variance of inter-sample contrasts depends on how close the samples appear in the loop

Evaluation of Designs for Class Discovery

- Generate data from two-varieties
- $\sigma_g^2 = \sigma_g^2$
- 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

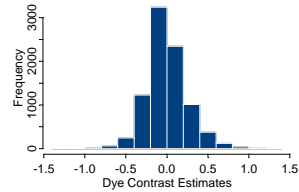


Myth

- 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



References

- 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
- Simon R, Radmacher MD, Dobbin K. Design of studies with DNA microarrays. *Genetic Epidemiology* 23:21-36, 2002
- Simon R, Dobbin K. Experimental design of DNA microarray experiments. *Biotechniques* 34:1-5, 2002

Dobbin, Shih, Simon ANOVA

$$r_{gadvf} = G_g + AG_{ga} + DG_{gd} + VG_{gv} + SG_{gf} + \varepsilon$$

r is background adjusted, normalized intensity

gene g

array a

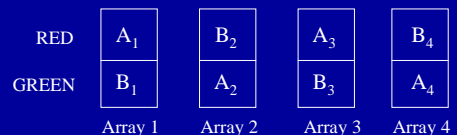
dye d

variety v (0=ref, 1=experimental)

sample s (individual)

- 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. Gene specific labeling bias does not effect class comparisons since specimens labeled with different dyes are never compared.

Balanced Block Design



A_i = i th specimen from class A

B_i = i th specimen from class B

4 Tables

Green	Array 1 Normal 1	Array 2 Normal 2	...	Array k Normal k	Array k+1 Normal k+1	...	Array n Normal n
Red	Cancer 1	Cancer 2	...	Cancer k	Cancer k+1	...	Cancer n
Green	Array n+1 Cancer 1	Array n+2 Cancer 2	...	Array n+k Cancer k			
Red	Normal 1	Normal 2	...	Normal k			

Table 1: Paired Samples Design: "Normal 1" indicates a sample of normal tissue from individual 1, and "Cancer 1" a sample of tumor tissue from individual 1. Array 1 represents a forward experiment for participant 1, and Array n + 1 a backward experiment for the same individual. k represents the number of samples that are run both forward and reverse. $n - k$ represents the number of samples that are run only once; these are assumed to be balanced, so that $\frac{n-k}{2}$ are run forward, and $\frac{n-k}{2}$ are run backward ($n - k$ is assumed even).

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

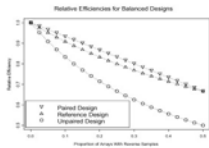


Figure 1: Balanced design comparison. Relative efficiency versus the proportion of arrays with reverse design, i.e. proportion of arrays with the same RNA sample on forward arrays (red) with the opposite channel. Relative Efficiency = $\frac{E_{\text{paired}}(n, k)}{E_{\text{unpaired}}(n, k)}$ where $E_{\text{paired}}(n, k)$ is the efficiency of the reverse design and $E_{\text{unpaired}}(n, k)$ is the efficiency of the forward design. Parameters include $\sigma^2 = \text{Var}(Y_{ij})$ for the paired and reference designs, and $\sigma_p^2 = \text{Var}(Y_{ij})$ for the unpaired design. In all three cases, the reverse design increases the efficiency.

Comparison of Experimental Specimens to Internal Reference Using a Reference Design

- Comparison to pooled normal tissue reference as a secondary objective
 - Inference limited to that pool
- Some reverse label array pairs are necessary to estimate gene specific dye bias; e.g. 5-10 pairs. Rest of arrays should be consistently labeled
- ANOVA model based comparison of specimen averages to internal reference pool adjusted for dye bias

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

Comparing 2 equal size classes

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

where δ = mean log-ratio difference between classes

σ = standard deviation

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

- Choose α small, e.g. $\alpha = .001$

- 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%

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)

Refinement

- Replace $z_{\alpha/2}$ for analytic strategy of using multivariate permutation test. In order to have probability $1 - \alpha$ that the number of false discoveries is no greater than k , use the $(1-\alpha)$ quantile of the permutation distribution of the k 'th smallest parametric t distribution p value computed from pilot or similar previous data.

References

- 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* (In Press).

Comparing 2 equal size classes with dual label arrays

- Total number of arrays for reference design:

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

$$\sigma^2 = b^2 + 2t$$

- Total number of arrays for balanced block design:

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

$$\tau^2 = 2b^2 + 2t^2$$

Number of Events Needed to Detect Gene Specific Effects on Survival

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

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

Number of Events Required to Detect Gene Specific Effects on Survival

$$\alpha = 0.001, \beta = 0.05$$

Hazard Ratio Ω	σ	Events Required
2	0.5	26
1.5	0.5	76