Design & Analysis of DNA Microarray Experiments in Translational Research

Richard Simon, D.Sc.

Chief, Biometric Research Branch, NCI

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

http://linus.nci.nih.gov

- Publications & technical reports
- Microarray myths
- BRB-ArrayTools

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, 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 (in press)

- 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

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 (In Press)

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 (In Press)

- Wright G.W. and Simon R. A random variance model for detection of differential gene expression in small microarray experiments. Bioinformatics (In Press)

BRB ArrayTools:
An integrated Package for the Analysis of DNA Microarray Data
Created by Statisticians for Biologists

BRB-ArrayTools

- Integrated software package using Excel-based user interface but state-of-the art analysis methods programmed in R, Java & Fortran
- Based on continuing evaluation of validity and usefulness of published methods
- Publicly available for non-commercial uses from BRB website:

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
- Class prediction models (6) with LOOCV prediction error and permutation analysis of LOOCV error rate
- Clustering tools for class discovery with reproducibility statistics on clusters
- Visualization tools including rotating 3D principal components plot exportable to Powerpoint with rotation controls
- Extensible via R plug-in feature
- Links genes to annotations in genomic databases
- Tutorials and datasets

- Effective microarray research requires clear objectives, careful planning and appropriate statistical analysis
- Clear objectives, but not gene specific mechanistic hypotheses

Design and Analysis Methods Should Be Tailored to Study Objectives

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

Class Comparison Examples

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

- 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

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.

Class Comparison Paradigm

- Evaluate extent to which each gene is differentially expressed among classes based on comparing means of independent biological samples from each class
  - e.g. t or F statistic
- Select the most differentially expressed genes in a manner that limits the false discovery rate to a specified level (e.g. 10%)
  - e.g. select threshold for t or F statistic

- 10,000 non-differentially expressed genes x 5% false positivity rate equals 500 false positives
- 10,000 x 0.1% = 10 false positives
- Multivariate permutation methods are the most powerful and robust methods for class comparison problems in microarray studies.
  - Available in BRB-ArrayTools

False Discovery Rate

- Proportion of the genes claimed to be differentially expressed among the classes that really are not

How many replicates should I do?

Levels of Replication

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

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.

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

- Reference Design
- Balanced Block Design
- Loop Design

Reference Design

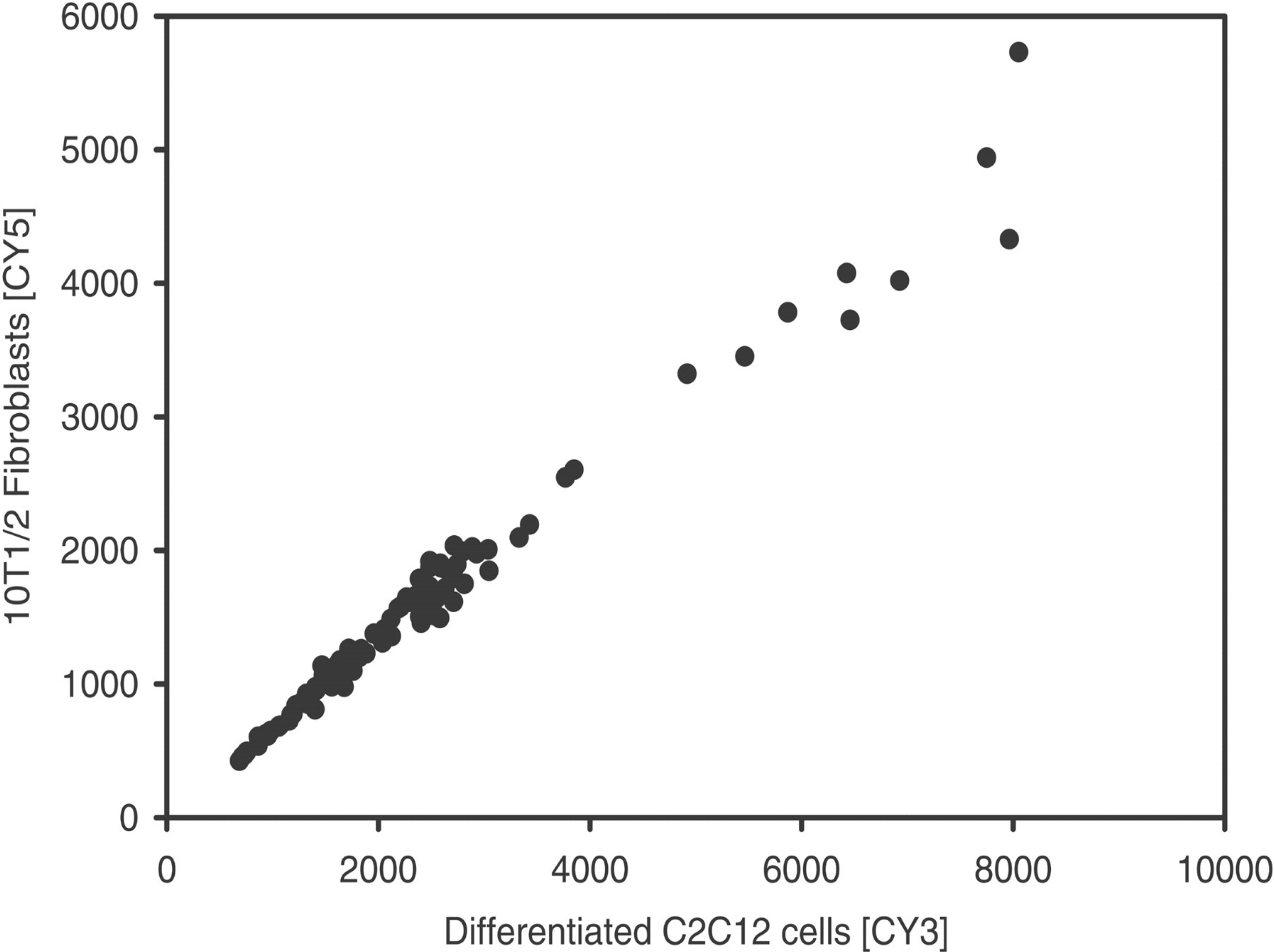
	Array 1	Array 2	Array 3	Array 4
RED	A_1	A_2	B_1	B_2
GREEN	R	R	R	R

A_i = *i*th specimen from class A

B_i = *i*th specimen from class B

R = aliquot from reference pool

- The 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.


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

Balanced Block Design

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

A_i = *i*th specimen from class A

B_i = *i*th specimen from class B

- 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.

- 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.

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

- 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

	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

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-β 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_β = standard normal percentiles

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

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)

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 Prediction

- Cluster analysis is generally not effective for class prediction
- Cluster analysis is frequently misleading when used for class prediction

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 10,000 genes measured, about 500 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

Class Prediction Paradigm

- Select features (F) to be included in predictive model using training data in which class membership of the samples is known
- Fit predictive model containing features F using training data
  - Diagonal linear discriminant analysis
  - Neural network
- Evaluate predictive accuracy of model on completely independent data not used in any way for development of the model

Leave-One-Out Cross-validation Paradigm for Evaluating Classification Error Rate

- Leave-out one specimen
  - Perform feature selection and model fitting on the training set consisting of the remaining specimens
  - Evaluate whether the model predicts correctly for the left-out specimen
- Repeat the above procedure leaving-out all specimens, one at a time, re-doing feature selection and model fitting for each training set separately
- Total the number of classification errors

Mis-conceptions About Cross Validation

- Too numerous to mention here
- Often used improperly in biomedical and bioinformatic literature

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

- 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

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