

Design And Analysis of 2-Channel Microarray Experiments

Jane Chang

Bowling Green State University

Joint With

Jason C. Hsu And Tao Wang

The Ohio State University

Graybill Conference 2003

Uses of Microarrays

■ Designer medicine

- ◆ Screen genes to build diagnostic or prognostic chip

■ Patient targeting

- ◆ Eliminate patient subgroup prone to serious side effects







■ Drug discovery

- ◆ Find proteins to synthesize or suppress to treat the disease

Research Goals

- **Biomedical:** Infer, from genes differentially expressed in normal vs diseased tissues, proteins involved in a disease (drug discovery).
- **Statistical:** Find a good design to efficiently and effectively identify the genes that are differentially expressed between 2 (or more) treatments.

2-Channel Microarray Data

		Array 1	Array 2	...	Array a
Green	1	3123.40	8143.71		1323.48
Dye				...	
	K	4522.56	4722.66		4522.86
Red	1	6123.27	4123.47		3126.44
Dye				...	
	K	9521.35	7522.54		4522.23

Outline of Presentation

- Overview
- Linear Model
- Simultaneous Confidence Intervals
- The Design ($t=2$)
- A- Optimality for gene allocation
- The Design ($t>2$)

Approaches to Gene Expressions Analysis

■ Presence of effect

- ◆ Randomization model:
random assignment of responses to treatment
- ◆ Test H_{0i} : no treatment effect
(e.g., permutation tests)

■ Magnitude of effect

- ◆ Model treatment effect θ_i on i^{th} gene
(e.g., location or scale)
- ◆ Test H_{0i} : $\theta_i = \delta$ or confidence interval on θ_i

Modeling Approach to Gene Expressions Analysis

1. Model response as appropriate, including blocking factors to remove nuisance effects (e.g, array and dye effects)
2. Estimate differential expressions
3. Obtain joint distribution of estimates
4. Adjust for multiplicity of multiple comparisons based on joint distribution

Control of Error Rate

- False Discovery Rate (FDR)
 - ◆ Control expected proportion of false positives
- Familywise Type I Error (FWER)
 - ◆ Control probability of at least one false positive
 - a. Tests of equality
 - ◆ Individual P-values
 - ◆ Multiplicity Adjusted P-values
 - b. Confidence intervals
 - ◆ Individual confidence intervals (*Kerr et al, 2000*)
 - ◆ Simultaneous confidence intervals

Model for Gene Expression Data

x_{ijklm} = observed gene expression

$$y_{ijklm} = \log(x_{ijklm}) = \mu + T_k + G_l + (TG)_{kl} + A_i + D_j + \varepsilon_{ijklm}$$

$$T_k + G_l + (TG)_{kl}$$

treatment effect

$$A_i + D_j$$

block effect

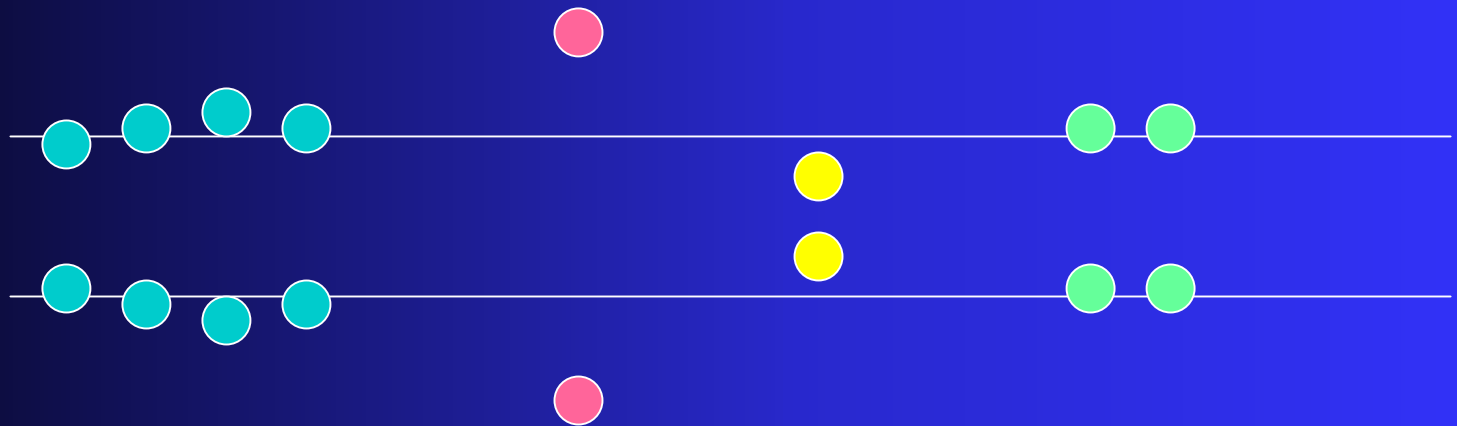
$$\varepsilon_{ijklm}$$

noise

Choice of “Control” Genes

- All genes
 - ◆ Average of all genes (Kerr, Churchill *et al*)
- Housekeeping genes (expressed in all cells)
 - ◆ Median of housekeeping genes (Amaratunga and Cabrera)
 - ◆ Average of housekeeping genes (Hsu, Chang, Wang)

Parameters of Interest



- If H denotes the set of housekeeping genes, then parameters of interest are

$$[(TG)_{2g} - (TG)_{1g}] - [\sum_{g^* \in H} (TG)_{2g^*} / |H| - \sum_{g^* \in H} (TG)_{1g^*} / |H|]$$

Microarray analysis essentials

Estimates need to be computed fast enough for data analysis and research!

1. Least squares (no iteration)
2. using designs giving explicit formulas (no matrix inversion)

Least squares: 1000 genes w/ 4 replications each

- Matrix inversion: ?
- Formulas (vectorized) < 8 minutes/20,000

$2 \times a$ Generalized Latin Square Design

Dye\Array	1	2	...	$a - 1$	a
Green	T_1	T_2	...	T_1	T_2
Red	T_2	T_1	...	T_2	T_1

- Each gene can be spotted unequal number of times on each array
- The number of replications of each gene remains constant across all arrays

Least Squares Estimation

- With *i.i.d.* normal errors, estimating $(TG)_{ig}$ by its sample mean results in BLUE for

$$(TG)_{2g} - (TG)_{1g} - [\sum_{g^* \in H} (TG)_{2g^*} / |H| - \sum_{g^* \in H} (TG)_{1g^*} / |H|]$$

Robust Design

- Orthogonality

$$y_{ijklm} = \mu + T_k + G_l + (TG)_{kl} + A_i + D_j + (AG)_{il} + (DG)_{jl} + \varepsilon_{ijklm}$$

- Mixed effect model

- ◆ Array is random

Optimal Allocation of Genes

K spots on each array (fixed)

g target genes (fixed)

h housekeeping genes (fixed)

n replications of each target gene

m replications of each housekeeping gene

A-optimality

Minimize average variance of differential estimates

$$m / n = g^{1/2} / h$$

$2 \times a$ Generalized Youden Design

Dye\Array	1	2	3	4	5	6
Green	T_1	T_2	T_1	T_3	T_2	T_3
Red	T_2	T_1	T_3	T_1	T_3	T_2

- Array - Balanced Incomplete Block Design
- Dye - Complete Block Design
- Each gene appears equal number of times on each array

Summary

- Linear Model Approach
- Simultaneous Confidence Intervals (FWER)
- Housekeeping genes for control
- $2 \times a$ Generalized Latin Square Design ($t=2$)
- A- Optimality for gene allocation
- $2 \times a$ Generalized Youden Design ($t>2$)