

Microarray Experimental Design and Analysis

**Short course: Practical Analysis of DNA
Microarray Data**
Instructors: Vince Carey & Sandrine Dudoit

KolleKolle, Denmark
October 26-28, 2003

Slides prepared jointly with
Yee Hwa (Jean) Yang
Biostatistics, UCSF.

Combining data across slides

Data on G genes for n hybridizations

→ $G \times n$ genes-by-arrays data matrix

		Array1	Array2	Array3	Array4	Array5	...
Genes	Gene1	0.46	0.30	0.80	1.51	0.90	...
	Gene2	-0.10	0.49	0.24	0.06	0.46	...
	Gene3	0.15	0.74	0.04	0.10	0.20	...
	Gene4	-0.45	-1.03	-0.79	-0.56	-0.32	...
	Gene5	-0.06	1.06	1.35	1.09	-1.09	...

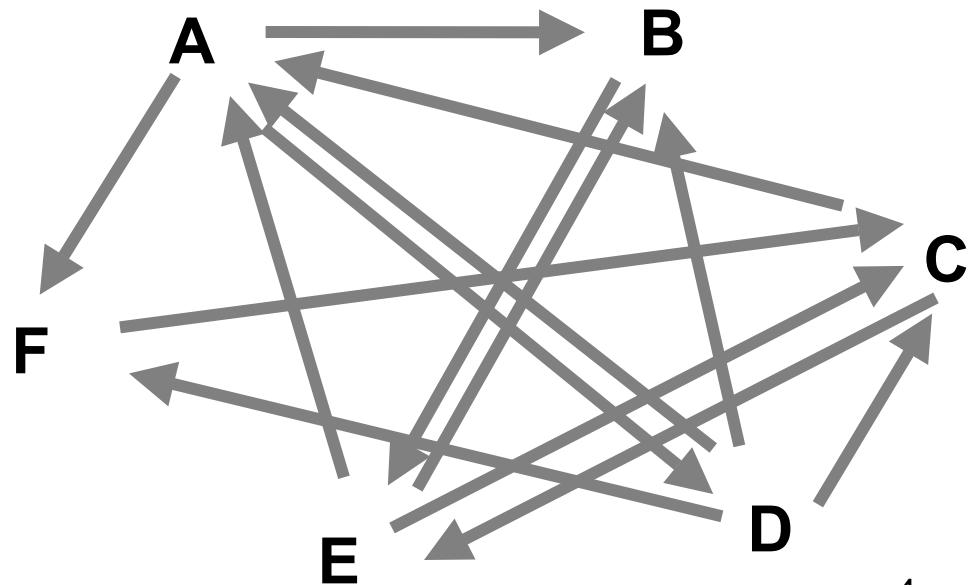
$M = \log_2(\text{Red intensity} / \text{Green intensity})$
expression measure, e.g, RMA

Combining data across slides

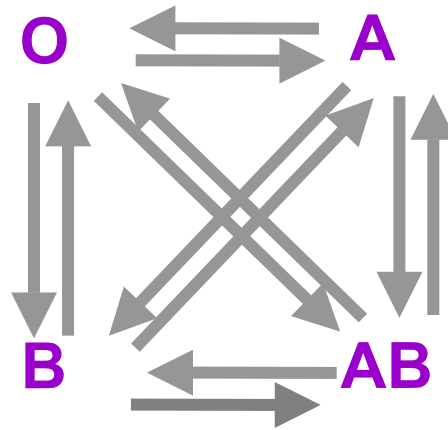
... but columns have **structure**

How can we design experiments and combine data across slides to provide accurate estimates of the effects of interest?

Experimental design
Regression analysis



Experimental design



Experimental design

Proper experimental design is needed to ensure that questions of interest **can** be answered and that this can be done **accurately**, given experimental constraints, such as cost of reagents and availability of mRNA.

Experimental design

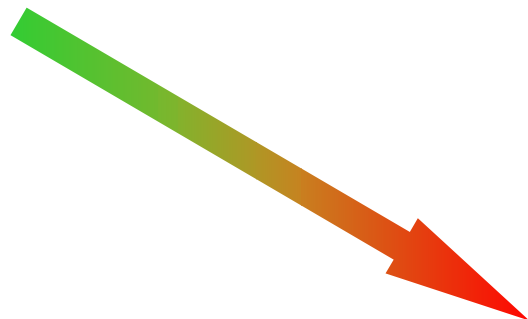
- Design of the array itself
 - which cDNA probe sequences to print;
 - whether to use replicated probes;
 - which control sequences;
 - how many and where these should be printed.
- Allocation of target samples to the slides
 - pairing of mRNA samples for hybridization;
 - dye assignments;
 - type and number of replicates.

Graphical representation

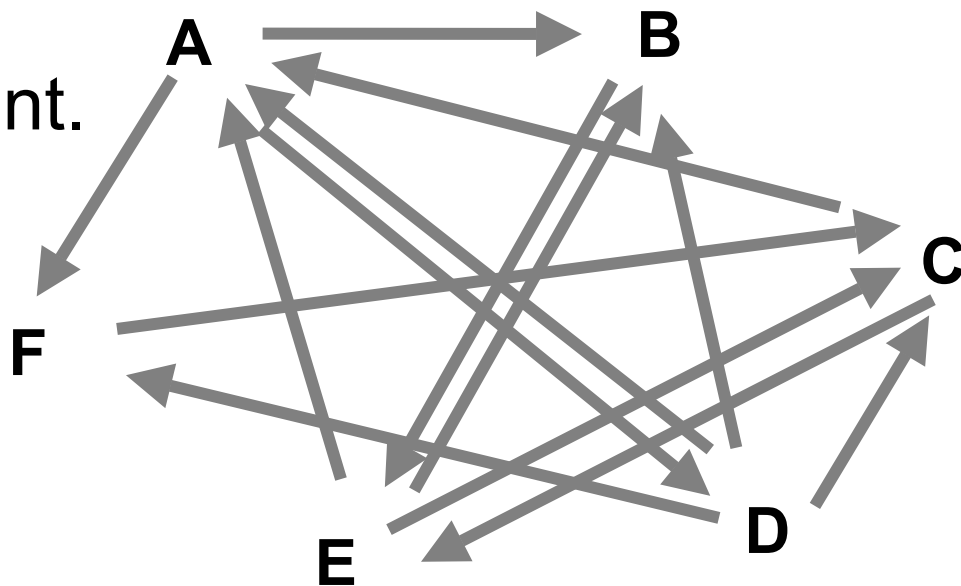
Multi-digraph

- *Vertices*: mRNA samples;
- *Edges*: hybridization;
- *Direction*: dye assignment.

Cy3 sample



Cy5 sample



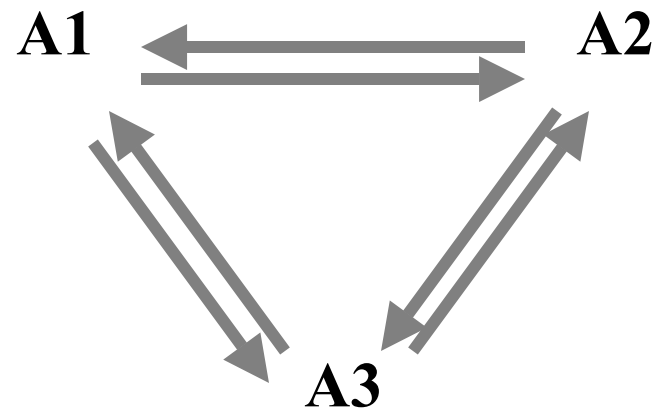
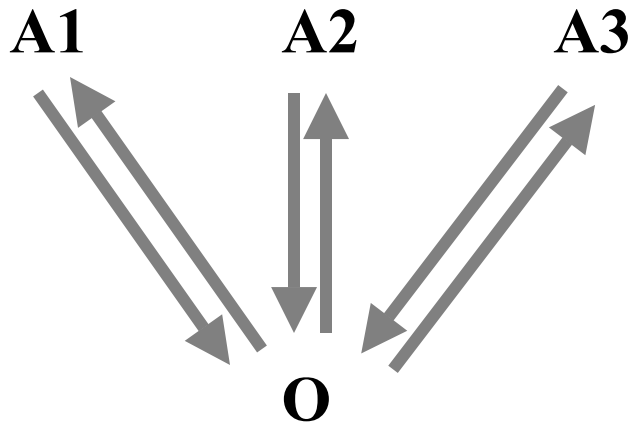
A design for 6 types of mRNA samples

Graphical representation

- The structure of the graph determines which effects can be estimated and the **precision** of the estimates.
 - Two mRNA samples can be compared only if there is a **path** joining the corresponding two vertices.
 - The precision of the estimated contrast then depends on the **number of paths** joining the two vertices and is inversely related to the **length of the paths**.
- Direct comparisons **within slides** yield more precise estimates than indirect ones between slides.

Comparing K treatments

(i) Common reference design (ii) All-pairs design



Question: Which design gives the most precise estimates of the contrasts $A1-A2$, $A1-A3$, and $A2-A3$?

Comparing K treatments

- **Answer:** The all-pairs design is better, because comparisons are done **within slides**.

For the same precision, the common reference design requires three times as many hybridizations or slides as the all-pairs design.

- In general, for K treatments

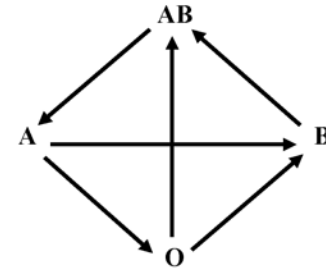
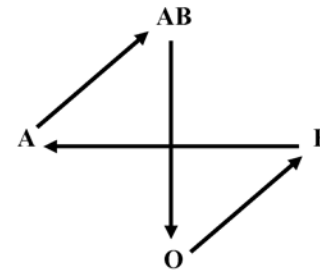
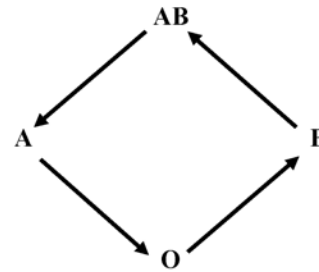
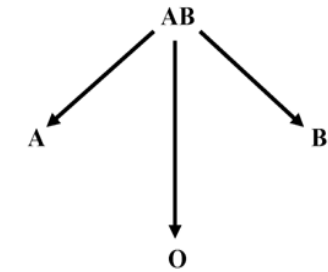
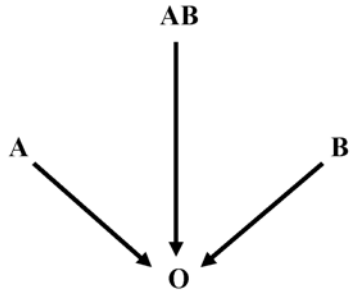
Relative efficiency

$$= 2K/(K-1) = 4, 3, 8/3, \dots \rightarrow 2.$$

For the same precision, the common reference design requires $2K/(K-1)$ times as many hybridizations as the all-pairs design.

2 x 2 factorial experiment

two factors, two levels each



(1) Common ref.

(2) Common ref.

(3) Connected

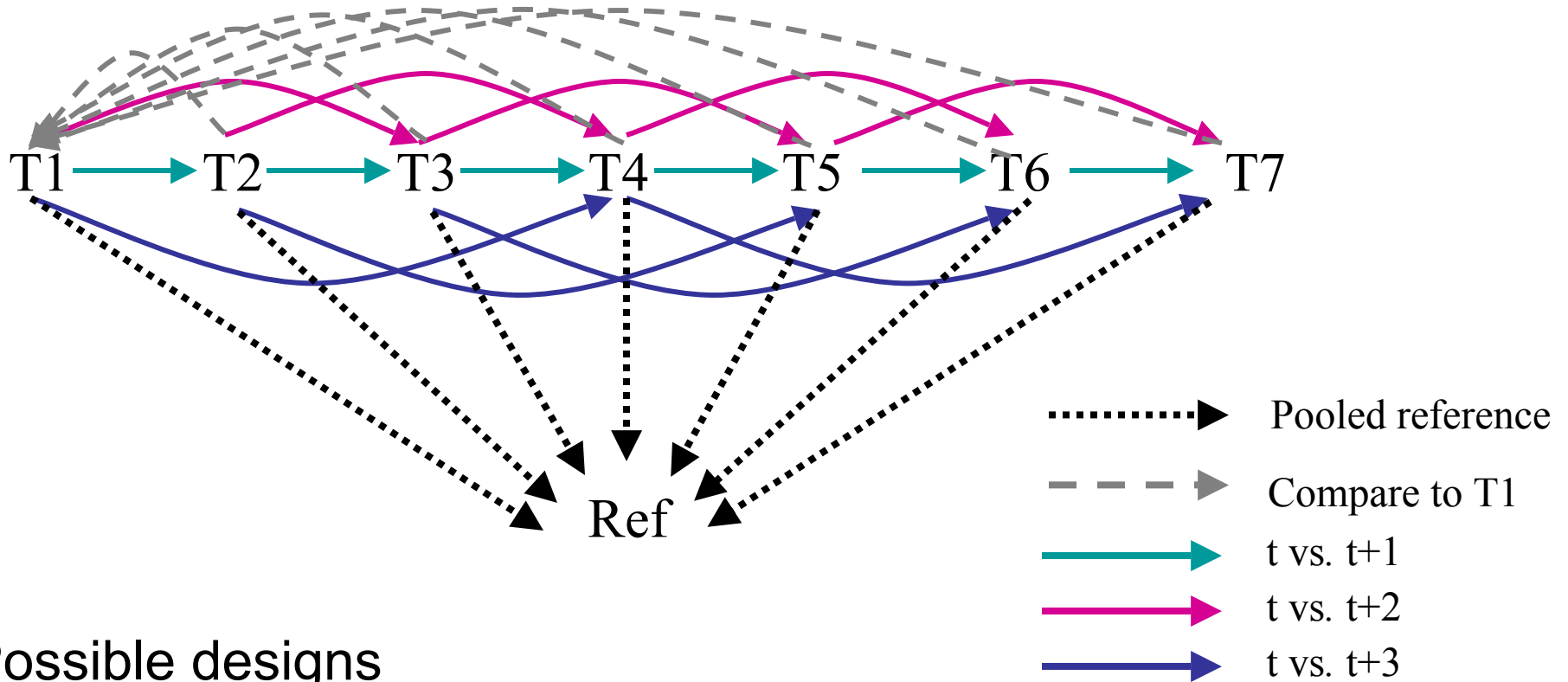
(4) Connected

(5) All-pairs

Scaled variances of estimated effects

	(1)	(2)	(3)	(4)	(5)
Main effect A	1	2	1	4/3	1
Main effect B	1	2	1	1	1
Interaction AB	3	3	4/3	8/3	2
Contrast A-B	2	2	4/3	1	1

Time course



Possible designs

- 1) All samples vs. common pooled reference
- 2) All samples vs. time 1
- 3) Direct hybridizations between timepoints

Design choices in time course experiments		t vs. t+1			t vs. t+2			
		T1T2	T2T3	T3T4	T1T3	T2T4	T1T4	Ave
N=3	A) T1 as common reference 	1	2	2	1	2	1	1.5
	B) Direct hybridization 	1	1	1	2	2	3	1.67
N=4	C) Common reference 	2	2	2	2	2	2	2
	D) T1 as common ref + more 	.67	.67	1.67	.67	1.67	1	1.06
	E) Direct hybridization choice 1 	.75	.75	.75	1	1	.75	.83
	F) Direct hybridization choice 2 	1	.75	1	.75	.75	.75	.83
Sandrine Dudoit 								14

Experimental design

- In addition to experimental constraints, design decisions should be guided by the knowledge of which effects are of greater interest to the investigator.
E.g. which main effects, which interactions.
- The experimenter should thus decide on the comparisons for which he wants the most precision and these should be made **within slides** to the extent possible.

Experimental design

- N.B. Efficiency can be measured in terms of different quantities
 - number of slides or hybridizations;
 - units of biological material, e.g. amount of mRNA for one channel.

Issues in experimental design

- Replication.
- Type of replication:
 - *within* or *between* slides replicates;
 - *biological* or *technical* replicates
i.e., different vs. same extraction:
generalizability vs. reproducibility.
- Sample size and power calculations.
- Dye assignments.
- Combining data across slides and sets of experiments:
regression analysis ... next.

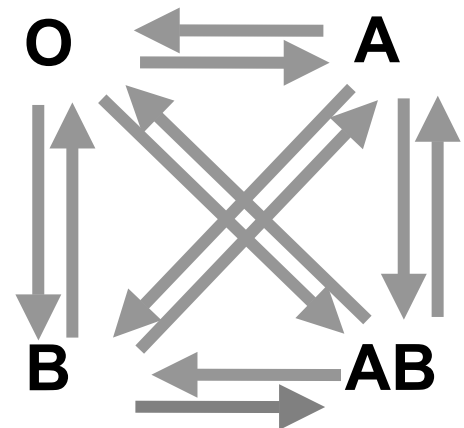
2 x 2 factorial experiment

two factors, two levels each

Study the **joint** effect of two treatments (e.g. drugs), A and B, say, on the gene expression response of tumor cells.

There are four possible treatment combinations

- AB: both treatments are administered;
- A : only treatment A is administered;
- B : only treatment B is administered;
- O : cells are untreated.



2 x 2 factorial experiment

For **each** gene, consider a linear model for the joint effect of treatments A and B on the expression response.

$$\mu_{AB} = \mu + \alpha + \beta + \gamma$$

$$\mu_A = \mu + \alpha$$

$$\mu_B = \mu + \beta$$

$$\mu_0 = \mu$$

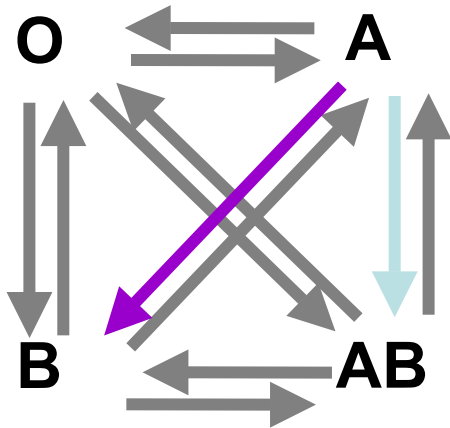
μ : baseline effect;

α : treatment A main effect;

β : treatment B main effect;

γ : interaction between treatments A and B.

2 x 2 factorial experiment



Log-ratio M for hybridization



estimates

$$\mu_{AB} - \mu_A = \beta + \gamma$$

Log-ratio M for hybridization



estimates

$$\mu_B - \mu_A = \beta - \alpha$$

+ 10 others.

Regression analysis

- For parameters $\theta = (\alpha, \beta, \gamma)$, define a **design matrix** X so that $E(M)=X\theta$.
- For each gene, compute **least squares estimates** of θ .

$$E \begin{pmatrix} M_{11} \\ M_{12} \\ M_{21} \\ M_{22} \\ M_{31} \\ M_{32} \\ M_{41} \\ M_{42} \\ M_{51} \\ M_{52} \\ M_{61} \\ M_{62} \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & -1 & 0 \\ 1 & 1 & 1 \\ -1 & -1 & -1 \\ 1 & 0 & 0 \\ -1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & -1 & -1 \\ 1 & 0 & 1 \\ -1 & 0 & -1 \\ -1 & 1 & 0 \\ 1 & -1 & 0 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix}$$



$$\hat{\theta} = (X'X)^{-1} X'M$$

Regression analysis

- Combine data across slides for **complex designs**
 - can “link” different sets of hybridizations.
- Obtain **unbiased** and **efficient** estimates of the effects of interest (BLUE).
- Obtain measures of **precision** for estimated effects.
- Perform **hypothesis testing**.
- Extensions of linear models
 - generalized linear models;
 - robust weighted regression, etc.

Regression analysis

- Use estimated effects in clustering and classification

genes x arrays matrix



genes x estimated effects matrix