DESIGN ISSUES FOR cDNA MICROARRAY EXPERIMENTS

- General Overview (CRD & RBD)
- Random and Fixed Effects
- Design Issues with Microarrays
  - Reference and loop designs
  - Biological and technical replications
- Additional Topics

INTRO TO EXPERIMENTAL DESIGN

- Most designs involve 2 or more factors.
- Generally two types of factors in an experiment:
  1) Treatment structure: consists of those factors that the experimenter has selected to study; e.g. diets, drugs, gender
  2) Design structure: consists of grouping of the experimental units into homogeneous groups or blocks; e.g. pens, litters, days (of assay), animals (repeated measures)

EXAMPLE 1

- One homogeneous set (block) of experimental units
  - assign treatments completely at random to experimental units.
  - Design: Completely Randomized Design (CRD)

Treatment 1	Treatment 2	Treatment 3
Pig 1	Pig 2	Pig 4
Pig 3	Pig 7	Pig 5
Pig 8	Pig 10	Pig 6
Pig 9	Pig 11	Pig 12

Limited design structure (i.e. no blocking factors!)

EXAMPLE 2

- Several blocks of experimental units
  - assign treatments at random to experimental units within blocks.
  - Design: Randomized Block Design (RBD)

Pen 1
Treatment 1 → Pig 1
Treatment 2 → Pig 2
Treatment 3 → Pig 3
Treatment 4 → Pig 4

Pen 2
Treatment 1 → Pig 1
Treatment 2 → Pig 2
Treatment 3 → Pig 3
Treatment 4 → Pig 4

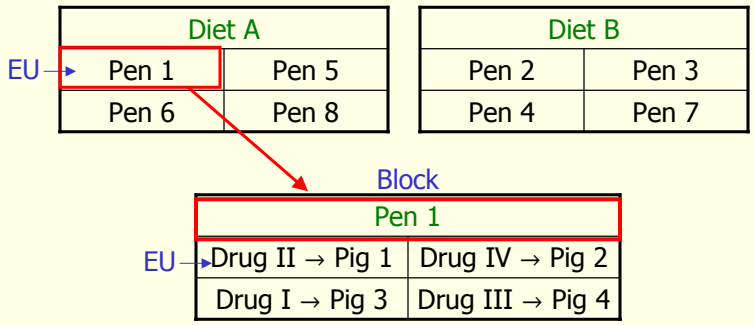
Pen 3
Treatment 1 → Pig 1
Treatment 2 → Pig 2
Treatment 3 → Pig 3
Treatment 4 → Pig 4

RCBD:

Block size = # of treatments

EXAMPLE 3

- Two different sizes of experimental units, each assigned at random to different levels of a respective factor.
  - Design: Split Plot Design



TYPICAL MODELING ASSUMPTIONS

1. The elements of the design structure are *random effects*.

2. There is *no interaction* among elements of the design structure and elements of the treatment structure.

- These assumptions aid in constructing an appropriate model.

MODEL CONSTRUCTION

Construct the model to describe data from a given treatment structure and design structure as:

Y = {treatment structure components}
+ {design structure components}
+ {error structure}

Error terms are typically derived from treatment structure by design structure interactions.

EXAMPLE

- One-way treatment structure in a randomized complete block design structure where t treatments are randomized to experimental units within each of b blocks.

Source	d.f.
Blocks	$b-1$
Treatments	$t-1$
Error = Block x Treatment	$(b-1)(t-1)$

Big tradeoff: Generally smaller standard errors in a block design compared to a completely randomized design...but

Less error df of test!!

Only block if blocking is effective

What if my block size is NOT the same as the number of treatments?

- Example, suppose wish to block on litters of size 3 but have 4 treatments to compare!

Pen 1
Treatment 1 → Pig 1
Treatment 2 → Pig 2
Treatment 3 → Pig 3

Pen 2
Treatment 2 → Pig 1
Treatment 3 → Pig 2
Treatment 4 → Pig 3

Pen 3
Treatment 3 → Pig 1
Treatment 4 → Pig 2
Treatment 1 → Pig 3

Pen 4
Treatment 4 → Pig 1
Treatment 1 → Pig 2
Treatment 2 → Pig 3

- Balanced incomplete block design (BIBD)
  - Each treatment occurs with each other treatment an equal number of times
- Partially balanced incomplete block designs (PBIBD)
  - Nearly as efficient as BIBD
  - Should be considered when BIBD design is not possible

TYPE OF EFFECTS

- A factor might be called a set of *random effects* if the levels of that factor are a random sample from a population of such levels; e.g. pens, litters, days, animals, microarrays (definition somewhat restrictive).
- A factor is called a set of *fixed effects* if the levels of that factor are selected by some nonrandom process. e.g. drugs, diets, gender

TYPES OF MODELS

- Fixed effects model: A model is called a fixed effects model if all of the factors in the model are fixed effects and it involves only one variance component.
- Random effects model: A model is called a random effects model if all of the factors in the model are random effects.
- Mixed effects model: A model is called a mixed effects model if some of the factors in the model are fixed effects and some are random effects or if all of the factors in the model are fixed effects and there is more than one variance component in the model.

Note: Most designs are mixed! Only a few designs; completely randomized designs: e.g. one-way, factorials, response surface) might be considered fixed.

Why are mixed model structures so popular in agricultural experimental designs?

- Because of the consideration of appropriate design structures!!
  - Error control (complete and incomplete block designs, nested designs)
  - Convenient to have different sizes of experimental units (split plots)

Design and Analysis of Comparative Microarrays Experiments

⇨ Most of the concepts and methods apply to all the most widely used types of microarrays:

- Radiolabeled cDNA arrays on nylon membranes.

our focus

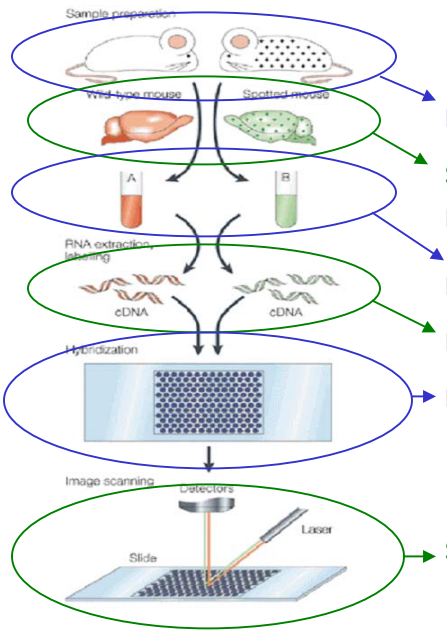
- Two color, fluorescently labeled cDNA arrays on nylon membranes.
- Two color, fluorescently labeled cDNA or oligonucleotide arrays on glass slides.

- Single color, fluorescently labeled, high-density short oligonucleotide arrays on silicon chips.

⇨ Microarray experimental design; two aspects:

- Designing of the array itself
  (controls, DNA probes, where to print, number of spots spots, etc.).
- Assignment of samples for competitive hybridization
  (treatments, replications, labeling, pooling, etc.).

Sources of Variability in Microarray Experiments



(Geschwind, 2001)

Biological heterogeneity

Specimen collection/Handling effects

Biological Heterogeneity within Specimen

RNA extraction/amplification

Fluor labeling

Hybridization

size/shape of spot

sample distribution across slide

Scanning (Voltage/power/software)

Random (blocking) effects or residual error??? Depends on the design!!!!

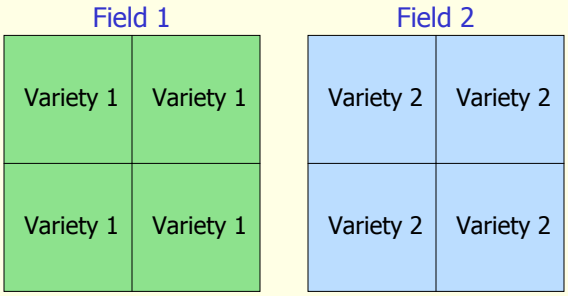
DESIGN ISSUES

- Should take sources of variation into consideration as fixed, random or residual effects!
  - Biological variability!!
  - Gene
  - Dye
  - Slide (array)
  - Patch or Print-tip within Slide
  - Spot within Patch effects
  - Print batch of slides
  - Sample handling
  - Any interactions thereof
    - Key interaction of interest → Treatment by Gene!!!

"It is possible, and indeed it is all too frequent, for an experiment to be so conducted that no valid estimate of error is available" (R. A. Fisher)

Down on the farm with experimental design…

- Say there are *four plots per field* and wish to compare *two corn varieties*



Field and Variety effects are confounded.

NO REPLICATION FOR VARIETY!

The ABC's of experimental design

- Blocking

Field 1

Variety 1	Variety 2
Variety 2	Variety 1

Field 2

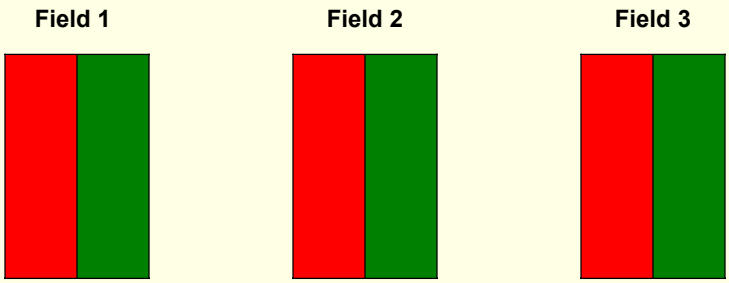
Variety 2	Variety 1
Variety 1	Variety 2

Field and Variety effects are separately estimable of each other.

Still, effectively only two replicates…not four!

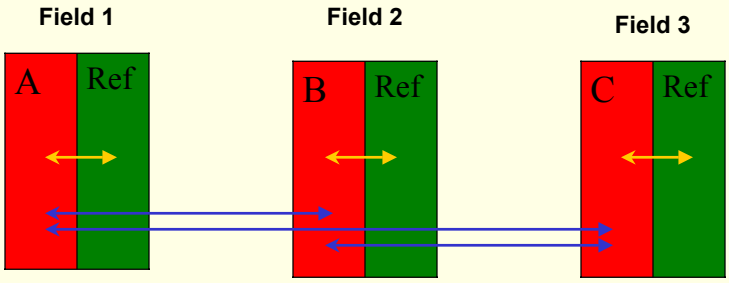
The ABC's of experimental design

Suppose you wish to compare 3 varieties in three different fields, but each field has enough room for only 2 plots. How would you allocate varieties to plots?



REFERENCE DESIGN

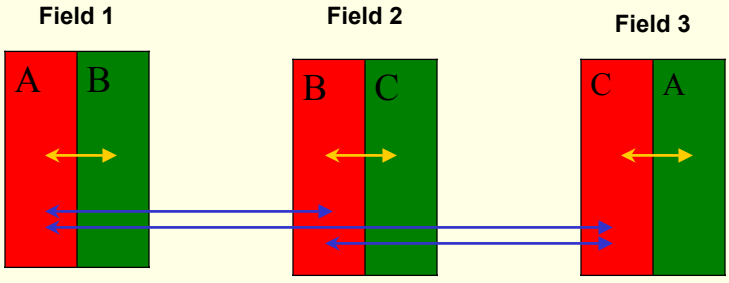
Introduce a 4th "reference" variety.



Direct comparisons between each variety with reference variety

Indirect comparison between each variety with each other.

Balanced incomplete block design (BIBD) "LOOP" design



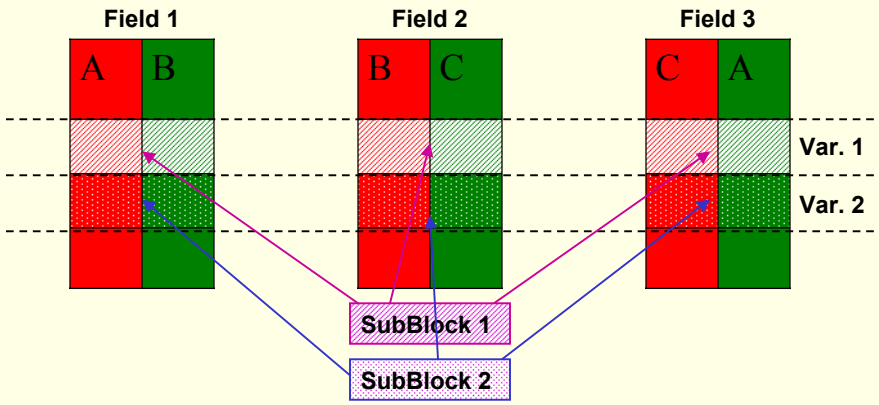
Direct comparisons between each variety with every other variety

Indirect comparison information utilized also.

Variance of treatment differences is 2/3 smaller relative to reference design

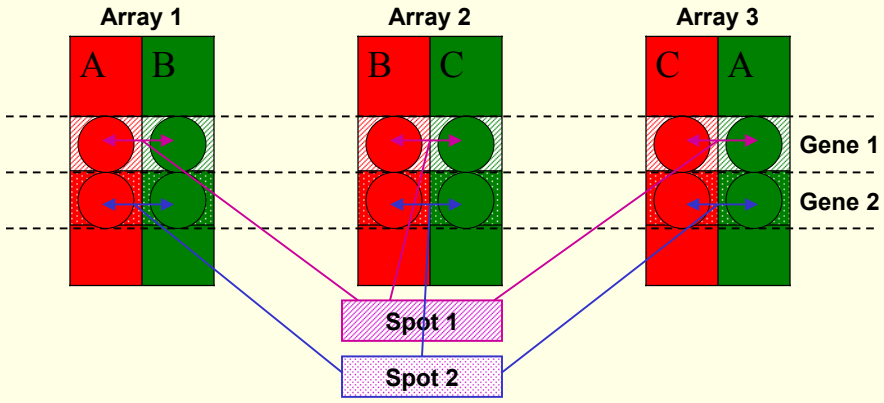
Adding an additional factor to the BIBD

A, B, & C: 3 different fertilizers



IN THE CONTEXT OF A MICROARRAY EXPERIMENT

A, B, & C: 3 different treatments (experimental conditions)

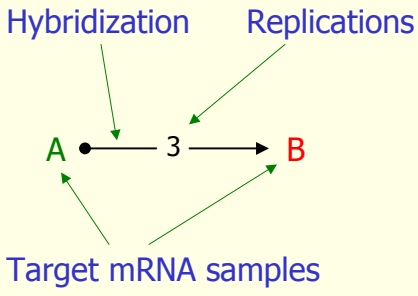


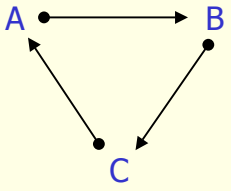
MORE ON REFERENCE DESIGN

- The reference should be such that
  - it is as *uniform* as possible (from same labeling reaction)
  - it *may not* need to be a biologically relevant sample
  - most genes are *expressed* (but not to the point of saturation) to avoid measurement error associated with low intensity transcripts
  - it is *labeled twice* (1 red-labeled, 1 green-labeled) to prevent potential dye biases. Controversial…

Actually preferred design by some statisticians (Dobbin and Simon, 2002); but depends on objectives and access to $$$

GRAPHICAL REPRESENTATION





⇨ Comparisons:

① Direct (more precise)

A ⟶ B $\log_2(A/B)$

② Indirect (less precise)

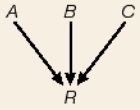
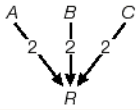
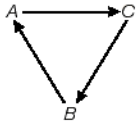
A ⟵ C ⟵ B

$\log_2(A/C) - \log_2(B/C)$

Cy3 (Green)

Cy5 (Red)

Table 1 | **Single-factor experiments**

Design choices	Number of slides	Units of material (number of samples)	Average variance
Indirect designs			
Design I 	3	$A = B = C = 1$	2.00
Design II 	6	$A = B = C = 2$	1.00
Direct design			
Design III 	3	$A = B = C = 2$	0.67

(Yang and Speed, 2002)

Variance of estimated effects for three different designs of single-factor experiments. σ^2 was set to 1 throughout.

Using PROC MIXED to compute s.e.

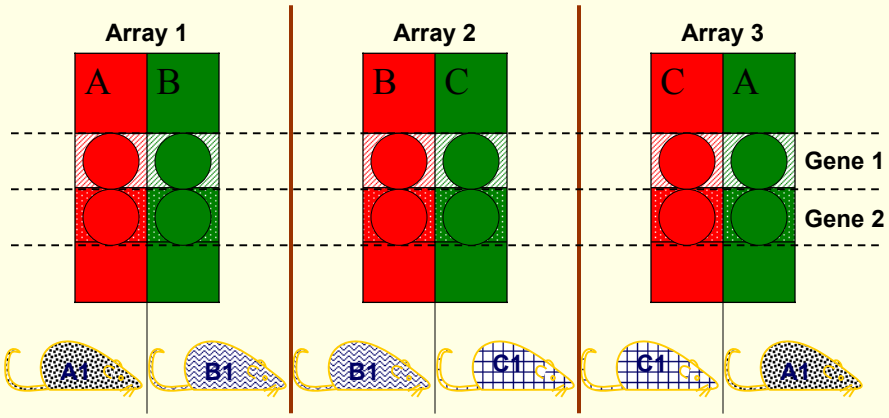
For Design I on previous slide

```
data new;
  input array trt $;
  y = 0;
  cards;
  1 A
  1 R
  2 B
  2 R
  3 C
  3 R
;

proc mixed noprofile;
  class array trt;
  model y = array trt;
  lsmeans trt /diff;
  parms (1) /hold = 1 noiter;
run;
```

NO REPLICATION!

Experimental units in the context of a microarray experiment



One mouse (or pooled sample) per treatment

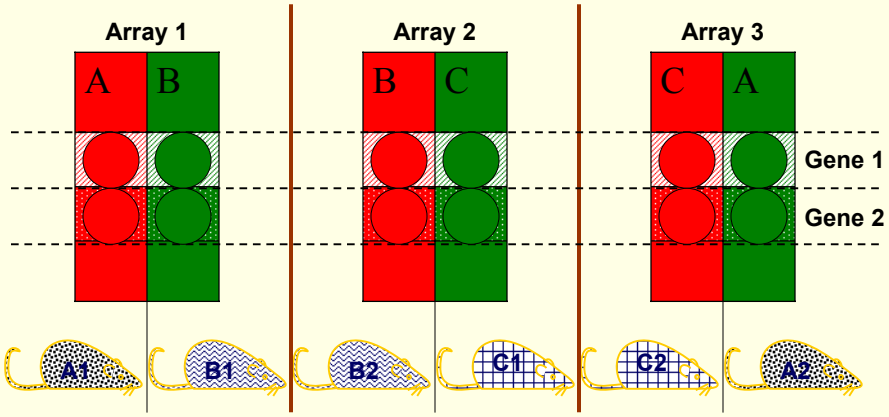
BIOLOGICAL REPLICATION

"One characteristic common to all biological material is that it varies." (Finney, 1953)

- "For comparing classes, replication of samples should generally be at the "subject" level because we want to make inference to the population of "subjects", not to the population of sub-samples of a single biological specimen" (Richard Simon, NCI)
- "Often the variation between individuals will be much larger than the other sources of variation and it will be inefficient to perform replicate arrays using specimens from a small number of individuals rather than performing single arrays from a larger number of individuals." (Richard Simon, NCI)

LITTLE REPLICATION!

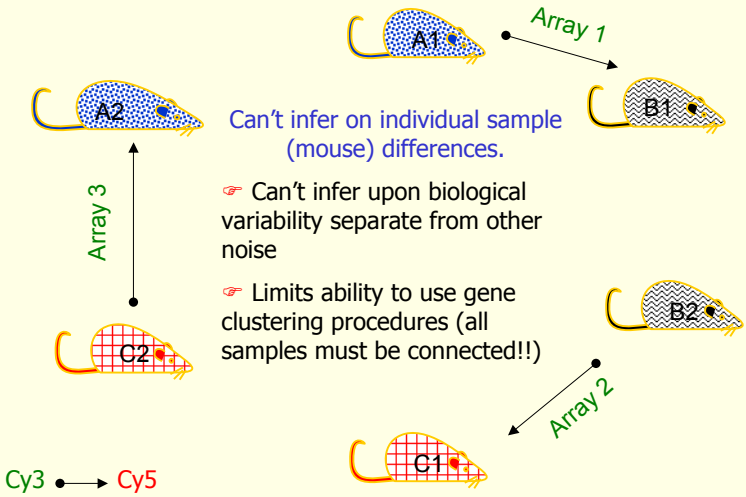
Experimental units in the context of a microarray experiment



Above…2 mice per treatment

…at least duplicate with 6 (or 12 or 24, etc.) more mice (unless large pools)

Minimally replicated block design for 3 treatments



Can't infer on individual sample (mouse) differences.

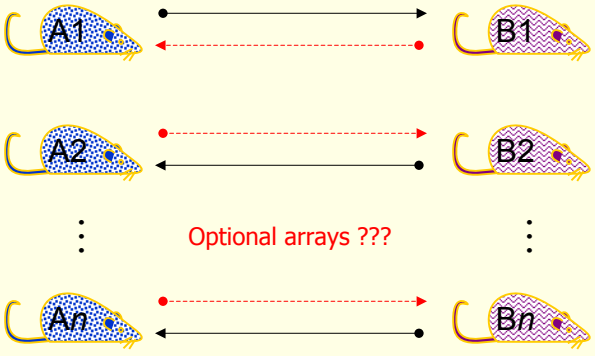
- Can't infer upon biological variability separate from other noise
- Limits ability to use gene clustering procedures (all samples must be connected!!)

(Simon et al., 2002)

2*n* mice and 2*n* arrays

Option 1: Block design

(Reverse fluor design replicated *n* times)

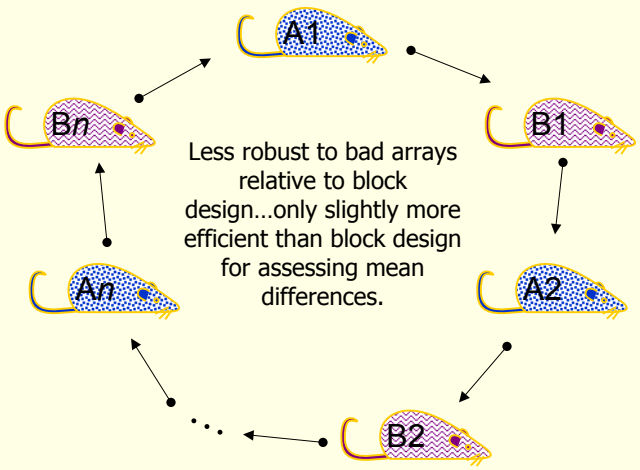


Cy3 ⟶ Cy5

(Kerr, 2002; ENAR-IBS spring meetings)

2*n* mice and 2*n* arrays

Option 2: Alternating loop design



Less robust to bad arrays relative to block design…only slightly more efficient than block design for assessing mean differences.

You have more treatments????

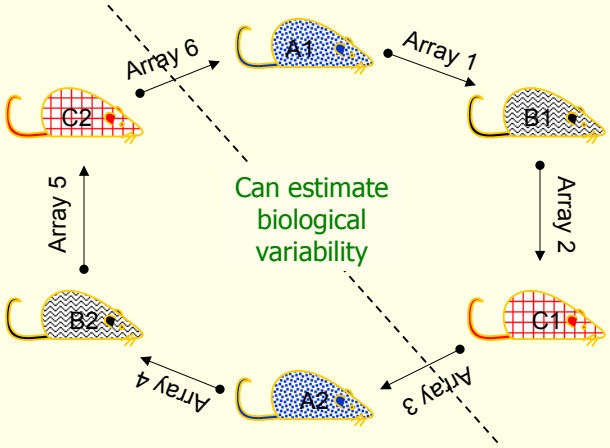
Consider alterations on Kerr and Churchill designs

(but see later)

Cy3 ⟶ Cy5

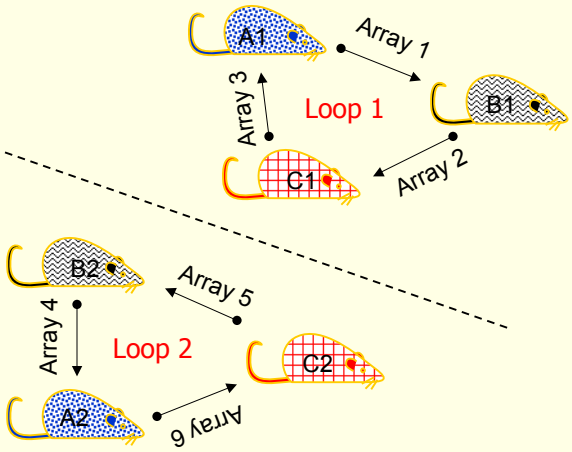
(Kerr, 2002)

Alternating Loop for 3 treatments



(deviation on Kerr's for 2 trts)

Block/loop design for 3 treatments



Might need more than 2 loops!!!!

Can estimate biological variability

Very slight disadvantage in power compared to alternating loop

Depends on:

1) Spot noise

2) Biological noise

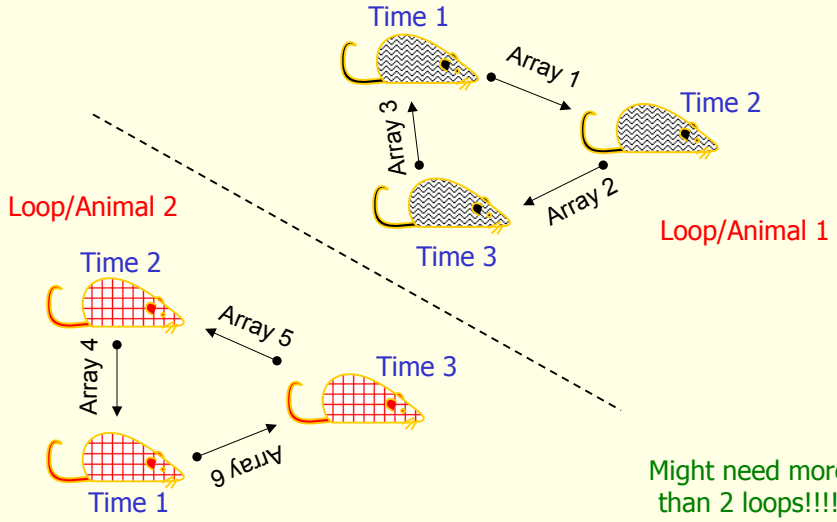
3) Residual variability

DYE SWAP

- Systematic differences in green and red intensities occur.
  - Can be intensity dependent!!!
  - Unlikely that "normalization" can be done equally well for every spot on every slide.
- Direct comparisons of replicates of slides with same labeling should be avoided because unadjusted color bias might persist and accumulate.

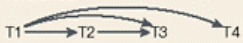
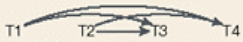
A Block Design for Time Course Study

(if mRNA sample retrieval doesn't involve destruction of animal)



Might need more than 2 loops!!!!

Table 2 | **Time-course experiments**

Design choices	Comparisons: *t* versus *t* + 1 t_1/t_2	t_2/t_3	t_3/t_4	*t* versus *t* + 2 t_1/t_3	t_2/t_4	*t* versus *t* + 3 t_1/t_4	Average variance
Design I — T1 as common reference 	1.00	2.00	2.00	1.00	2.00	1.00	1.5
Design II — direct: sequential 	1.00	1.00	1.00	2.00	2.00	3.00	1.67
Design III — common reference 	2.00	2.00	2.00	2.00	2.00	2.00	2.00
Design IV — T1 as common reference 	0.67	0.67	1.67	0.67	1.67	1.00	1.06
Design V — direct: loop 	0.75	0.75	0.75	1.00	1.00	0.75	0.83
Design VI — direct: mixed 	1.00	0.75	1.00	0.75	0.75	0.75	0.83

Variance of estimated effects for six different designs of time-course experiments. Designs I and II involve only three slides and the remaining designs involve four. σ^2 was set to 1 throughout.

(Yang and Speed, 2002)

How about factorial experiments?

Example 2 x 2 factorial

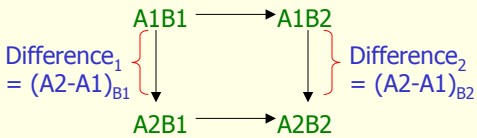
Yang and Speed (2002)

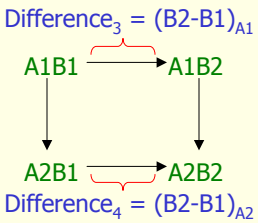
Two factors A and B

Two levels per factor • A1, A2 (e.g. diseased vs. not diseased)

• B1, B2 (e.g. Drug A vs. Drug B)

Four different "treatment" combinations: • A1B1, A1B2, A2B1, A2B2



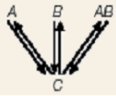
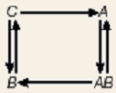
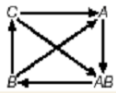


Main effect of A = 0.5 (Difference$_1$ + Difference$_2$)

Main effect of B = 0.5 (Difference$_3$ + Difference$_3$)

Interaction = Difference$_1$ – Difference$_2$ = Difference$_3$ – Difference$_4$

Table 3 | **2 × 2 factorial experiments**

Design choices	Main effect *A*	Main effect *B*	Interaction *A.B*
Indirect design			
Design I 	0.50	0.50	1.50
A balance of direct and indirect design			
Design II 	0.67	0.43	0.67
Design III 	0.50	0.50	1.00
Design IV 	N/A	0.30	0.67

Variance of estimated effects for four different designs of 2 × 2 factorial experiments. σ^2 was set to 1 throughout.

C = A1B1

A = A2B1

B = A2B2

AB = A2B2

Which comparisons would be of greatest interest to you??

(Yang and Speed, 2002)

Choice of design

- Up to you to decide which contrasts are of greatest interest.
- Combination of direct and indirect comparisons is often the best practical solution to a design problem.

(Yang and Speed, 2002)

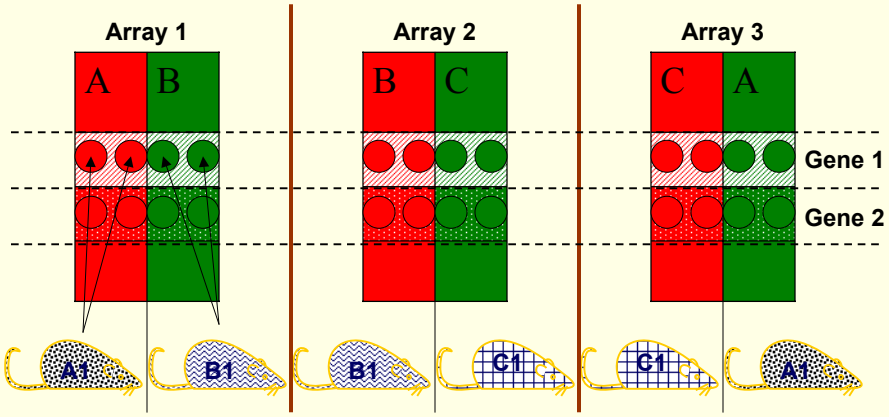
The case for multiple aliquots from the same specimen

- Quality control
- Reproducibility and sampling variability (not biological variability!) assessment.
- Within animal: Multiple specimens versus multiple independent labelings (aliquots) of one RNA sample??
  - Different phenomena when tissue is inhomogeneous

(Simon et al., 2002)

NO REPLICATION!

How about multiple (say, duplicate) spots per gene??



One mice per treatment…

Another analogy

- Treatment (Strain) A vs. B; one mouse per strain
  - Suppose you weigh each mouse 3 times

20.1
20.0
20.1

22.1
22.2
22.2

Would a *t*-test comparing the means of the three measures on each mouse be a valid test of the Treatment A vs Treatment B difference??

NO!!!!!

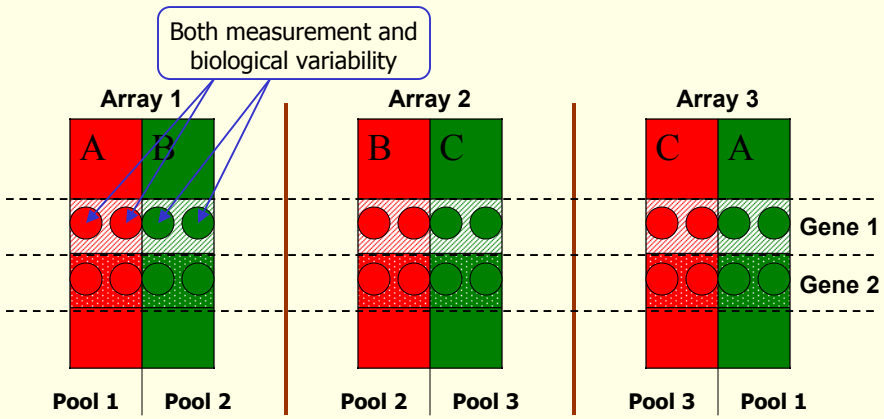
DIFFERENT TYPES OF "REPLICATION"

- Assessing measurement error
  - Two levels
    - Multiple spots per gene on an array
      - Spread out around the array if possible
    - Multiple arrays to study the same sample (e.g. one subject assigned to Treatment A is arrayed twice)
  - USEFUL FOR QUALITY CONTROL ASSESSMENT BUT NOT TOO USEFUL FOR INFERENCE ON TREATMENT EFFECTS!
- Accounting for biological variability
  - Multiple subjects-arrays per treatment
  - TRUE REPLICATION
  - Random assignment of individuals to treatment groups

Useful to have spots that are well spaced apart and not adjacent to give better reflection of variability across slide

REPLICATION??

How about pooling samples?
(mRNA sample from 2 or more animals)



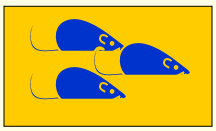
"Pool" (handling and some biological) effects are still confounded with treatment!!

At the very least duplicate the above with 3 other pools and reversing the fluors.

Another analogy

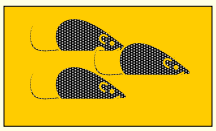
- Pens of animals, 1 pen per treatment

Treatment A



Pen is adjacent to entrance of barn

Treatment B



Pen is at the back of the barn

Record pen feed intake 3 times

Do you have replication??? NO!!!

PEN IS THE EXPERIMENT AL UNIT FOR TREATMENT

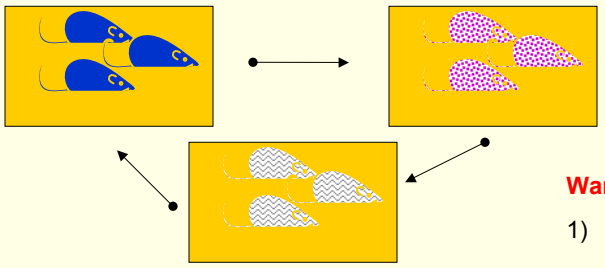
Comments by Churchill and Oliver (2001)

"Ideally, one would assay several independent pools but if this is prohibitive, then one may consider using *multiple assays* of a single pool."

however…

"One must assume that the variability between different pools of animals would be negligible. Whereas this seems reasonable when many hundreds of flies can be pooled, it is less obvious that pools will be homogeneous when samples from smaller numbers of animals are combined."

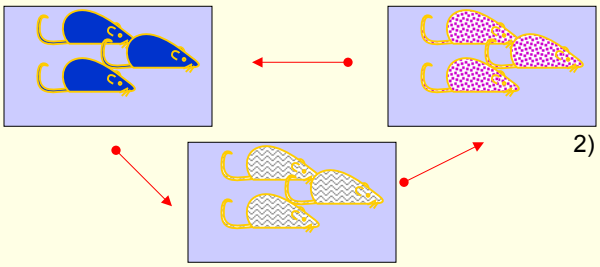
Barebones pooled sample block design for three treatments



6 pools
(pens)

Warning!

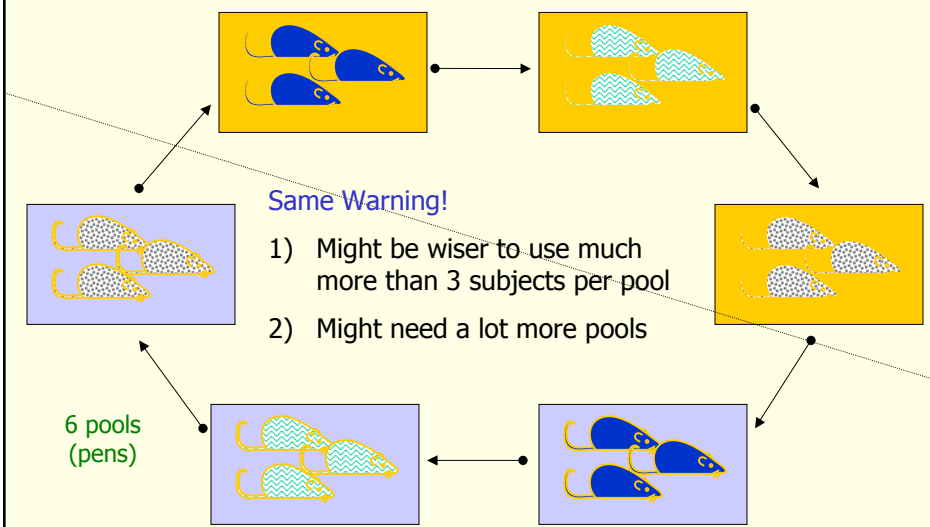
1) Might be wiser to use much more than 3 subjects per pool



Reverse fluor setup

2) Might need a lot more pools to get sufficient power

Barebones alternating loop pooled sample design for three treatments



Same Warning!

1) Might be wiser to use much more than 3 subjects per pool

2) Might need a lot more pools

6 pools
(pens)

Issues surrounding pooled samples

- Advantages:
  - Reduces variability of response; potentially substantially lowering standard errors of mean (ratio) inferences.
  - $$$ (depending on cost of animal relative to cost of array)
    - Cost/Std error tradeoff (Kendziorski et al., 2002)
  - When there is not enough RNA sample from individual subjects
    - Amplification protocols may distort expression profiles
- Potential disadvantage:
  - Inability to assess biological variability in response (?)
  - Masked effect of potentially outlying subjects
- Generally best suited to populations with highly homogeneous experimental units.
  - e.g. inbred lines of mice
  - The less homogeneous a population, the greater the number of subjects that should be represented in a pool and/or the greater the number of pools!!!!
    - Power formulas provided by Kendziorski et al. (2002) for Affymetrix designs.

How many replicates???

- Answer may be not pretty if you're genuinely concerned about biological variability.
- Required sample size (Reference design) from Simon et al. (2002) to compare two treatments:

α: Type I error rate (recommend α = 1/N) such that expected number of false positives is 1

β: proportion of differentially expressed genes not detected

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

δ: expected difference in $\log_2$ intensities

($\delta = 1 \rightarrow$ Fold difference = 2)

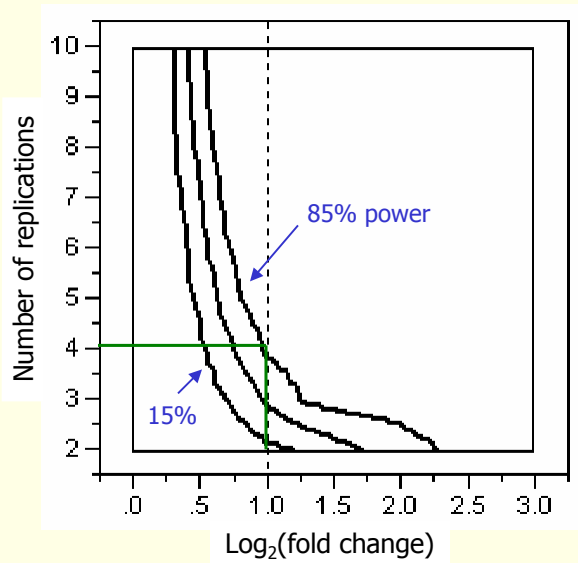
σ: based on degree of variability of expression among similar biological tissue samples (0.25-1)

Example calculation

- Reference design for two treatment comparison
- To detect statistical significance for ratio = 2 ($\delta = 1$) and "typical" gene expression biological variability, would require 28 tissue samples (each sample from a different subject) per treatment (based on $\alpha = .001$, $\beta = 0.05$, $\delta = 1$, $\sigma = 0.75$)
  - Two treatments = 56 arrays!!

(Simon et al., 2002)

Power assessment in a mixed model analysis for a loop design



LOOP DESIGN ON 12 TREATMENTS

False positive rate of 1/20000

Spot variance = 1

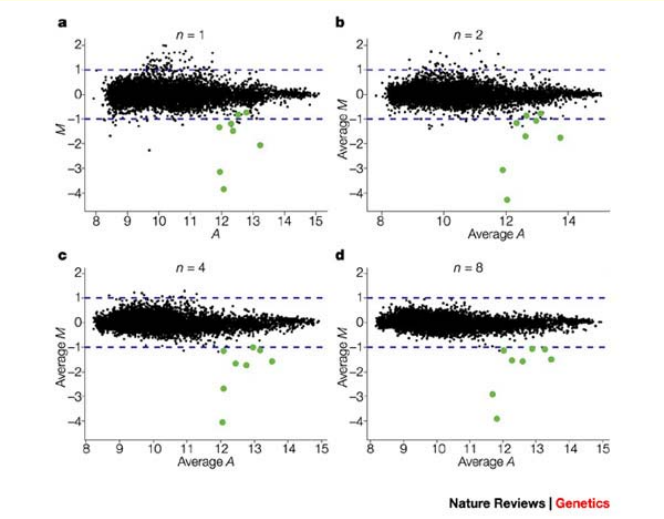
Residual variance = 0.25

4 reps = 2 loops

(24 arrays)

(Wolfinger et al., 2001)

Replication reduces variability!!

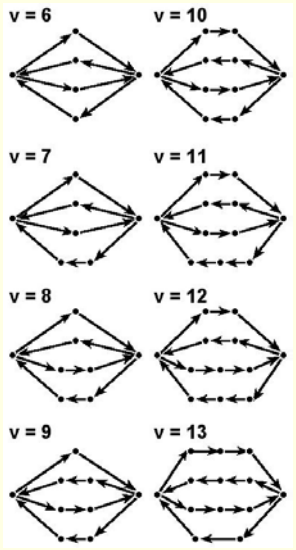


(Yang and Speed, 2002)

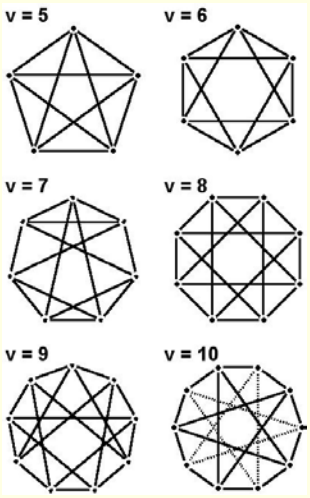
A-optimal designs

(Kerr and Churchill, 2001)

$n = v + 2$



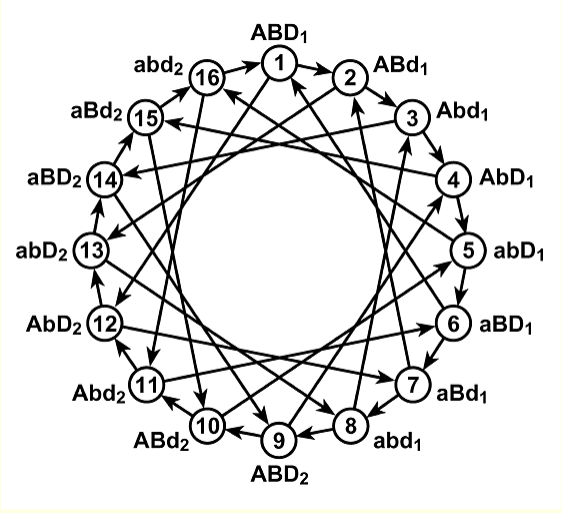
$n = 2v$



But don't forget biological replication!!!

A 2 x 2 x 2 *A-optimal* design allowing for biological replication

2 ages (A and a) x 2 sexes (B and b) x 2 strains (D and d) = 8 groups



- 2 biological / group
- 16 samples (pools)
- 32 arrays

Power considerations may dictate need of another 16 samples with reverse fluor.

(Churchill and Oliver, 2001)

"Biologists interested in gene expression profiling should feel free to match experimental design to their particular situation; there is no universal microarray design.

A careful grounding in the principles of experimental design will help to ensure that we will accumulate knowledge and not just enormous amounts of data".

Churchill and Oliver (2001)

References

Churchill, G.A., and B. Oliver. 2001. Sex, flies, and microarrays. Nature Genetics, 29, 355-356.

Gary Churchill's web site (www.jax.org/research/churchill)

Dobbin, K. and R. Simon. 2002. Comparison of Microarray Designs for Class Comparison and Class Discovery. National Cancer Institute Technical Report 006.

Geschwind, Nature Reviews – Neuroscience 2001

Kendziorski, C.M., H. Lan, and A.D. Attie. 2002. The efficiency of pooling mRNA in microarray experiments. Technical Report # 168, Department of Biostatistics, University of Wisconsin-Madison (http://www.biostat.wisc.edu)

Simon, R., M.D. Radmacher, and K. Dobbin. 2002. Design of studies using DNA microarrays. Genetic Epidemiology 23:21-36

Wolfinger, R.D., G. Gibson, E. D. Wolfinger, L. Bennett, H. Hamadeh, P. Bushel, C. Afshari, and R.S. Paules. 2001. Assessing gene significance from cDNA microarray expression data via mixed models. Journal of Computational Biology, 8: 625-637.

Yang, Y.H. and T. Speed. 2002. Design issues for cDNA microarray experiments. Nature Reviews Genetics 3:579-587.