Factorial and time course designs for cDNA microarray experiments

Patty Solomon

Department of Applied Mathematics and
Centre for the Molecular Genetics of Development
University of Adelaide

11 July 2002

http://www.maths.adelaide.edu.au/MAG

patty.solomon@adelaide.edu.au

Microarray Analysis Group

- University of Adelaide
  - Patty Solomon
  - Gary Glonek*
  - Jonas Lloyd (Ph.D.)
  - Michael Calvert (Grad. Dip.)
  - Adam Kister (Grad. Dip.)
- Hanson Institute/IMVS
  - Anna Tsykin

Microarray Analysis Group

- University of Adelaide
  - Patty Solomon
  - Gary Glonek*
  - Jonas Lloyd (Ph.D.)
  - Michael Calvert (Grad. Dip.)
  - Adam Kister (Grad. Dip.)
- Hanson Institute/IMVS
  - Anna Tsykin

- *Acknowledgement:* Terry Speed, UC Berkeley and WEHI.

What are microarrays?

- *Powerful new tools for surveying the expression levels of many thousands of genes simultaneously.*

What are microarrays?

- *Powerful new tools for surveying the expression levels of many thousands of genes simultaneously.*
- There are many different technologies:
  - ⋆ High-density nylon membrane arrays.
  - ⋆ Short oligonucleotide arrays (Affymetrix).
  - ⋆ *Spotted cDNA arrays: Brown & Botstein (1999).
  - ⋆ *Spotted long oligonucleotide arrays.
  - ⋆ ...

What are microarrays?

- *Powerful new tools for surveying the expression levels of many thousands of genes simultaneously.*
- There are many different technologies:
  - ⋆ High-density nylon membrane arrays.
  - ⋆ Short oligonucleotide arrays (Affymetrix).
  - ⋆ *Spotted cDNA arrays: Brown & Botstein (1999).
  - ⋆ *Spotted long oligonucleotide arrays.
  - ⋆ ...
- *There are common themes to all these technologies.*

What are microarrays used for?

- Mutational analysis, genetic mapping studies, molecular profiling of disease states, analysing cellular responses to biological stimuli, . . .

What are microarrays used for?

- Mutational analysis, genetic mapping studies, molecular profiling of disease states, analysing cellular responses to biological stimuli, ...
- *Simplest experiments* seek to identify changes in gene expression between
  - ◇ different tissue types e.g. normal vs tumour
  - ◇ different drugs e.g. treatment vs control
  - ◇ different locations within an organ: spatial effects.

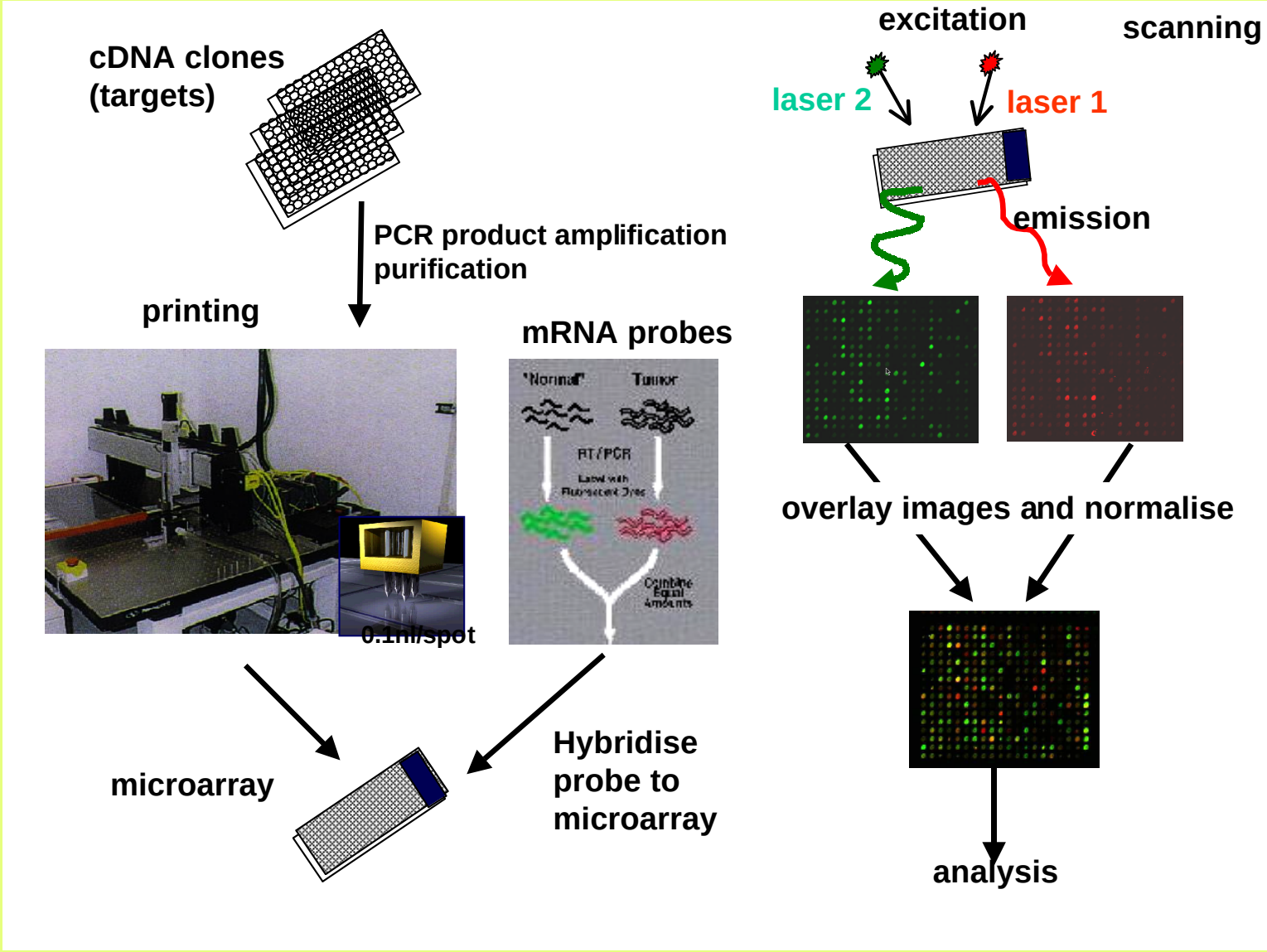
What are microarrays used for?

- Mutational analysis, genetic mapping studies, molecular profiling of disease states, analysing cellular responses to biological stimuli, ...
- *Simplest experiments* seek to identify changes in gene expression between
  - ◇ different tissue types e.g. normal vs tumour
  - ◇ different drugs e.g. treatment vs control
  - ◇ different locations within an organ: spatial effects.
- *More complex experiments* seek to identify *patterns in groups of genes* or monitor *expression profiles* over time.

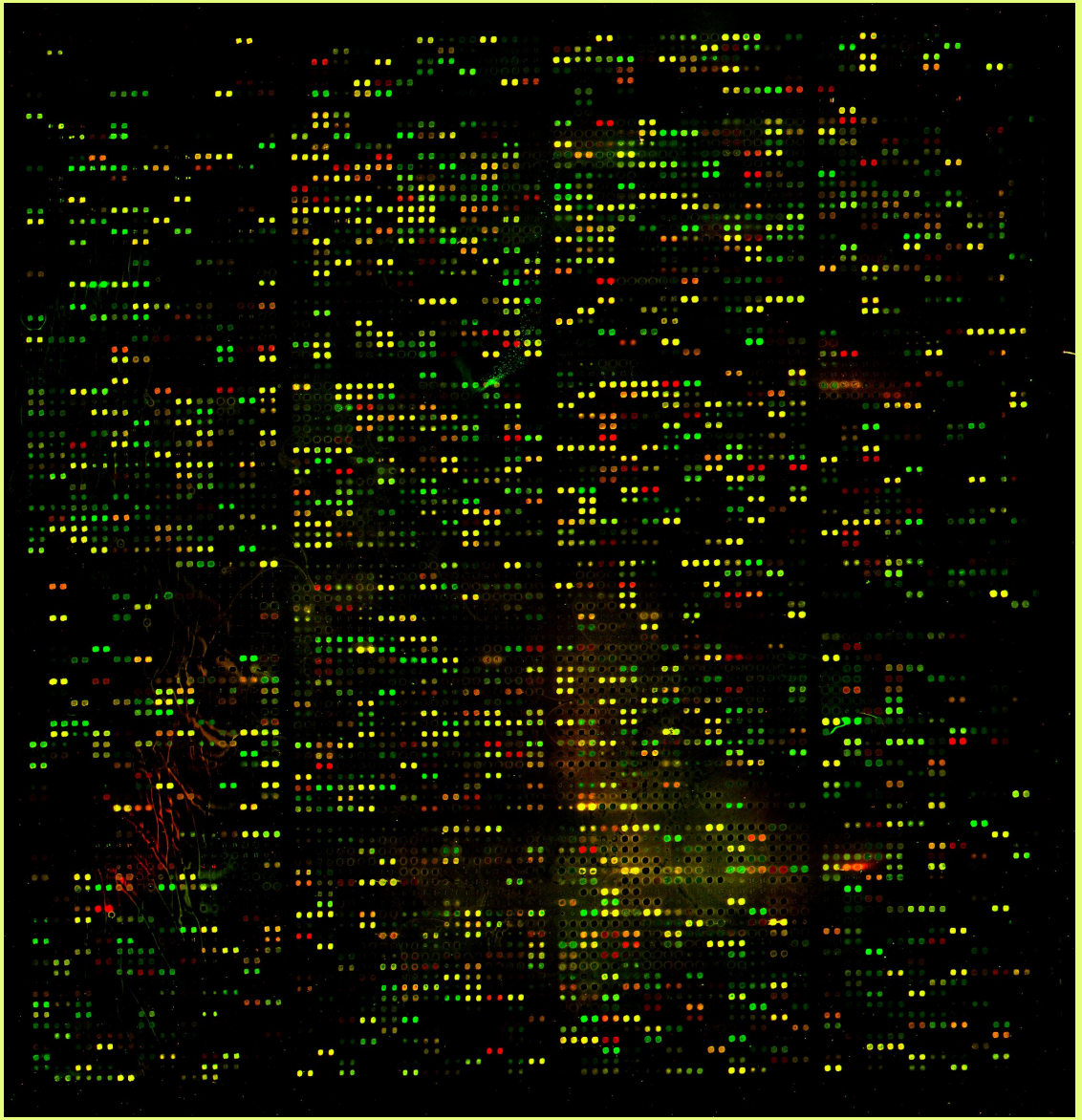
What are microarrays used for?

- Mutational analysis, genetic mapping studies, molecular profiling of disease states, analysing cellular responses to biological stimuli, . . .
- *Simplest experiments* seek to identify changes in gene expression between
  - ◇ different tissue types e.g. normal vs tumour
  - ◇ different drugs e.g. treatment vs control
  - ◇ different locations within an organ: spatial effects.
- *More complex experiments* seek to identify *patterns in groups of genes* or monitor *expression profiles* over time.
- A single experiment can involve 1 to 100's of slides.

cDNA microarray and scanner process



Human clones with MCF-7 and Jurkatt probes



Quantifying expression

For each spot on the array, calculate the background-adjusted intensities:

red intensity $R \rightarrow R - Rb$

green intensity $G \rightarrow G - Gb$

and combine them in the log base 2 ratio:

$$M = \log_2(R/G) = \log_2 R - \log_2 G$$

Quantifying expression

For each spot on the array, calculate the background-adjusted intensities:

red intensity $R \rightarrow R - Rb$

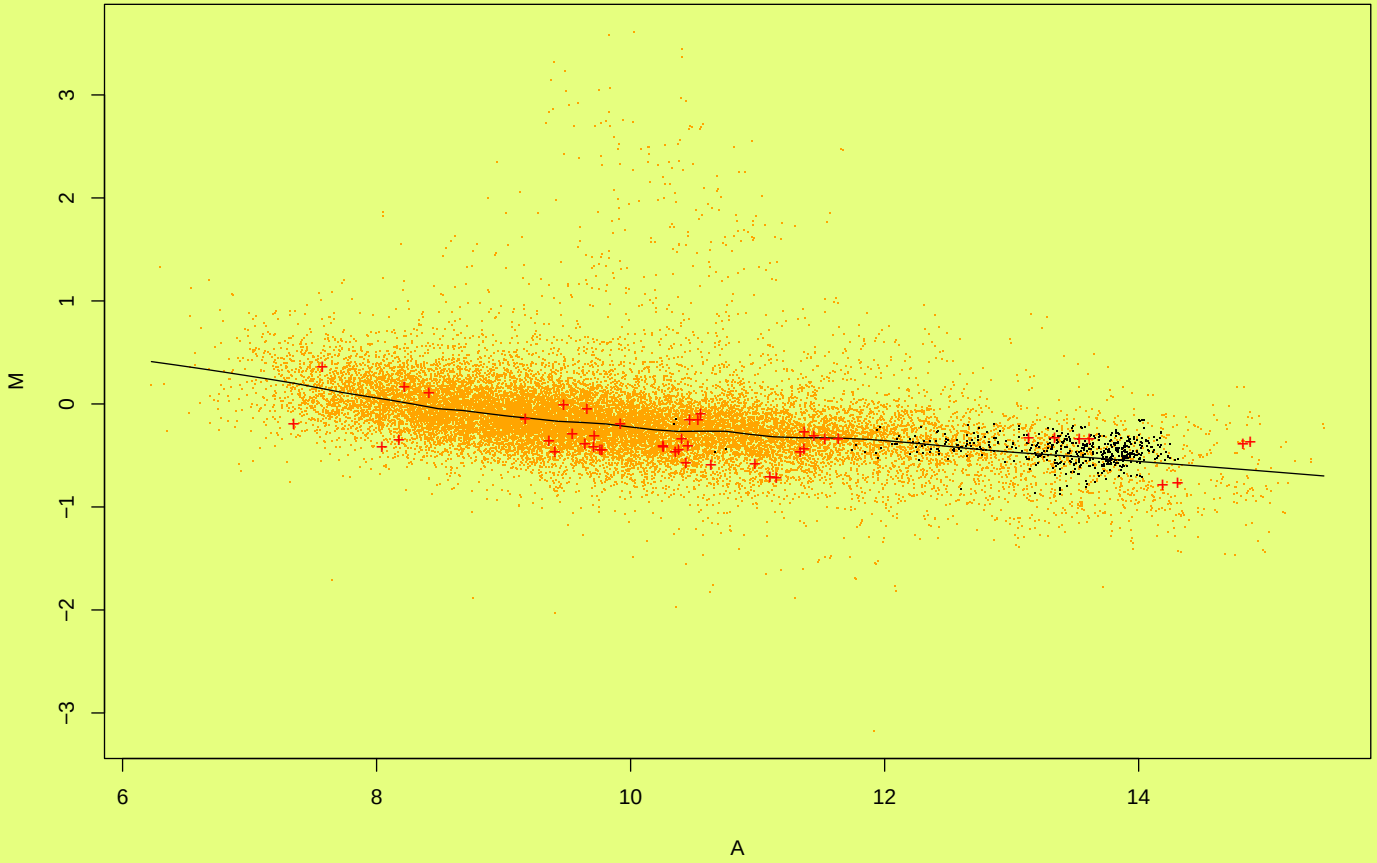
green intensity $G \rightarrow G - Gb$

and combine them in the log base 2 ratio:

$$M = \log_2(R/G) = \log_2 R - \log_2 G$$

- *Background is a huge issue.*

Angiogenesis: time 0 versus .5 hour



Why is statistics important?

- *The interface between statistics, biology, computer science and medicine has gone from information-poor to information-mega-rich*.
  - ◇ Statistics has a central role to play in processing that information and making it intelligible.

Why is statistics important?

- *The interface between statistics, biology, computer science and medicine has gone from information-poor to information-mega-rich*.
  - ◇ Statistics has a central role to play in processing that information and making it intelligible.
- Statistics is an *enabling* discipline.
  - ★ Statisticians have by training the skills of synthesis, empirical investigation, modelling and interpretation which are crucial to application areas such as microarrays.

Why is statistics important?

- *The interface between statistics, biology, computer science and medicine has gone from information-poor to information-mega-rich*.
  - ◇ Statistics has a central role to play in processing that information and making it intelligible.
- Statistics is an *enabling* discipline.
  - ⋆ Statisticians have by training the skills of synthesis, empirical investigation, modelling and interpretation which are crucial to application areas such as microarrays.
- *Challenges*: design, analysis and interpretation of data from microarray experiments.

What do we mean by ‘design’ for microarrays?

- Which mRNA samples should be competitively hybridised together on the same slide?*

What do we mean by ‘design’ for microarrays?

- Which mRNA samples should be competitively hybridised together on the same slide?*
  - Should samples from individual animals or people be compared *directly* or via a common *reference* mRNA sample?*

What do we mean by 'design' for microarrays?

- Which mRNA samples should be competitively hybridised together on the same slide?*
  - Should samples from individual animals or people be compared *directly* or via a common *reference* mRNA sample?*
- Should tissue samples from animals be *pooled* then compared, or should different animals be hybridised to different slides?
  - If pooled samples are to be compared, what is the optimal number of pools?

- Which sample should be labelled with one dye and which with the other?
  - Should dye-swapped replicates be made on different amplifications or the same?

- Which sample should be labelled with one dye and which with the other?
  - Should dye-swapped replicates be made on different amplifications or the same?
- How many replicates should there be of each gene within an array?
- How many times should each array be replicated?*
- ...

- Which sample should be labelled with one dye and which with the other?
  - Should dye-swapped replicates be made on different amplifications or the same?
- How many replicates should there be of each gene within an array?
- How many times should each array be replicated?*
- ...

  - Kerr & Churchill 2001
  - Speed & Yang 2002, Yang 2002
  - Jin et al. 2001, Wolfinger et al. 2001, Pan et al. 2002

    ...

There are practical constraints

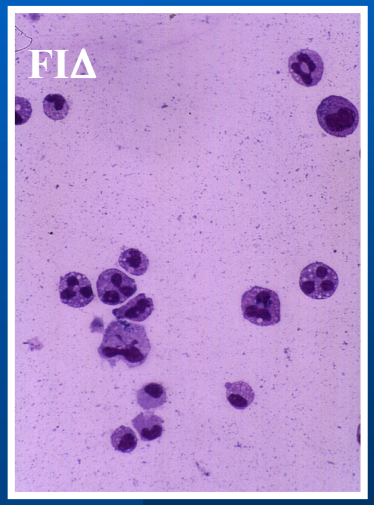
- Limitations on the available mRNA.
- Cost: the number of slides available for experimentation is often fixed in advance.
  - *There is typically not enough information available for standard sample size calculations.*
- Risk of failed hybridisations.

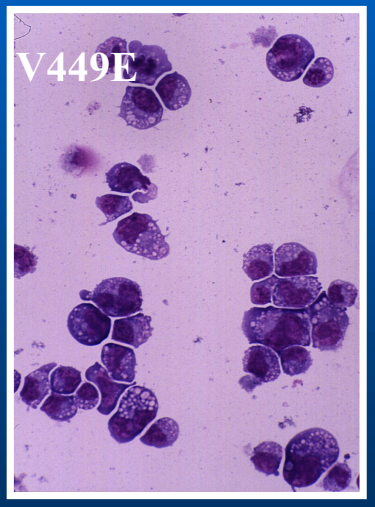
There are practical constraints

- Limitations on the available mRNA.
- Cost: the number of slides available for experimentation is often fixed in advance.
  - *There is typically not enough information available for standard sample size calculations.*
- Risk of failed hybridisations.
- *We contend that the most appropriate way to find differentially expressed genes is to prescribe a design subject to*
  - the key contrasts and parameters of interest
  - and the practical constraints of the problem.

Case Study 1

Biological function of the activated mutants in FDB cells





⇒ Different signals generated by the two mutants

- *Aim: to identify genes that play an important role in receptor signalling and leukaemogenesis in mutant mice.*

A 2x2 factorial design

- It is anticipated that measuring changes over 24 hours will distinguish genes involved in promoting or blocking differentiation, or that suppress or enhance growth, as genes potentially involved in leukaemia.
  - ◇ Compare two cell populations: FIΔ and V449E at two times 0 hours and 24 hours
    - ⋆ $\longrightarrow$ 2×2 factorial design of block size 2.

A 2x2 factorial design

- It is anticipated that measuring changes over 24 hours will distinguish genes involved in promoting or blocking differentiation, or that suppress or enhance growth, as genes potentially involved in leukaemia.
  - ◇ Compare two cell populations: FIΔ and V449E at two times 0 hours and 24 hours
    - ⋆ $\longrightarrow$ 2×2 factorial design of block size 2.
- *Interaction of primary interest:*
  e.g. genes that are differentially expressed in the two samples at time 24 hours, but not at time 0 hours.

A 2x2 factorial design

- It is anticipated that measuring changes over 24 hours will distinguish genes involved in promoting or blocking differentiation, or that suppress or enhance growth, as genes potentially involved in leukaemia.
  - ◇ Compare two cell populations: FIΔ and V449E at two times 0 hours and 24 hours
    - ⋆ ⟶ 2×2 factorial design of block size 2.
- *Interaction of primary interest:*
  e.g. genes that are differentially expressed in the two samples at time 24 hours, but not at time 0 hours.
- *Sample effects of interest too.*

Notation

- Samples: FIΔ and V449E
- Times: 0 hours and 24 hours

$\longrightarrow$ 4 combinations $F0, V0, F24, V24$.

Notation

- Samples: FIΔ and V449E
- Times: 0 hours and 24 hours

$\longrightarrow$ 4 combinations $F0, V0, F24, V24$.

- Level of expression of each gene is described by four parameters:
  - μ: Overall mean (baseline)
  - α: Main effect for sample
  - β: Main effect for time
  - $\alpha\beta$: Interaction

- On the log scale, the interaction is the difference between the F, V differences at times $0, 24$:

$$(F24 - V24) - (F0 - V0)$$

- On the log scale, the interaction is the difference between the F, V differences at times $0, 24$:

$$(F24 - V24) - (F0 - V0)$$

- To get it, experimenters could do 2 hybridisations:

$F24$ versus $V24$ AND $F0$ versus $V0$

OR

$F24$ versus $F0$ AND $V24$ versus $V0$

- On the log scale, the interaction is the difference between the F, V differences at times $0, 24$:

$$(F24 - V24) - (F0 - V0)$$

- To get it, experimenters could do 2 hybridisations:

$F24$ versus $V24$ AND $F0$ versus $V0$

OR

$F24$ versus $F0$ AND $V24$ versus $V0$

- The latter is the 'usual' experiment.
- *If replication is possible, do both.*

- Similarly for the main effects:

$$\text{Sample}: \quad (F24 + F0) - (V24 + V0)$$

$F24$ versus $V24$ AND $F0$ versus $V0$

OR

$F24$ versus $V0$ AND $F0$ versus $V24$

- Similarly for the main effects:

$$\text{Sample}: \quad (F24 + F0) - (V24 + V0)$$

$F24$ versus $V24$ AND $F0$ versus $V0$

OR

$F24$ versus $V0$ AND $F0$ versus $V24$

$$\text{Time}: \quad (F24 + V24) - (F0 + V0)$$

$F24$ versus $F0$ AND $V24$ versus $V0$

OR

$F24$ versus $V0$ AND $V24$ versus $F0$

- Similarly for the main effects:

$$\text{Sample}: \quad (F24 + F0) - (V24 + V0)$$

$F24$ versus $V24$ AND $F0$ versus $V0$

OR

$F24$ versus $V0$ AND $F0$ versus $V24$

$$\text{Time}: \quad (F24 + V24) - (F0 + V0)$$

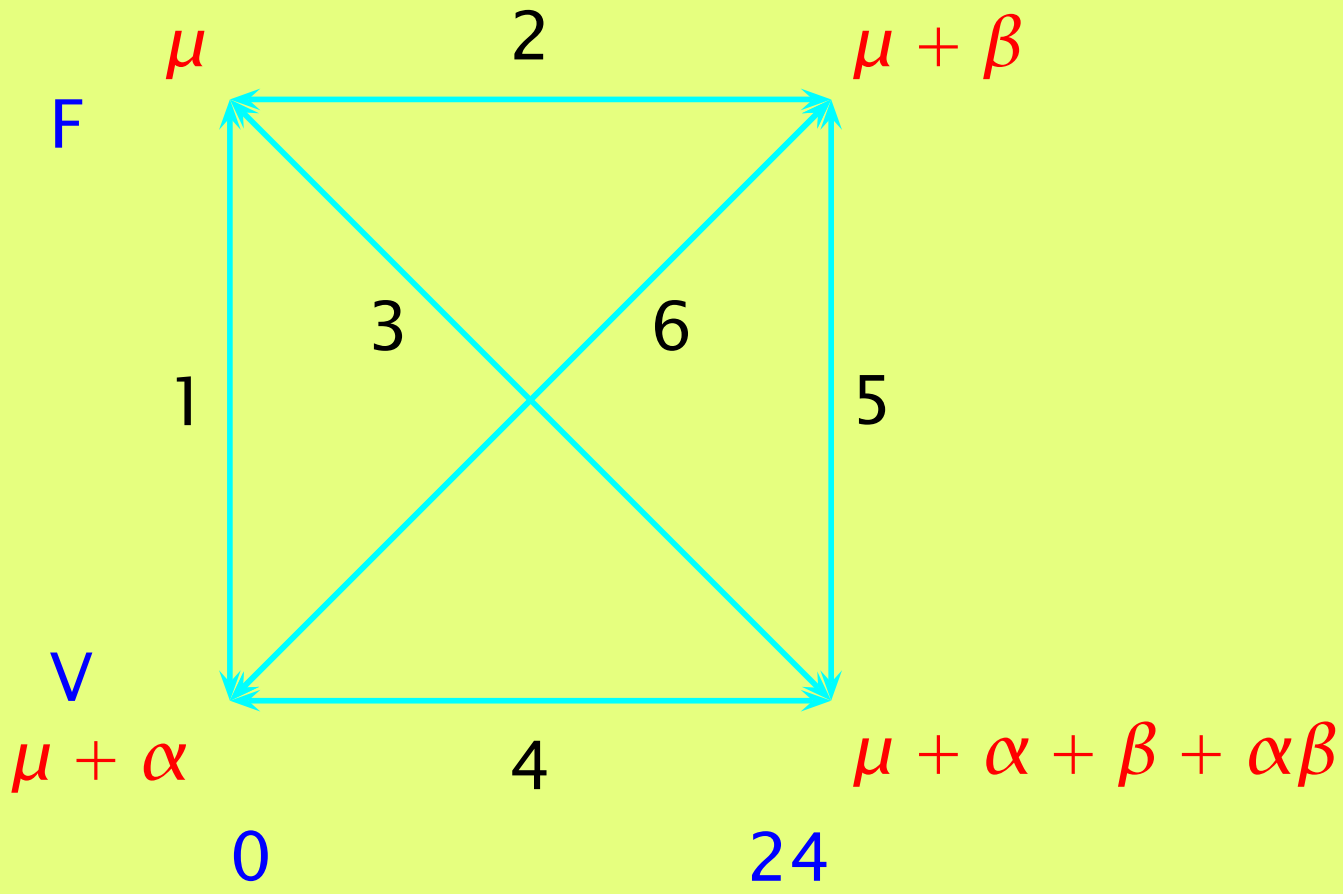
$F24$ versus $F0$ AND $V24$ versus $V0$

OR

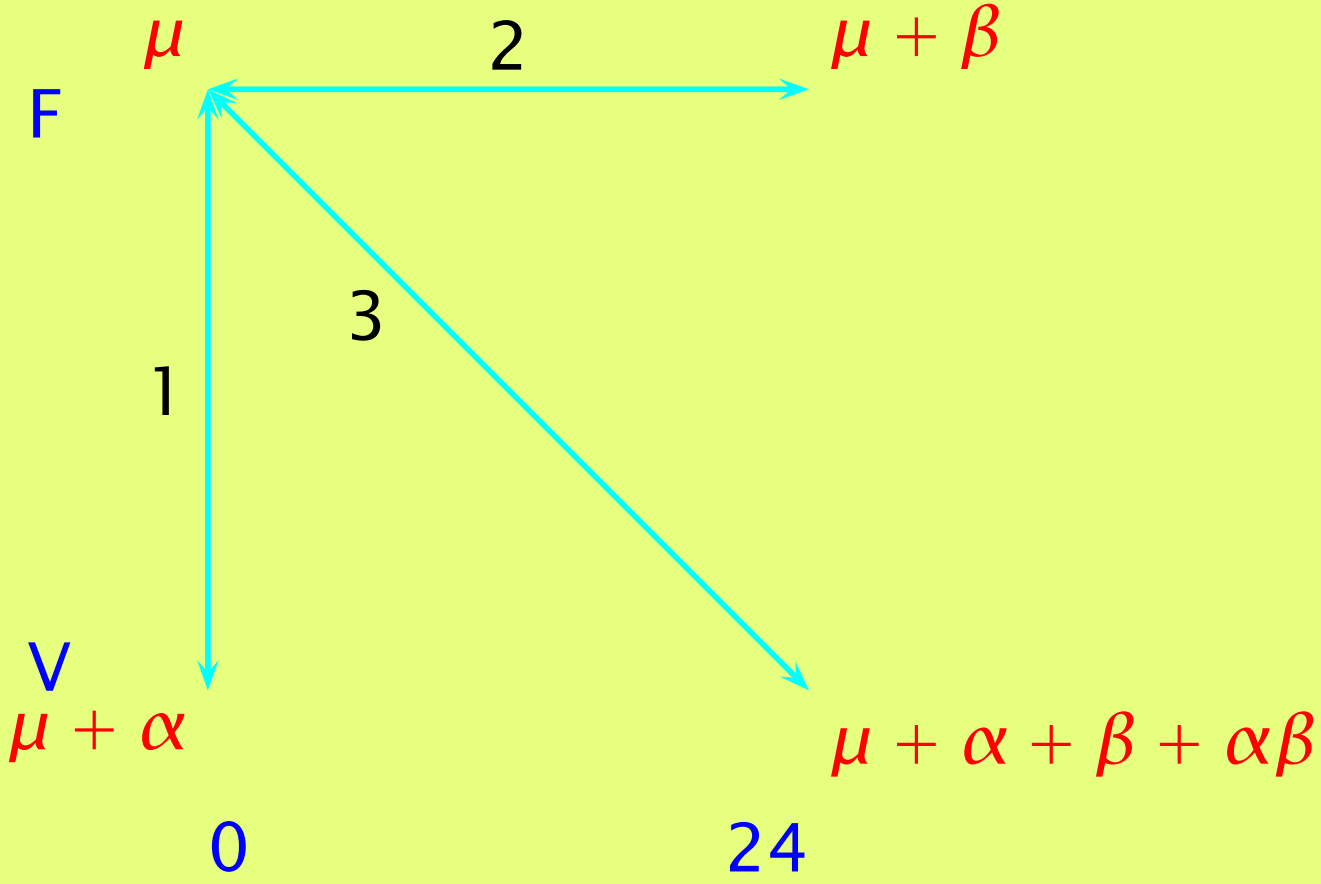
$F24$ versus $V0$ AND $V24$ versus $F0$

- *Again, if replicates are available, do both.*

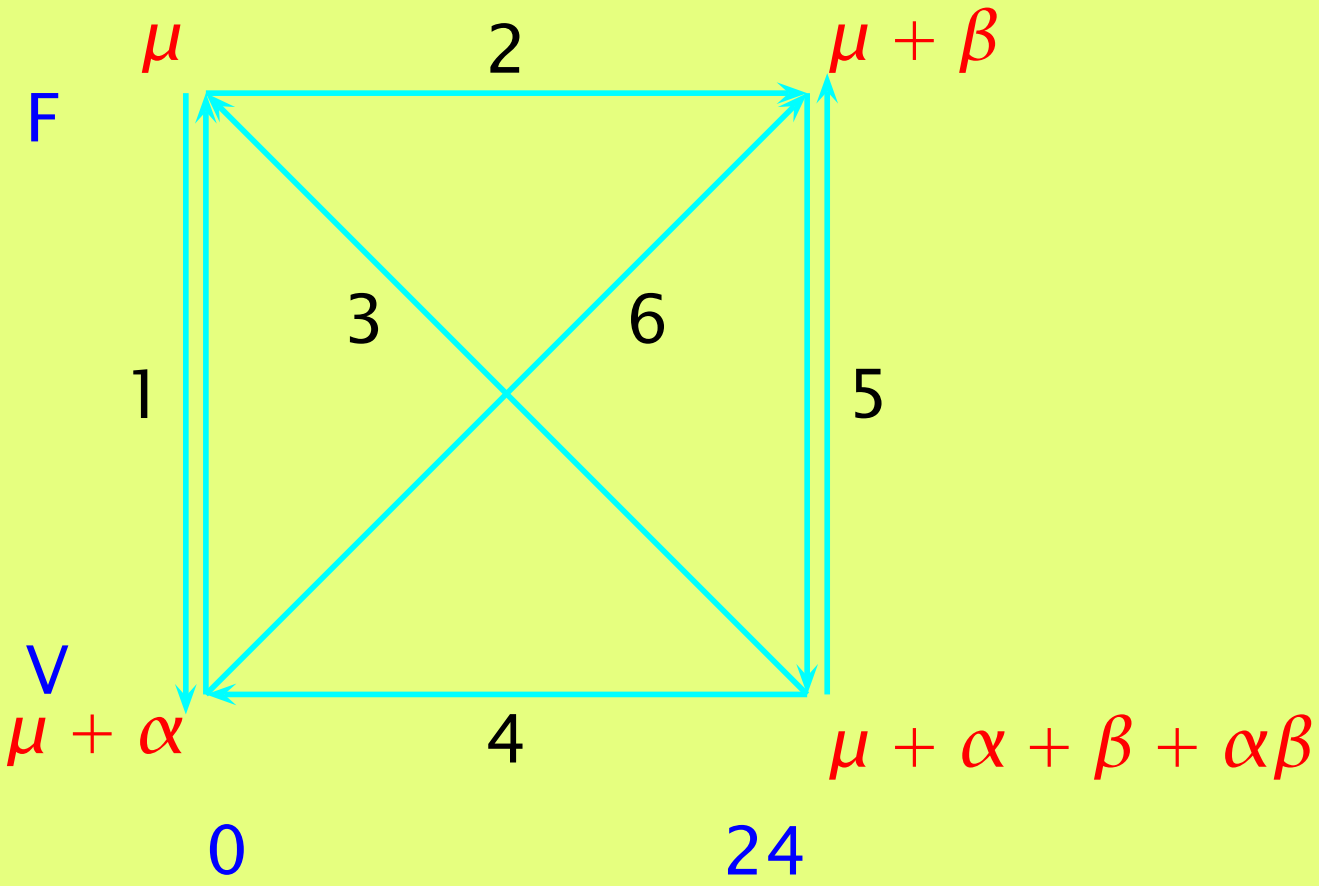
The ‘optimal’ design?



The reference design



Case study: 8 slide limit



- 6 pairwise comparisons with dye-swaps on cell line comparisons at times 0 and 24 hours.

Two snags

Two snags

- Biologists a bit non-plussed by the ‘cross-hybridisations’.
- It’s not quite optimal either.

Two snags

- Biologists a bit non-plussed by the 'cross-hybridisations'.
- It's not quite optimal either.
  - *Designs incorporating all possible multiple direct comparisons rarely, if ever, optimal by the criterion of statistical efficiency.*

Two snags

- Biologists a bit non-plussed by the 'cross-hybridisations'.
- It's not quite optimal either.
  - *Designs incorporating all possible multiple direct comparisons rarely, if ever, optimal by the criterion of statistical efficiency.*
- *Reference design*: poor statistical properties.

Two snags

- Biologists a bit non-plussed by the 'cross-hybridisations'.
- It's not quite optimal either.
  - *Designs incorporating all possible multiple direct comparisons rarely, if ever, optimal by the criterion of statistical efficiency.*
- *Reference design*: poor statistical properties.
- Of course, statistical efficiency is not the only criterion for good experimental design.

The design question

- For a fixed number of slides and effects of interest, the question we should be asking is . . .
- *How many slides of each configeration should be produced?*

The design question

- For a fixed number of slides and effects of interest, the question we should be asking is . . .
- *How many slides of each configeration should be produced?*
  - *The standard answer*: prescribe a threshold for M and require experiment to have a pre-determined level of *power* against any such alternatives.

The design question

- For a fixed number of slides and effects of interest, the question we should be asking is . . .
- *How many slides of each configeration should be produced?*
  - *The standard answer*: prescribe a threshold for M and require experiment to have a pre-determined level of *power* against any such alternatives.
  - *Choose a design such that the standard error for each parameter of interest falls below a certain value.*

Complications for microarrays

- Factorial designs are multi-dimensional.

Complications for microarrays

- Factorial designs are multi-dimensional.
- The standard error of a given parameter estimate is $\sigma\sqrt{c}$, where σ is the standard deviation between slides for a particular gene, and c is derived from the design matrix.
- However, a single experiment typically involves $> 10,000$ genes in which σ varies greatly from gene to gene, and is usually unknown.

Complications for microarrays

- Factorial designs are multi-dimensional.
- The standard error of a given parameter estimate is $\sigma\sqrt{c}$, where σ is the standard deviation between slides for a particular gene, and c is derived from the design matrix.
- However, a single experiment typically involves $> 10,000$ genes in which σ varies greatly from gene to gene, and is usually unknown.
- *The good news: the design with the smallest standard error and thus the highest power will be that which has the smallest value of c and this applies equally to every gene.*

Least squares analysis

- We know the *least squares estimates* are given by

$$(X^T X)^{-1} X^T M$$

- and the s.e. of the ith parameter estimate is

$$\sigma \sqrt{c_i}$$

where c_i is the ith diagonal element of $(X^T X)^{-1}$.

Least squares analysis

- We know the *least squares estimates* are given by

$$(X^T X)^{-1} X^T M$$

- and the s.e. of the ith parameter estimate is

$$\sigma \sqrt{c_i}$$

  where c_i is the ith diagonal element of $(X^T X)^{-1}$.

- *It makes sense, all other things being equal, to choose a design that makes the* c_i's *as small as possible.*

Least squares analysis

- We know the *least squares estimates* are given by

$$(X^T X)^{-1} X^T M$$

- and the s.e. of the ith parameter estimate is

$$\sigma \sqrt{c_i}$$

  where c_i is the ith diagonal element of $(X^T X)^{-1}$.

- *It makes sense, all other things being equal, to choose a design that makes the* c_i's *as small as possible.*
  - ⋆ Unfortunately, this criterion is not straightforward.

Admissible designs

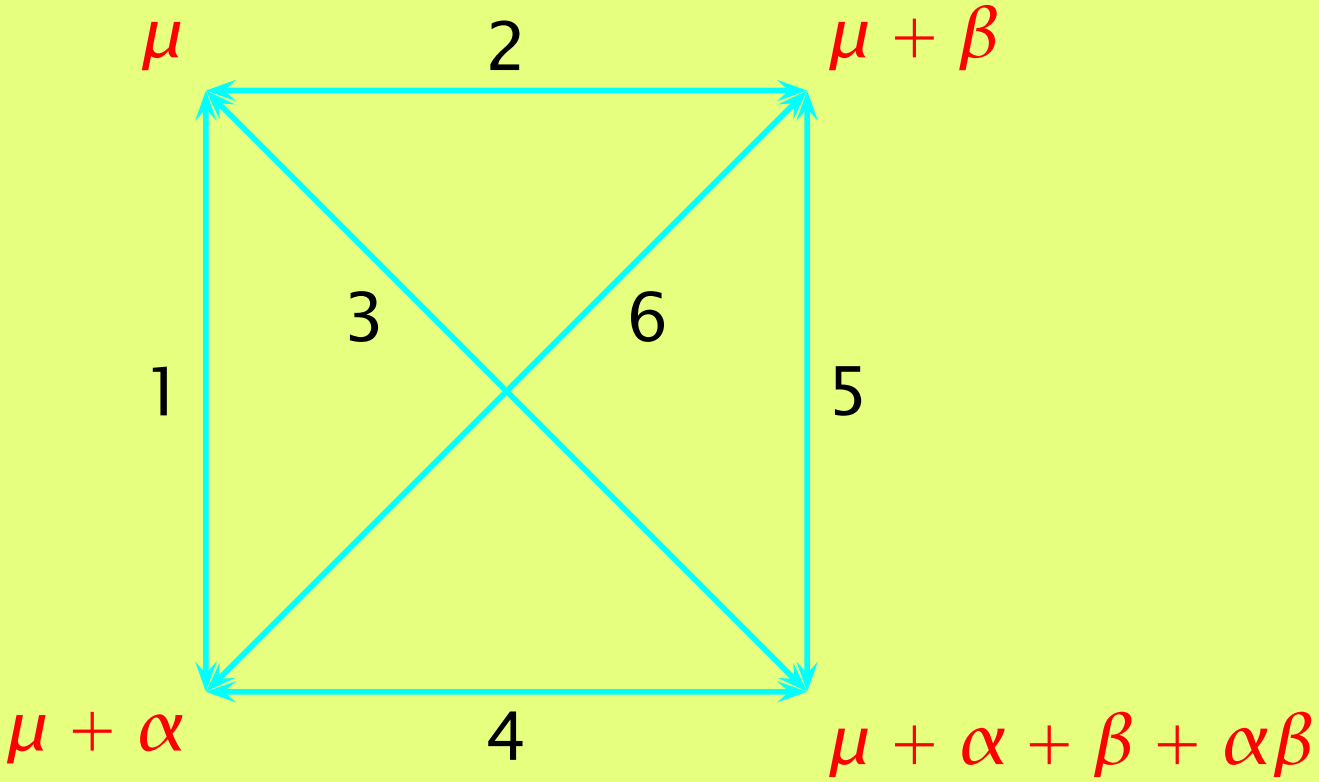
A design with a total of n slides and design matrix X is said to be *admissible* if there exists no other design with n slides and design matrix X_* such that

$$c_i \geq c_i^*$$

for all i *with strict inequality for at least one* i, where c_i, c_i^* are respectively the diagonal elements of $(X^T X)^{-1}$ and $(X_*^T X_*)^{-1}$. A design that is not admissible is said to be *inadmissible*.

- *There is no simple way to identify the set of admissible designs.*
  - e.g. for 24 slides, there are 118,755 possible ways to allocate them to the 6 slide types.
- *However, for small problems, we can simply enumerate the possibilities.*

‘Simplest’ case

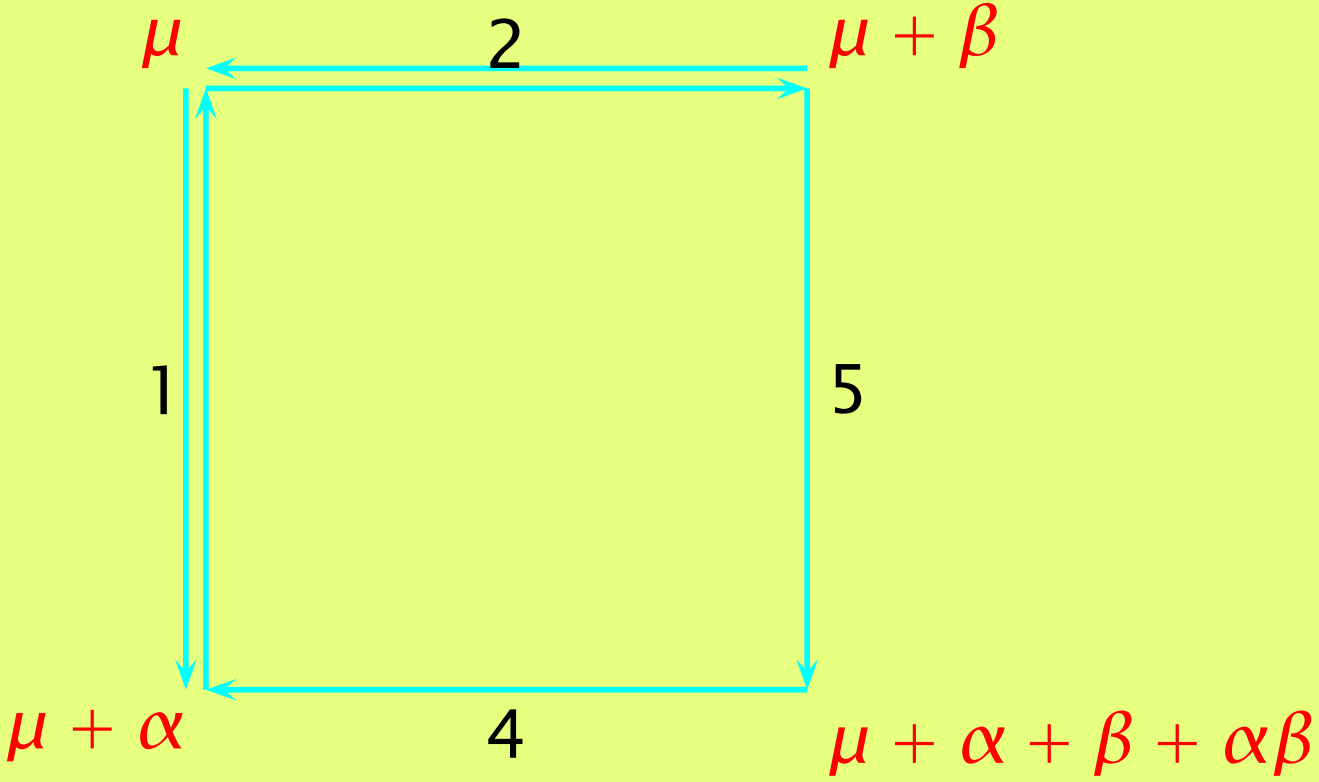


Forward

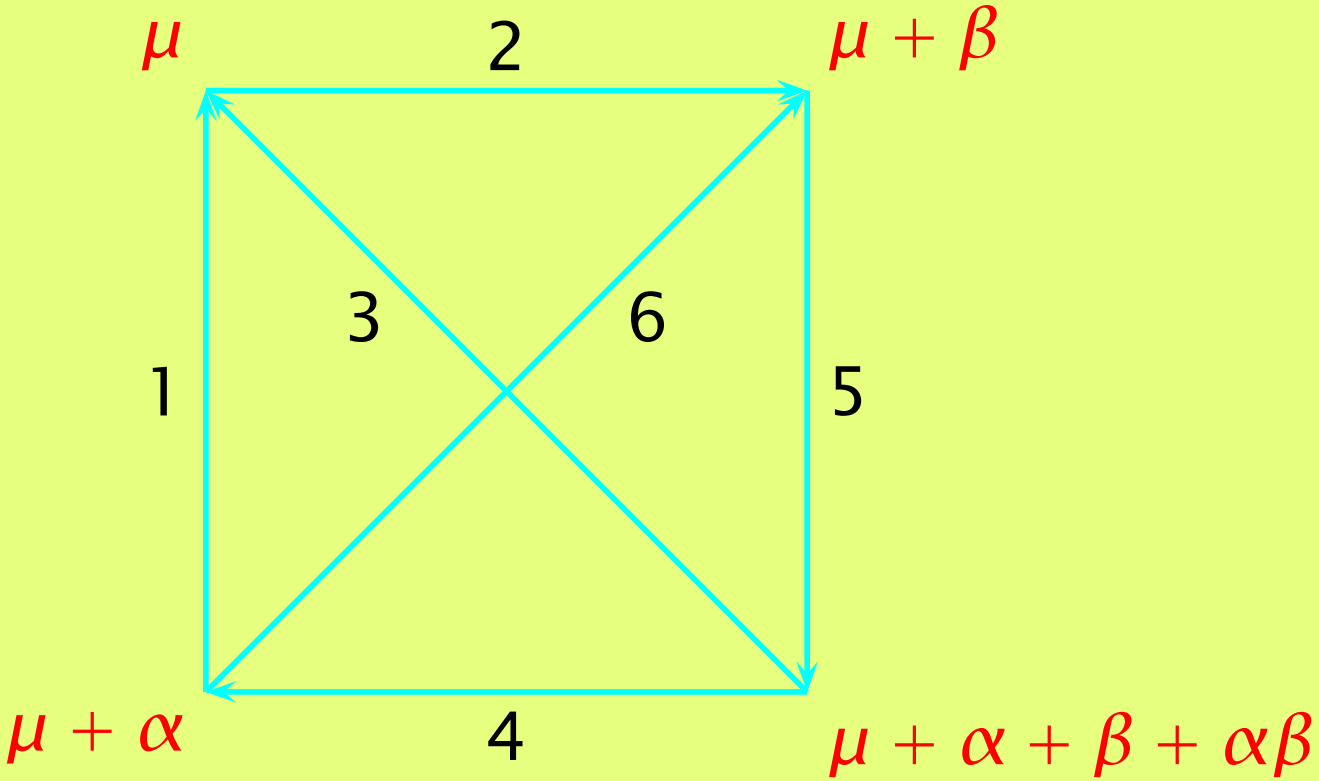
```
macsolomon2[psolomon]% ./a.out

% parameters a b ab;
% comparisons a vs baseline; b vs baseline;
Comparison % a+b+ab vs baseline;
Comparison % a+b+ab vs a;
Comparison % a+b+ab vs b;
Comparison % b vs a;;
% slides 6
% show
1 0 0
0 1 0
1 1 1
0 1 1
1 0 1
-1 1 0
```

```
% generate
Total: 462
Admissible: 21

% equal 1 2;
Admissible: 3
0.4167 0.4167 0.7500 |2 2 0 1 1 0
0.3750 0.3750 1.3750 |2 2 0 1 0 1
0.3750 0.3750 1.3750 |2 2 0 0 1 1

% better 1 1 1 1 1 1
Your Design:
0.5000 0.5000 1.0000 |1 1 1 1 1 1
Better Designs:
0.4167 0.4167 0.7500 |2 2 0 1 1 0
```

6 slides: admissible design 1

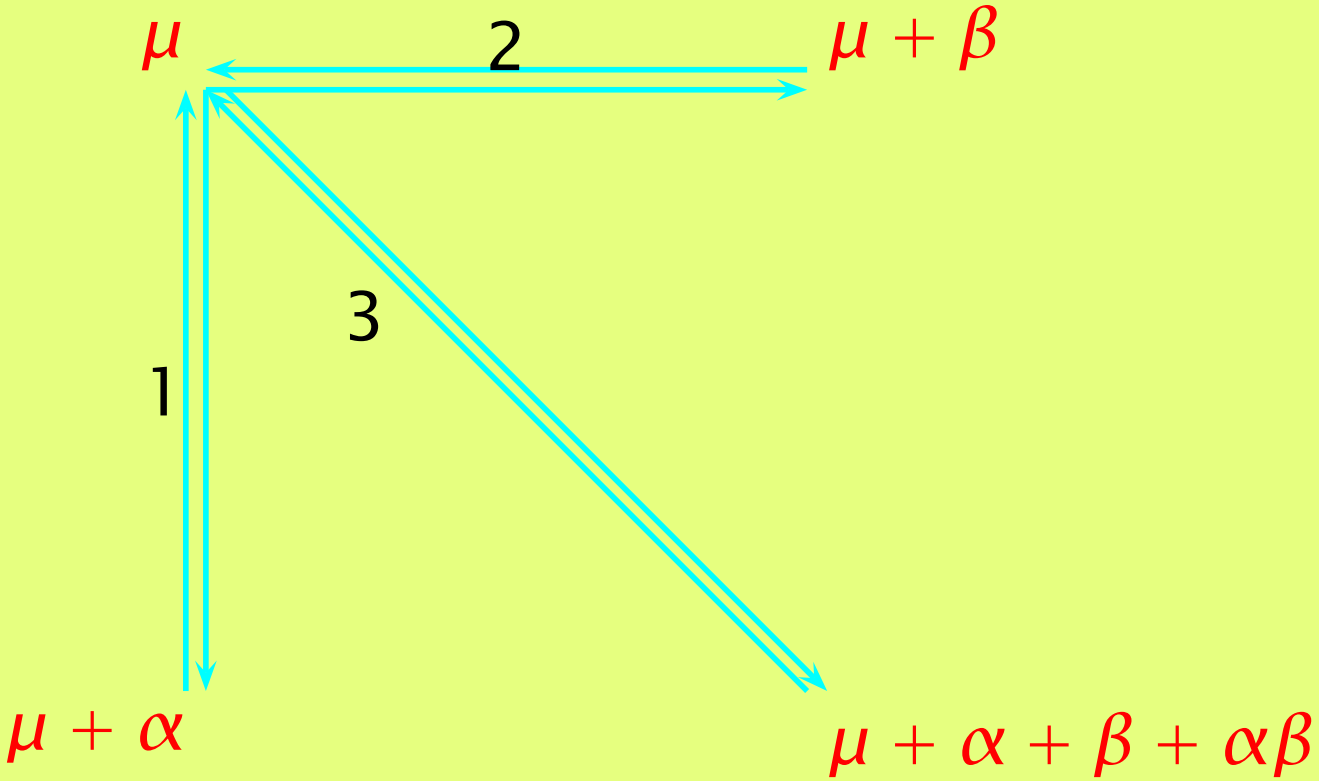


6 slides: inadmissible design



All pairwise comparisons.

6 slides: inadmissible design



The common reference design.

Designs with 6 slides

Des	Replication Configuration 1	2	3	4	5	6	C_α	C_β	$C_{(\alpha\beta)}$
A 1	2	2	0	1	1	0	0.4167	0.4167	0.7500
A 2	2	2	0	1	0	1	0.3750	0.3750	1.3750
A 3	2	2	0	0	1	1	0.3750	0.3750	1.3750
Ref	2	2	2	0	0	0	0.5000	0.5000	1.5000
All	1	1	1	1	1	1	0.5000	0.5000	1.0000

Designs with 6 slides

Des	Replication Configuration 1	2	3	4	5	6	C_α	C_β	$C_{(\alpha\beta)}$
A 1	2	2	0	1	1	0	0.4167	0.4167	0.7500
A 2	2	2	0	1	0	1	0.3750	0.3750	1.3750
A 3	2	2	0	0	1	1	0.3750	0.3750	1.3750
Ref	2	2	2	0	0	0	0.5000	0.5000	1.5000
All	1	1	1	1	1	1	0.5000	0.5000	1.0000

- *Large gains in efficiency can be obtained using admissible designs.*

Case study 2: a 2x3 factorial design

- Aim: to study the role in human disease of cell-surface receptors for a group of signalling molecules.

  - Two cell lines: mutant (M) versus wildtype (W).
  - Three time points: 0, 6 and 12 hours.
  - 9 slides.

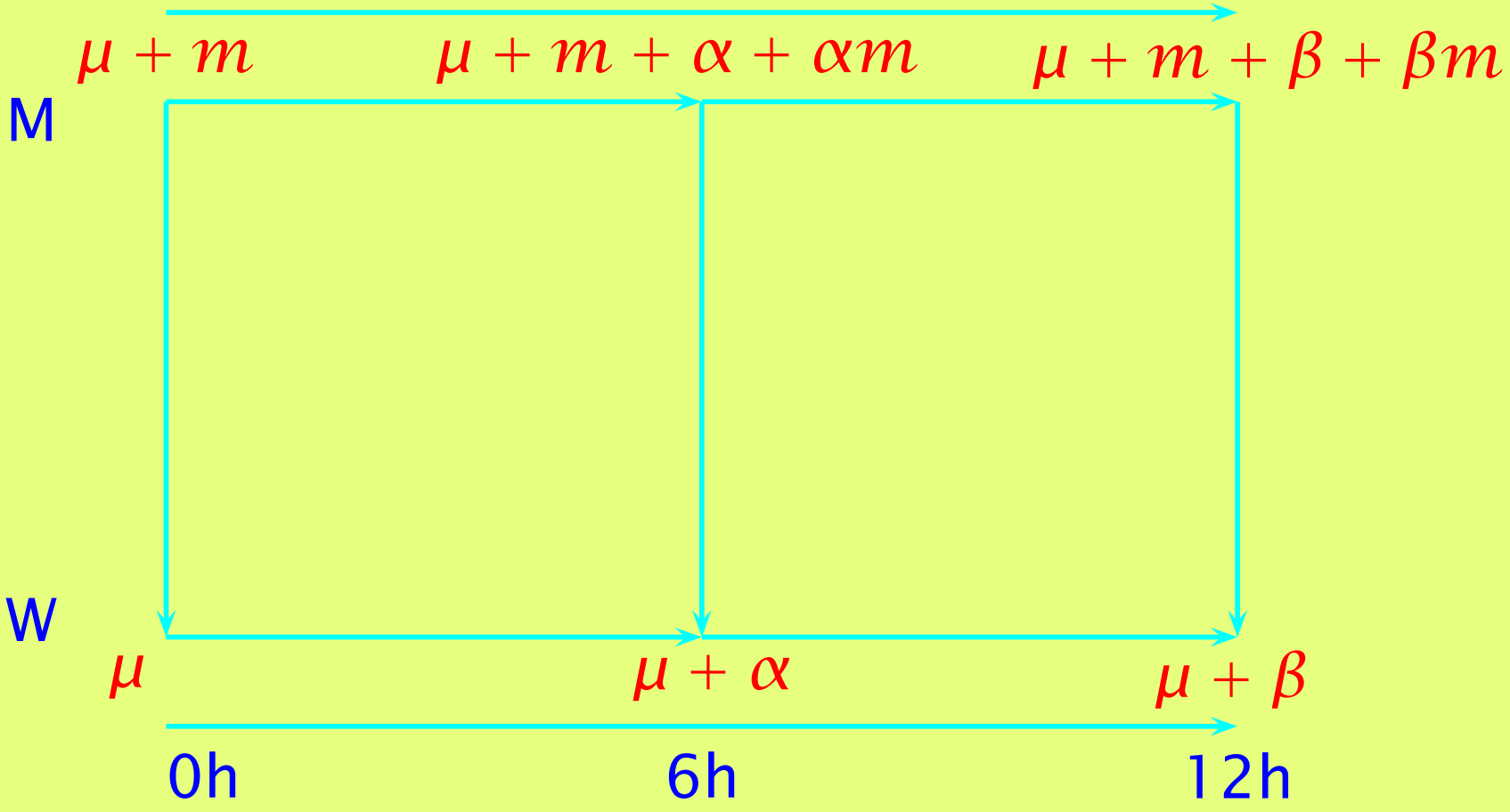
Case study 2: a 2x3 factorial design

- Aim: to study the role in human disease of cell-surface receptors for a group of signalling molecules.

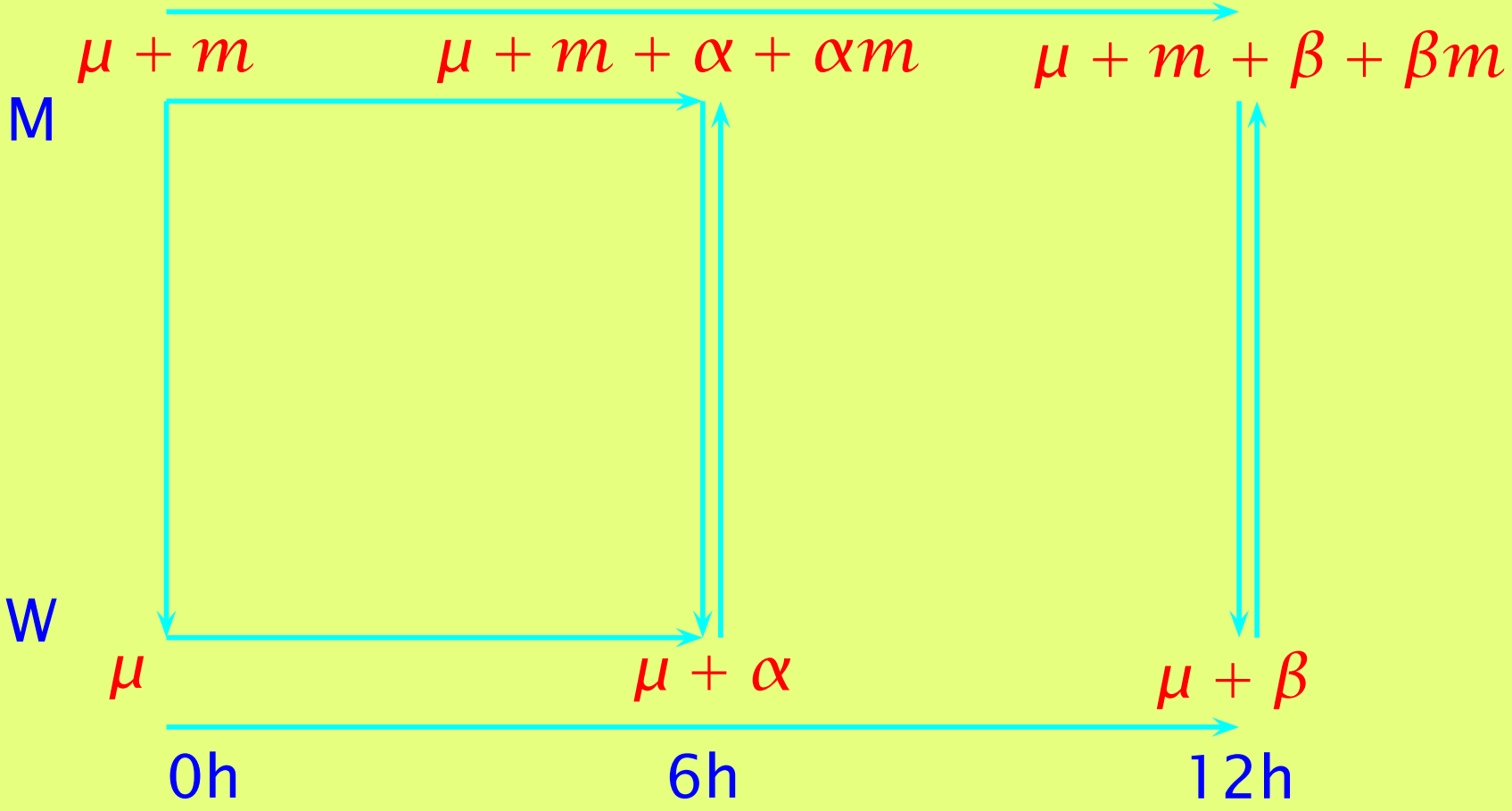
  - Two cell lines: mutant (M) versus wildtype (W).
  - Three time points: 0, 6 and 12 hours.
  - 9 slides.

- *Researchers are particularly interested in the cell line comparisons at 6 and 12 hours.*

a 2x3 factorial design



Best admissible design: 9 slides



Time course experiments

- Single mRNA sample studied at times $0, 1, \ldots, r$.
- For a single gene, μ_t denotes expression level at time t.
- Current recommendation: always co-hybridise samples with a common reference.

Time course experiments

- Single mRNA sample studied at times $0, 1, \ldots, r$.
- For a single gene, μ_t denotes expression level at time t.
- Current recommendation: always co-hybridise samples with a common reference.

- *But 'best' design will be experiment dependent.*

Time course experiments

- Single mRNA sample studied at times $0, 1, \ldots, r$.
- For a single gene, μ_t denotes expression level at time t.
- Current recommendation: always co-hybridise samples with a common reference.

- *But 'best' design will be experiment dependent.*
  - There are $r(r+1)/2$ possible slides, so, how many of *each type* should be made?

Time course experiments

- Single mRNA sample studied at times $0, 1, \ldots, r$.
- For a single gene, μ_t denotes expression level at time t.
- Current recommendation: always co-hybridise samples with a common reference.

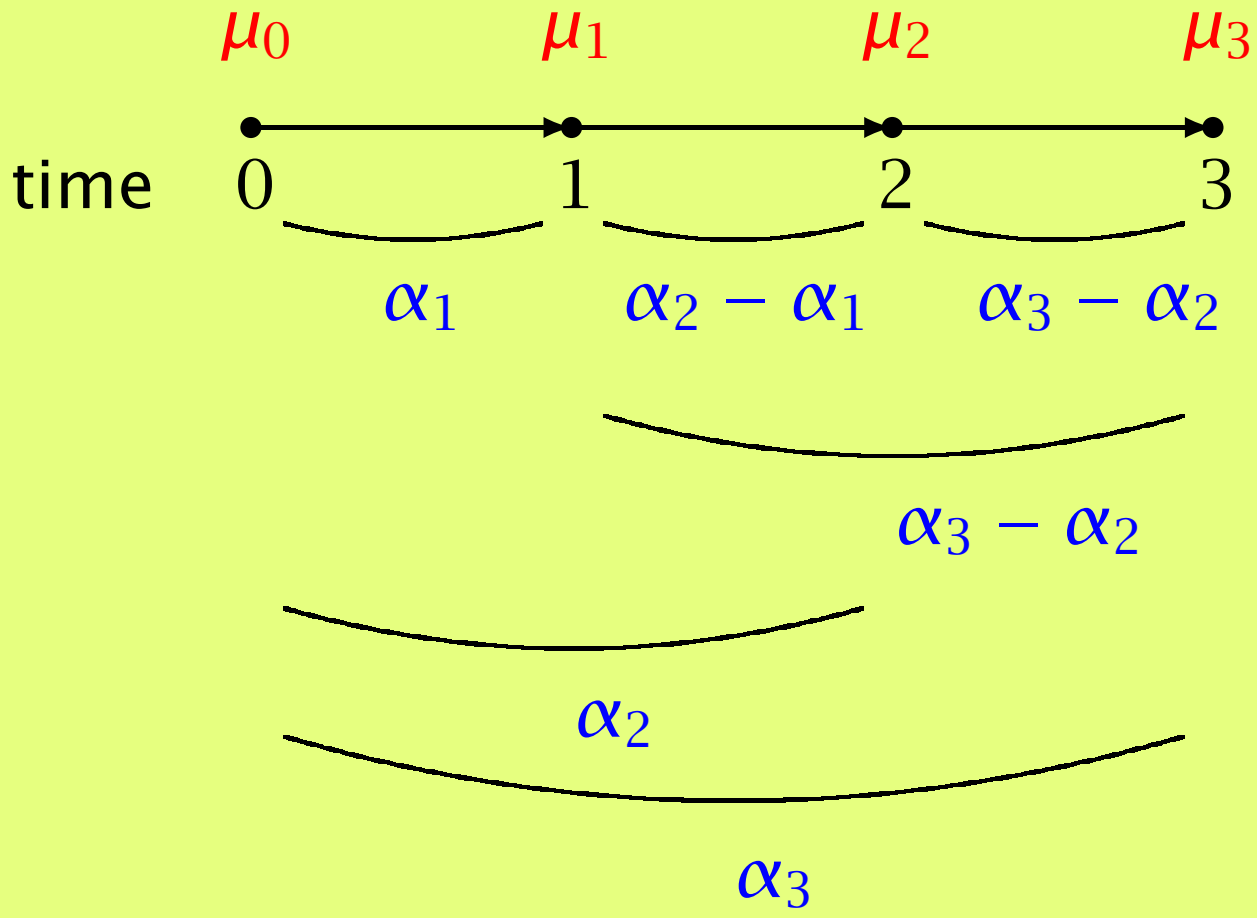
- *But 'best' design will be experiment dependent.*
  - There are $r(r+1)/2$ possible slides, so, how many of *each type* should be made?
- *Key step is to identify the parameters of interest.*
  - e.g. monitoring smooth changes, or
  - e.g. detecting sudden changes or unusual patterns in expression.

- One approach is to take the *differences* relative to time 0

$$\alpha_t = \mu_t - \mu_0 \quad \text{for} \quad t = 1, 2, \ldots, r$$

as the parameters of interest.
In the case $r = 3$, there are six possible slide types:

	Configuration		Expected Log Ratio
	Red	Green	
1	Time 0	Time 1	α_1
2	Time 0	Time 2	α_2
3	Time 0	Time 3	α_3
4	Time 1	Time 2	$\alpha_2 - \alpha_1$
5	Time 1	Time 3	$\alpha_3 - \alpha_1$
6	Time 2	Time 3	$\alpha_3 - \alpha_2$


μ_0
μ_1
μ_2
μ_3
time
0
1
2
3
α_1
$\alpha_2 - \alpha_1$
$\alpha_3 - \alpha_2$
$\alpha_3 - \alpha_2$
α_2
α_3

6 slides: admissible designs

Replication								
Configuration								
1	2	3	4	5	6	c_{α_1}	c_{α_2}	c_{α_3}
2	2	2	0	0	0	0.500	0.500	0.500
1	1	1	1	1	1	0.500	0.500	0.500

6 slides: admissible designs

Replication								
Configuration								
1	2	3	4	5	6	c_{α_1}	c_{α_2}	c_{α_3}
2	2	2	0	0	0	0.500	0.500	0.500
1	1	1	1	1	1	0.500	0.500	0.500

- Covariance matrices:

$$\begin{pmatrix} 0.5 & 0 & 0 \\ 0 & 0.5 & 0 \\ 0 & 0 & 0.5 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 0.5 & 0.25 & 0.25 \\ 0.25 & 0.5 & 0.25 \\ 0.25 & 0.25 & 0.5 \end{pmatrix}$$

A useful tool

- Use admissible designs when time course patterns pre-specified and parameterised in advance
- $\longrightarrow$ *choose best design for allocating a given number of slides to the possible slide types to optimise detection of patterns or profiles of interest.*

Some concluding 'truths'

- '*When we entered the era of high technology, we entered the era of mathematical technology*' [1].

Some concluding 'truths'

- '*When we entered the era of high technology, we entered the era of mathematical technology*' [1].
  - ◇ Innovation in statistical thinking and methods is best driven by substantive applications.

Some concluding 'truths'

- '*When we entered the era of high technology, we entered the era of mathematical technology*' [1].
  - ◇ Innovation in statistical thinking and methods is best driven by substantive applications.
  - ◇ Statistical science must be strong for other disciplines to make effective use of the methods and ideas.

Some concluding 'truths'

- '*When we entered the era of high technology, we entered the era of mathematical technology*' [1].
  - ◇ Innovation in statistical thinking and methods is best driven by substantive applications.
  - ◇ Statistical science must be strong for other disciplines to make effective use of the methods and ideas.
- *Biology looks set to dominate statistics at the beginning of this century, just as it did at the beginning of the last one.*

---

[1] Ad hoc Committee on Resources for the Mathematical Sciences, US National Research Council, 1981.

Further reading and web sites

- Bioconductor Project: Open source bioinformatics using R. http://www.bioconductor.org
- MJ Buckley. The Spot user's guide. CSIRO Mathematical and Information Sciences, August 2000. http://www.cmis.csiro.au/IAP/Spot.htm
- G Glonek & PJ Solomon (2002). Factorial designs for microarray experiments. Submitted. http://maths.adelaide.edu.au/MAG
- R. Ihaka & R. Gentleman. R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics* **5**, 299–314, 1996.
- M Kerr & G Churchill (2001). Experimental design for gene expression microarrays. *Biostatistics* **2**, 183–201.

- Microarray Analysis Group. The University of Adelaide http://maths.adelaide.edu.au/MAG.
- sma v. 0.5.6 (November 2001), B Bolstad, S Dudoit, YH Yang. *http://www.stat.berkeley.edu/users/terry/zarray/Software/smacode.html*.
- G Smyth, et al. (2002) Statistical issues in cDNA microarray data analysis. Research Report, WEHI.
- Statistical Science Web http://www.statsci.org/micrarra/index/html.
- The Chipping Forecast. *Supplement to Nature Genetics*, **21**, 1999.
- TP Speed. Preprints, information and software (sma) www.stat.berkeley.edu/users/terry/zarray
- TP Speed & YH Yang (2002). Direct versus indirect designs for cDNA microarray experiments. Preprint.

THE END