

# Microarray Data Analysis using R and Bioconductor

28-30th, August 2013


University of Cambridge, Cambridge, UK

## Linear Models for Microarrays: *from Experimental Design to Model*

(most slides by *Nuno L. Barbosa Morais and Natalie Thorne*)

*Oscar M. Rueda*

Breast Cancer Functional Genomics Group.  
CRUK Cambridge Research Institute (a.k.a. Li Ka Shing Centre)

 `Oscar.Rueda@cruk.cam.ac.uk`



UNIVERSITY OF  
CAMBRIDGE



CANCER  
RESEARCH  
UK

*To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of.*



Sir Ronald Fisher (1890-1962)

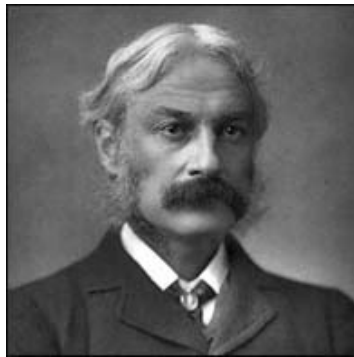
[evolutionary biologist, geneticist and statistician]

*An approximate answer to the right problem is worth a good deal more than an exact answer to an approximate problem.*



John Tukey (1915-2000)  
[Statistician]

*An unsophisticated forecaster uses statistics as a drunken man uses lamp-posts - for support rather than for illumination.*



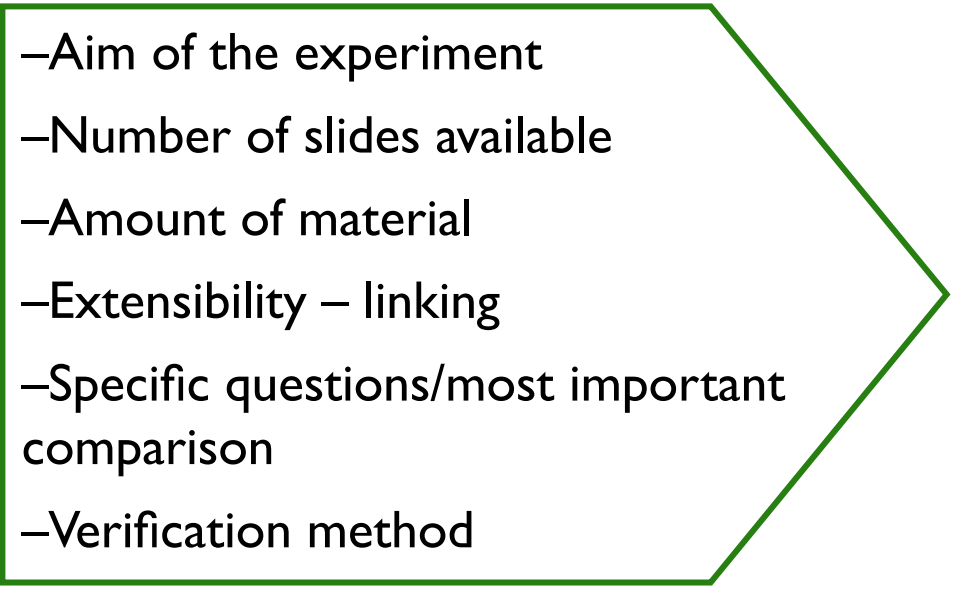
Andrew Lang (1844-1912)

[Poet, novelist and literary critic]

# 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.

# Which design is best?

- 
- Aim of the experiment
  - Number of slides available
  - Amount of material
  - Extensibility – linking
  - Specific questions/most important comparison
  - Verification method

## **Allocation of samples to the slides**

### **– Design layout**

- T vs C or T & C vs Ref
- Multiple treatments
- Factorial design
- Time series


### **– Replication**

### **– Pooling**

# Allocation of samples to the slides

## A Types of Samples

- Replication – technical, biological.
- Pooled vs individual samples.
- Pooled vs amplification samples.



This relates to both one color and two color arrays.

## B Different design layout

- Scientific aim of the experiment.
- Robustness.
- Extensibility.
- Efficiency.

Taking physical limitations or cost into consideration:

- the number of slides;
- the amount of material.

# Avoidance of bias

- Conditions of an experiment; mRNA extraction and processing, the reagents, the operators, the scanners and so on can leave a “global signature” in the resulting expression data.
- Randomization.
- Local control is the general term used for arranging experimental material.



# Confounding

- Consider the following hypothesis:  
“Drinking tea has a different gene expression in men than in women”
- If we have only two two-colour arrays available, which design to use?

Design 1

Tea male  
No tea male

Tea female  
No tea female

Design 2

Tea male  
Tea female

No tea male  
No tea female

Design 3

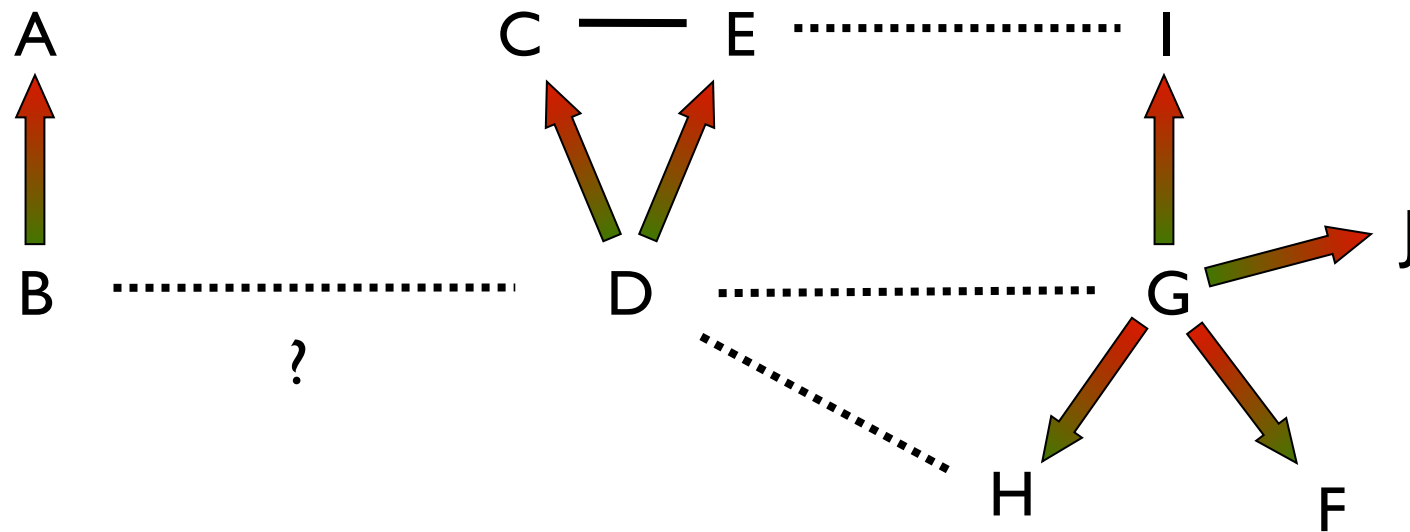
Tea male  
No tea female

No tea male  
Tea female

# Design : two-colour

- How do we allocate two samples to each chip?
  - Comparisons within chip : Direct
  - Comparisons between  $M'$  s between chips : Indirect
  - Comparisons between chips : Single-channel

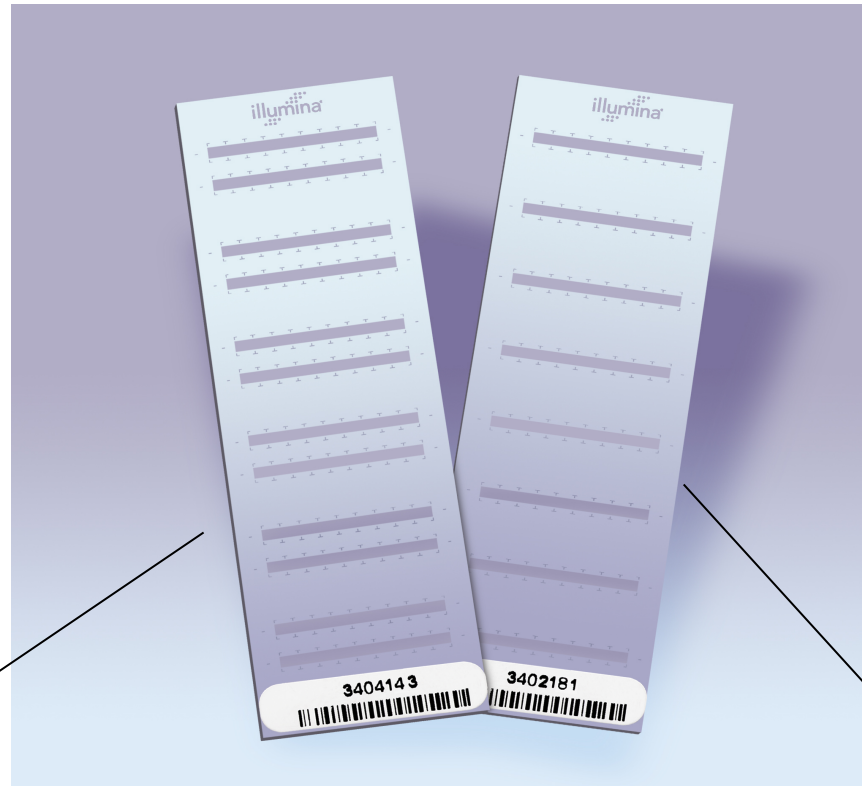
# 2-colour design : connectivity



Direct comparison, Indirect comparison, single-channel comparison



# Illumina BeadChips



Whole Genome

6 arrays per chip: 2 strips = 1 array

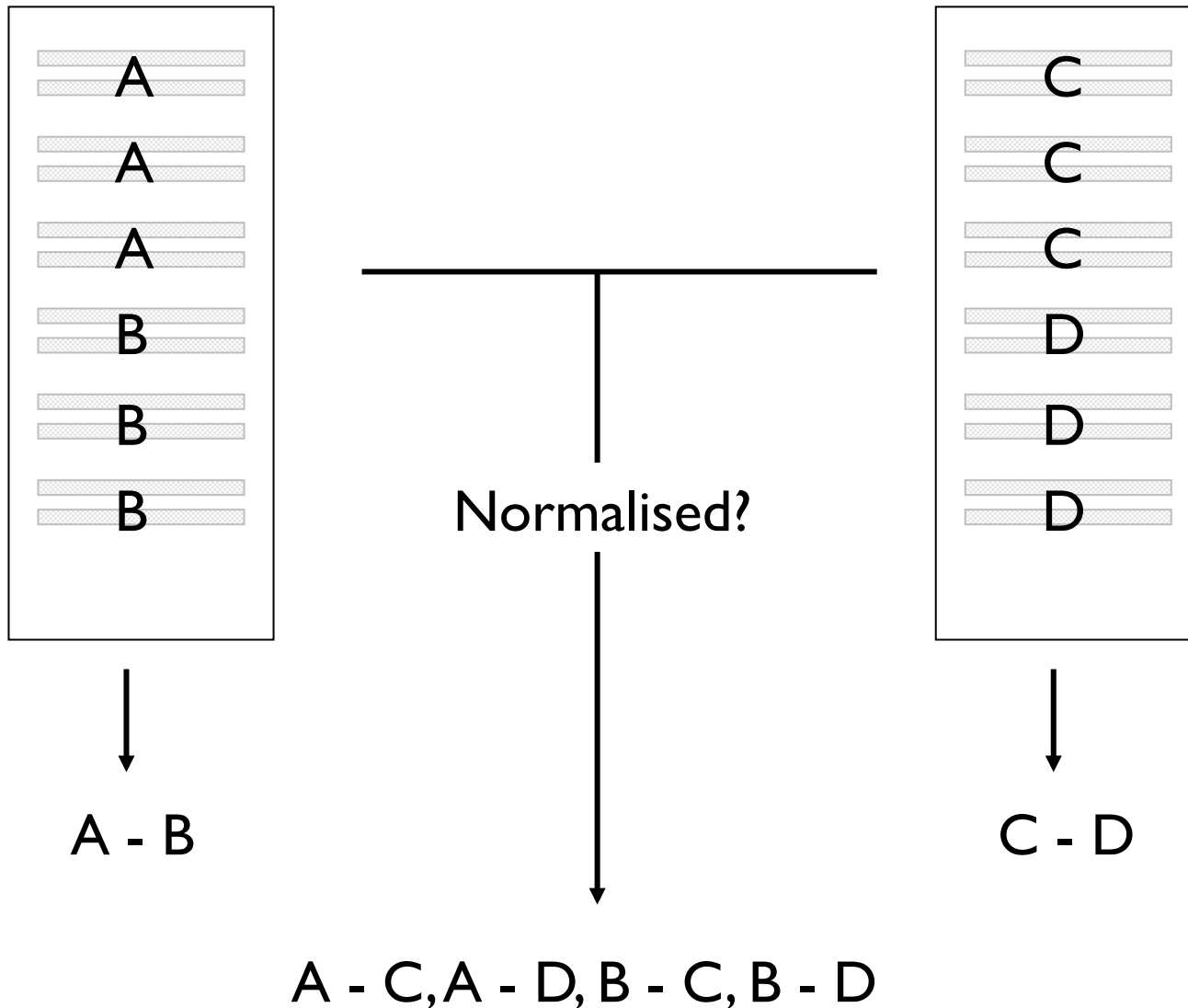
RefSeq BeadChip

8 arrays per chip 1 strip = 1 array

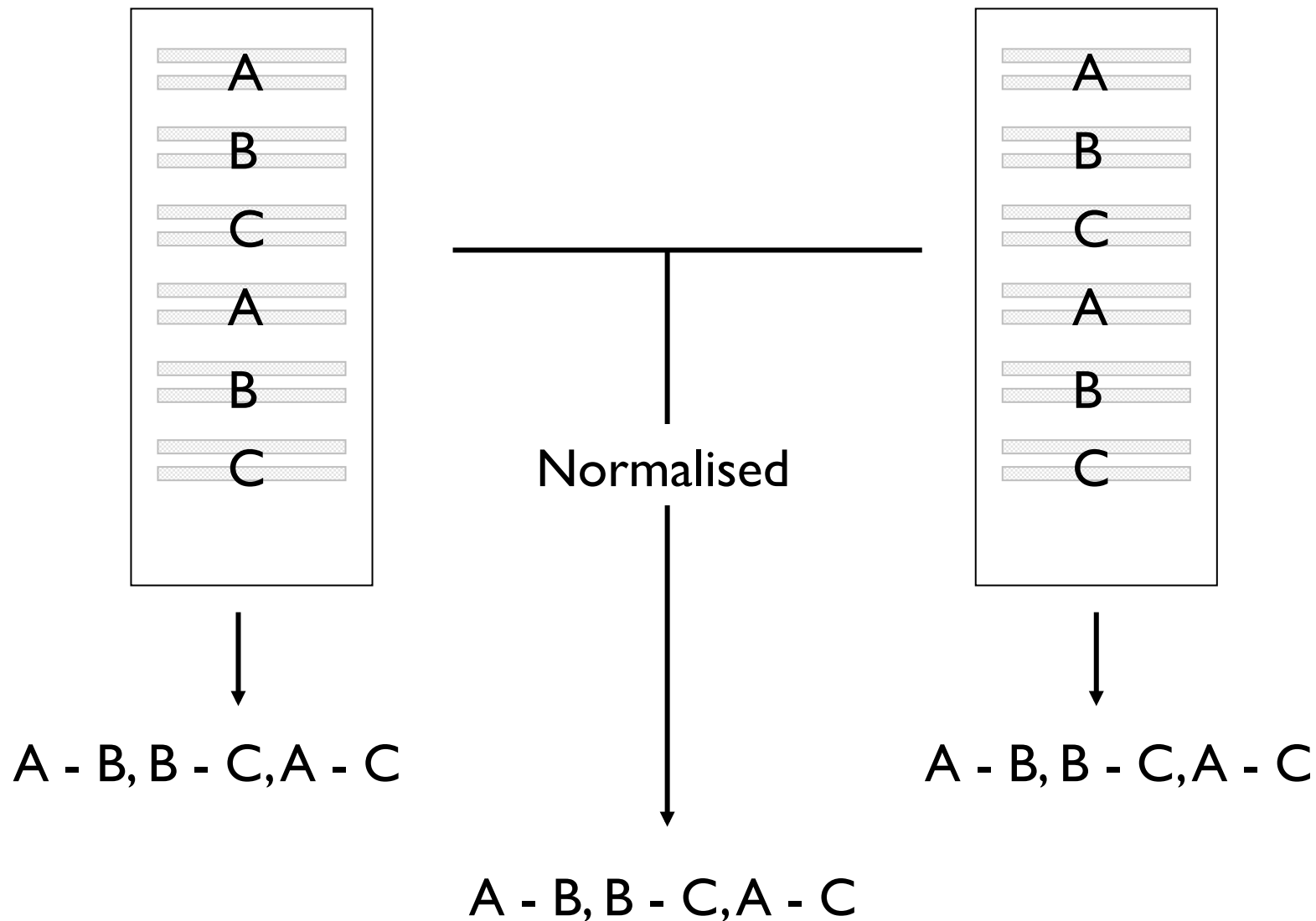
# Design : Illumina

- How do we allocate six samples to each chip?
  - Comparisons **within-chips** more precise than **between-chips**
  - Normalisation assumes distributions of samples from different chips are roughly the same

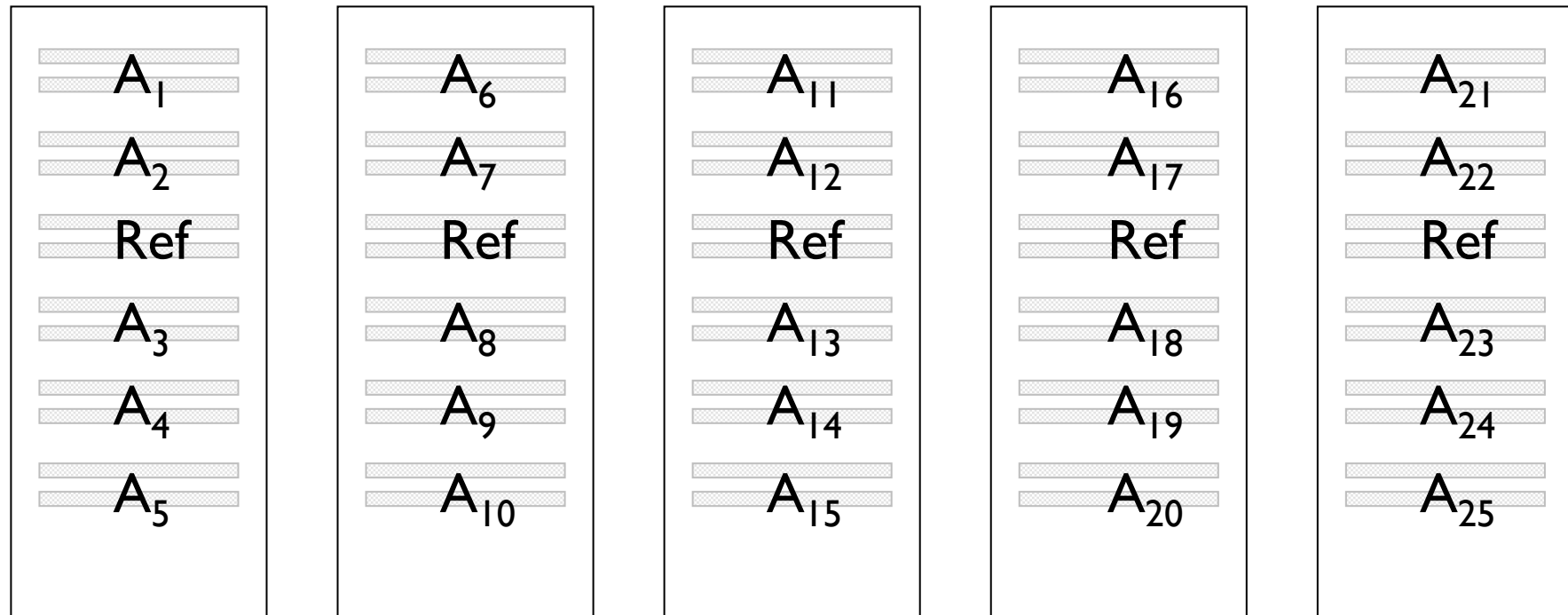
# Illumina design : connectivity / ability to normalise



# Illumina design : connectivity / ability to normalise



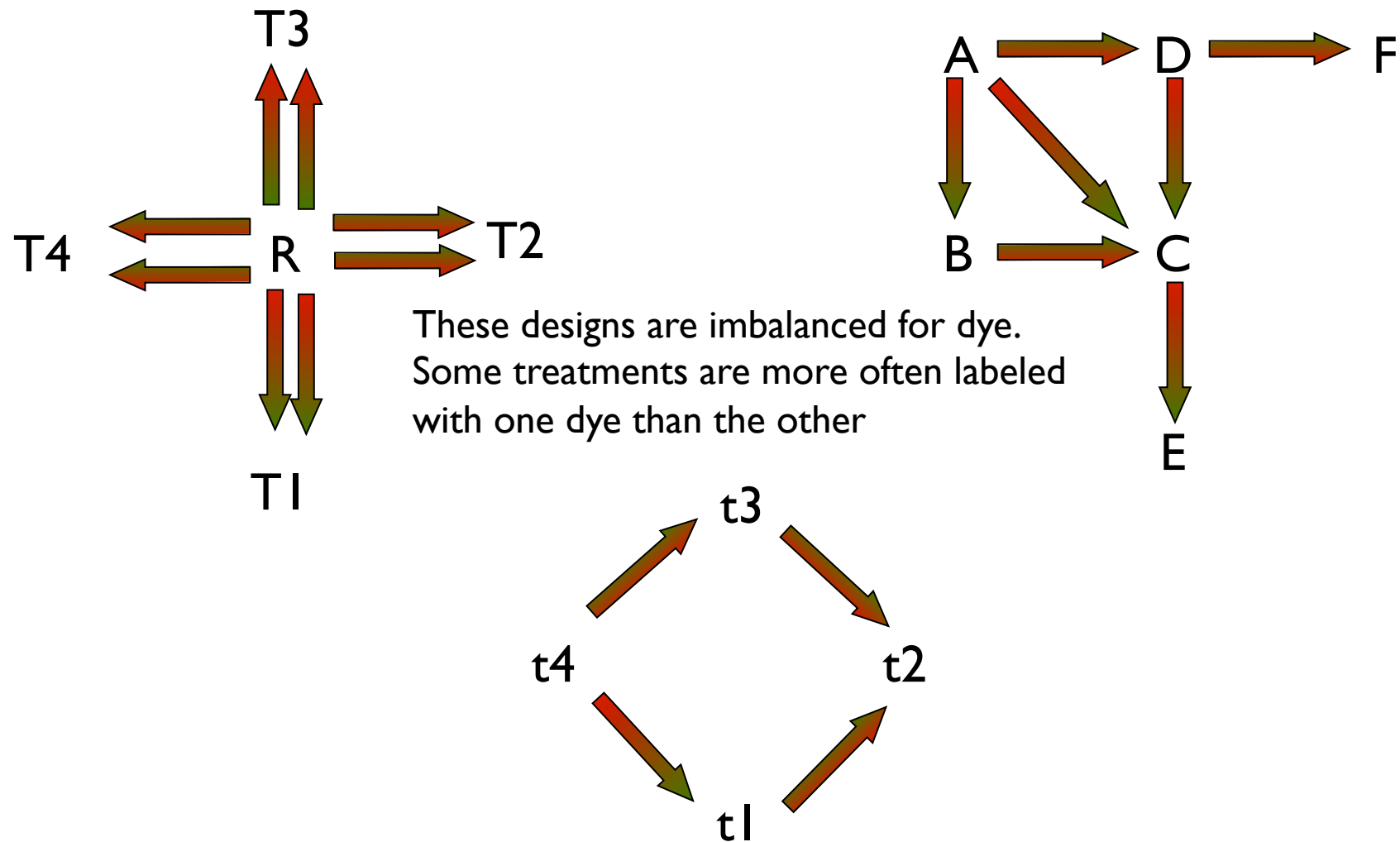
# Illumina design : ability to normalise



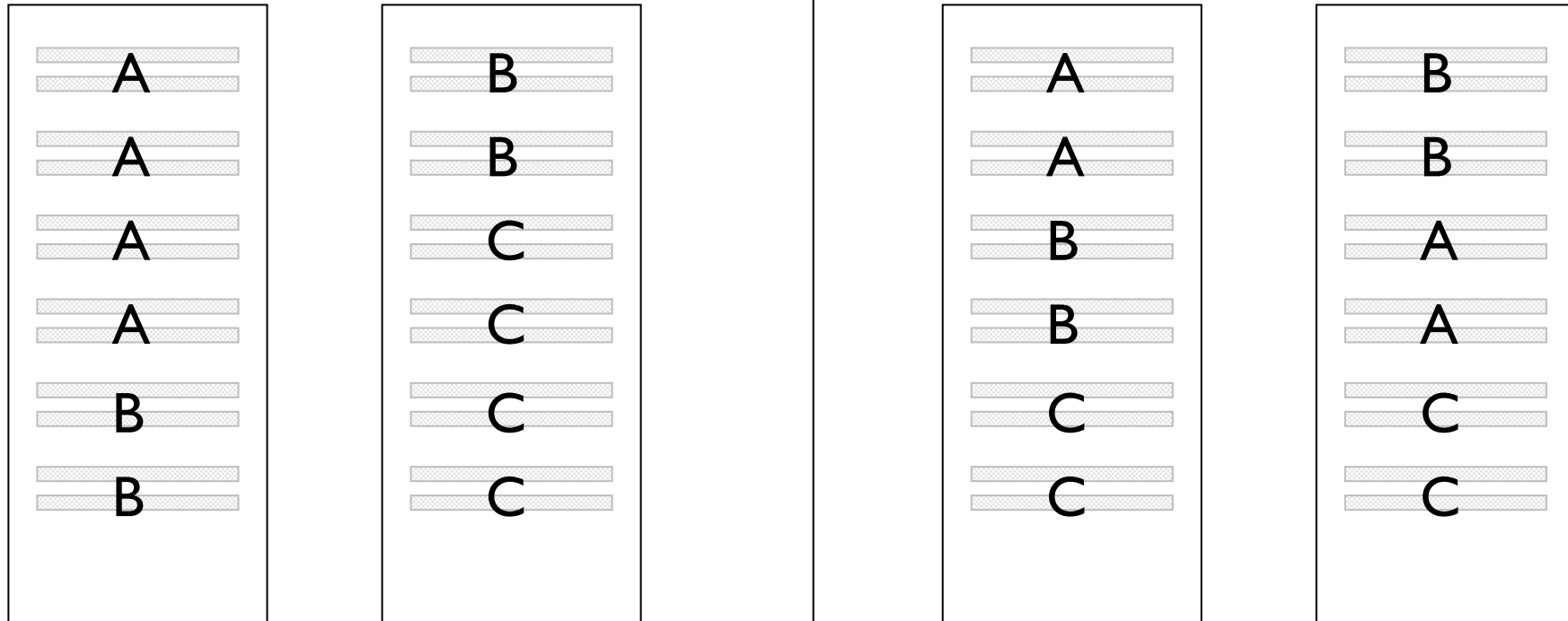
Does Ref act as a common “normalisation reference”? This might be particularly useful for large studies involving many treatments/sample types.



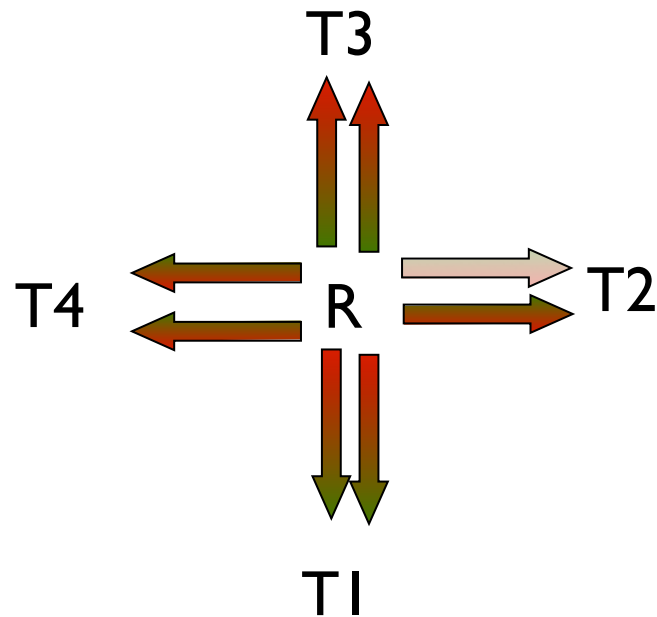
# 2-colour design : dye balance



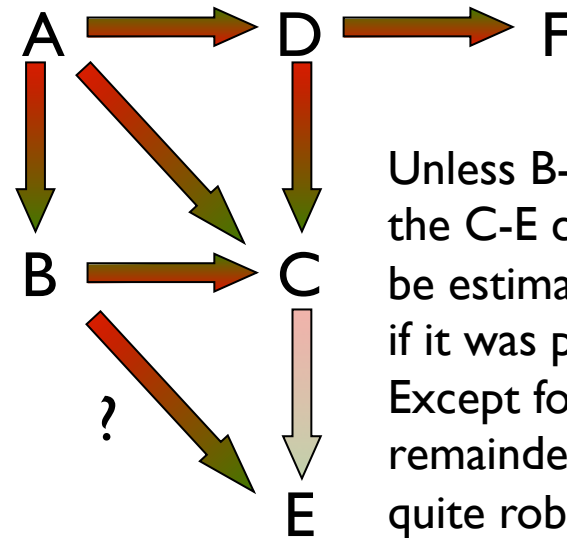
# Illumina design : balance



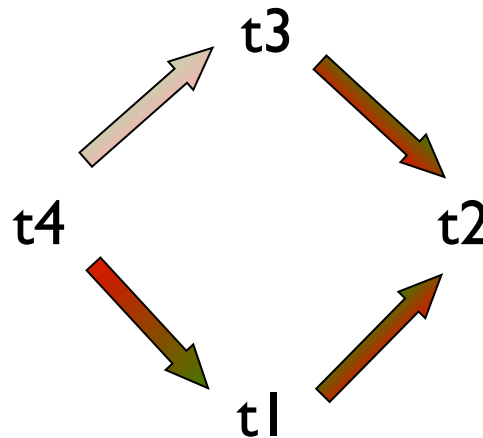
# 2-colour design : robustness



The reference design can only be made robust by increasing replication

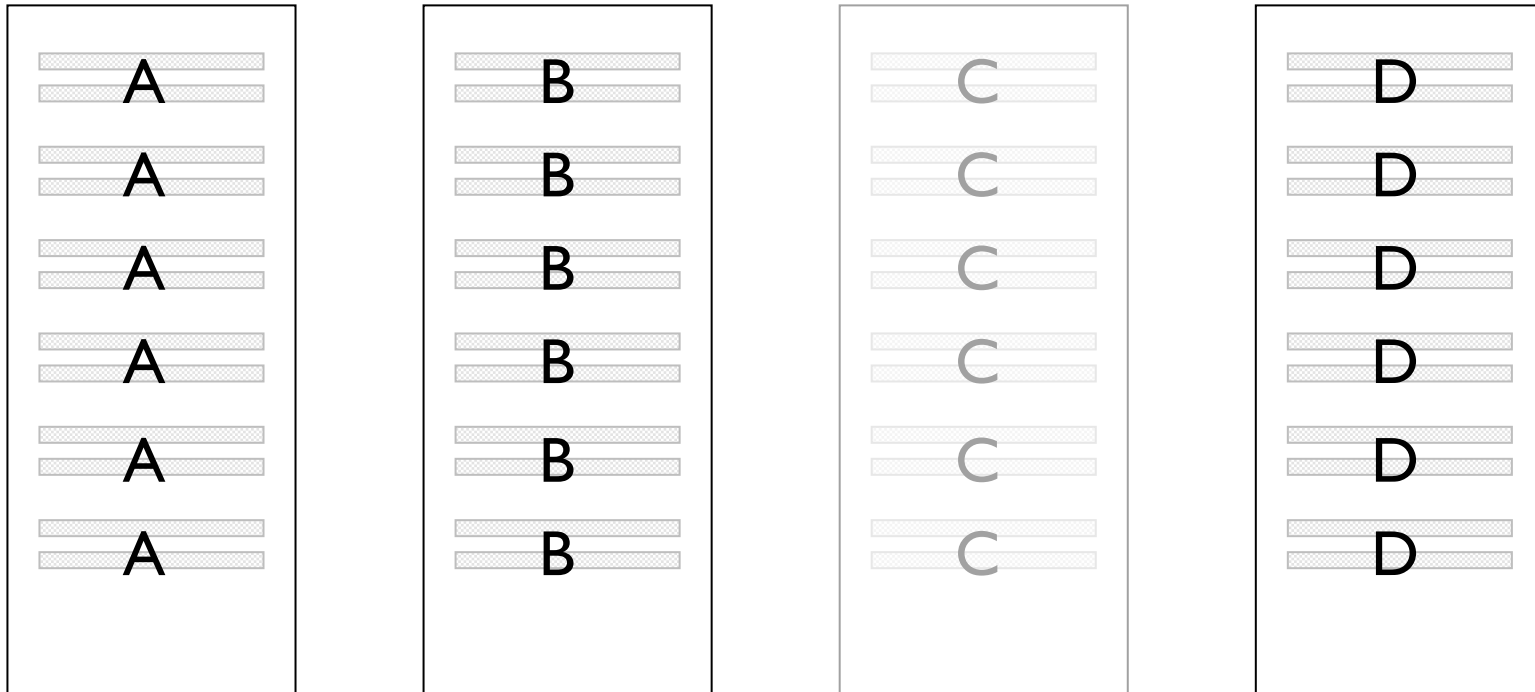


Unless B-E array is added, the C-E comparison cannot be estimated any other way if it was poor quality. Except for F-D, the remainder of the design is quite robust.

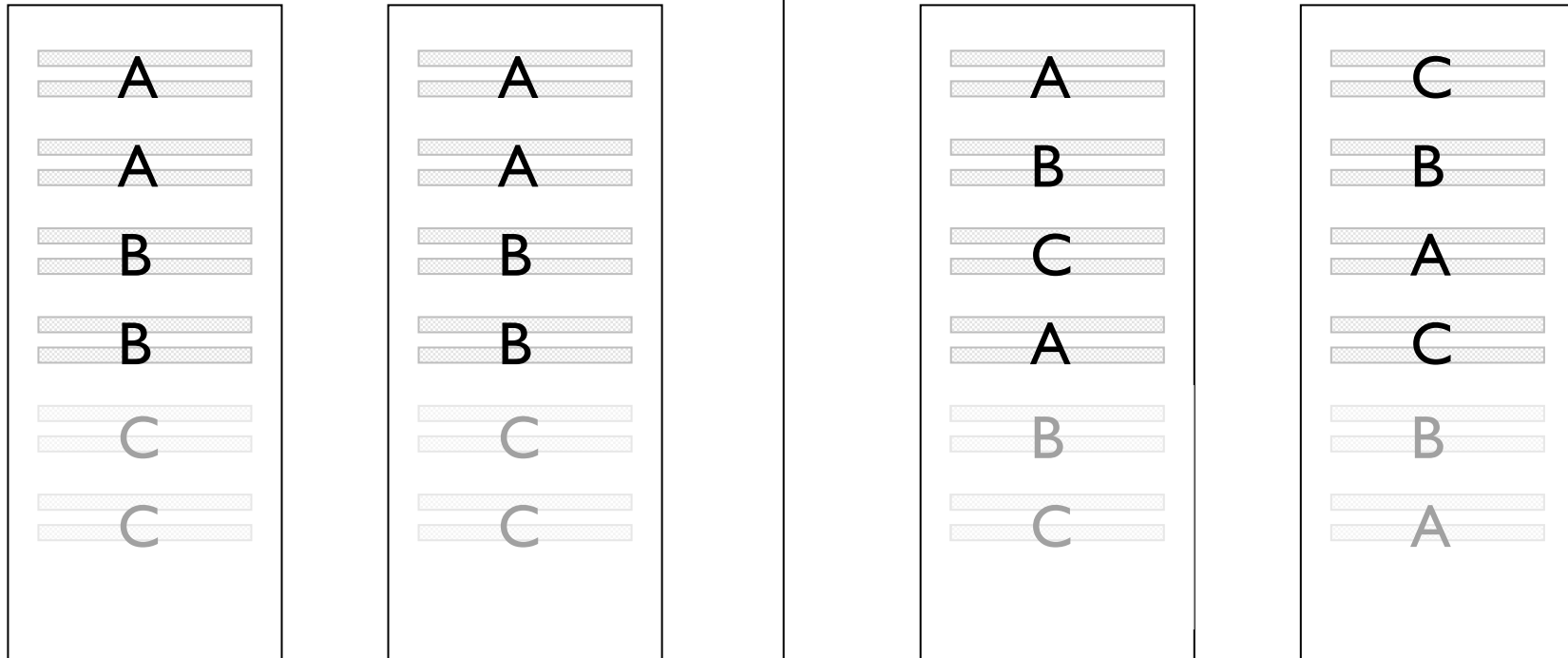


Loop design is not very robust, as loop gets bigger, little way to recover precision for any poor quality arrays

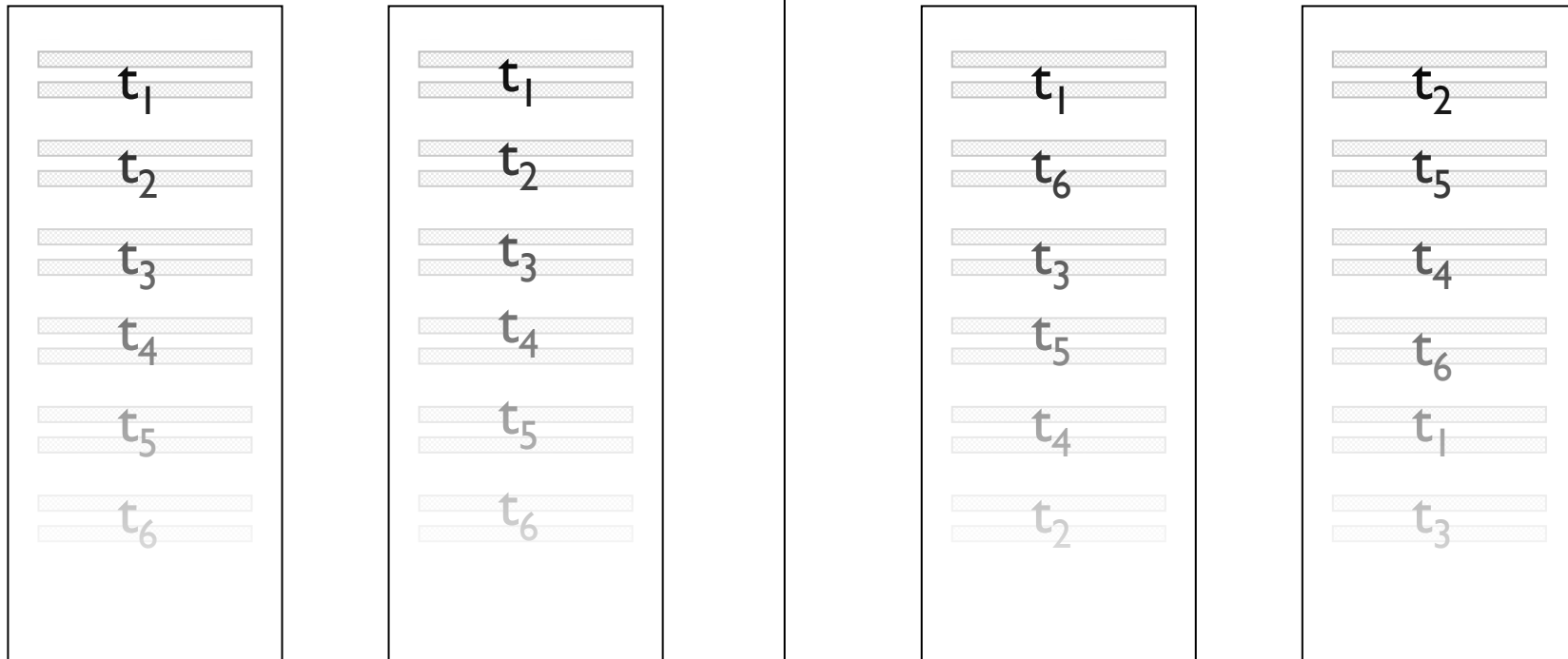
# Illumina design : robustness



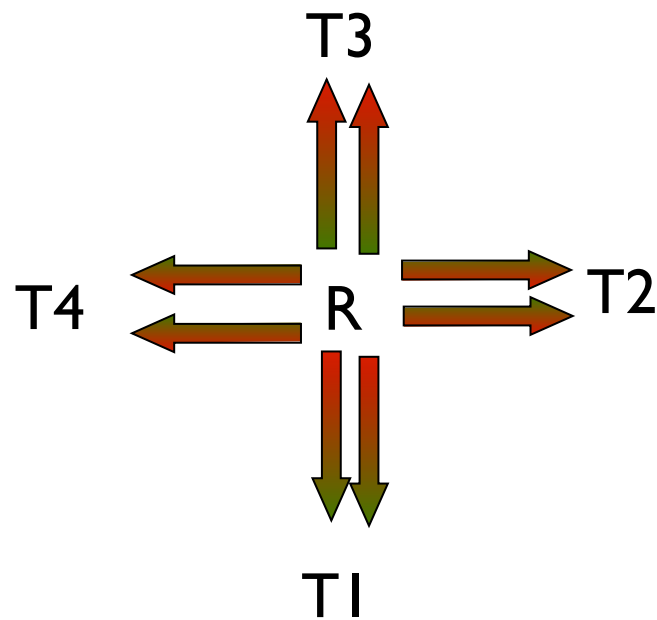
# Illumina design : robustness



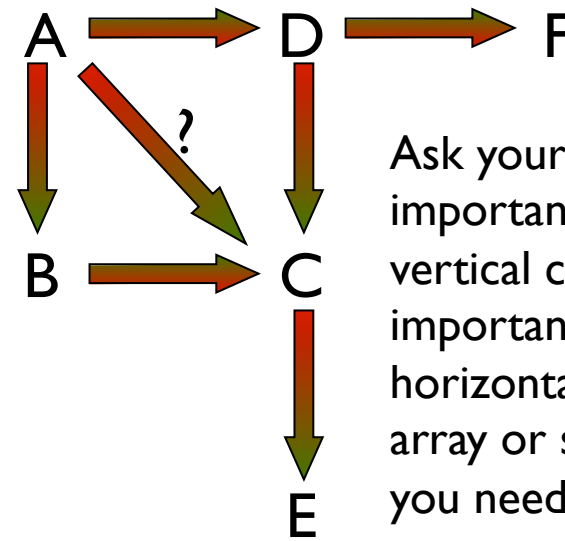
# Illumina design : normalisation and spatial trends



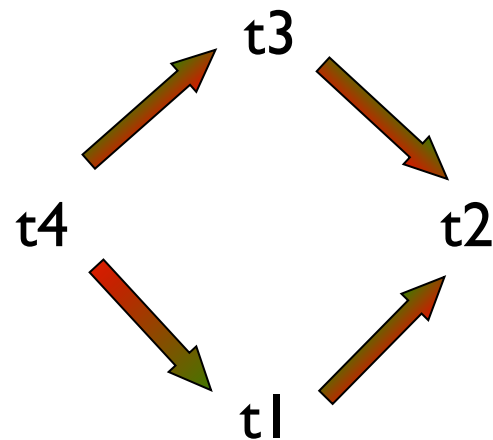
# 2-colour design : important questions



The reference design is useful if no particular treatment comparisons are important, but overall comparison of treatments is important

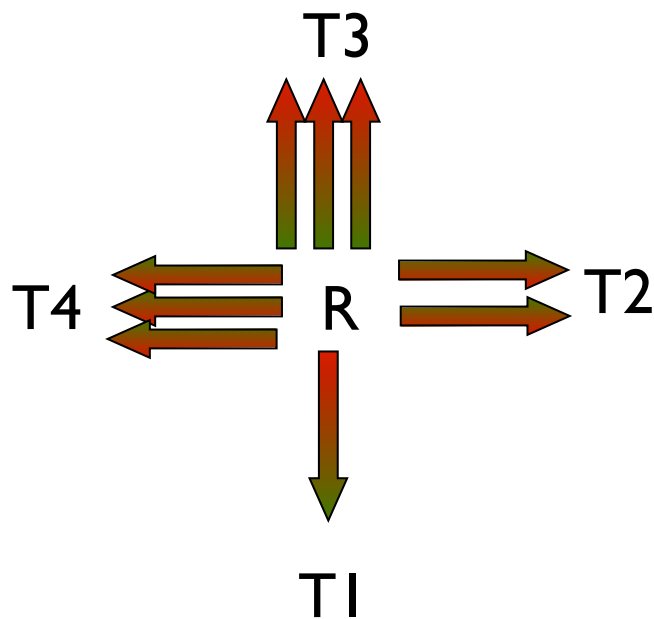


Ask yourself, is A-C more important than B-D, or are the vertical comparisons more important to you than the horizontal ones. With money, array or sample limitations, you need to decide which comparisons are most important.

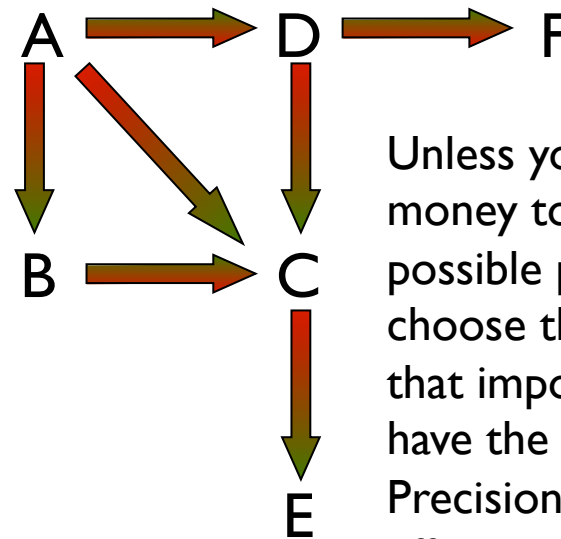


Loop design is useful if it is important to estimate consecutive differences. Usually this is only the case for some time course experiments.

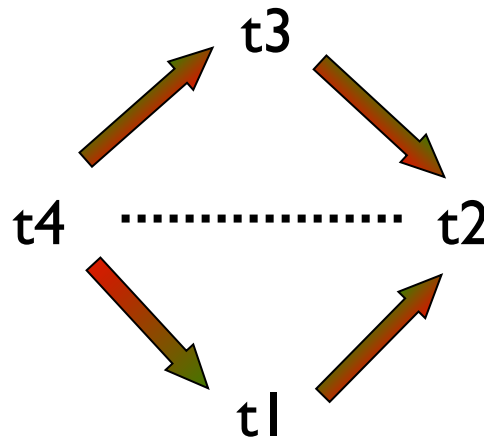
# 2-colour design : precision



Comparisons with more replication will have more precision. This means you will be able to detect smaller changes in T3 and T4 and have less ability to detect changes in T2 and especially T1 (compared to reference)



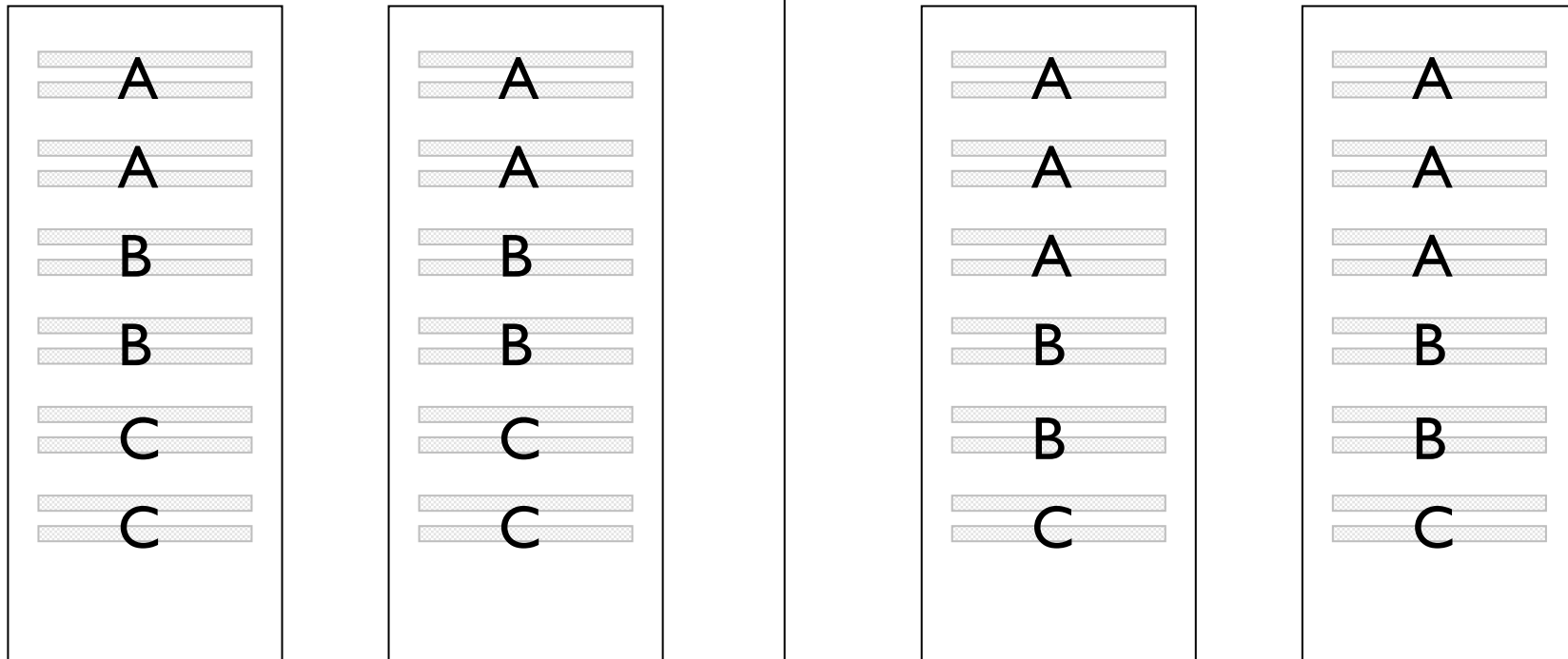
Unless you have the time and money to perform every possible pairwise comparison, choose the design carefully so that important comparisons have the most precision. Precision is determined by the effective replication



Loop design has high precision for estimating consecutive effects, but the precision to estimate comparisons between treatments on opposite sides of the loop is very very low.



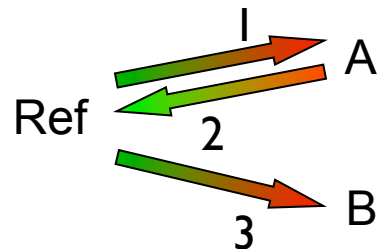
# Illumina design : important questions/precision



# Specifying the design

## Sample types

I. Represent the effect measured by each sample type.



Single-channel  
representation

## Possible parameters

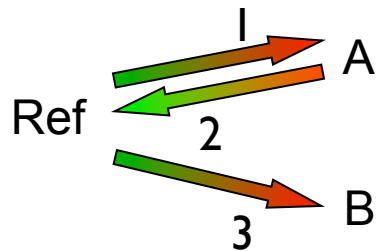
I. What differences are important?

Log-ratio  
representation

# Specifying the design

## Sample types

I. Represent the effect measured by each sample type.



A = baseline + a

B = baseline + b

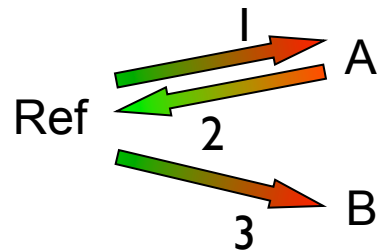
Ref = baseline

## Possible parameters

I. What differences are important?

Based on your idea about what is measured in each sample type (in relation to the other sample types), we do this to help you understand and interpret results from your log-ratios

# Specifying the design



## Sample types

I. Represent the effect measured by each sample type.

$$A = \text{baseline} + a$$

$$B = \text{baseline} + b$$

$$\text{Ref} = \text{baseline}$$

## Possible parameters

I. What differences are important?

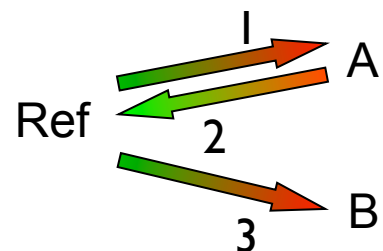
$$A - \text{Ref} = \text{baseline} + a - (\text{baseline}) = a$$

$$B - \text{Ref} = \text{baseline} + b - (\text{baseline}) = b$$

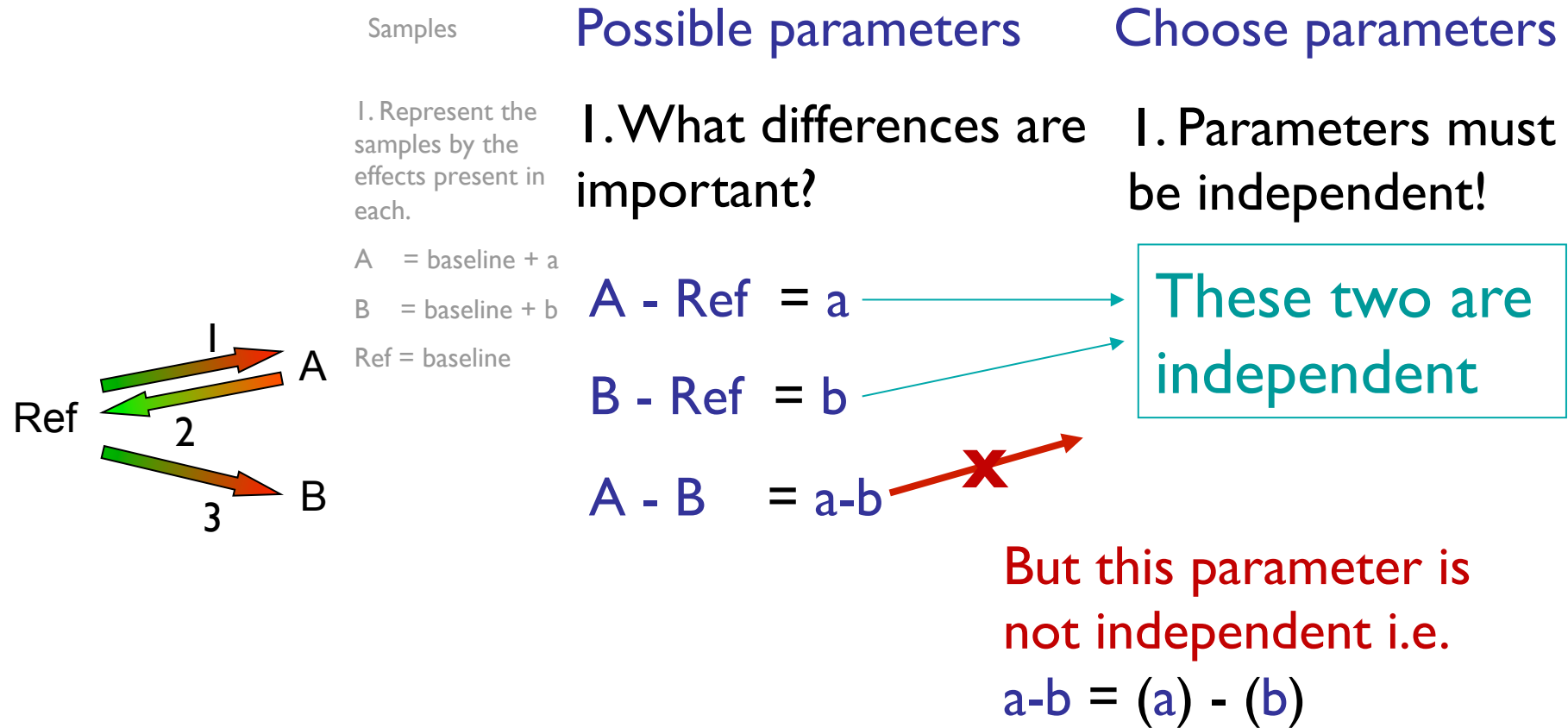
$$A - B = \text{baseline} + a - (\text{baseline} + b) = a - b$$

Parameters in two-colour experiments need to be representative of log-ratio comparisons (this is the data that you would typically model). Once you've defined the effects in each sample type, it is fairly easy to define possible parameters for your model.

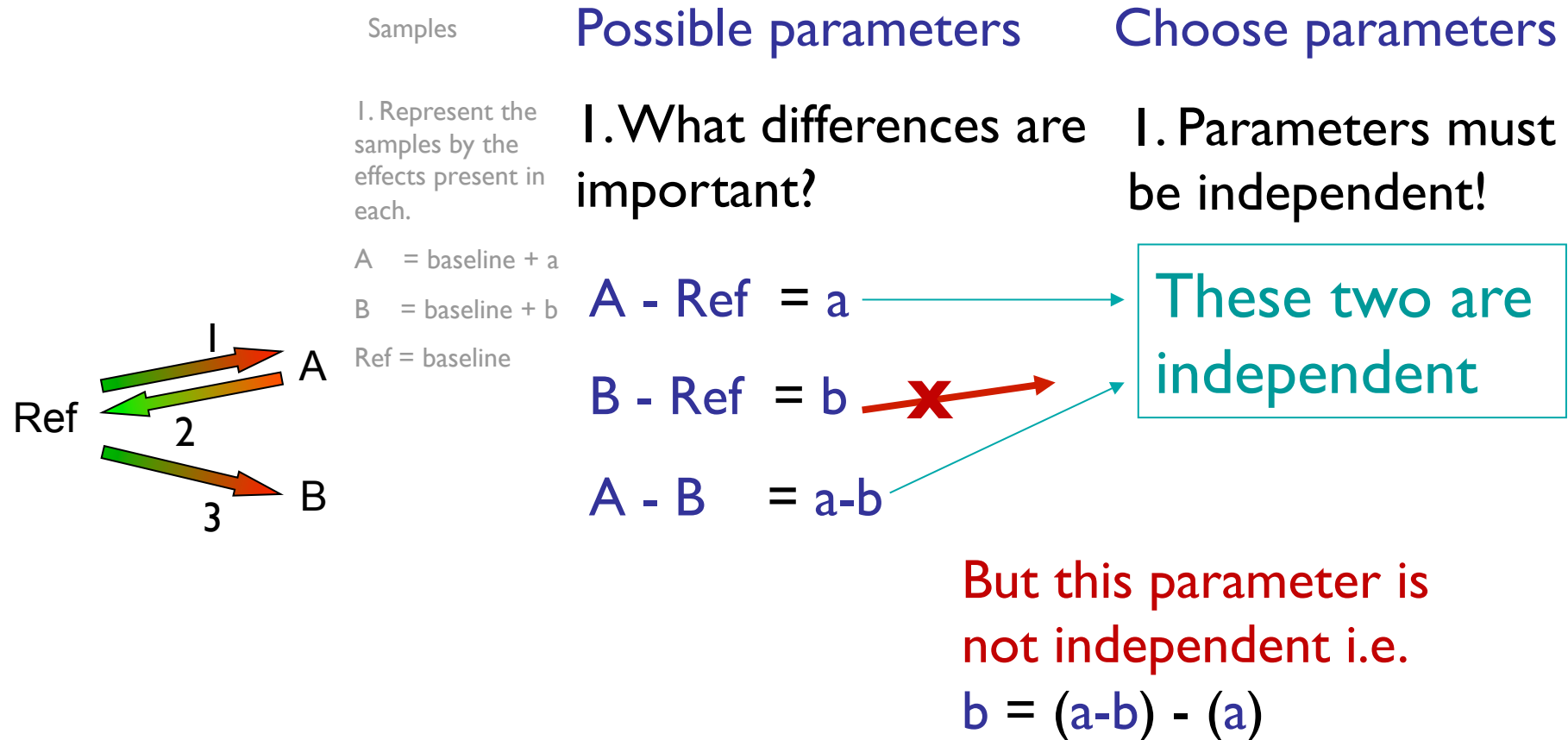
# Specifying the design

Samples	Possible parameters	Choose parameters
I. Represent the samples by the effects present in each.  A = baseline + a B = baseline + b Ref = baseline 	I. What differences are important?  $A - \text{Ref} = a$  $B - \text{Ref} = b$  $A - B = a - b$	I. Parameters must be independent!

# Specifying the design

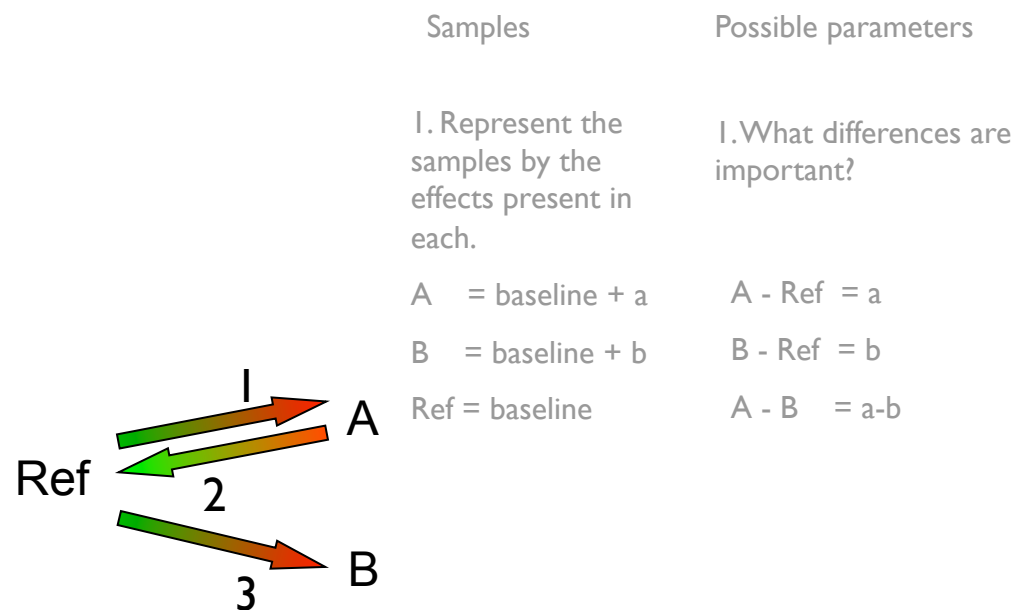


# Specifying the design



A definition of independent parameters : No combination of the parameters can equal any of the parameters

# Specifying the design



## Choose parameters

I. Parameters must be independent!

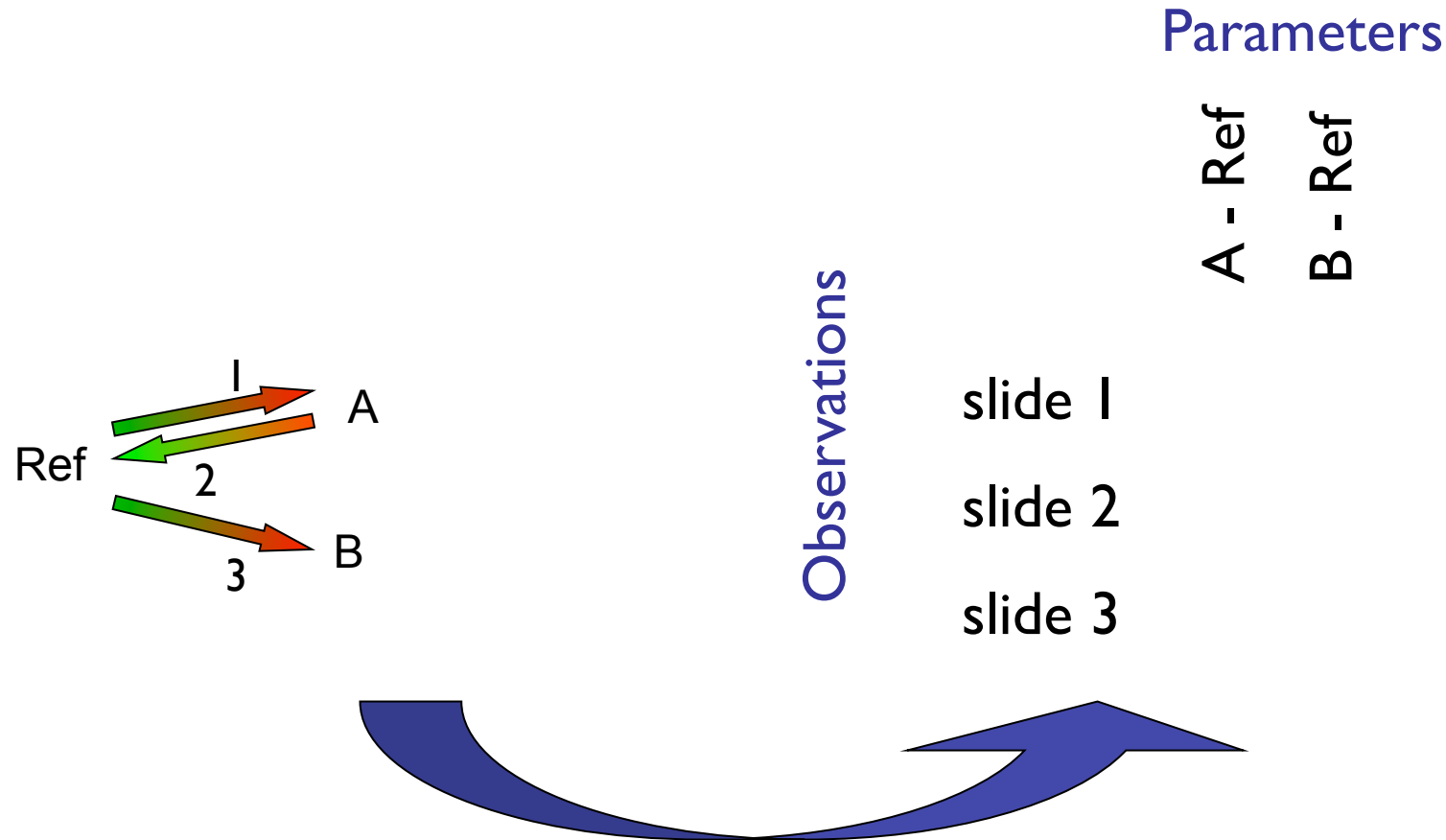
$A - \text{Ref}, a$

$B - \text{Ref}, b$

Now, specify the design and the parameters that you've chosen as a matrix

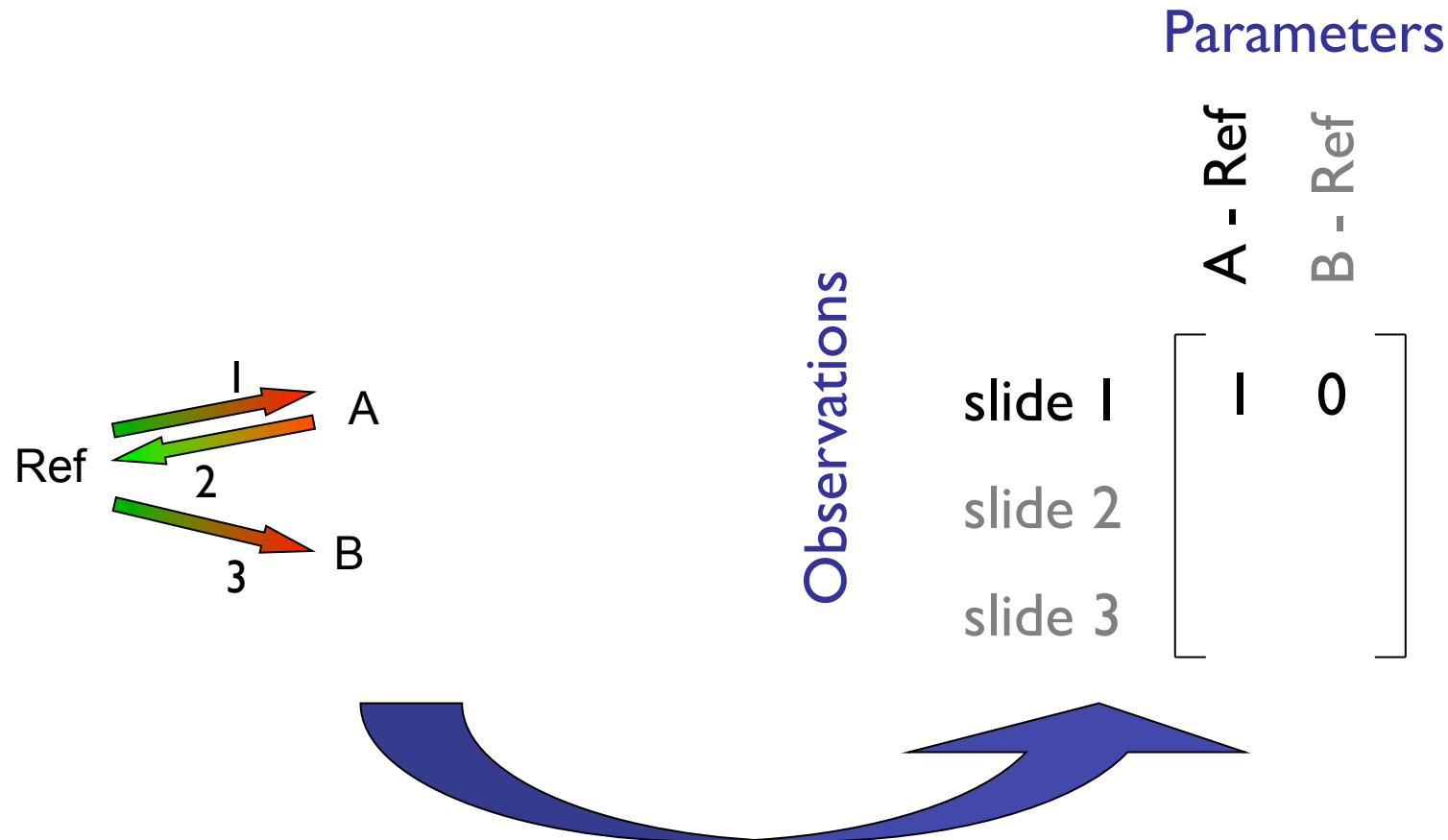


# Specifying the design as a matrix



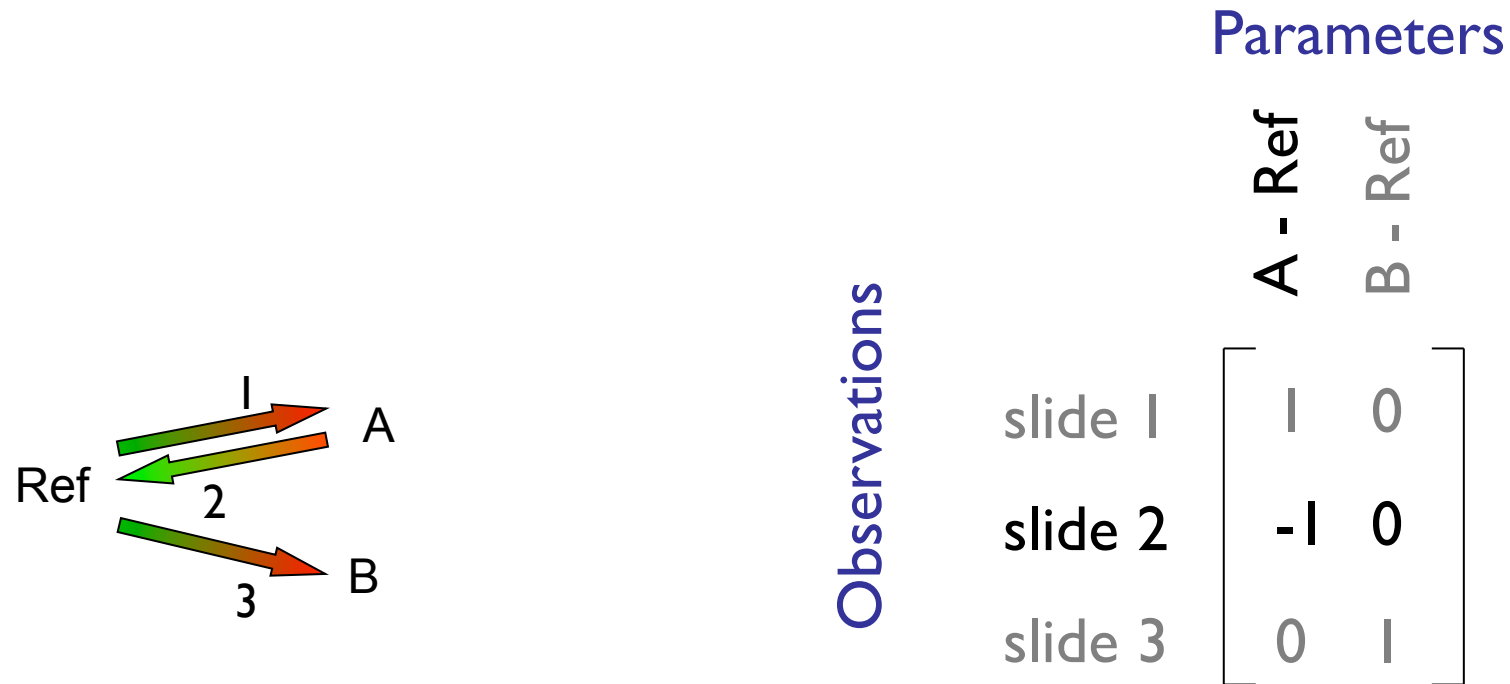
Represent log-ratio from each slide by a parameter => specify the model for your data

# Specifying the design as a matrix



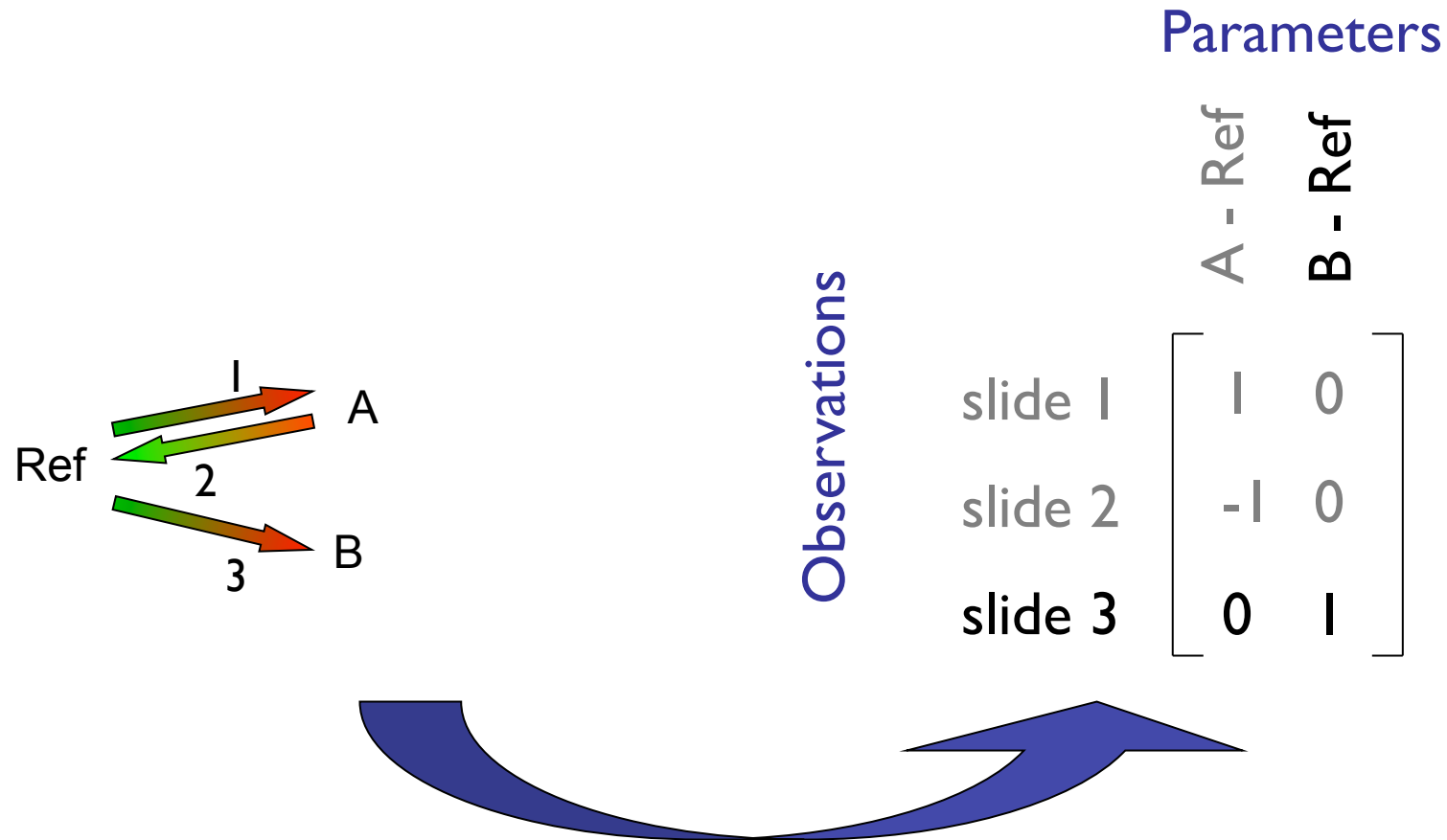
Represent log-ratio from each slide by a parameter => specify the model for your data

# Specifying the design as a matrix



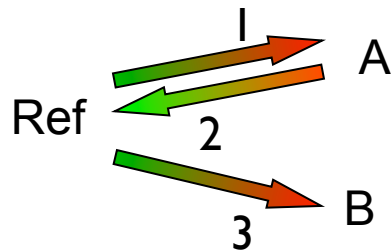
Represent log-ratio from each slide by a parameter => specify the model for your data

# Specifying the design as a matrix



Represent log-ratio from each slide by a parameter => specify the model for your data

# Specifying the design as a matrix



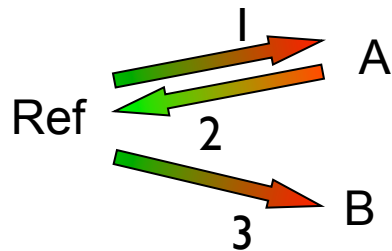
Observations

Parameters

	A - Ref	B - Ref
slide 1	1	0
slide 2	-1	0
slide 3	0	1

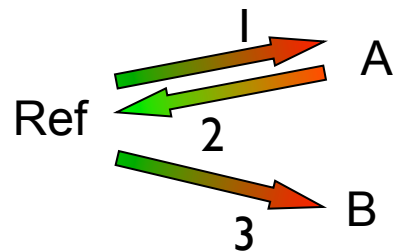
This is called the design matrix

# Write the model using design matrix



	A - Ref	B - Ref
slide 1	1	0
slide 2	-1	0
slide 3	0	1

# Write the model using design matrix



slide 1

slide 2

slide 3

$$\begin{bmatrix} 1 & 0 \\ -1 & 0 \\ 0 & 1 \end{bmatrix}$$

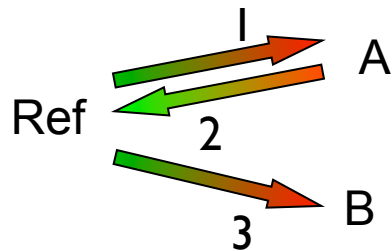
A - Ref

B - Ref

# Write the model using design matrix

Observed data modelled by these parameters

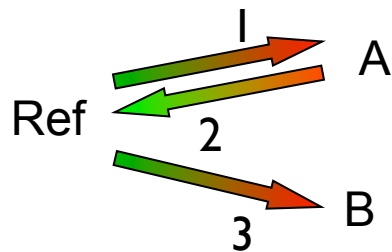
## Matrix notation



$$\begin{bmatrix} \text{slide 1} \\ \text{slide 2} \\ \text{slide 3} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -1 & 0 \\ 0 & 1 \end{bmatrix} \times \begin{bmatrix} A - \text{Ref} \\ B - \text{Ref} \end{bmatrix}$$



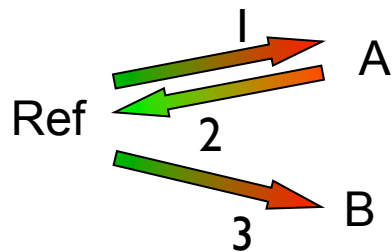
# Write the model using design matrix



$$E \begin{bmatrix} y1 \\ y2 \\ y3 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -1 & 0 \\ 0 & 1 \end{bmatrix} \mathbf{X} \begin{bmatrix} a \\ b \end{bmatrix}$$

# Write the model using design matrix

## Matrix multiplication

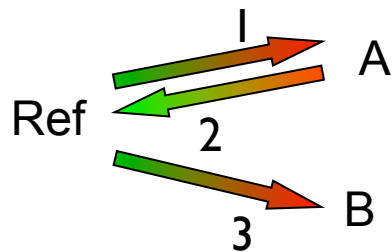


$$E \begin{bmatrix} y1 \\ y2 \\ y3 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -1 & 0 \\ 0 & 1 \end{bmatrix} \times \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 1 \times a + 0 \times b \\ -1 \times a + 0 \times b \\ 0 \times a + 1 \times b \end{bmatrix}$$

$$= \begin{bmatrix} a \\ -a \\ b \end{bmatrix}$$

# Write the model using design matrix

## Matrix multiplication

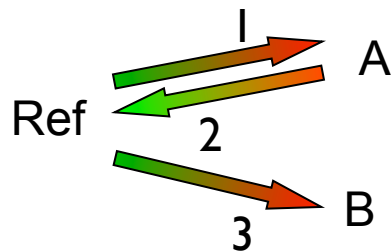


$$E \begin{bmatrix} y1 \\ y2 \\ y3 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -1 & 0 \\ 0 & 1 \end{bmatrix} \times \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 1 \times a + 0 \times b \\ -1 \times a + 0 \times b \\ 0 \times a + 1 \times b \end{bmatrix}$$

$$= \begin{bmatrix} a \\ -a \\ b \end{bmatrix}$$

# Write the model using design matrix

## Matrix multiplication



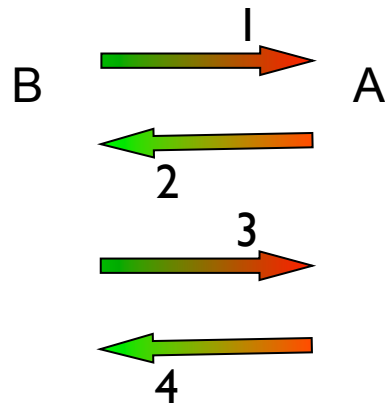
$$E \begin{bmatrix} y1 \\ y2 \\ y3 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -1 & 0 \\ 0 & 1 \end{bmatrix} \times \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 1 \times a + 0 \times b \\ -1 \times a + 0 \times b \\ 0 \times a + 1 \times b \end{bmatrix}$$

$$= \begin{bmatrix} a \\ -a \\ b \end{bmatrix}$$

# Specifying the design

## Sample types

I. Represent the effect measured by each sample type.



Single-channel representation  
(2 types of samples)

$$A = a$$

$$B = b$$

## Possible parameters

I. What differences are important?

Log-ratio representation (choose  $2 - 1 = 1$  parameters)

$$A - B = a - b$$

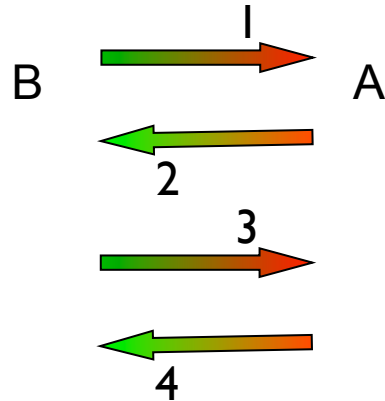
$$B - A = b - a$$

Choose one parameter to model your data.  
Write the design matrix for this experiment.

# Strategy to get design matrix

1. Write what is estimated in each array in terms of what is **measured in each sample type**.
2. Write what is estimated in each array in terms of a **combination of the parameters**.
3. Write the **multipliers** for the combinations of parameters as columns in a matrix.

# Write the model using design matrix



Samples = 2

Parameter = 1

$$A = a$$

$$B - A = b - a$$

$$B = b$$

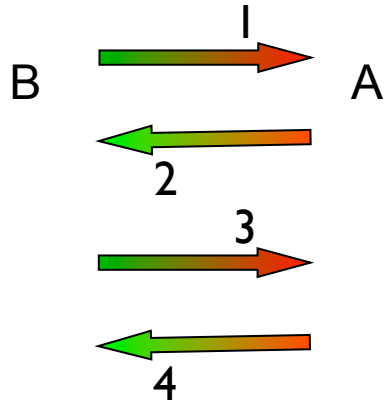
$$E \begin{bmatrix} y1 \\ y2 \\ y3 \\ y4 \end{bmatrix} = \begin{bmatrix} \\ \\ \\ \end{bmatrix} \times \begin{bmatrix} b-a \end{bmatrix}$$

Y

X

$\beta$

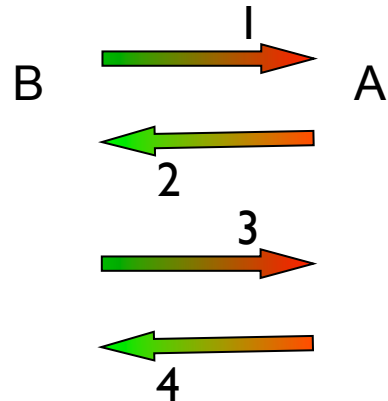
# Write the model using design matrix



$$\begin{array}{c}
 \text{E} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} \\
 \text{Y}
 \end{array}
 =
 \begin{array}{c}
 \begin{bmatrix} -1 \\ 1 \\ -1 \\ 1 \end{bmatrix} \\
 \text{X}
 \end{array}
 \times
 \begin{array}{c}
 \boxed{b-a} \\
 \beta
 \end{array}
 =
 \begin{array}{c}
 \begin{bmatrix} -1 \times (b-a) \\ 1 \times (b-a) \\ -1 \times (b-a) \\ 1 \times (b-a) \end{bmatrix} \\
 \end{array}
 =
 \begin{array}{c}
 \begin{bmatrix} a-b \\ b-a \\ a-b \\ b-a \end{bmatrix}
 \end{array}$$

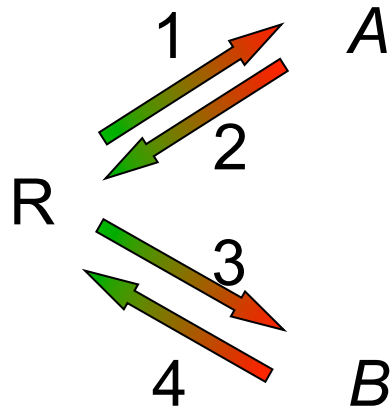


# Write the model using design matrix



$$\begin{array}{c}
 \mathbf{E} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} \\
 \mathbf{Y}
 \end{array}
 =
 \begin{array}{c}
 \begin{bmatrix} 1 \\ -1 \\ 1 \\ -1 \end{bmatrix} \\
 \mathbf{X}
 \end{array}
 \times
 \begin{array}{c}
 \boxed{a-b} \\
 \beta
 \end{array}
 =
 \begin{array}{c}
 \begin{bmatrix} 1 \times (a - b) \\ -1 \times (a - b) \\ 1 \times (a - b) \\ -1 \times (a - b) \end{bmatrix} \\
 \end{array}
 =
 \begin{array}{c}
 \begin{bmatrix} a-b \\ b-a \\ a-b \\ b-a \end{bmatrix}
 \end{array}$$

# Specifying the design



## Sample types

I. Represent the effect measured by each sample type.

Single-channel representation  
(3 types of samples)

$$A = \text{base} + a$$

$$B = \text{base} + b$$

$$R = \text{base}$$

## Possible parameters

I. What differences are important?

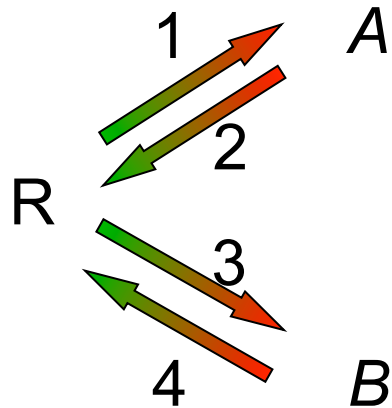
Log-ratio representation (choose 3-1 = 2 parameters)

$$A - B = \text{base} + a - (\text{base} + b)$$

$$A - R = a$$

$$B - R = b \quad + \text{ more}$$

# Specifying the design (alternative)



## Sample types

I. Represent the effect measured by each sample type.

## Possible parameters

I. What differences are important?

Single-channel representation  
(3 types of samples)

$$A = a$$

$$B = b$$

$$R = r$$

Log-ratio representation (choose 3-1 = 2 parameters)

$$A - B = a - b$$

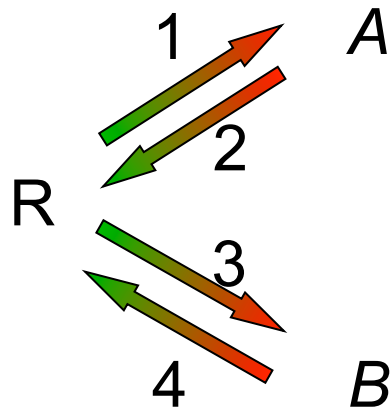
$$A - R = a - r$$

$$B - R = b - r \quad + \text{ more}$$

Choose two parameters to model your data.

Write the design matrix for this experiment.

# Write the model using design matrix



Samples = 3

Parameters = 2

$$A = a$$

$$B = b$$

$$R = r$$

$$A - R = a - r$$

$$A - B = a - b$$

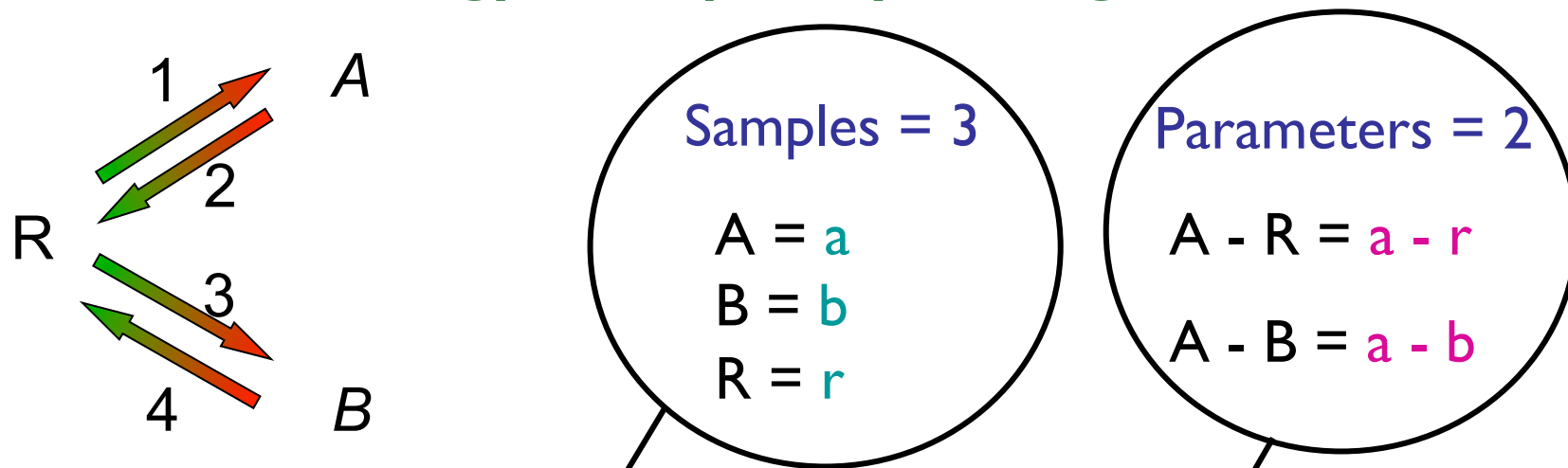
$$E \begin{bmatrix} y1 \\ y2 \\ y3 \\ y4 \end{bmatrix} = \begin{bmatrix} \phantom{0} \\ \phantom{0} \\ \phantom{0} \\ \phantom{0} \end{bmatrix} \times \begin{bmatrix} a-r \\ a-b \end{bmatrix} =$$

Y                      X                      β

# Strategy to get design matrix

1. Write what is estimated in each array in terms of what is **measured in each sample type**.
2. Write what is estimated in each array in terms of a **combination of the parameters**.
3. Write the **multipliers** for the combinations of parameters as columns in a matrix.

# Strategy to specify design matrix



Step 1.

Step 2.

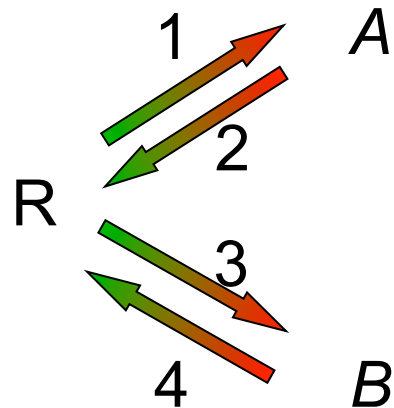
Step 3.

array

1	$(a) - (r) = a - r$	$= 1x(a-r) + 0x(a-b)$
2	$(r) - (a) = r - a$	$= -1x(a-r) + 0x(a-b)$
3	$(b) - (r) = b - r$	$= 1x(a-r) + -1x(a-b)$
4	$(r) - (b) = r - b$	$= -1x(a-r) + 1x(a-b)$

$$\begin{bmatrix} 1 & 0 \\ -1 & 0 \\ 1 & -1 \\ -1 & 1 \end{bmatrix}$$

# Write the model using design matrix



Samples = 3

Parameters = 2

$$A = a$$

$$B = b$$

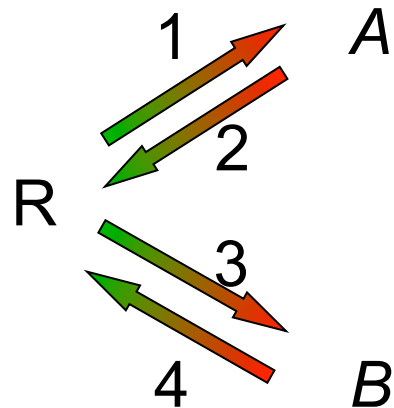
$$R = r$$

$$A - R = a - r$$

$$A - B = a - b$$

$$\begin{array}{c} E \\ Y \end{array} \begin{bmatrix} y1 \\ y2 \\ y3 \\ y4 \end{bmatrix} = \begin{array}{c} \\ \\ \\ X \end{array} \begin{bmatrix} 1 & 0 \\ -1 & 0 \\ 1 & -1 \\ -1 & 1 \end{bmatrix} \times \begin{array}{c} \beta \\ \\ \\ \end{array} \begin{bmatrix} a-r \\ a-b \end{bmatrix} = \begin{bmatrix} (a-r)+0 \\ -1(a-r)+0 \\ (a-r)-(a-b) \\ -1(a-r)+(a-b) \end{bmatrix} = \begin{bmatrix} a-r \\ r-a \\ b-r \\ r-b \end{bmatrix}$$

# design matrix with alternative parameterisation



Samples = 3

Parameters = 2

$$A = r + a$$

$$B = r + b$$

$$R = r$$

$$A - R = a$$

$$B - R = b$$

$$E \begin{bmatrix} y1 \\ y2 \\ y3 \\ y4 \end{bmatrix} = \begin{bmatrix} \phantom{0} \\ \phantom{0} \\ \phantom{0} \\ \phantom{0} \end{bmatrix} \times \begin{bmatrix} a \\ b \end{bmatrix}$$

Y

X

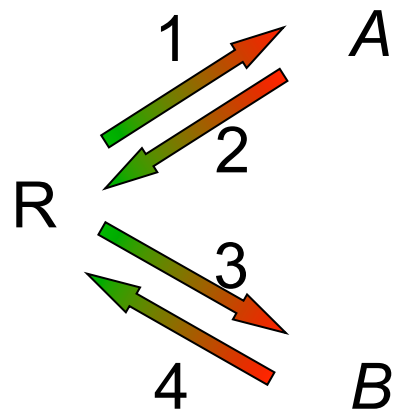
$\beta$



# Strategy to follow

1. Write what is estimated in each array in terms of what is **measured in each sample type**.
2. Write what is estimated in each array in terms of a **combination of the parameters**.
3. Write the **multipliers** for the combinations of parameters as columns in a matrix.

# Strategy to specify design matrix



Samples = 3

$$\begin{aligned} A &= r+a \\ B &= r+b \\ R &= r \end{aligned}$$

Parameters = 2

$$\begin{aligned} A - R &= a \\ B - R &= b \end{aligned}$$

Step 1.

Step 2.

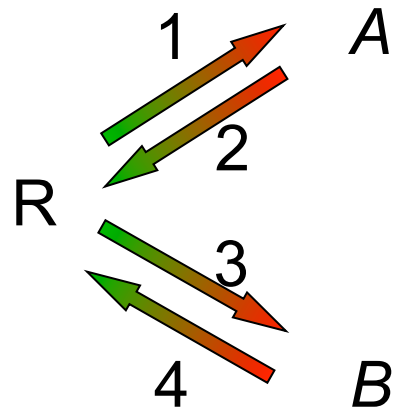
Step 3.

array

1	$(r+a) - r = a$	$= 1x(a) + 0x(b)$
2	$r - (r+a) = -a$	$= -1x(a) + 0x(b)$
3	$(r+b) - r = b$	$= 0x(a) + 1x(b)$
4	$r - (r+b) = -b$	$= 0x(a) + -1x(b)$

$$\begin{bmatrix} 1 & 0 \\ -1 & 0 \\ 0 & 1 \\ 0 & -1 \end{bmatrix}$$

# design matrix with alternative parameterisation



Samples = 3

Parameters = 2

$$A = r + a$$

$$B = r + b$$

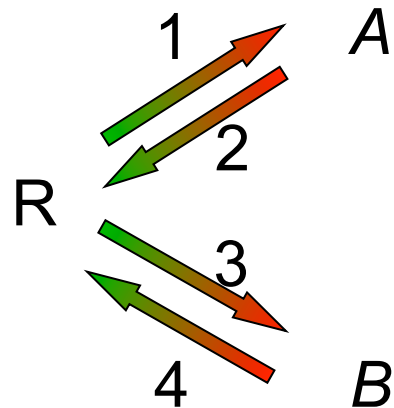
$$R = r$$

$$A - R = a$$

$$B - R = b$$

$$\begin{array}{c} E \\ Y \end{array} \begin{bmatrix} y1 \\ y2 \\ y3 \\ y4 \end{bmatrix} = \begin{array}{c} \\ X \end{array} \begin{bmatrix} 1 & 0 \\ -1 & 0 \\ 0 & 1 \\ 0 & -1 \end{bmatrix} \times \begin{array}{c} \beta \\ \begin{bmatrix} a \\ b \end{bmatrix} \end{array} = \begin{bmatrix} a+0 \\ -a+0 \\ 0+b \\ 0-b \end{bmatrix} = \begin{bmatrix} a \\ -a \\ b \\ -b \end{bmatrix}$$

# Specifying a contrast matrix



Samples = 3

Parameters = 2

$$A = r + a$$

$$A - R = a$$

$$B = r + b$$

$$B - R = b$$

$$R = r$$

$$E \begin{bmatrix} y1 \\ y2 \\ y3 \\ y4 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -1 & 0 \\ 0 & 1 \\ 0 & -1 \end{bmatrix} \times \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} a \\ -a \\ b \\ -b \end{bmatrix}$$

Y

X

$\beta$

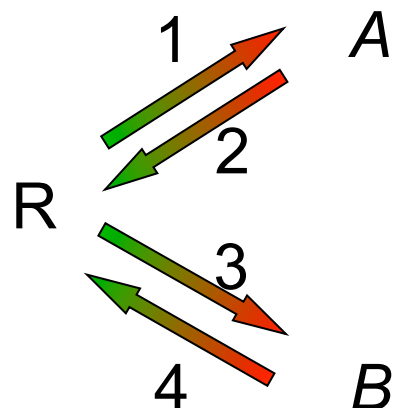
Linear model  
estimates of  
parameters



$$\begin{bmatrix} \hat{a} \\ \hat{b} \end{bmatrix}$$

Parameter  
estimates  
(called coefficients  
in limma)

# Specifying a contrast matrix



Samples = 3

Parameters = 2

$$A = r + a$$

$$A - R = a$$

$$B = r + b$$

$$B - R = b$$

$$R = r$$

$$E \begin{bmatrix} y1 \\ y2 \\ y3 \\ y4 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -1 & 0 \\ 0 & 1 \\ 0 & -1 \end{bmatrix} \times \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} a \\ -a \\ b \\ -b \end{bmatrix}$$

Y                      X                      β

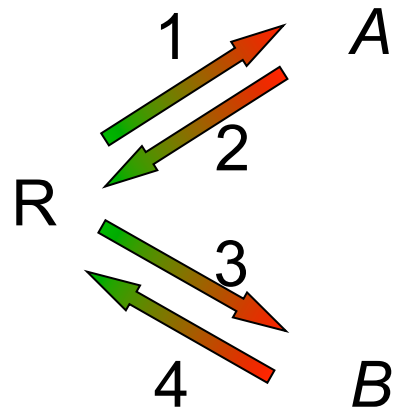
$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & -1 \\ .5 & .5 \end{bmatrix} \times \begin{bmatrix} \hat{a} \\ \hat{b} \end{bmatrix}$$

Parameter estimates

(called coefficients in limma)

Contrast matrix

# Specifying a contrast matrix



Samples = 3

Parameters = 2

$$A = r + a$$

$$A - R = a$$

$$B = r + b$$

$$B - R = b$$

$$R = r$$

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & -1 \\ .5 & .5 \end{bmatrix}$$

$$\times \begin{bmatrix} \hat{a} \\ \hat{b} \end{bmatrix}$$

=

$$\begin{bmatrix} \hat{a} \\ \hat{b} \\ \hat{a} - \hat{b} \\ .5(\hat{a} + \hat{b}) \end{bmatrix}$$

→

A

→

B

→

A - B

→

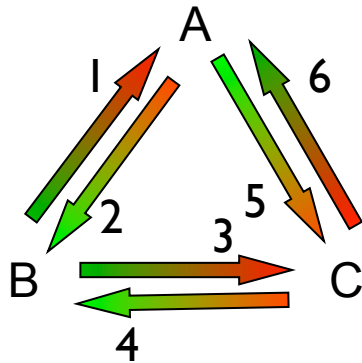
1/2(A + B)

Contrast  
matrix

Parameter  
estimates  
(called coefficients  
in limma)

Contrasts  
of interest

# Specifying the design



## Sample types

I. Represent the effect measured by each sample type.

Single-channel representation  
(3 types of samples)

$$A = a$$

$$B = b$$

$$C = c$$

## Possible parameters

I. What differences are important?

Log-ratio representation (choose  $3 - 1 = 2$  parameters)

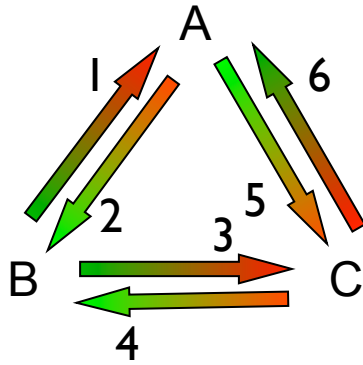
$$A - B = a - b$$

$$B - C = b - c$$

$$C - A = c - a \quad + \text{ more}$$

Choose two parameters to model your data.  
Write the design matrix for this experiment.

# Write the model using design matrix



Samples = 3

A = a

B = b

C = c

Parameters = 2

A - B = a - b

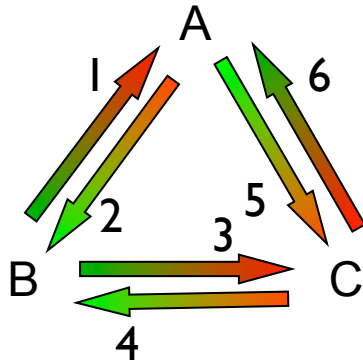
B - C = b - c

$$E \begin{bmatrix} y1 \\ y2 \\ y3 \\ y4 \\ y5 \\ y6 \end{bmatrix} = \begin{bmatrix} \\ \\ \\ \\ \\ \end{bmatrix} \times \begin{bmatrix} a-b \\ b-c \end{bmatrix}$$

Y                      X                      β



# Write the model using design matrix



Samples = 3

Parameters = 2

A = a

A - B = a - b

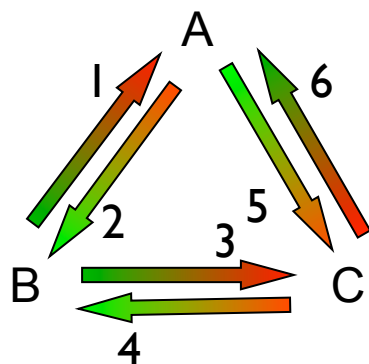
B = b

B - C = b - c

C = c

$$\begin{array}{c} E \\ \begin{bmatrix} y1 \\ y2 \\ y3 \\ y4 \\ y5 \\ y6 \end{bmatrix} \\ Y \end{array} = \begin{array}{c} \begin{bmatrix} 1 & 0 \\ -1 & 0 \\ 0 & -1 \\ 0 & 1 \\ -1 & -1 \\ 1 & 1 \end{bmatrix} \\ X \end{array} \times \begin{array}{c} \begin{bmatrix} a-b \\ b-c \end{bmatrix} \\ \beta \end{array} = \begin{array}{c} \begin{bmatrix} (a-b)+0 \\ 0-1(a-b) \\ 0-1(b-c) \\ 0+1(b-c) \\ -1(a-b)-(b-c) \\ (a-b)+(b-c) \end{bmatrix} \\ \end{array} = \begin{array}{c} \begin{bmatrix} a-b \\ b-a \\ c-b \\ b-c \\ c-a \\ a-c \end{bmatrix} \end{array}$$

# Specify contrasts of interest



Samples = 3

A = a

B = b

C = c

Parameters = 2

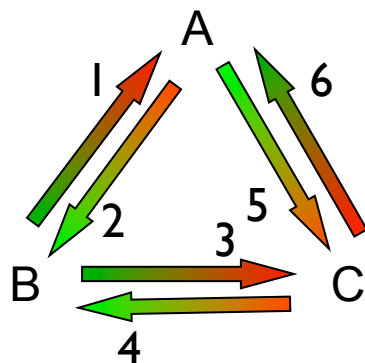
A - B = a - b

B - C = b - c

$$\begin{array}{c}
 E \\
 \begin{bmatrix} y1 \\ y2 \\ y3 \\ y4 \\ y5 \\ y6 \end{bmatrix} \\
 Y
 \end{array}
 =
 \begin{array}{c}
 \begin{bmatrix} 1 & 0 \\ -1 & 0 \\ 0 & -1 \\ 0 & 1 \\ -1 & -1 \\ 1 & 1 \end{bmatrix} \\
 X
 \end{array}
 \times
 \begin{array}{c}
 \begin{bmatrix} a-b \\ b-c \end{bmatrix} \\
 \beta
 \end{array}
 \begin{array}{c}
 \begin{bmatrix} \\ \\ \\ \\ \\ \end{bmatrix} \\
 \text{Contrast} \\
 \text{matrix}
 \end{array}
 \times
 \begin{array}{c}
 \begin{bmatrix} \hat{a}-b \\ \hat{b}-c \end{bmatrix} \\
 \text{Parameter} \\
 \text{estimates} \\
 \text{(called coefficients} \\
 \text{in limma)}
 \end{array}
 =
 \begin{array}{c}
 \begin{bmatrix} \hat{a}-b \\ \hat{b}-c \\ \hat{a}-c \end{bmatrix} \\
 \text{Contrasts} \\
 \text{of interest}
 \end{array}$$

In limma, the contrast matrix is the transpose of the above!!  
 i.e. rows become the columns and the columns become the rows

# Specify contrasts of interest



Samples = 3

A = a

B = b

C = c

Parameters = 2

A - B = a - b

B - C = b - c

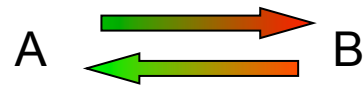
$$\begin{array}{c}
 \begin{bmatrix} y1 \\ y2 \\ y3 \\ y4 \\ y5 \\ y6 \end{bmatrix} \\
 \text{Y}
 \end{array}
 =
 \begin{array}{c}
 \begin{bmatrix} 1 & 0 \\ -1 & 0 \\ 0 & -1 \\ 0 & 1 \\ -1 & -1 \\ 1 & 1 \end{bmatrix} \\
 \text{X}
 \end{array}
 \times
 \begin{array}{c}
 \begin{bmatrix} a-b \\ b-c \end{bmatrix} \\
 \beta
 \end{array}
 \\
 \\
 \begin{array}{c}
 \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{bmatrix} \\
 \text{Contrast} \\
 \text{matrix}
 \end{array}
 \times
 \begin{array}{c}
 \begin{bmatrix} \hat{a}-b \\ \hat{b}-c \end{bmatrix} \\
 \text{Parameter} \\
 \text{estimates} \\
 \text{(called coefficients} \\
 \text{in limma)}
 \end{array}
 =
 \begin{array}{c}
 \begin{bmatrix} \hat{a}-b \\ \hat{b}-c \\ \hat{a}-c \end{bmatrix} \\
 \text{Contrasts} \\
 \text{of interest}
 \end{array}
 \end{array}$$

In limma, the contrast matrix is the transpose of the above!!  
 i.e. rows become the columns and the columns become the rows

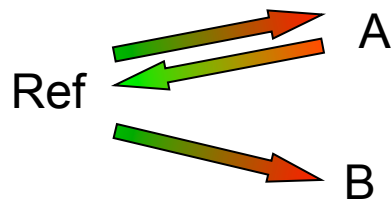
# Linear models for differential expression



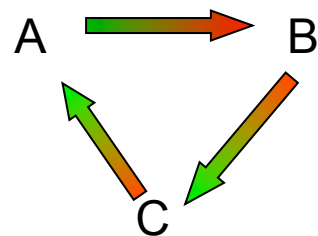
$$y = \log_2(R) - \log_2(G) \equiv B - A$$



$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} 1 \\ -1 \end{pmatrix} \beta \quad \beta \equiv B - A$$



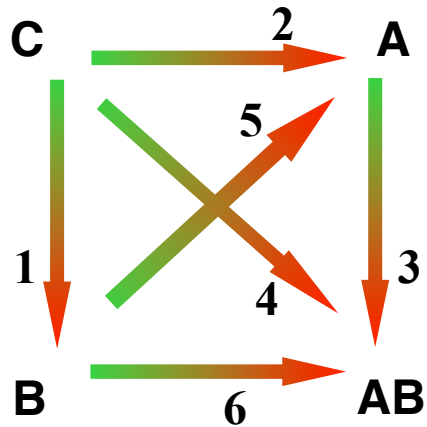
$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ -1 & 0 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \quad \begin{aligned} \beta_1 &\equiv A - \text{Ref} \\ \beta_2 &\equiv B - A \end{aligned}$$



$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ -1 & 1 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \quad \begin{aligned} \beta_1 &\equiv B - A \\ \beta_2 &\equiv C - A \end{aligned}$$

Allows all comparisons to be estimated simultaneously

## Factorial experiment: one sample as a common reference



Samples = 4

Parameters = 3

$$A = \text{base} + a$$

$$A - C = a$$

$$B = \text{base} + b$$

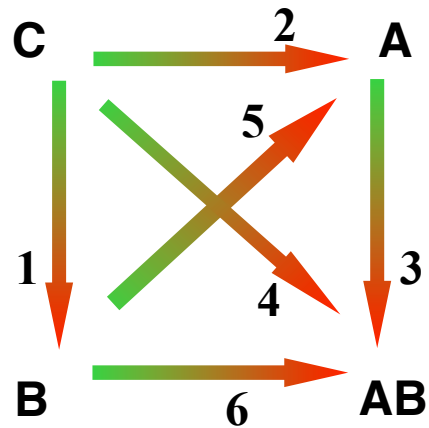
$$B - C = b$$

$$C = \text{base}$$

$$AB = \text{base} + ab$$

$$AB - C = ab$$

## Factorial experiment: one sample as a common reference



Samples = 4

Parameters = 3

$$A = \text{base} + a$$

$$A - C = a$$

$$B = \text{base} + b$$

$$B - C = b$$

$$C = \text{base}$$

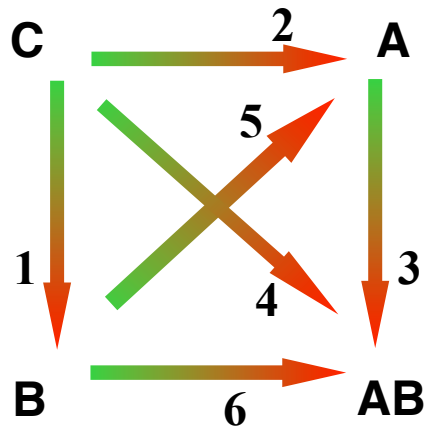
$$AB = \text{base} + ab$$

$$AB - C = ab$$

$$E \begin{bmatrix} y1 \\ y2 \\ y3 \\ y4 \\ y5 \\ y6 \end{bmatrix} = \begin{bmatrix} \\ \\ \\ \\ \\ \end{bmatrix} \times \begin{bmatrix} a \\ b \\ ab \end{bmatrix} =$$

Y                      X                      β

## Factorial experiment: one sample as a common reference



Samples = 4

Parameters = 3

$$A = \text{base} + a$$

$$A - C = a$$

$$B = \text{base} + b$$

$$B - C = b$$

$$C = \text{base}$$

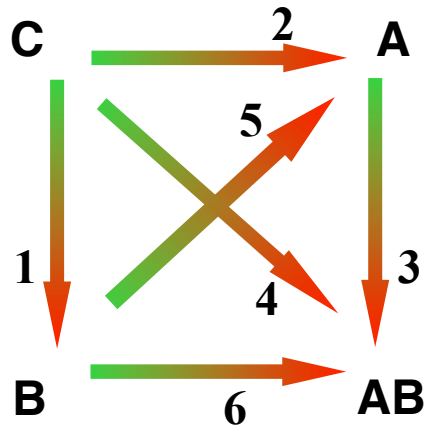
$$AB = \text{base} + ab$$

$$AB - C = ab$$

$$E \begin{bmatrix} y1 \\ y2 \\ y3 \\ y4 \\ y5 \\ y6 \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ -1 & 0 & 1 \\ 0 & 0 & 1 \\ 1 & -1 & 0 \\ 0 & -1 & 1 \end{bmatrix} \times \begin{bmatrix} a \\ b \\ ab \end{bmatrix} = \begin{bmatrix} 0 + b + 0 \\ a + 0 + 0 \\ -a + 0 + ab \\ 0 + 0 + ab \\ a - b + 0 \\ 0 - b + ab \end{bmatrix} = \begin{bmatrix} b \\ a \\ -a + ab \\ 0 + 0 + ab \\ a - b \\ -b + ab \end{bmatrix}$$

$Y$ 
 $X$ 
 $\beta$

# Design for factorial experiment with interaction



Samples = 4

Parameters = 3

$$A = c + a$$

$$A - C = a$$

$$B = c + b$$

$$B - C = b$$

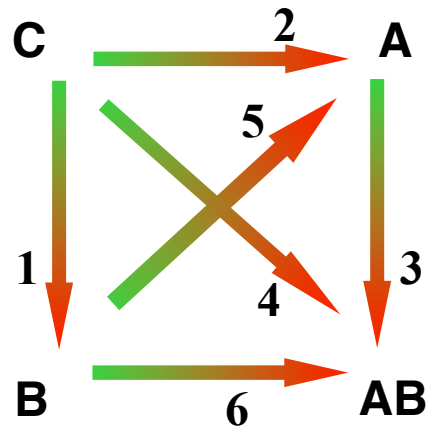
$$C = c$$

$$AB = c + a + b + ab$$

$$AB - A - B + C = ab$$



# Design for factorial experiment with interaction



Samples = 4

Parameters = 3

$$A = c + a$$

$$A - C = a$$

$$B = c + b$$

$$B - C = b$$

$$C = c$$

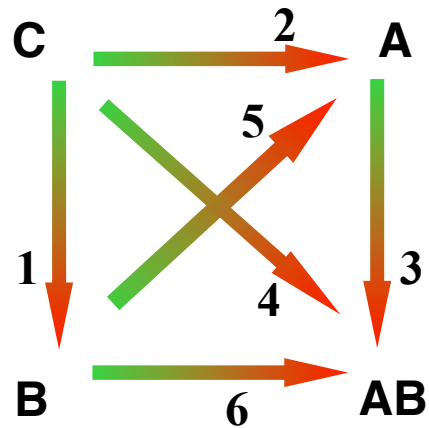
$$AB = c + a + b + ab$$

$$AB - A - B + C = ab$$

$$E \begin{bmatrix} y1 \\ y2 \\ y3 \\ y4 \\ y5 \\ y6 \end{bmatrix} = \begin{bmatrix} \\ \\ \\ \\ \\ \end{bmatrix} \times \begin{bmatrix} a \\ b \\ ab \end{bmatrix} =$$

Y                      X                      β

# Design for factorial experiment with interaction



Samples = 4

Parameters = 3

$$A = c + a$$

$$A - C = a$$

$$B = c + b$$

$$B - C = b$$

$$C = c$$

$$AB = c + a + b + ab$$

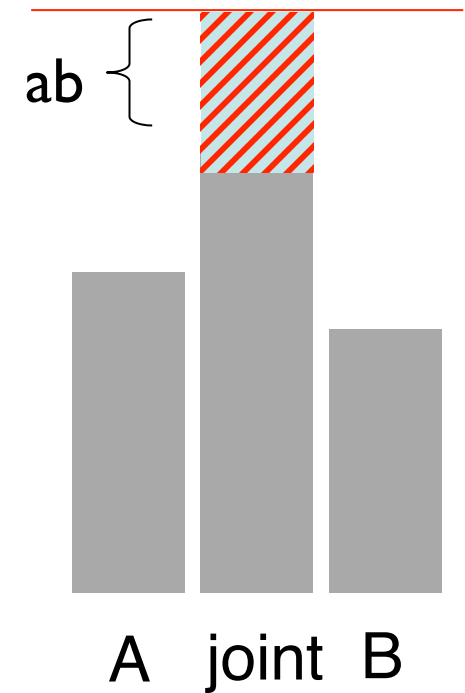
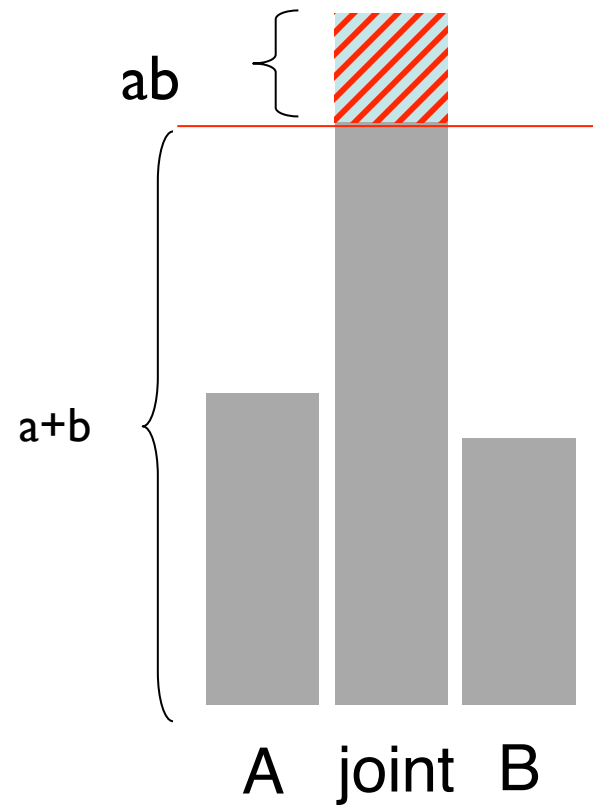
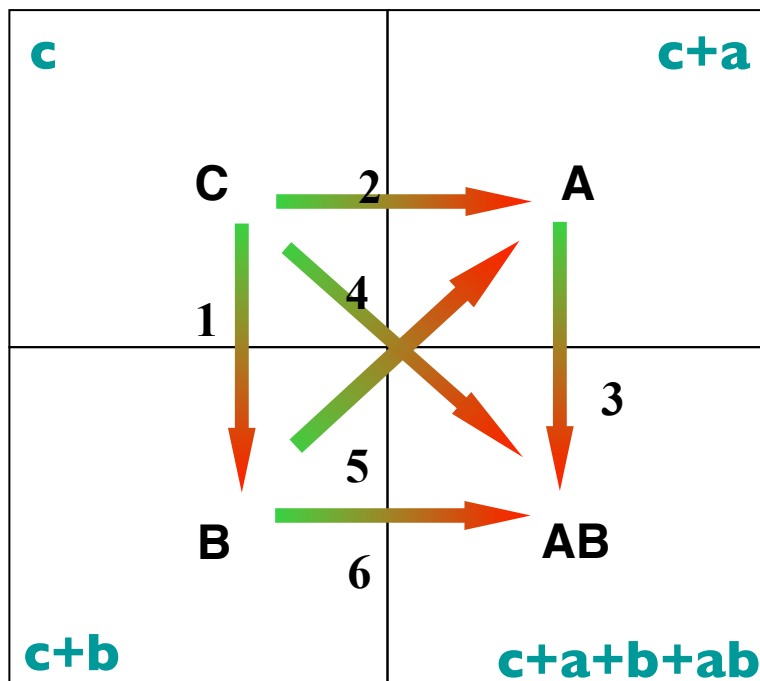
$$AB - A - B + C = ab$$

$$\begin{array}{c}
 E \\
 \begin{bmatrix} y1 \\ y2 \\ y3 \\ y4 \\ y5 \\ y6 \end{bmatrix} \\
 Y
 \end{array}
 =
 \begin{array}{c}
 \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & -1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \\
 X \quad \beta
 \end{array}
 \times
 \begin{array}{c}
 \begin{bmatrix} a \\ b \\ ab \end{bmatrix}
 \end{array}
 =
 \begin{array}{c}
 \begin{bmatrix} 0 + b + 0 \\ a + 0 + 0 \\ 0 + b + ab \\ a + b + ab \\ a - b + 0 \\ a + 0 + ab \end{bmatrix}
 \end{array}
 =
 \begin{array}{c}
 \begin{bmatrix} b \\ a \\ b+ab \\ a+b+ab \\ a-b \\ a+ab \end{bmatrix}
 \end{array}$$

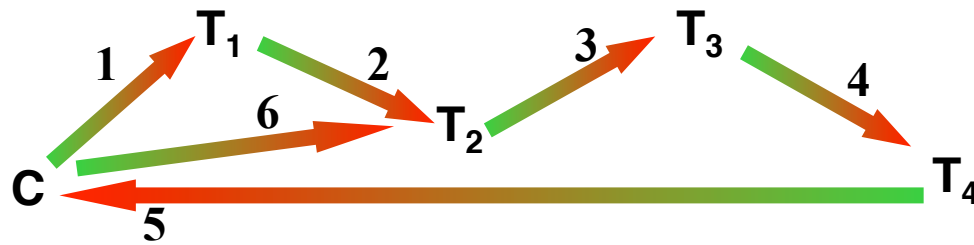
# Interaction

ab positive

ab negative



## Trend analysis



Samples = 5

Parameters = 1

$$\begin{aligned} C &= \text{base} \\ T_1 &= \text{base} + a \\ T_2 &= \text{base} + 2a \\ T_3 &= \text{base} + 3a \\ T_4 &= \text{base} + 4a \end{aligned}$$

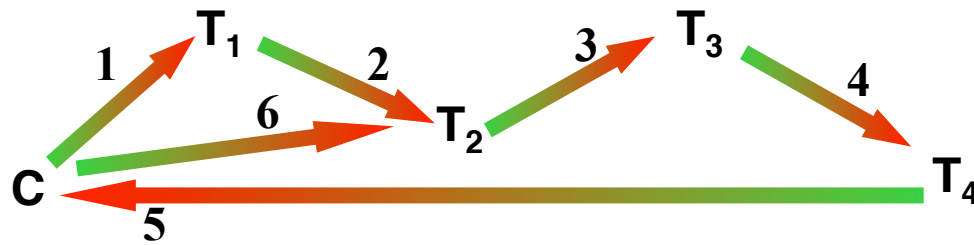
$$T_2 - T_1 = a$$

$$E \begin{bmatrix} y1 \\ y2 \\ y3 \\ y4 \\ y5 \\ y6 \end{bmatrix} = \begin{bmatrix} \phantom{0} \\ \phantom{0} \\ \phantom{0} \\ \phantom{0} \\ \phantom{0} \\ \phantom{0} \end{bmatrix} \times \begin{bmatrix} a \end{bmatrix} =$$

Y                      X                      β

Note: the possible number of parameters is 4, but we choose here to use only one parameter

## Trend analysis



Samples = 5

Parameters = 1

$C$  = base

$T_2 - T_1 = a$

$T_1$  = base + a

$T_2$  = base + 2a

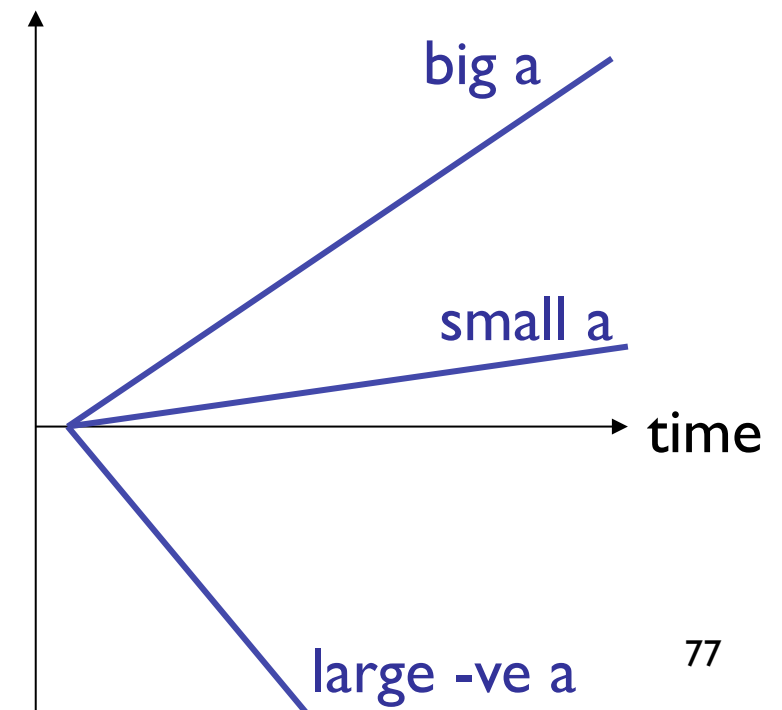
$T_3$  = base + 3a

$T_4$  = base + 4a

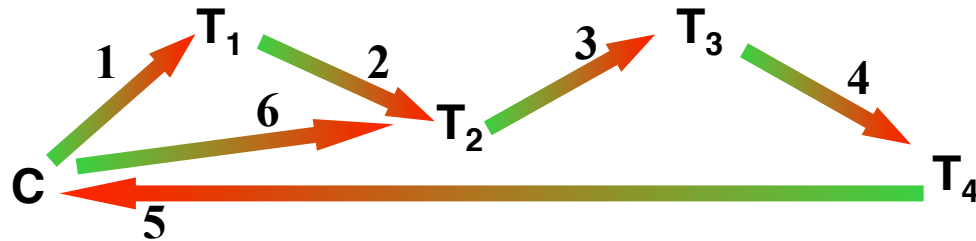
$$E \begin{bmatrix} y1 \\ y2 \\ y3 \\ y4 \\ y5 \\ y6 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ -4 \\ 2 \end{bmatrix} \times \begin{bmatrix} a \end{bmatrix} =$$

$Y \quad X \quad \beta$

straight line model is fitted



# Trend analysis



Samples = 5

Parameters = 1

$$C = \text{base} \quad T_1 - C = a$$

$$T_1 = \text{base} + a$$

$$T_2 = \text{base} + 4a$$

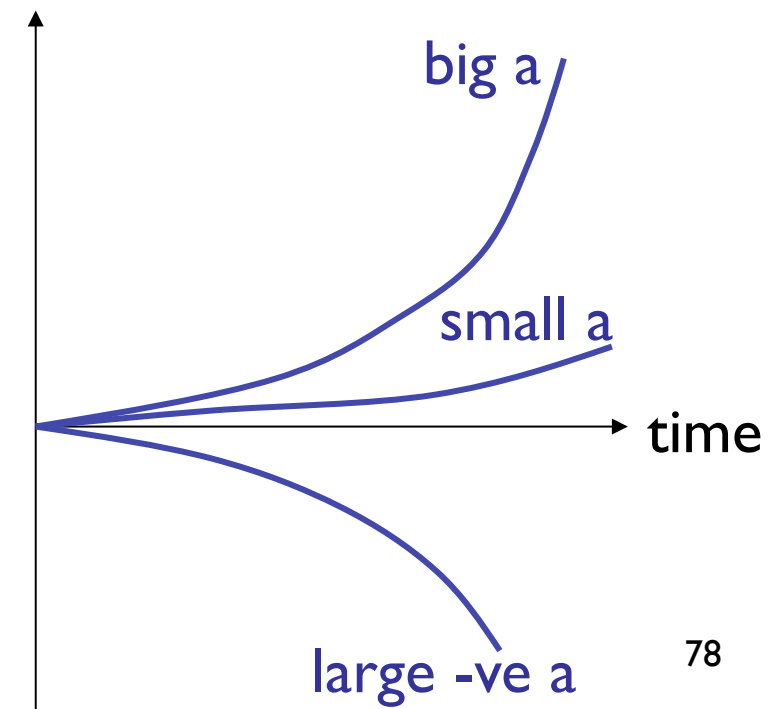
$$T_3 = \text{base} + 9a$$

$$T_4 = \text{base} + 16a$$

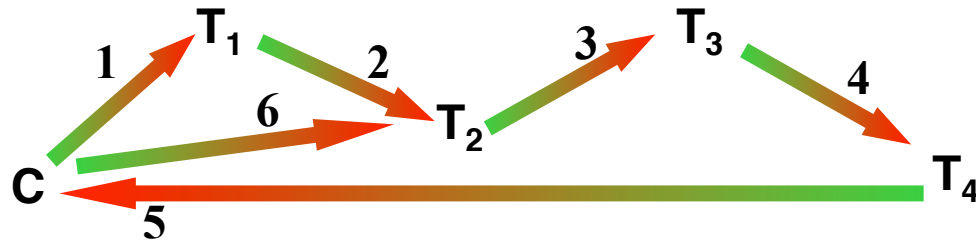
$$E \begin{bmatrix} y1 \\ y2 \\ y3 \\ y4 \\ y5 \\ y6 \end{bmatrix} = \begin{bmatrix} 1 \\ 3 \\ 5 \\ 7 \\ -16 \\ 4 \end{bmatrix} \times \begin{bmatrix} a \end{bmatrix} =$$

$\beta$

quadratic model is fitted



# Trend analysis



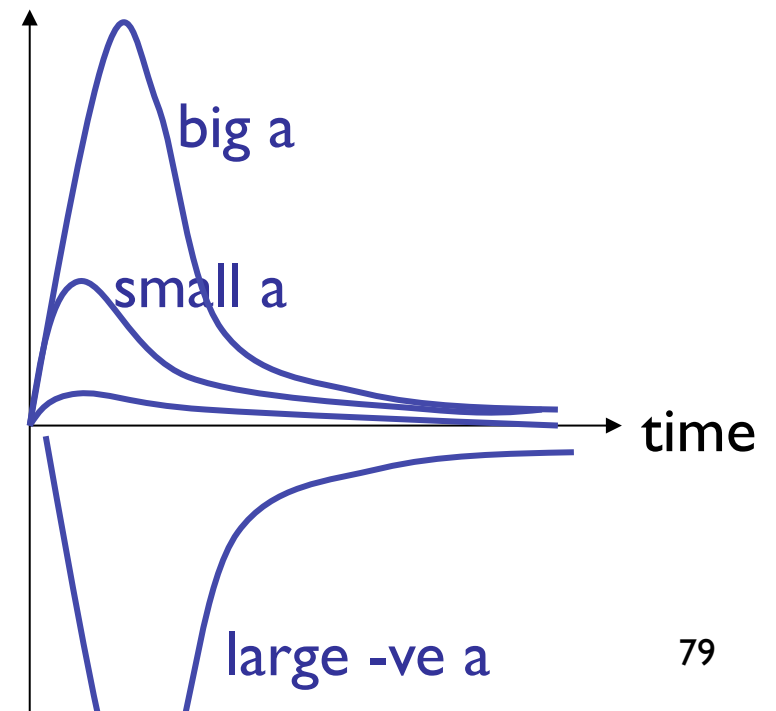
Samples = 5      Parameters = 1

$$\begin{aligned} C &= \text{base} & T_4 - C &= a \\ T_1 &= \text{base} + 16a \\ T_2 &= \text{base} + 9a \\ T_3 &= \text{base} + 4a \\ T_4 &= \text{base} + a \end{aligned}$$

$$E \begin{bmatrix} y1 \\ y2 \\ y3 \\ y4 \\ y5 \\ y6 \end{bmatrix} = \begin{bmatrix} 16 \\ -7 \\ -5 \\ -3 \\ -1 \\ 9 \end{bmatrix} \times \begin{bmatrix} a \end{bmatrix} =$$

$\begin{matrix} Y & X & \beta \end{matrix}$

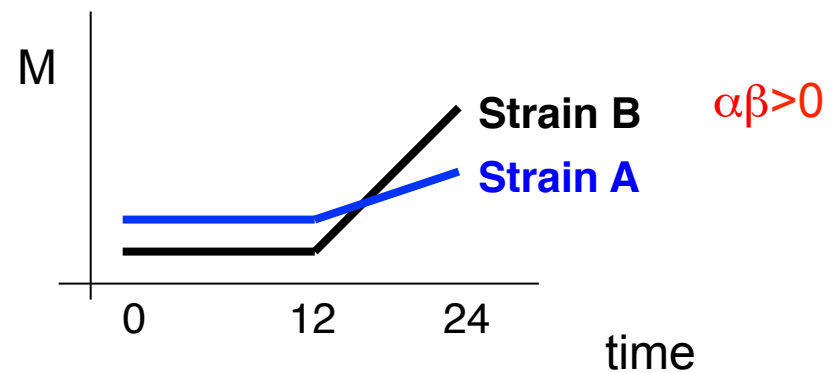
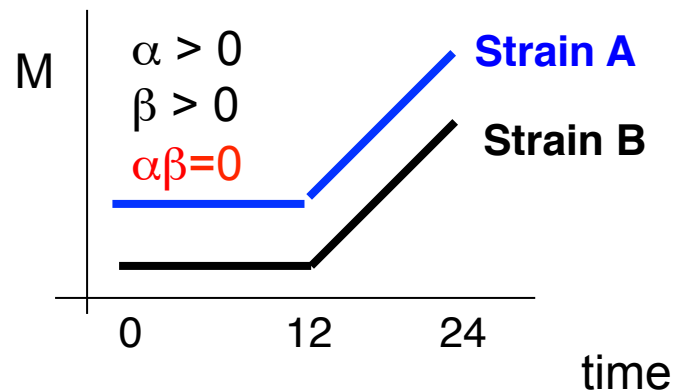
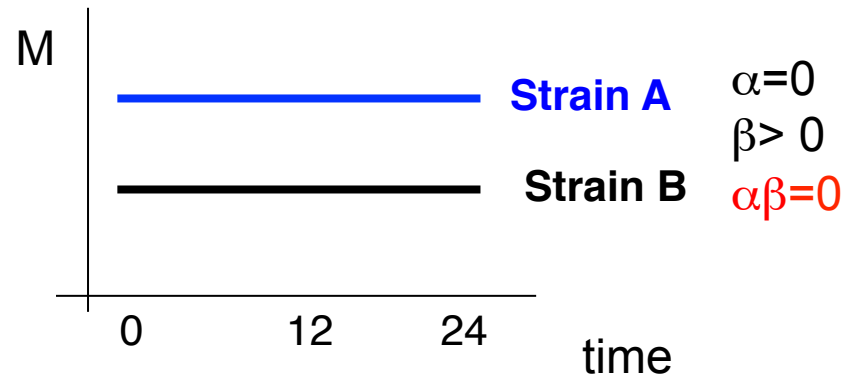
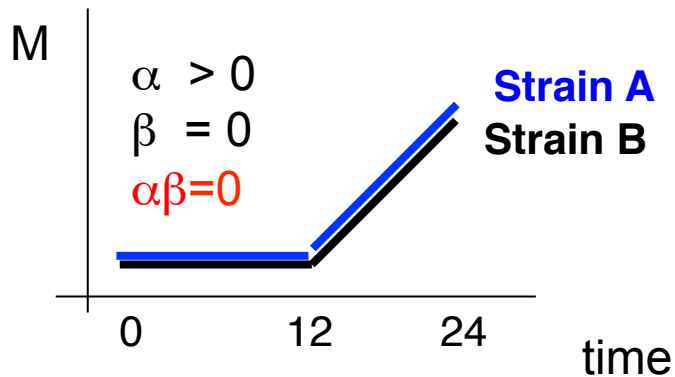
quadratic model is fitted



# 2 by 3 factorial experiment

- Identify DE genes that have different time profiles between different mutants.

$\alpha$  = time effect,  $\beta$  = strains,  $\alpha\beta$  = interaction effect





# Design matrix for single-colour arrays

		Samples			
		→			
		A	B	C	D
Replicates	1	1 ■	4 ■	7 ■	9 ■
	2	2 ■	5 ■	8 ■	10 ■
	3	3 ■	6 ■		

Squares represent single-colour arrays,  
numbers represent the array number.  
Observations (data) are log-intensities.

# Samples = # Parameters

I. Represent the  
effect measured by  
each sample.

Single-channel  
representation  
(4 types of samples)

A = a

B = b

C = c

D = d

A	B	C	D
1 ■	4 ■	7 ■	9 ■
2 ■	5 ■	8 ■	10 ■
3 ■	6 ■		

$$\begin{array}{c} \text{E} \end{array} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ y_8 \\ y_9 \\ y_{10} \end{bmatrix} = \begin{bmatrix} \phantom{y_1} \\ \phantom{y_2} \\ \phantom{y_3} \\ \phantom{y_4} \\ \phantom{y_5} \\ \phantom{y_6} \\ \phantom{y_7} \\ \phantom{y_8} \\ \phantom{y_9} \\ \phantom{y_{10}} \end{bmatrix} \times \begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix}$$

Y
X
 $\beta$

# Samples = # Parameters

A = a

B = b

C = c

D = d

Data are  
log-intensities  
NOT log-ratios

A	B	C	D
1	4	7	9
2	5	8	10
3	6		

# Samples = # Parameters

A = a

B = b

C = c

D = d

$$\begin{array}{c} \text{E} \end{array} \begin{bmatrix} y1 \\ y2 \\ y3 \\ y4 \\ y5 \\ y6 \\ y7 \\ y8 \\ y9 \\ y10 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} a \\ b \\ c \\ d \end{bmatrix} = \begin{bmatrix} a \\ a \\ a \\ b \\ b \\ b \\ c \\ c \\ d \\ d \end{bmatrix}$$

$\begin{matrix} Y & X & \beta \end{matrix}$

Data are  
log-intensities  
NOT log-ratios  
  
Design matrix is  
EASY!!!

A	B	C	D
1 ■	4 ■	7 ■	9 ■
2 ■	5 ■	8 ■	10 ■
3 ■	6 ■		

Contrast matrix

$$\begin{bmatrix} 1 & -1 & 0 & 0 \\ 1 & 0 & -1 & 0 \end{bmatrix} \times \begin{bmatrix} \hat{a} \\ \hat{b} \\ \hat{c} \\ \hat{d} \end{bmatrix} = \begin{bmatrix} \hat{a}-\hat{b} \\ \hat{a}-\hat{c} \\ \hat{a}-\hat{d} \\ \hat{b}-\hat{c} \\ \hat{b}-\hat{d} \end{bmatrix}$$

In limma, the contrast matrix is the transpose of the above!!  
 i.e. rows become the columns and the columns become the rows

C	A	B	AB
1 ■	4 ■	7 ■	9 ■
2 ■	5 ■	8 ■	10 ■
3 ■	6 ■		

# Interaction v I

$C = c$   
 $A = a$   
 $B = b$   
 $AB = ab$

$$\begin{array}{c} E \\ \begin{bmatrix} y1 \\ y2 \\ y3 \\ y4 \\ y5 \\ y6 \\ y7 \\ y8 \\ y9 \\ y10 \end{bmatrix} \\ Y \end{array} = \begin{array}{c} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \\ X \end{array} \times \begin{array}{c} \begin{bmatrix} c \\ a \\ b \\ ab \end{bmatrix} \\ \beta \end{array} = \begin{array}{c} \begin{bmatrix} c \\ c \\ c \\ a \\ a \\ a \\ b \\ b \\ ab \\ ab \end{bmatrix} \\ 85 \end{array}$$

	C	A	B	AB
1	■	4 ■	7 ■	9 ■
2	■	5 ■	8 ■	10 ■
3	■	6 ■		

$C = c$   
 $A = a$   
 $B = b$   
 $AB = ab$

Contrast matrix

$$\begin{bmatrix} -1 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 1 & -1 & -1 & 1 \end{bmatrix} \times \begin{bmatrix} \hat{c} \\ \hat{a} \\ \hat{b} \\ \hat{ab} \end{bmatrix} = \begin{bmatrix} \hat{a}-\hat{c} \\ \hat{b}-\hat{c} \\ ab-\hat{a}-\hat{b}+\hat{c} \end{bmatrix}$$

$\longrightarrow$  *Treatment A*  
 $\longrightarrow$  *Treatment B*  
 $\longrightarrow$  *Interaction*

C	A	B	AB
1 ■	4 ■	7 ■	9 ■
2 ■	5 ■	8 ■	10 ■
3 ■	6 ■		

## Interaction v2

$$C = c$$

$$A = c+a$$

$$B = c+b$$

$$AB = c+a+b+ab$$

$$\begin{array}{c} E \\ \begin{bmatrix} y1 \\ y2 \\ y3 \\ y4 \\ y5 \\ y6 \\ y7 \\ y8 \\ y9 \\ y10 \end{bmatrix} \\ Y \end{array} = \begin{array}{c} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix} \\ X \end{array} \times \begin{array}{c} \begin{bmatrix} c \\ a \\ b \\ ab \end{bmatrix} \\ \beta \end{array} = \begin{array}{c} \begin{bmatrix} c \\ c \\ c \\ c+a \\ c+a \\ c+a \\ c+b \\ c+b \\ c+a+b+ab \\ c+a+b+ab \end{bmatrix} \end{array}$$

	C	A	B	AB
1	■	4 ■	7 ■	9 ■
2	■	5 ■	8 ■	10 ■
3	■	6 ■		

$$C = c$$

$$A = c+a$$

$$B = c+b$$

$$AB = c+a+b+ab$$

Contrast matrix

$$\begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \times \begin{bmatrix} \hat{c} \\ \hat{a} \\ \hat{b} \\ \hat{ab} \end{bmatrix} = \begin{bmatrix} \hat{a} \\ \hat{b} \\ \hat{ab} \end{bmatrix}$$

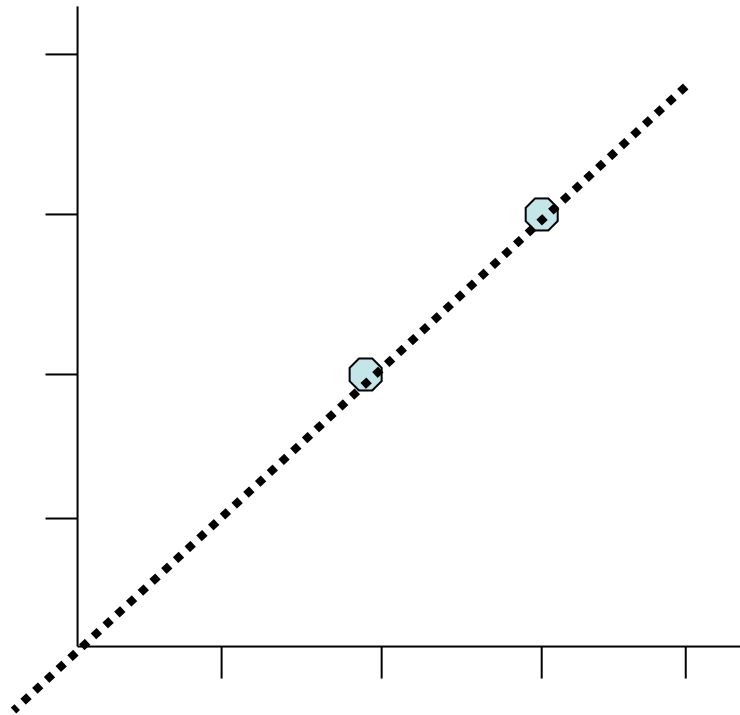
$\hat{a}$  → Treatment A  
 $\hat{b}$  → Treatment B  
 $\hat{ab}$  → Interaction

Red arrows indicate:
 

- $\hat{c}$  is labeled "log-intensity".
- $\hat{a}$ ,  $\hat{b}$ , and  $\hat{ab}$  are collectively labeled "log-ratios".

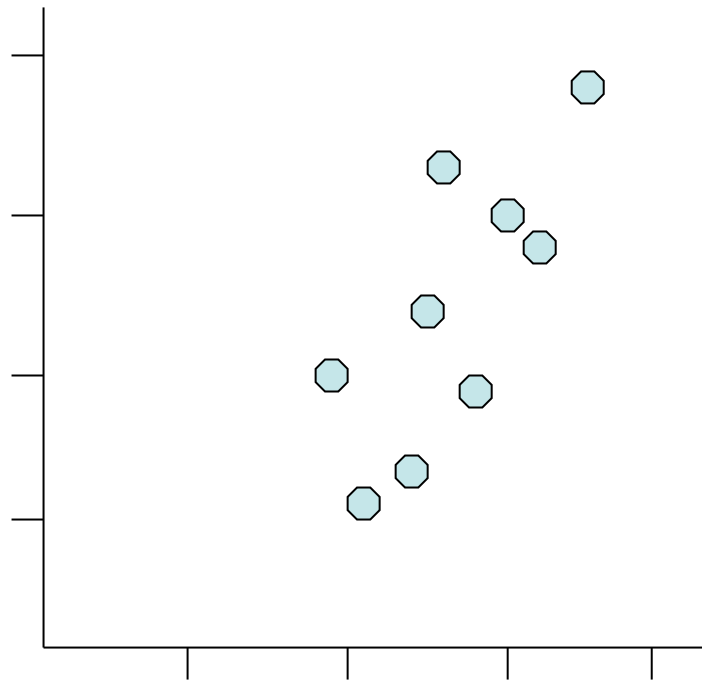


# Modelling data



With two observations the line is  
**DETERMINED!!!**

# Modelling data

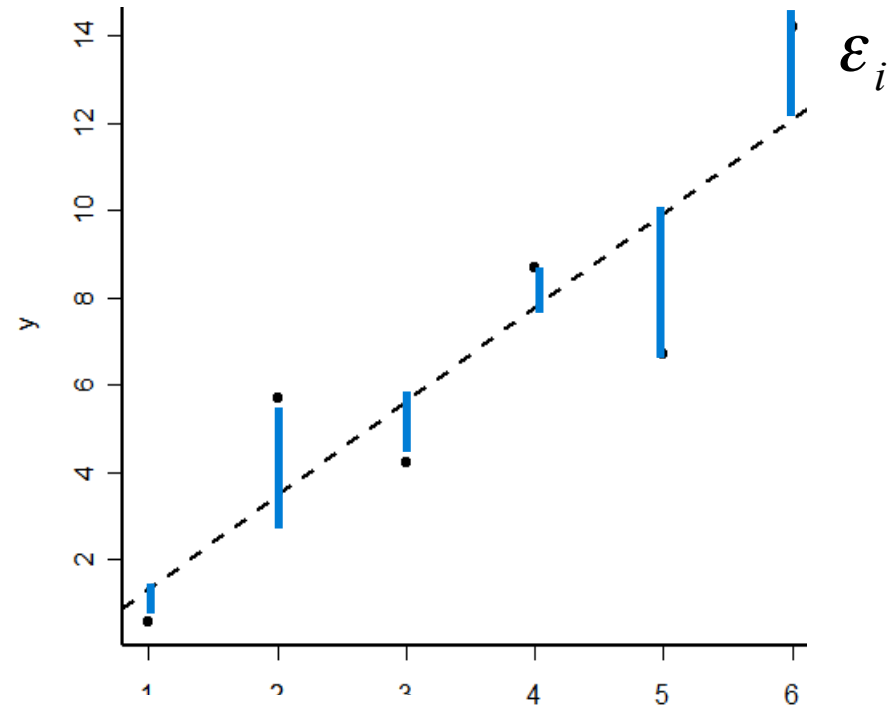


With many observations the line is NOT  
DETERMINED - we must estimate it!!!!

# Simple linear regression

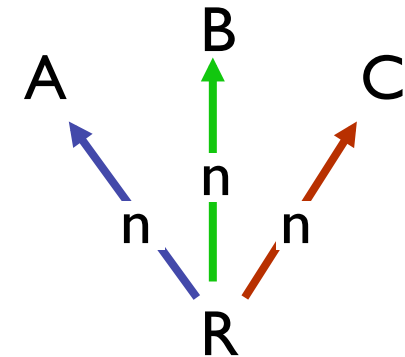
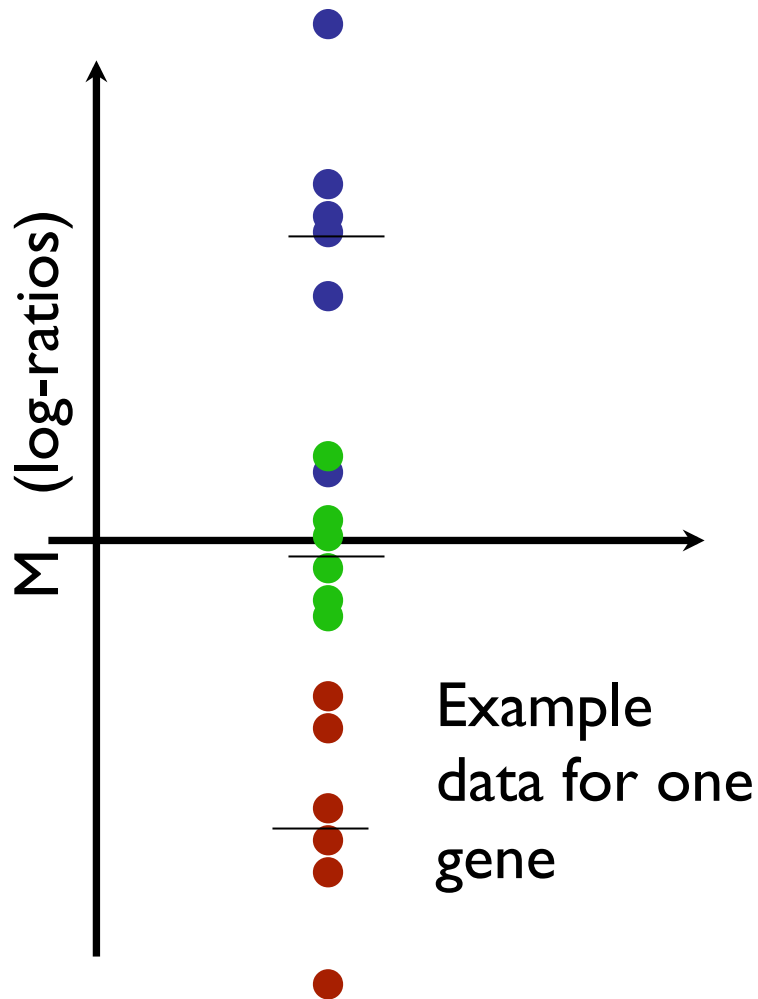
**Minimize the difference between the observation and its prediction according to the line.**

Method of least squares find the line which minimizes the sum of square errors.



$$\min \left\{ \sum_{i=1}^n (y_i - (\alpha + \beta x_i))^2 \right\} = \min \sum_{i=1}^n (\epsilon_i^2)$$

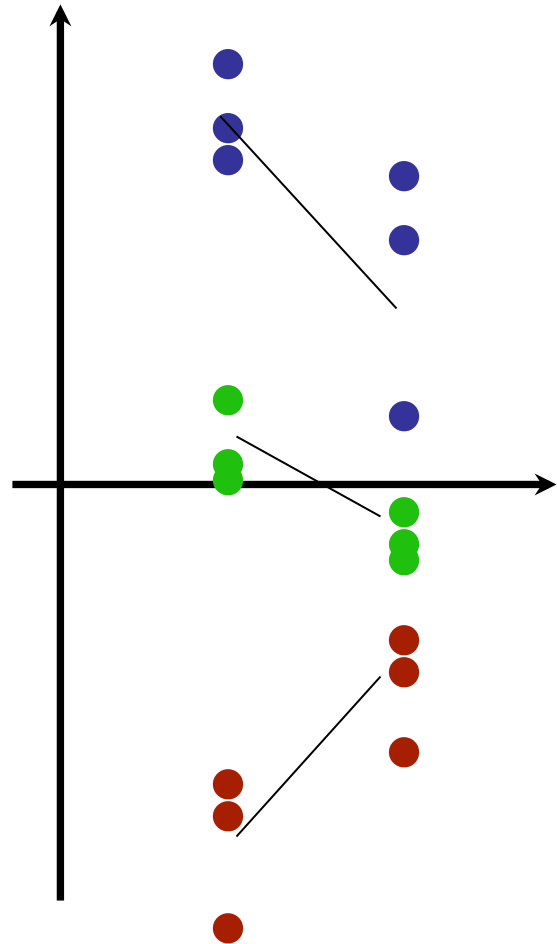
# Linear model for different groups



Experiment that might result in this data

Minimise the errors around the means of each group. Notice, that only with replication in each group, can we estimate a mean for each group and fit a statistical model to the data.

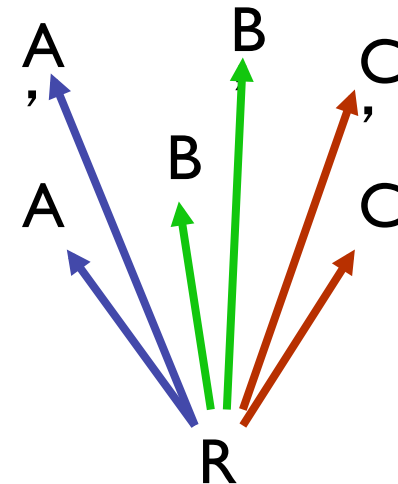
# Linear model for different groups at two levels



Example data for one gene

Drugs result in different transcriptional response (main effects)

Effect over time is also different for different drugs (interaction)



Experiment that might result in this data

i.e. A = treatment with drug A 1hr later

A' = treatment with drug A 24hrs later

# Linear models

$$y_{ijk} = \mu + s_i + t_j + st_{ij} + \varepsilon_{ijk};$$

$$i = wt, mt, j = time$$

s = effect of treatment

t = effect of time

$$Y = X\beta + \varepsilon$$

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix} \cdot \begin{pmatrix} \alpha \\ \beta \\ \alpha\beta \end{pmatrix} + \varepsilon$$

Linear model  
relationship

Matrix  
notation

Questions + design matrix

- One factor with multiple levels
- 2 by 2 factorial

# Definition

$\begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$  Parameter of interest

$\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix}$  Estimates

$s.e\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix}$  Standard error of our estimates

# Least square estimation

$$\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = (X^T X)^{-1} X^T Y \quad \text{Cov} \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \sigma^2 (X^T X)^{-1}$$

where  $X$  represents the design matrix

variance  
between slides

and  $Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}$  ← observed  
log-ratios or  
log-intensities

$$s.e.(\hat{\beta}_i) = \sigma \sqrt{c_i} \longrightarrow \text{sqrt}(\text{diag}(\text{solve}(\text{t}(X) \%* \% X)))$$

where  $c_i$  is the  $i^{\text{th}}$  diagonal element of  $(X^T X)^{-1}$



# Linear model and array platforms

- Linear modelling approach applies to both single channel (Affymetrix) and two-colour spotted arrays.
- Two colour with common reference is virtually equivalent to single channel from an analysis point of view
- Need to cover some special features of two-colour arrays using direct comparisons.

# Limma

- Especially for the application of linear models for analysing designed experiments and the assessment of differential expression
- Includes processing capabilities for two-colour spotted arrays and affymetrix chip data. The differential expression methods treat two-colour, affymetrix and single-colour experiments in a unified way.

<http://bioinf.wehi.edu.au/limma> (home page)

<http://www.bioconductor.org/packages/2.6/bioc/html/limma.html>

(link to userguide)

Smyth, G. K. (2005). Limma: linear models for microarray data. In: *Bioinformatics and Computational Biology Solutions using R and Bioconductor*, R. Gentleman, V. Carey, S. Dudoit, R. Irizarry, W. Huber (eds.), Springer, New York, pages 397- 420

# The simplest design question: Direct versus indirect comparisons

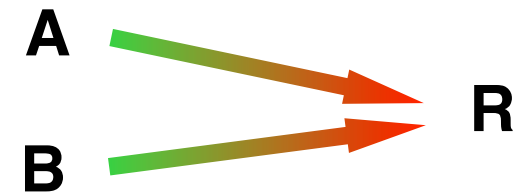
Two samples (A vs B)

e.g. KO vs. WT or mutant vs. WT

**Direct**



**Indirect**



$$\hat{M}_{A-B} : \quad \text{average} (\log (A/B))$$

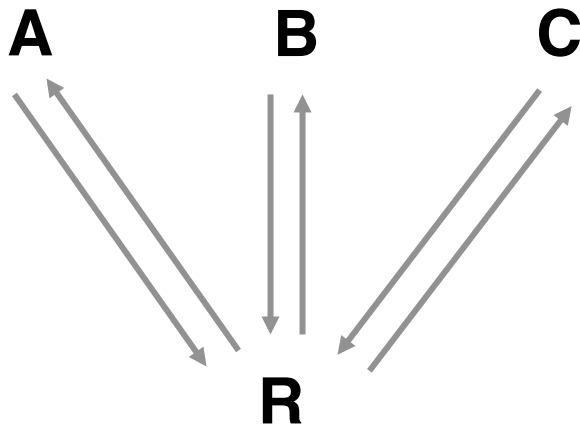
$$\log (A / R) - \log (B / R)$$

$$se.(\hat{M}_{A-B}) : \quad \sigma / \sqrt{2}$$

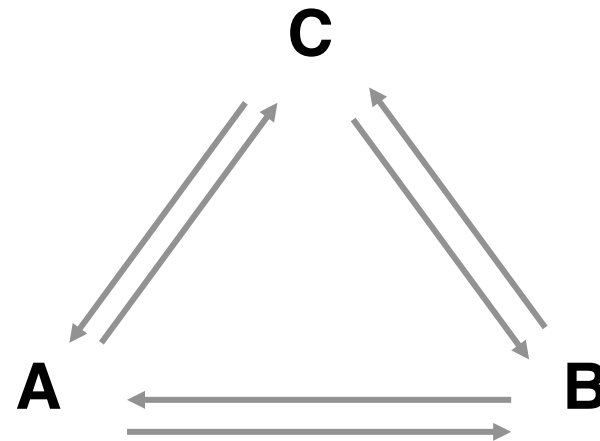
$$\sqrt{2} . \sigma$$

# Comparing K treatments

(i) Indirect design



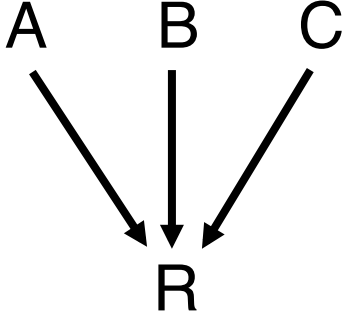
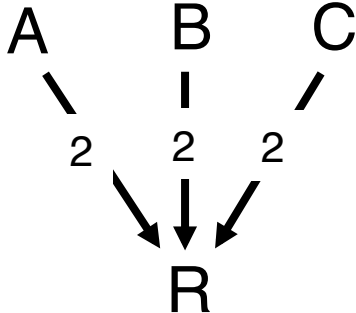
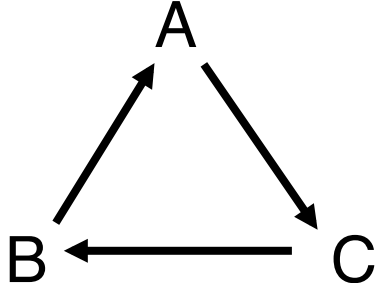
(ii) Direct design

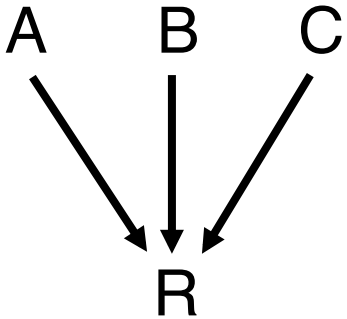
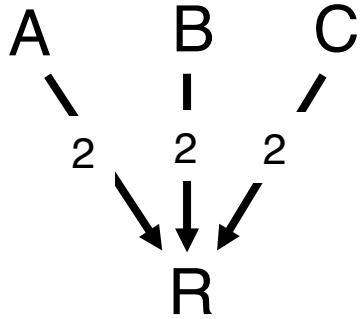
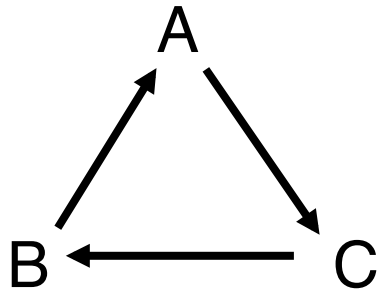


**Question:** Which design gives the most precise estimates of the contrasts A-B, A-C, and B-C?

# Experimental design

- 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.

	I (a) Common reference	I (b) Common reference	II Direct comparison
			
Number of Slides	3	6	3
Units of material	$A = B = C = 1$	$A = B = C = 2$	$A = B = C = 2$
Standard error [ $\times \sigma$ ] (A-B, etc)	$\sqrt{2} = 1.414$	1	$\sqrt{2/3} = 0.816$

	I (a) Common reference	I (b) Common reference	II Direct comparison
			
Number of Slides	3	6	3
Units of material	$A = B = C = 1$	$A = B = C = 2$	$A = B = C = 2$
Standard error [ $\times \sigma$ ] (A-B, etc)	$\sqrt{2} = 1.414$	1	$\sqrt{2/3} = 0.816$

# 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.



## 2 x 2 factorial

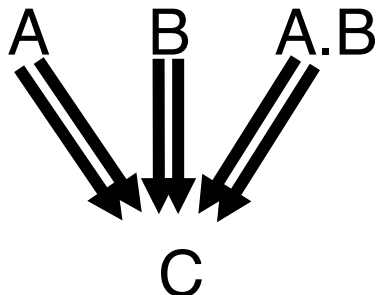
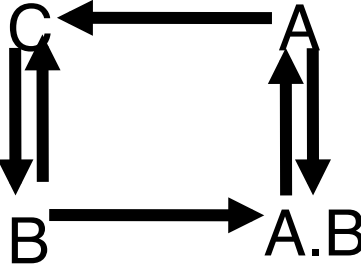
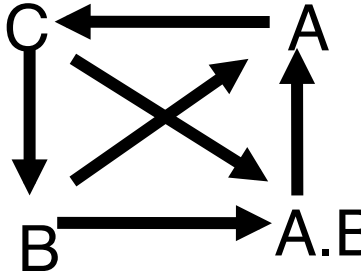
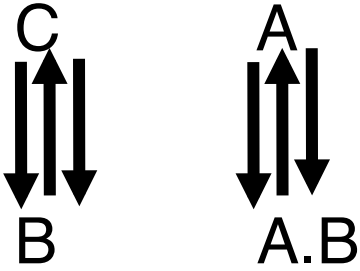
	Indirect	A balance of direct and indirect		
	I) 	II) 	III) 	IV) 
# Slides	N = 6			
Main effect <b>A</b>	0.71	0.82	0.71	NA
Main effect <b>B</b>	0.71	0.65	0.71	0.58
Interaction <b>A.B</b>	1.22	0.82	1.00	0.82

Table entry: standard error

Ref: Glonek & Solomon (2002)

# The simplest design question: Direct versus indirect comparisons

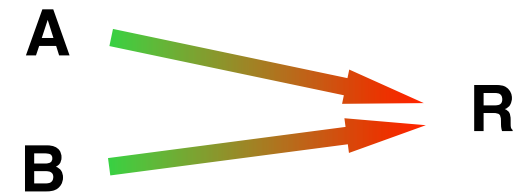
Two samples (A vs B)

e.g. KO vs. WT or mutant vs. WT

**Direct**



**Indirect**



$$\hat{M}_{A-B} : \quad \text{average} (\log (A/B))$$

$$\log (A / R) - \log (B / R)$$

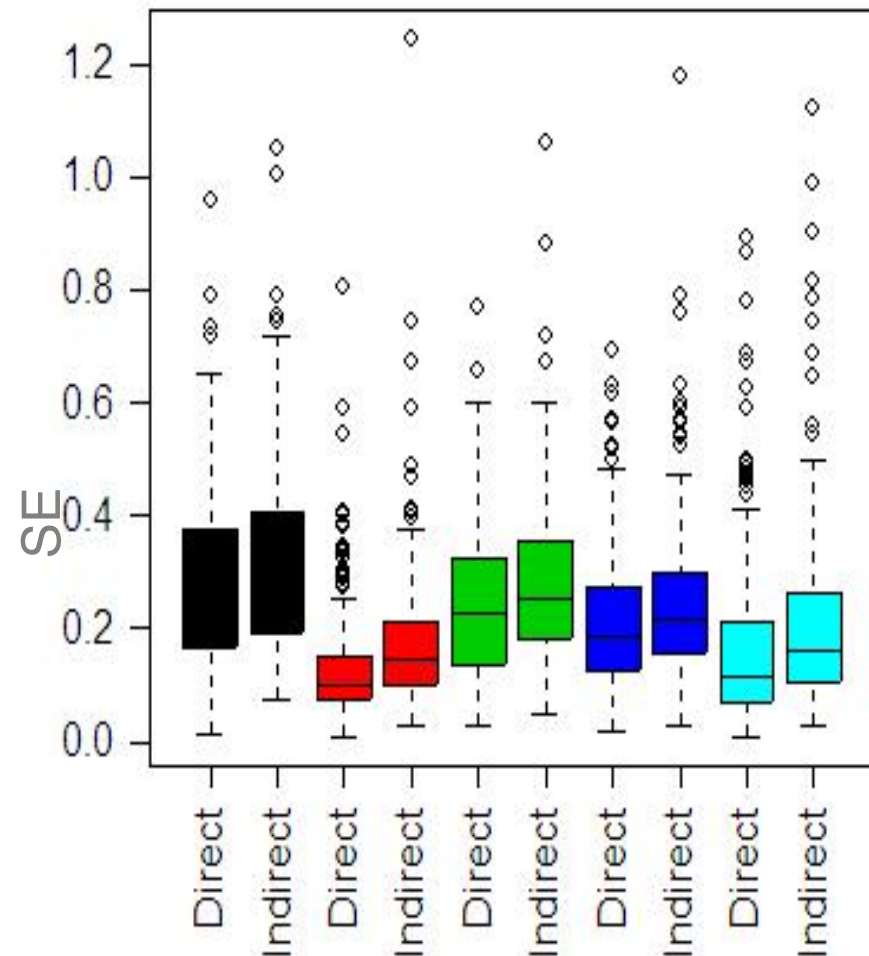
$$se.(\hat{M}_{A-B}) : \quad \sigma / \sqrt{2}$$

$$\sqrt{2} . \sigma$$

These calculations assume independence of replicates: the reality is not so simple.

# Experimental results

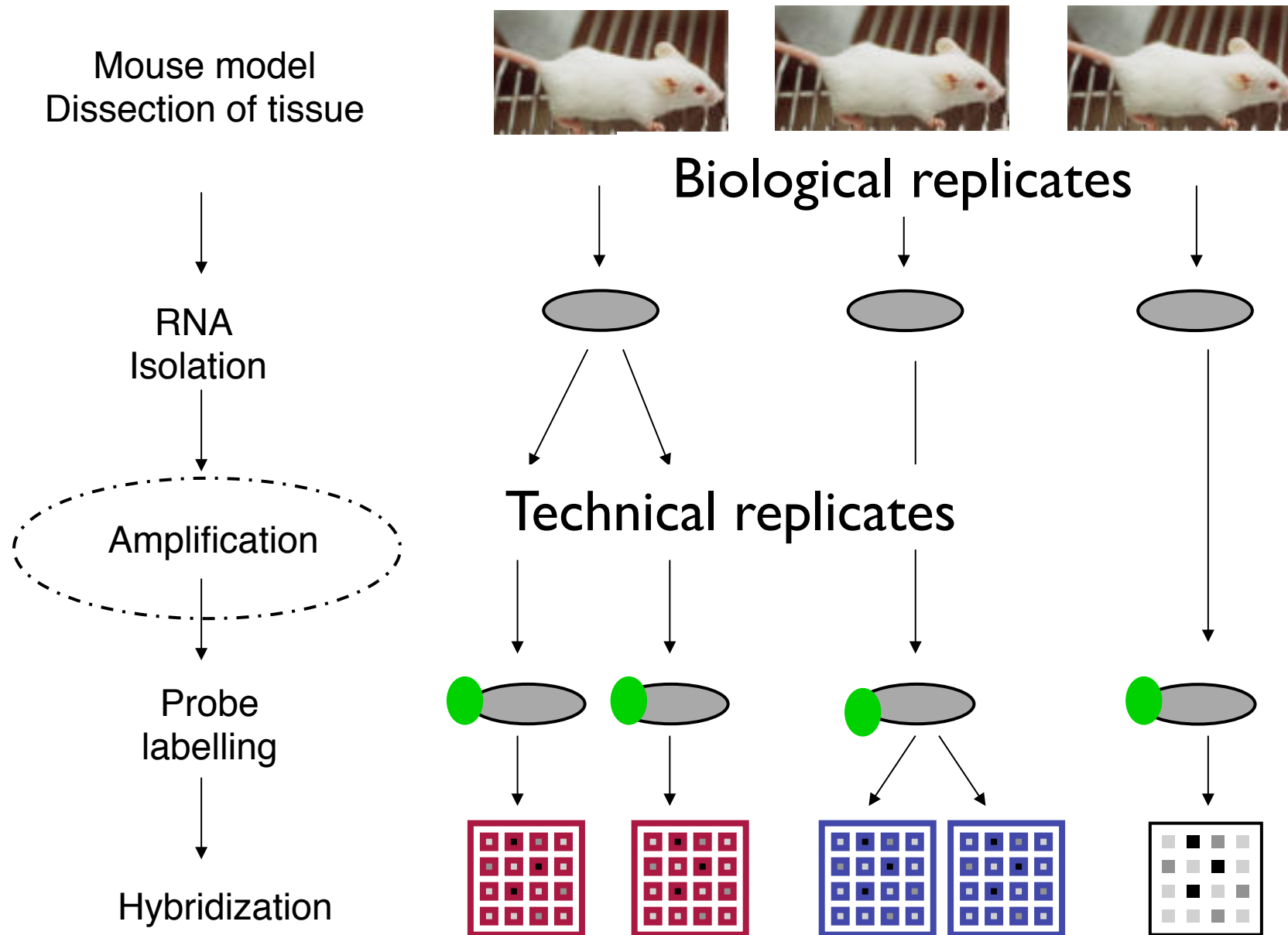
- 5 sets of experiments with similar structure.
- Compare (y axis)
  - A) SE for  $\text{aveM}_{\text{mt}}$
  - B) SE for  $\text{aveM}_{\text{mt}} - \text{aveM}_{\text{wt}}$
- Theoretical ratio of direct vs indirect SE is 1.6
- Experimental observation is 1.1 to 1.4.



# Caveats

- Variability decreases with replication. But the effective replication is not equal to the number of replicates.
- Why? Because replicates are correlated and therefore not independent sources of information.
- The **more independent** (uncorrelated) the replicates, the **higher** the **effective replication**. Biological replicates preferable to technical replicates.
- The advantage of direct over indirect comparisons is real. However the difference is not a factor of 2, as theory predicts.
- Why? Possibly because mRNA from the same extractions - and pools of controls or reference material are the norm - give correlated expression levels. In other words, the assumption of independence between  $\log(A/R)$  and  $\log(B/R)$  is not valid.

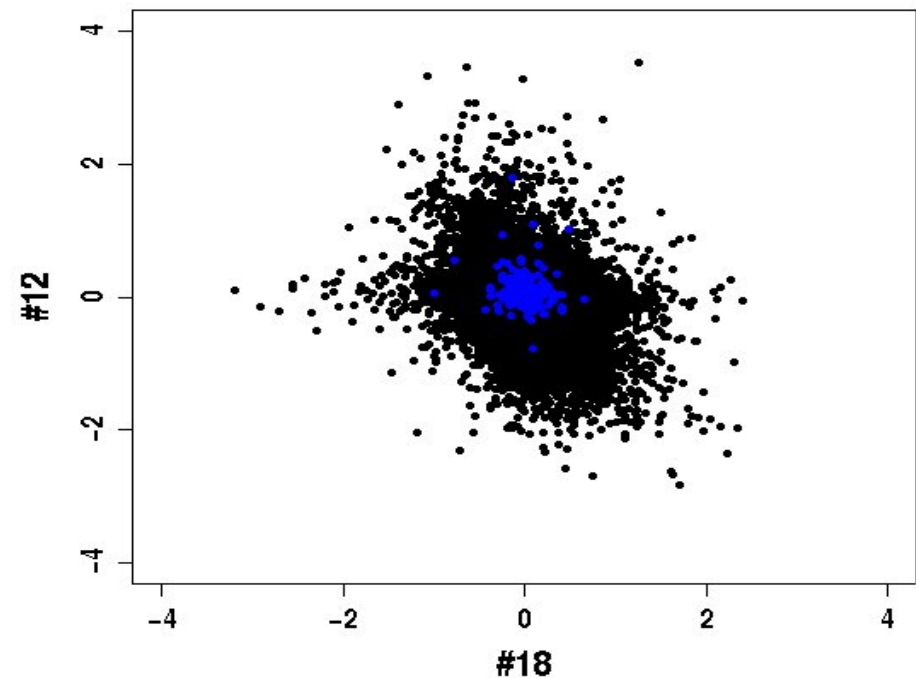
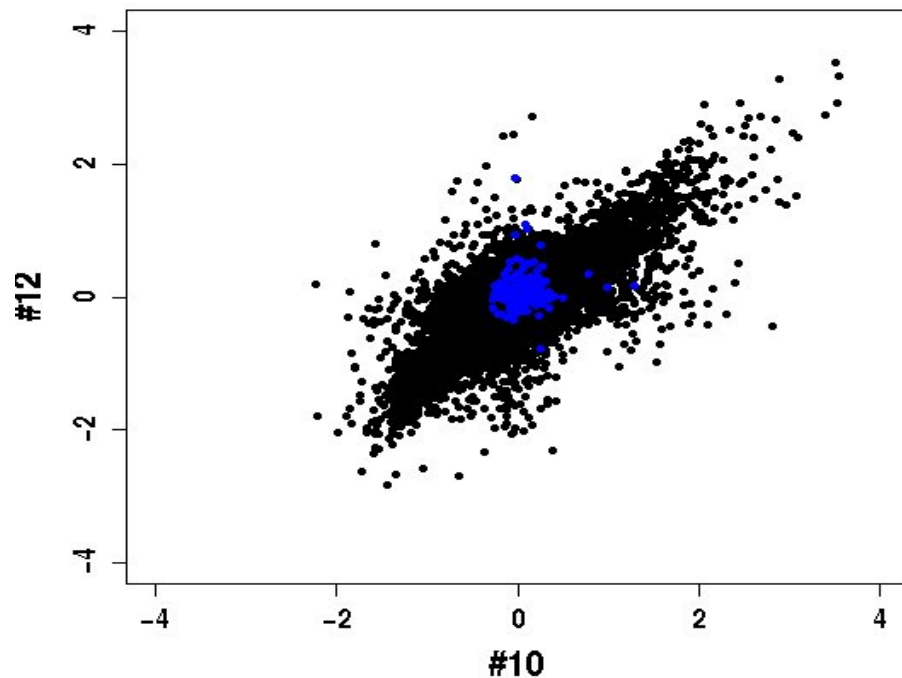
# Preparing mRNA samples



# Technical replication - amplification

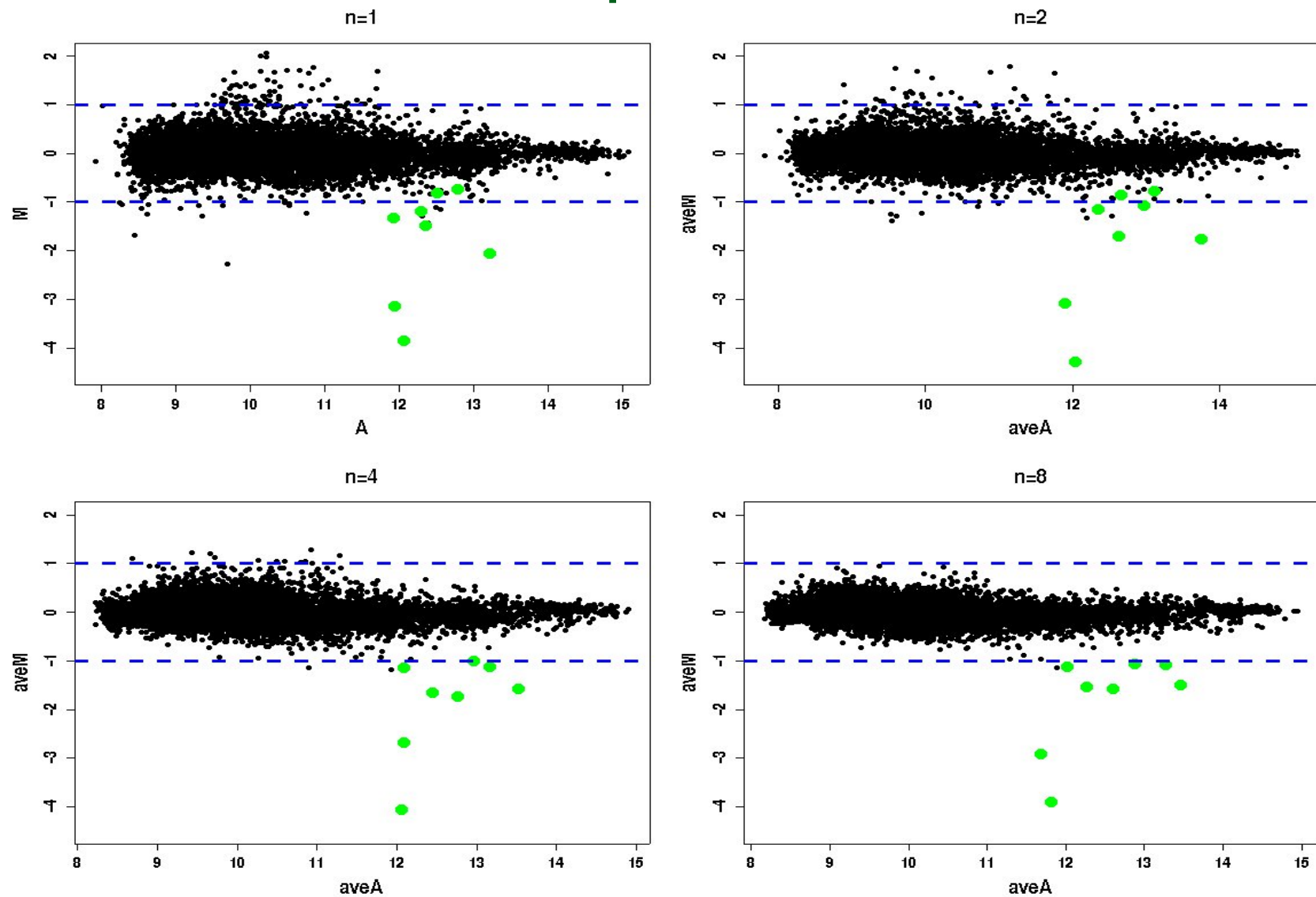
Olfactory bulb experiment:

- 3 sets of Anterior vs Dorsal performed on different days
- #10 and #12 were from the same RNA isolation and amplification
- #12 and #18 were from different dissections and amplifications
- All 3 data sets were labeled separately before hybridization



Data provided by  
Dave Lin (Cornell)<sup>110</sup>

# Sample size



Data provided by Matt Callow

# Technical replicates

- It is not entirely correct to treat technical replicates like independent biological replicates: technical replicates are not independent, they are likely to be highly correlated.
- *limma* allows you to estimate a consensus or average correlation within blocks of technical replicates for different biological replicates. Obviously, if the consensus correlation estimate turns out to be negative, then we can proceed by regarding them as independent replicates. However, you need to be careful. If the technical replicates are in dye-swap, then a negative correlation is expected!

`duplicateCorrelation(...)`

- If the dye-swapping is not even between one set of technical reps and another, then it is difficult to estimate this correlation with the approach as outlined in *limma*.



# Summary

- Create highly correlated reference samples to overcome inefficiency in common reference design.
- **Not** advocating the use of technical replicates in place of biological replicates for samples of interest.

# References

- T. P. Speed and Y. H Yang (2002). **Direct versus indirect designs for cDNA microarray experiments**. *Sankhya : The Indian Journal of Statistics*, Vol. 64, Series A, Pt. 3, pp 706-720
- Y.H. Yang and T. P. Speed (2003). **Design and analysis of comparative microarray Experiments** In T. P Speed (ed) *Statistical analysis of gene expression microarray data*, Chapman & Hall.
- R. Simon, M. D. Radmacher and K. Dobbin (2002). **Design of studies using DNA microarrays**. *Genetic Epidemiology* 23:21-36.
- F. Bretz, J. Landgrebe and E. Brunner (2003). **Efficient design and analysis of two color factorial microarray experiments**. *Biostatistics*.
- G. Churchill (2003). **Fundamentals of experimental design for cDNA microarrays**. *Nature genetics review* 32:490-495.
- G. Smyth, J. Michaud and H. Scott (2003) **Use of within-array replicate spots for assessing differential experssion in microarray experiments**. Technical Report In WEHI.
- Glonek, G. F. V., and Solomon, P. J. (2002). **Factorial and time course designs for cDNA microarray experiments**. Technical Report, Department of Applied Mathematics, University of Adelaide. 10/2002

# Pooled vs Individual samples

- Pooling is seen as “biological averaging”.
- Trade off between
  - Cost of performing a hybridisation.
  - Cost of the mRNA samples.

Cost of mRNA samples  $\ll$  Cost per hybridisation

Pooling can assist in reducing the number of hybridisations.

# Pooled vs Amplified samples

- In the cases where we do not have enough material from one biological sample to perform one array (chip) hybridisations. Pooling or Amplification is necessary.
- Amplification
  - Introduces more noise.
  - Non-linear amplification (??), different genes amplified at different rate.
  - Able to perform more hybridisations.
- Pooling
  - Less replicate hybridisations.

# References

- Pooling vs Non-Pooling
  - Han, E.-S., Wu, Y., Bolstad, B., and Speed, T. P. (2003). [A study of the effects of pooling on gene expression estimates using high density oligonucleotide array data](#). Department of Biological Science, University of Tulsa, February 2003.
  - Kendzierski, C.M., Y. Zhang, H. Lan, and A.D. Attie. (2003). [The efficiency of mRNA pooling in microarray experiments](#). *Biostatistics* **4**, 465-477. 7/2003
  - Xuejun Peng, Constance L Wood, Eric M Blalock, Kuey Chu Chen, Philip W Landfield, Arnold J Stromberg (2003). [Statistical implications of pooling RNA samples for microarray experiments](#). *BMC Bioinformatics* 4:26. 6/2003