UNIVERSITY OF CAMBRIDGE

MPhil in Computational Biology 2016/17
Course: **Functional Genomics**

Lectures 2 and 3 (October 12th, 14th, 2016)

From experimental design to the analysis of genomic data

Oscar M. Rueda

Oscar.Rueda@cruk.cam.ac.uk

Contributions by ***Nuno L. Barbosa-Morais and Natalie Thorne, Benilton Carvalho, Alex Lewin, Ernest Turro, Paul O'Reilly, Andy Lynch, Terry Speed, Gordon Smyth, Jean Yang, Ingrid Lonnstedt, Matt Ritchie.***

Outline

- Experimental Design
- Building the model
- Parameter estimation and hypothesis testing.

Experimental Design

Experimental design

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

Principles of Statistical Design of Experiments

- R. A. Fisher:
  - Replication
  - Randomization
  - Blocking.

- They have been used in microarray studies from the beginning.
- Bar coding makes easy to adapt them to Next Generation Sequencing studies.

Avoidance of bias

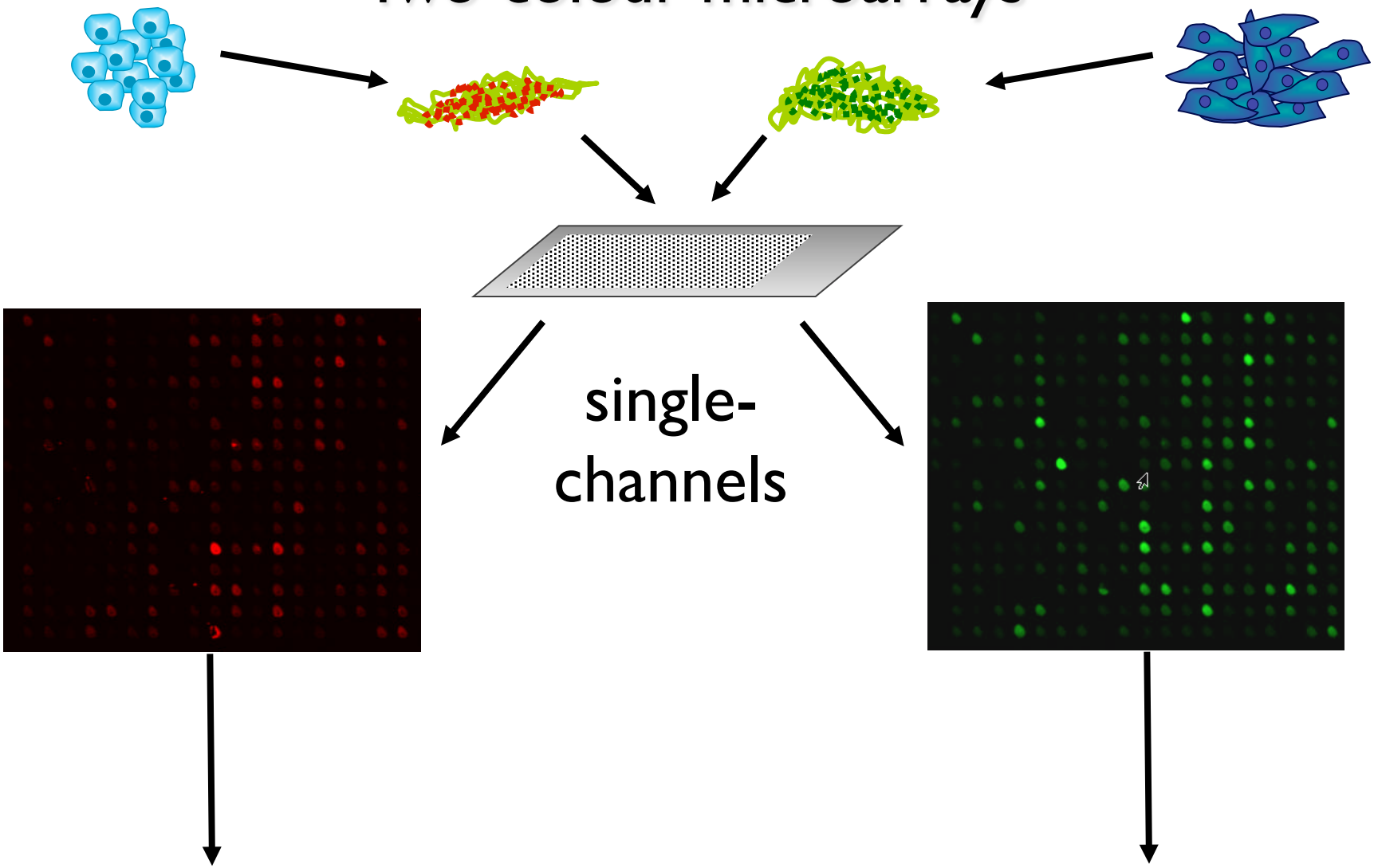
- Conditions of an experiment; DNA/mRNA extraction and processing, library preparation, the reagents, the operators, the scanners and so on can leave a "global signature" in the resulting expression data.

- Randomization.

- Local control is the general term used for arranging experimental material.

Experimental Design in microarray studies

Two-colour microarrays



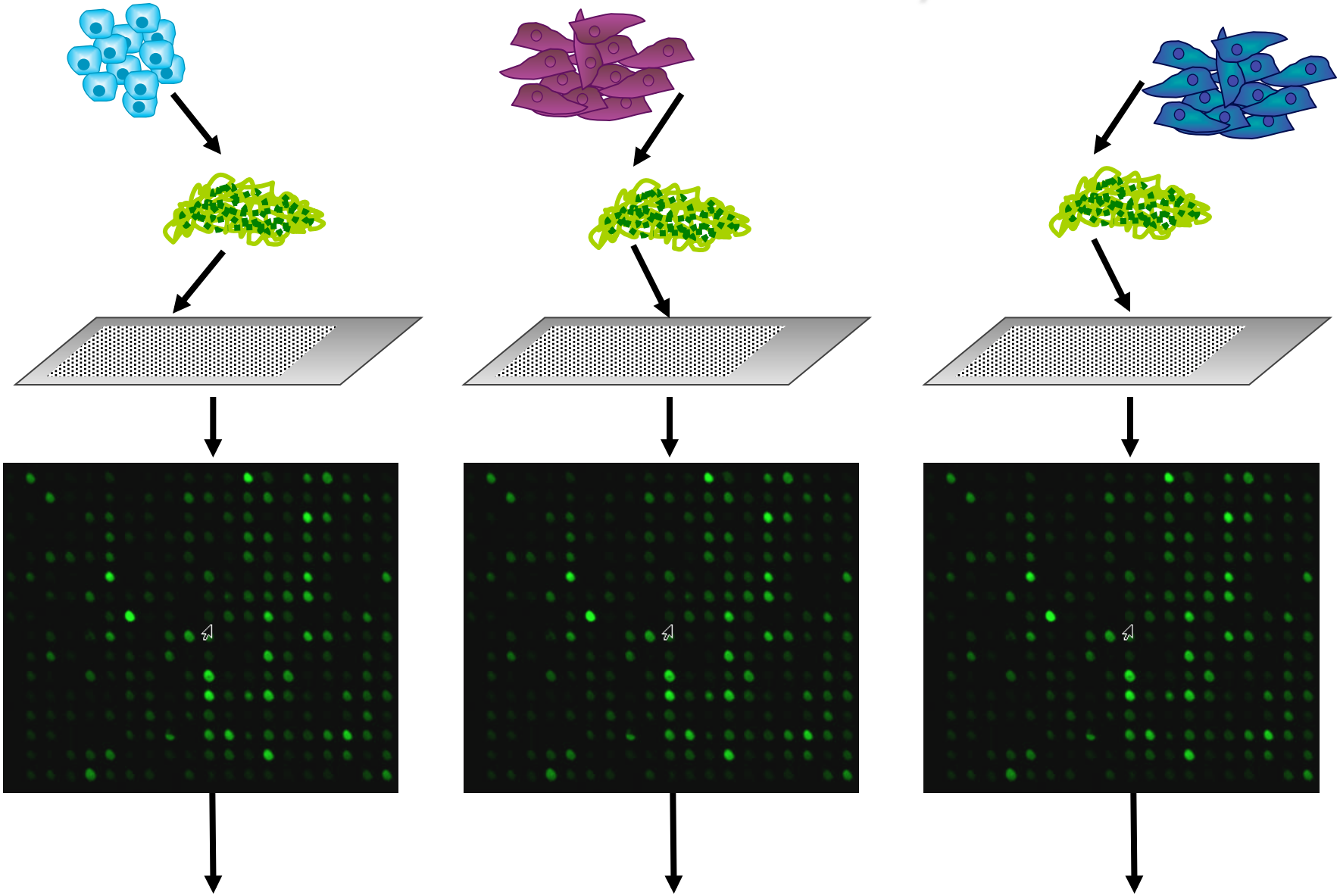
Two-colour microarray statistics

$\log_2 R$ $\log_2 G$

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

$$A = {}^1/_2 (\log_2 R + \log_2 G)$$

One-colour microarrays

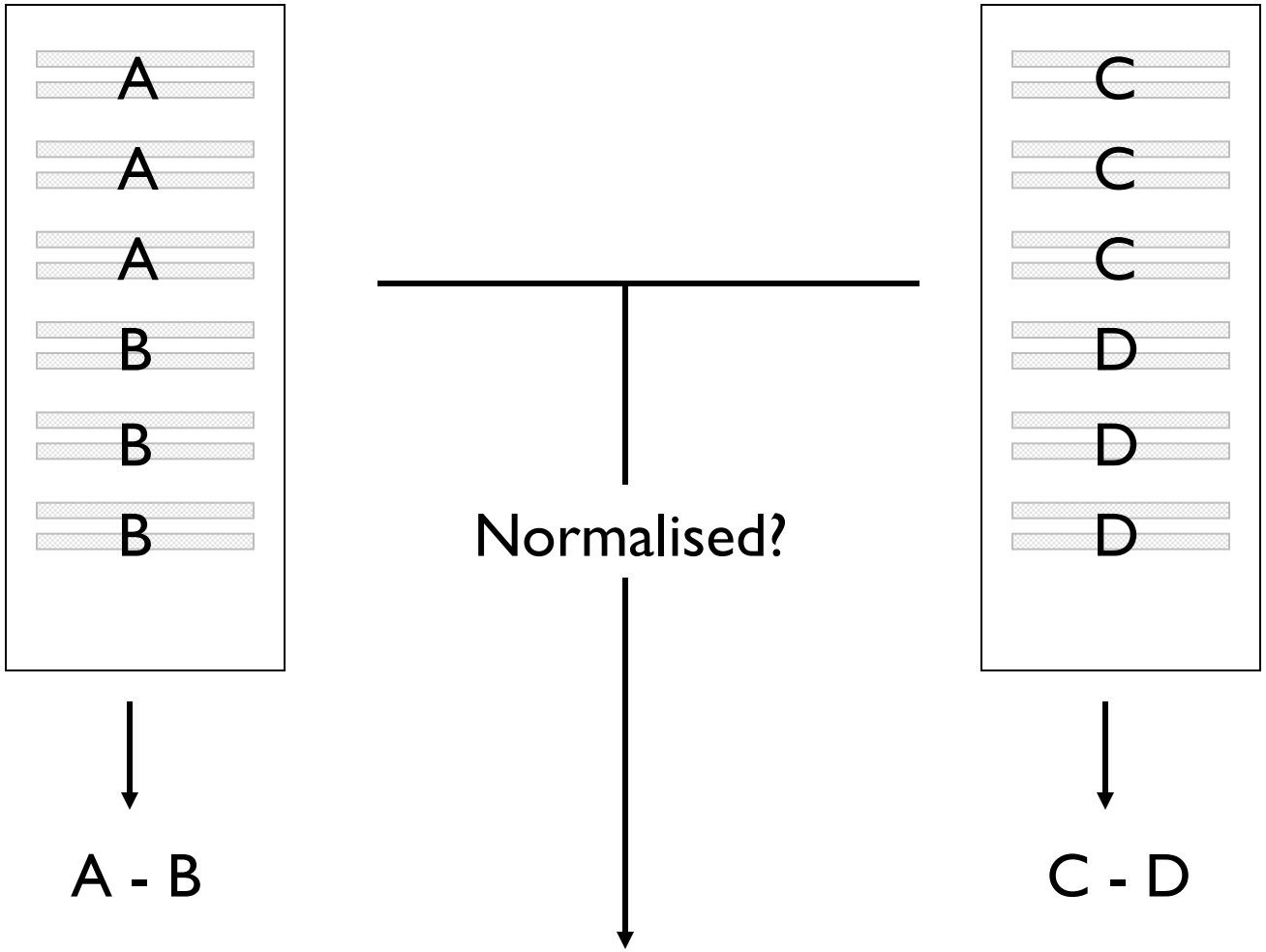


One color microarray design : connectivity / ability to normalise

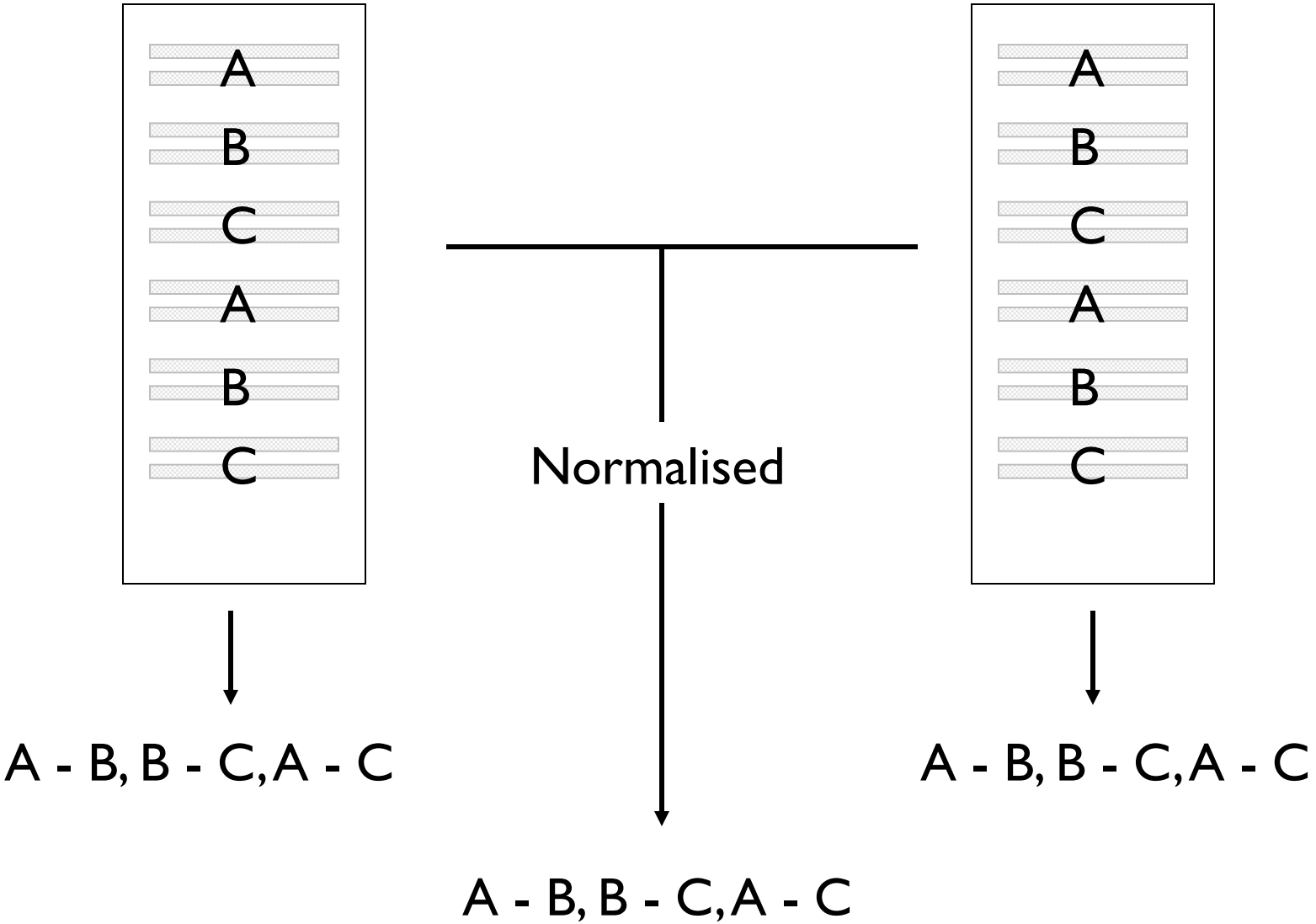
- How do we allocate six samples to each chip?
  - Comparisons within-chips more precise than between-chips
  - Normalisation assumes distributions of samples from different chips are roughly the same

In the examples we will discuss Illumina arrays

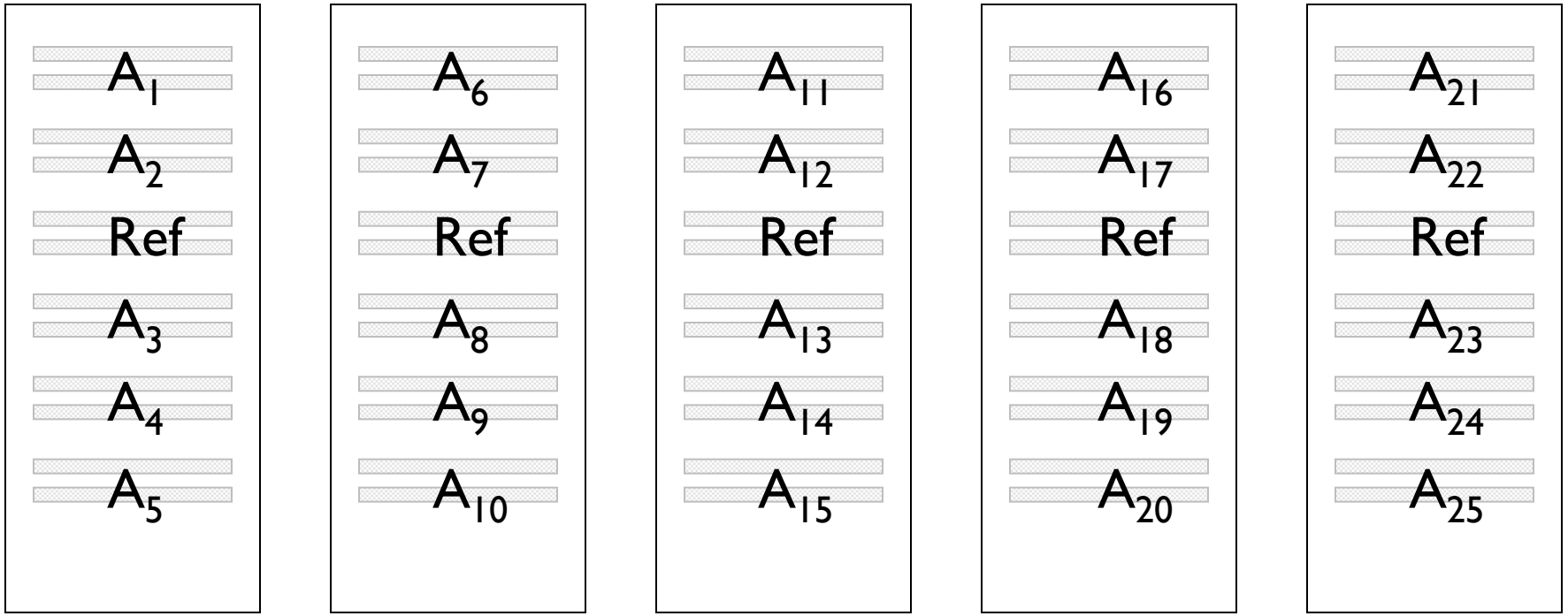
Illumina design : connectivity / ability to normalise



Illumina design : connectivity / ability to normalise

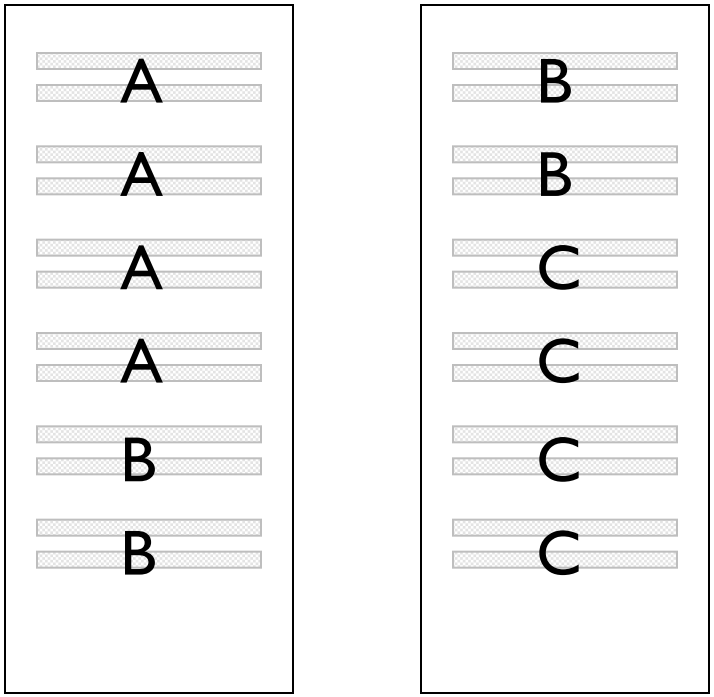


Illumina design : ability to normalise

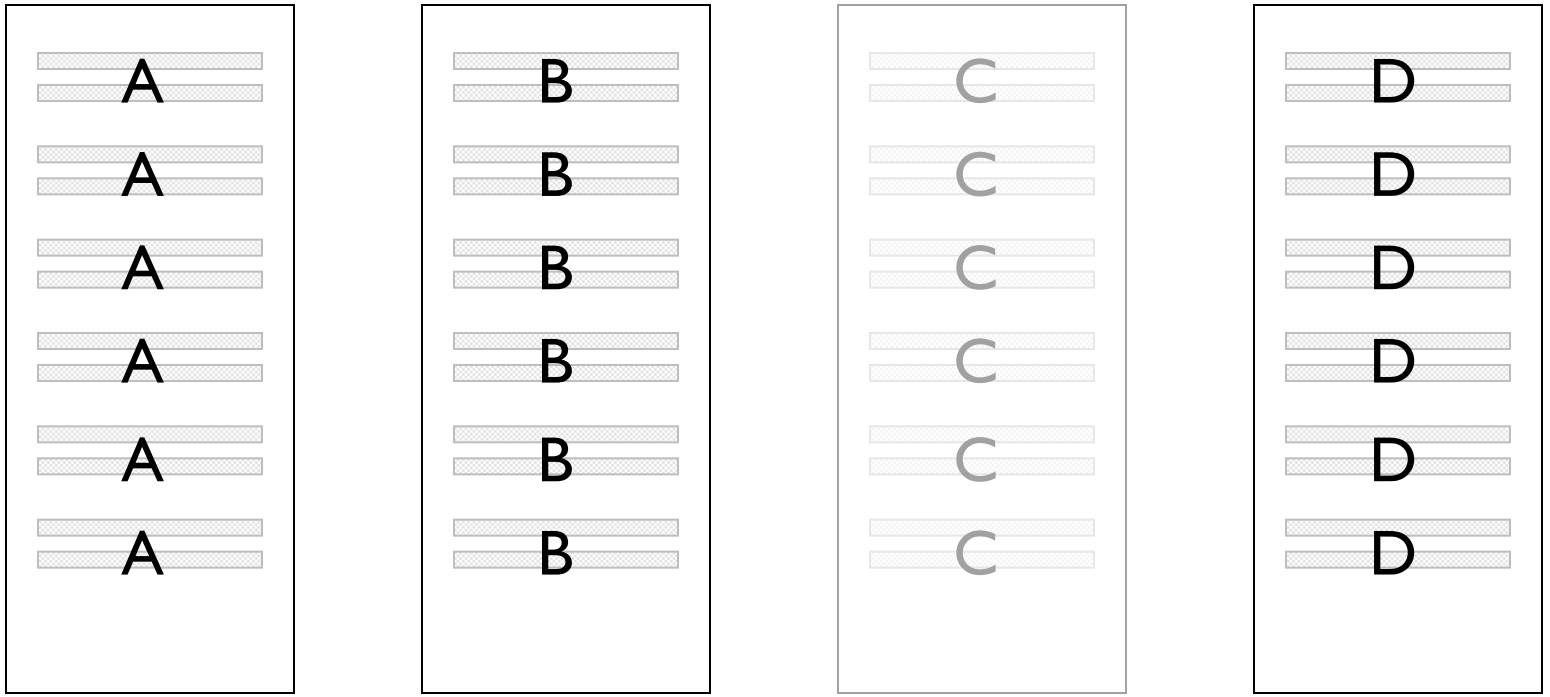


Does Ref act as a common "normalisation reference"? This might be particularly useful for large studies involving many treatments/sample types.

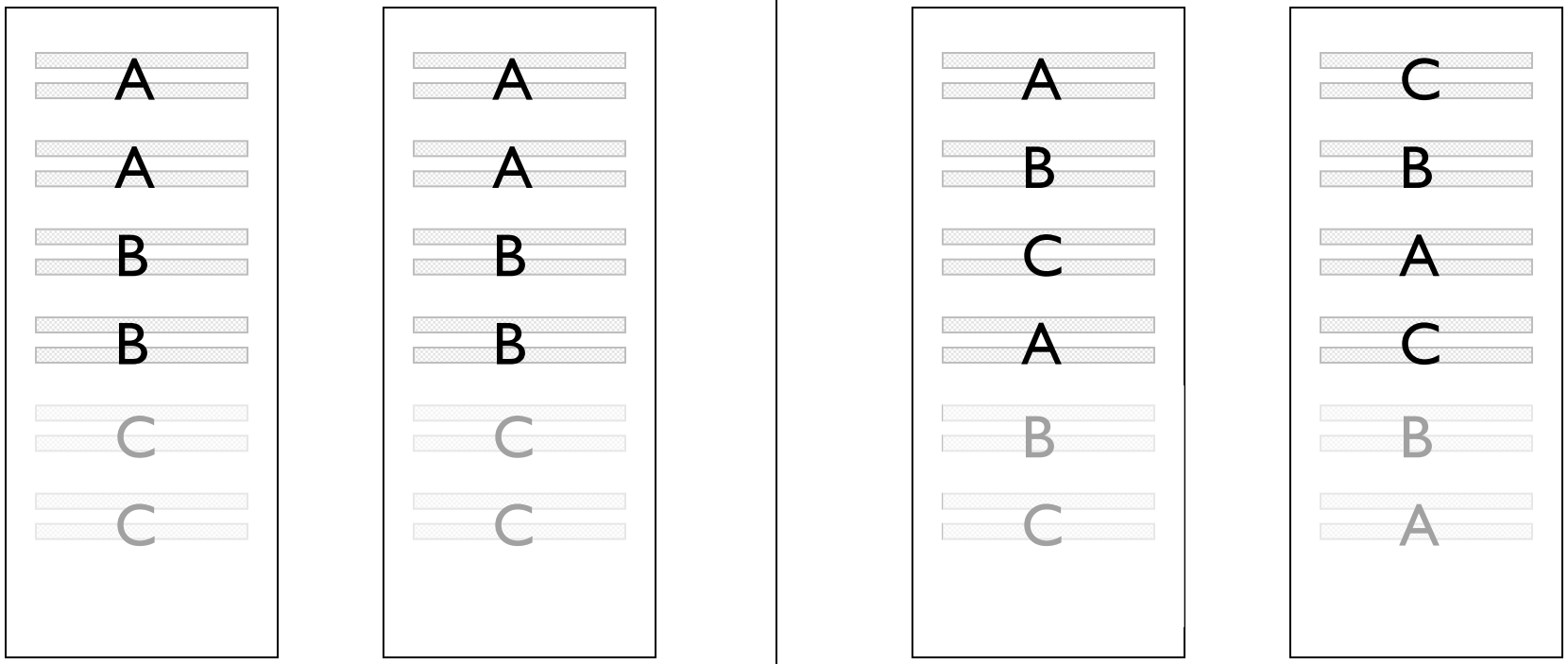
Illumina design : balance



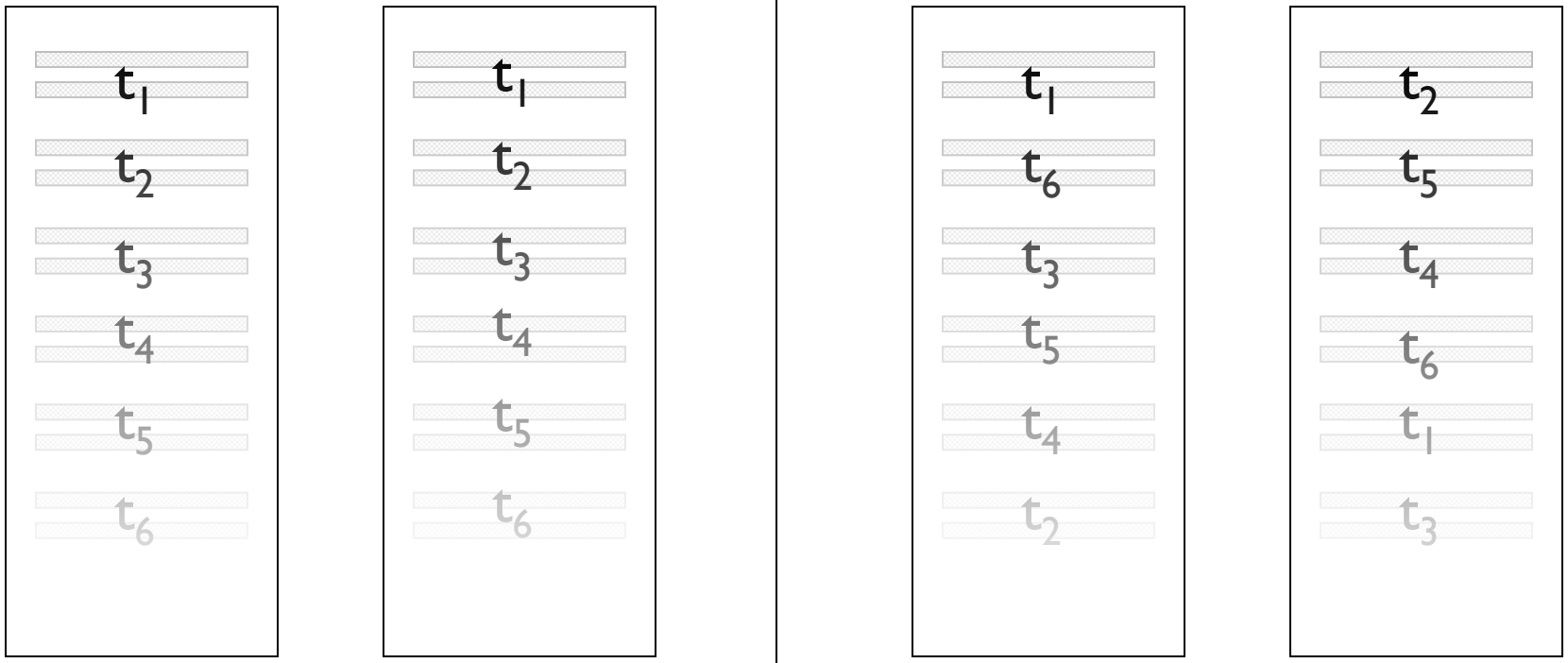
Illumina design : robustness



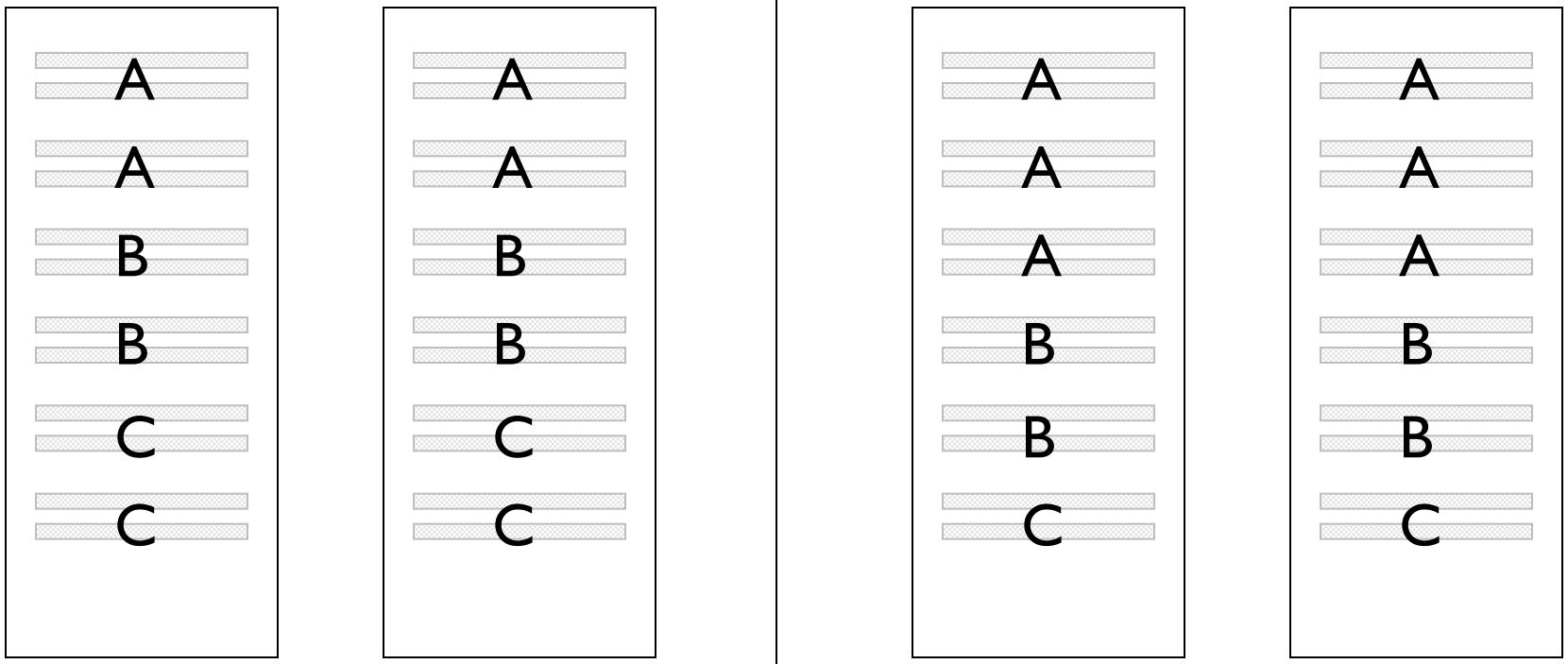
Illumina design : robustness



Illumina design : normalisation and spatial trends



Illumina design : important questions/precision

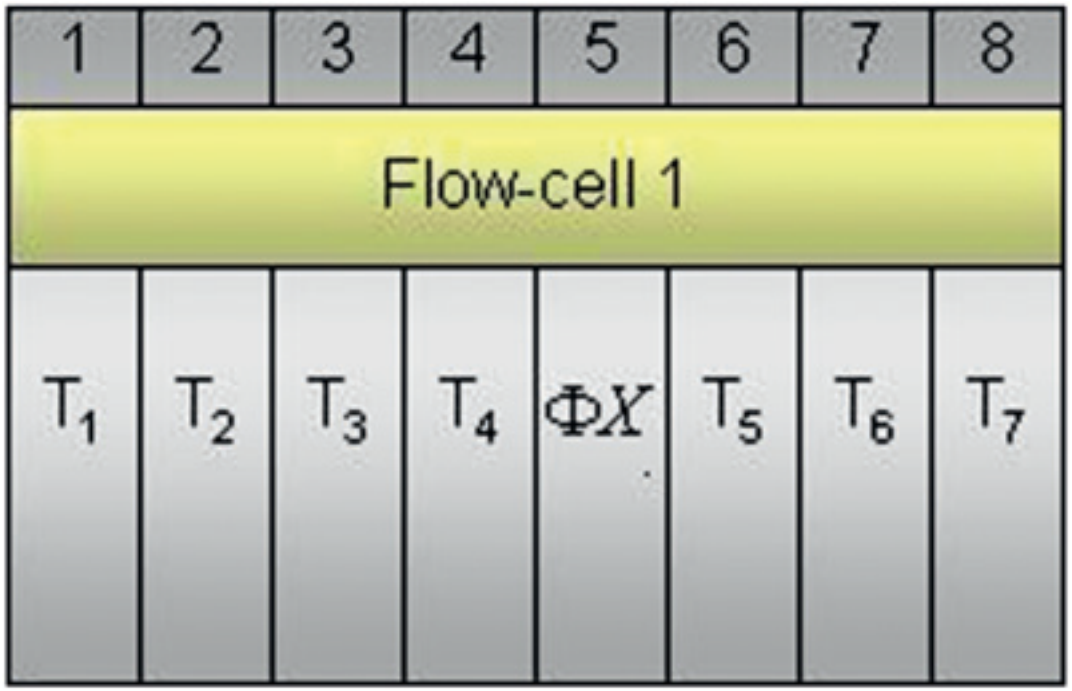


Design Principles

- randomisation: don't allocate all replicate samples of the same type to the one chip. Mix up the order in which samples occur on a chip
- replication, replication, replication…
- blocking: how to allocate 6, 8, 12 or 96 samples to each chip to ensure comparisons of interest can be estimated precisely?

Experimental Design in sequencing studies

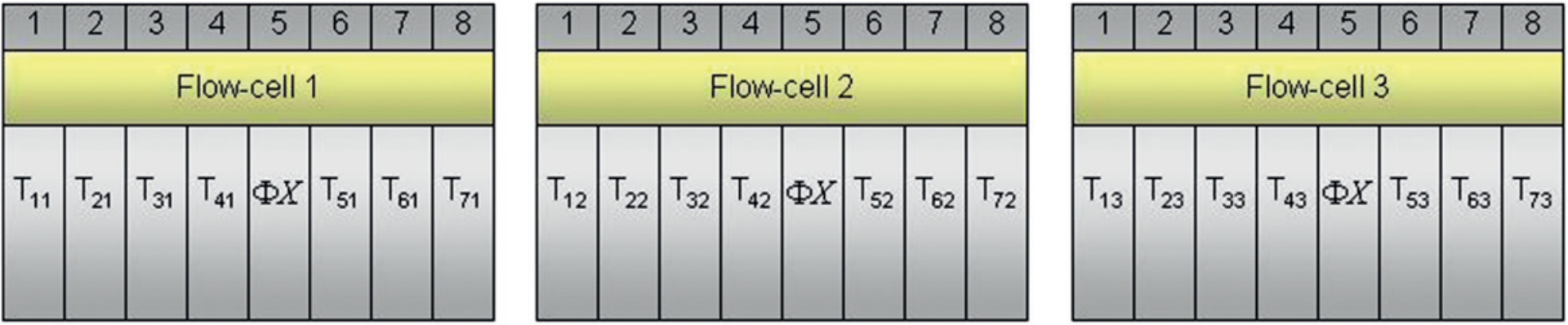
Unreplicated Data



Inferences for RNA and fragment-level can be obtained through Fisher's test. But they don't reflect biological variability.

Auer and Doerge. Genetics 185:405-416(2010)

Replicated Data



Inferences for treatment effect using generalized linear models (more on this later).

Is this a good design? We should randomize within block!

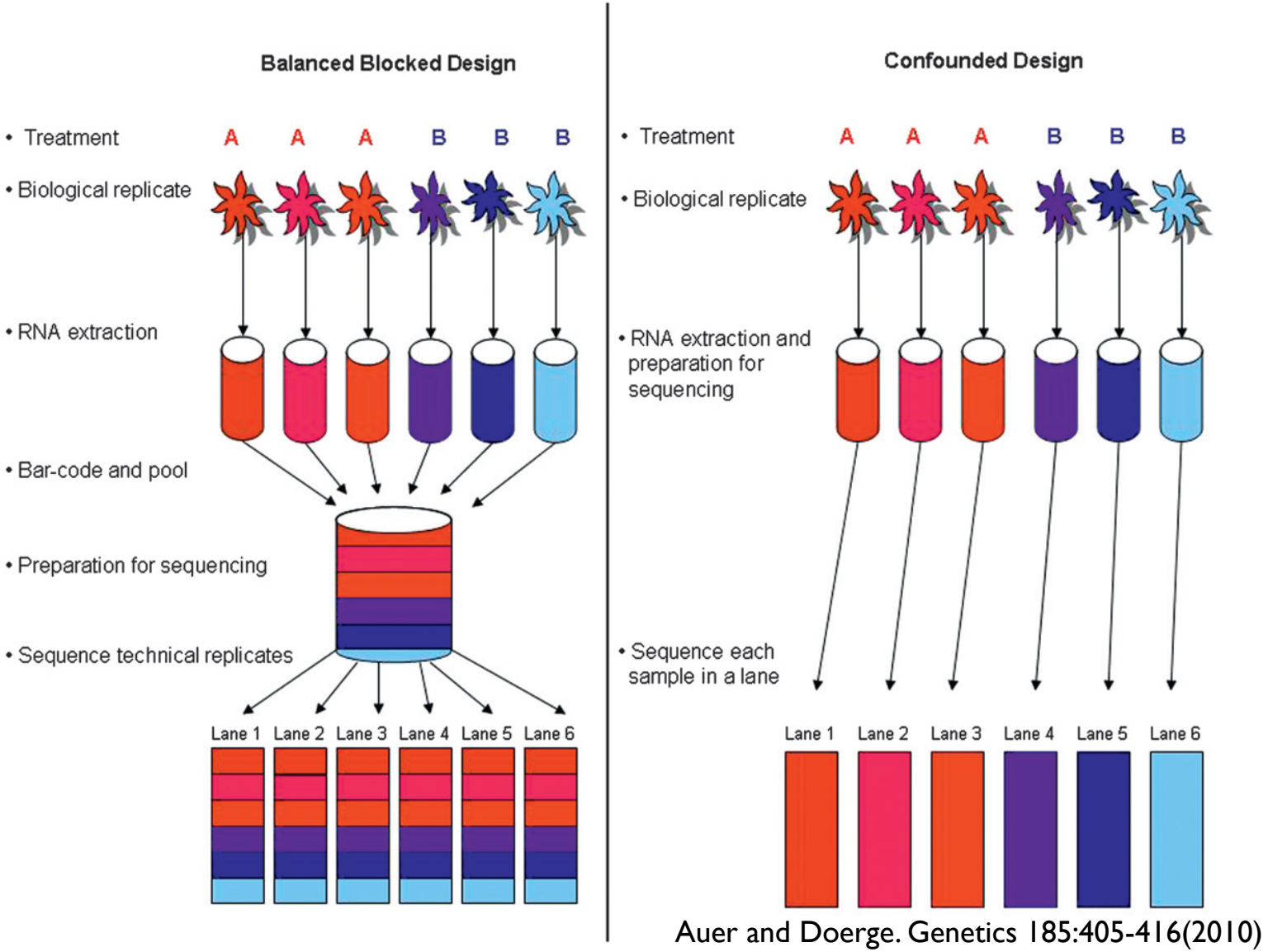
Auer and Doerge. Genetics 185:405-416(2010)

Balanced Block Designs

- Avoids confounding effects:
  - Batch effects (any errors after random fragmentation of the RNA until it is input to the flow cell). Examples: PCR amplification, reverse transcription artifacts…
  - Lane effects (any errors from the point where the sample is input to the flow cell until the data output). Examples: systematically bad sequencing cycles, errors in base calling…
  - Other effects non related to treatment.

Auer and Doerge. Genetics 185:405-416(2010)

Balanced blocks by multiplexing



Auer and Doerge. Genetics 185:405-416(2010)

Benefits of a proper design in NGS studies

- NGS is benefited with design principles
- Technical replicates can not replace biological replicates
- It is possible to avoid multiplexing with enough biological replicates and sequencing lanes
- The advantages of multiplexing are bigger than the disadvantages (cost, loss of sequencing depth, bar-code bias…)

Building the model

Statistical models

– We want to model the expected result of an outcome (dependent variable) under given values of other variables (independent variables)

Arbitrary function (any shape)

Expected value of variable Y

A set of k independent variables (also called factors)

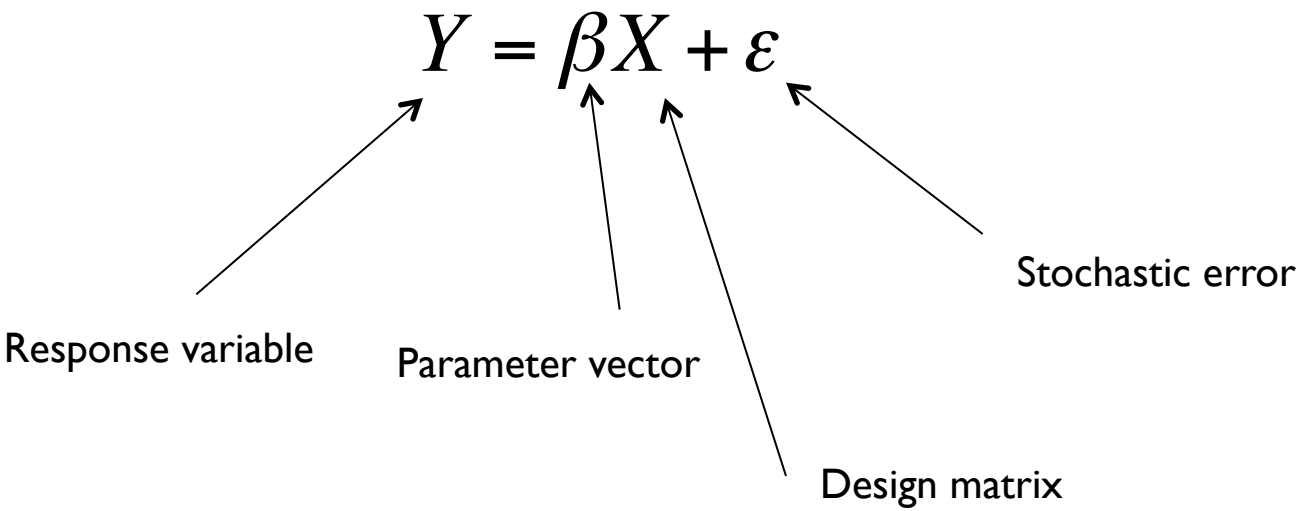
$$E(Y) = f(X)$$

$$Y = f(X) + \varepsilon$$

This is the variability around the expected mean of y

Design matrix

- Represents the independent variables that have an influence in the response variable, but also the way we have coded the information and the design of the experiment.
- For now, let's restrict to models



Common designs in functional genomic studies

- Models with 1 factor
  - Models with two treatments
  - Models with several treatments
- Models with 2 factors
  - Interactions
- Paired designs
- Models with categorical and continuous factors
- TimeCourse Experiments
- Multifactorial models.

The models are built for each probe!

Models with 1 factor, 2 levels

Sample	Treatment
Sample1	Treatment A
Sample 2	Control
Sample 3	Treatment A
Sample 4	Control
Sample 5	Treatment A
Sample 6	Control

Number of samples: 6
Number of factors: 1
Treatment: Number of levels: 2

Possible parameters (What differences are important)?

- Effect of Treatment A
- Effect of Control

Design matrix for models with 1 factor, 2 levels

Sample	Treatment
Sample1	Treatment A
Sample 2	Control
Sample 3	Treatment A
Sample 4	Control
Sample 5	Treatment A
Sample 6	Control

$$\begin{matrix} \text{Sample 1} \\ \text{Sample 2} \\ \text{Sample 3} \\ \text{Sample 4} \\ \text{Sample 5} \\ \text{Sample 6} \end{matrix} \begin{bmatrix} S1 \\ S2 \\ S3 \\ S4 \\ S5 \\ S6 \end{bmatrix} = \begin{pmatrix} \text{Treat.A} & \text{Control} \\ & \end{pmatrix} \begin{bmatrix} T \\ C \end{bmatrix}$$

Parameters (coefficients, levels of the variable)

Design Matrix

C is the mean expression of the control
T is the mean expression of the treatment

Equivalent to a t-test

Design matrix for models with 1 factor, 2 levels

Sample	Treatment
Sample1	Treatment A
Sample 2	Control
Sample 3	Treatment A
Sample 4	Control
Sample 5	Treatment A
Sample 6	Control

$$\begin{array}{l}\text{Sample 1}\\ \text{Sample 2}\\ \text{Sample 3}\\ \text{Sample 4}\\ \text{Sample 5}\\ \text{Sample 6}\end{array}\begin{bmatrix} S1 \\ S2 \\ S3 \\ S4 \\ S5 \\ S6 \end{bmatrix} = \begin{array}{c}\begin{matrix}\text{Treat.A} & \text{Control}\end{matrix}\\ \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}\end{array} \begin{bmatrix} T \\ C \end{bmatrix}$$

Parameters (coefficients, levels of the variable)

Design Matrix

Equivalent to a t-test

Intercepts

Different parameterization: using intercept

Sample	Treatment
Sample1	Treatment A
Sample 2	Control
Sample 3	Treatment A
Sample 4	Control
Sample 5	Treatment A
Sample 6	Control

Let's now consider this parameterization:

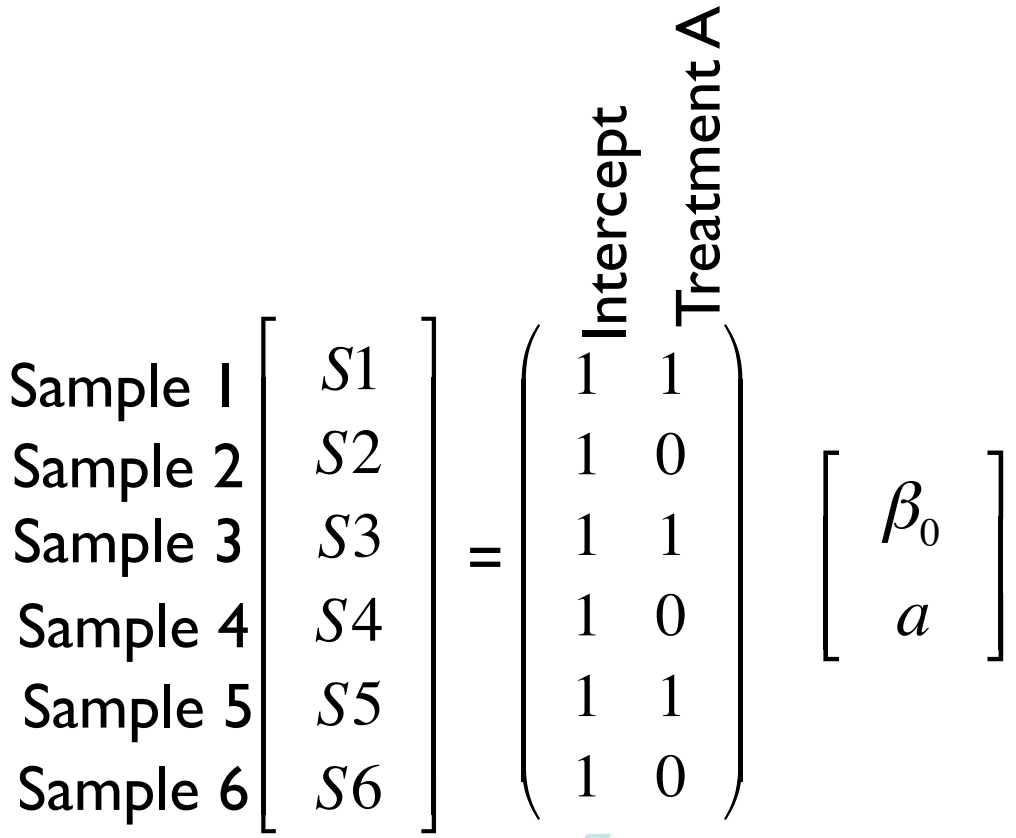
C= Baseline expression
T_A= Baseline expression + effect of treatment

So the set of parameters are:

C = Control (mean expression of the control)
a = T_A – Control (mean change in expression under treatment

Intercept

Different parameterization: using intercept



Parameters (coefficients, levels of the variable)

Intercept measures the baseline expression.

a measures now the differential expression between Treatment A and Control

Design Matrix

Contrast matrices

Are the two parameterizations equivalent?

$$\begin{bmatrix} 1 & -1 \end{bmatrix} \begin{bmatrix} \hat{T} \\ \hat{C} \end{bmatrix} = \widehat{T - C}$$

Contrast matrix

Contrast matrices allow us to estimate (and test) linear combinations of our coefficients.

Models with 1 factor, more than 2 levels

Sample	Treatment
Sample1	Treatment A
Sample 2	Treatment B
Sample 3	Control
Sample 4	Treatment A
Sample 5	Treatment B
Sample 6	Control

Number of samples: 6

Number of factors: 1

Treatment: Number of levels: 3

Possible parameters (What differences are important)?

- Effect of Treatment A
- Effect of Treatment B
- Effect of Control
- Differences between treatments?

Design matrix for ANOVA models

Sample	Treatment
Sample1	Treatment A
Sample 2	Treatment B
Sample 3	Control
Sample 4	Treatment A
Sample 5	Treatment B
Sample 6	Control

$$\begin{bmatrix} S1 \\ S2 \\ S3 \\ S4 \\ S5 \\ S6 \end{bmatrix} = \begin{pmatrix} \quad \end{pmatrix} \begin{bmatrix} T_A \\ T_B \\ C \end{bmatrix}$$

$$\begin{bmatrix} S1 \\ S2 \\ S3 \\ S4 \\ S5 \\ S6 \end{bmatrix} = \begin{pmatrix} \quad \end{pmatrix} \begin{bmatrix} \beta_0 \\ a \\ b \end{bmatrix}$$

Design matrix for ANOVA models

Sample	Treatment
Sample1	Treatment A
Sample 2	Treatment B
Sample 3	Control
Sample 4	Treatment A
Sample 5	Treatment B
Sample 6	Control

$$\begin{bmatrix} S1 \\ S2 \\ S3 \\ S4 \\ S5 \\ S6 \end{bmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{bmatrix} T_A \\ T_B \\ C \end{bmatrix}$$

$$\begin{bmatrix} S1 \\ S2 \\ S3 \\ S4 \\ S5 \\ S6 \end{bmatrix} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} \begin{bmatrix} \beta_0 \\ a \\ b \end{bmatrix}$$

Control = Baseline

T_A = Baseline + a

T_B = Baseline + b

Baseline levels

The model with intercept always take one level as a baseline:

The baseline is treatment A, the coefficients are comparisons against it!

By default, R uses the first level as baseline

$$\begin{bmatrix} S1 \\ S2 \\ S3 \\ S4 \\ S5 \\ S6 \end{bmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix} \begin{bmatrix} \beta_0 \\ b \\ c \end{bmatrix}$$

R code

R code:

```
> Treatment <- rep(c("TreatmentA", "TreatmentB", "Control"), 2)
> design.matrix <- model.matrix(~ Treatment)  (model with intercept)
> design.matrix <- model.matrix(~ -1 + Treatment)  (model without intercept)
> design.matrix <- model.matrix(~ 0 + Treatment)  (model without intercept)
```

Exercise

Build contrast matrices for all pairwise comparisons for this design:

$$\begin{bmatrix} S1 \\ S2 \\ S3 \\ S4 \\ S5 \\ S6 \end{bmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{bmatrix} T_A \\ T_B \\ C \end{bmatrix} \qquad \begin{pmatrix} \quad \end{pmatrix} \begin{bmatrix} \hat{T}_A \\ \hat{T}_B \\ \hat{C} \end{bmatrix}$$

Exercise

Build contrast matrices for all pairwise comparisons for these designs:

$$\begin{bmatrix} S1 \\ S2 \\ S3 \\ S4 \\ S5 \\ S6 \end{bmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{bmatrix} T_A \\ T_B \\ C \end{bmatrix}$$

$$\begin{pmatrix} 1 & 0 & -1 \\ 0 & 1 & -1 \\ 1 & -1 & 0 \end{pmatrix} \begin{bmatrix} \hat{T}_A \\ \hat{T}_B \\ \hat{C} \end{bmatrix}$$

Exercise

Build contrast matrices for all pairwise comparisons for these designs:

$$\begin{bmatrix} S1 \\ S2 \\ S3 \\ S4 \\ S5 \\ S6 \end{bmatrix} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \\ 1 & 0 & 0 \end{pmatrix} \begin{bmatrix} \beta_0 \\ a \\ b \end{bmatrix} \qquad \begin{pmatrix} \quad \end{pmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{a} \\ \hat{b} \end{bmatrix}$$

Exercise

Build contrast matrices for all pairwise comparisons for these designs:

$$\begin{bmatrix} S1 \\ S2 \\ S3 \\ S4 \\ S5 \\ S6 \end{bmatrix} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \\ 1 & 0 & 0 \end{pmatrix} \begin{bmatrix} \beta_0 \\ a \\ b \end{bmatrix}$$

$$\begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & -1 \end{pmatrix} \begin{bmatrix} \hat{\beta}_0 \\ \hat{a} \\ \hat{b} \end{bmatrix}$$

Models with 2 factors

Sample	Treatment	ER status
Sample1	Treatment A	+
Sample 2	No Treatment	+
Sample 3	Treatment A	+
Sample 4	No Treatment	+
Sample 5	Treatment A	-
Sample 6	No Treatment	-
Sample 7	Treatment A	-
Sample 8	No Treatment	-

Number of samples: 8

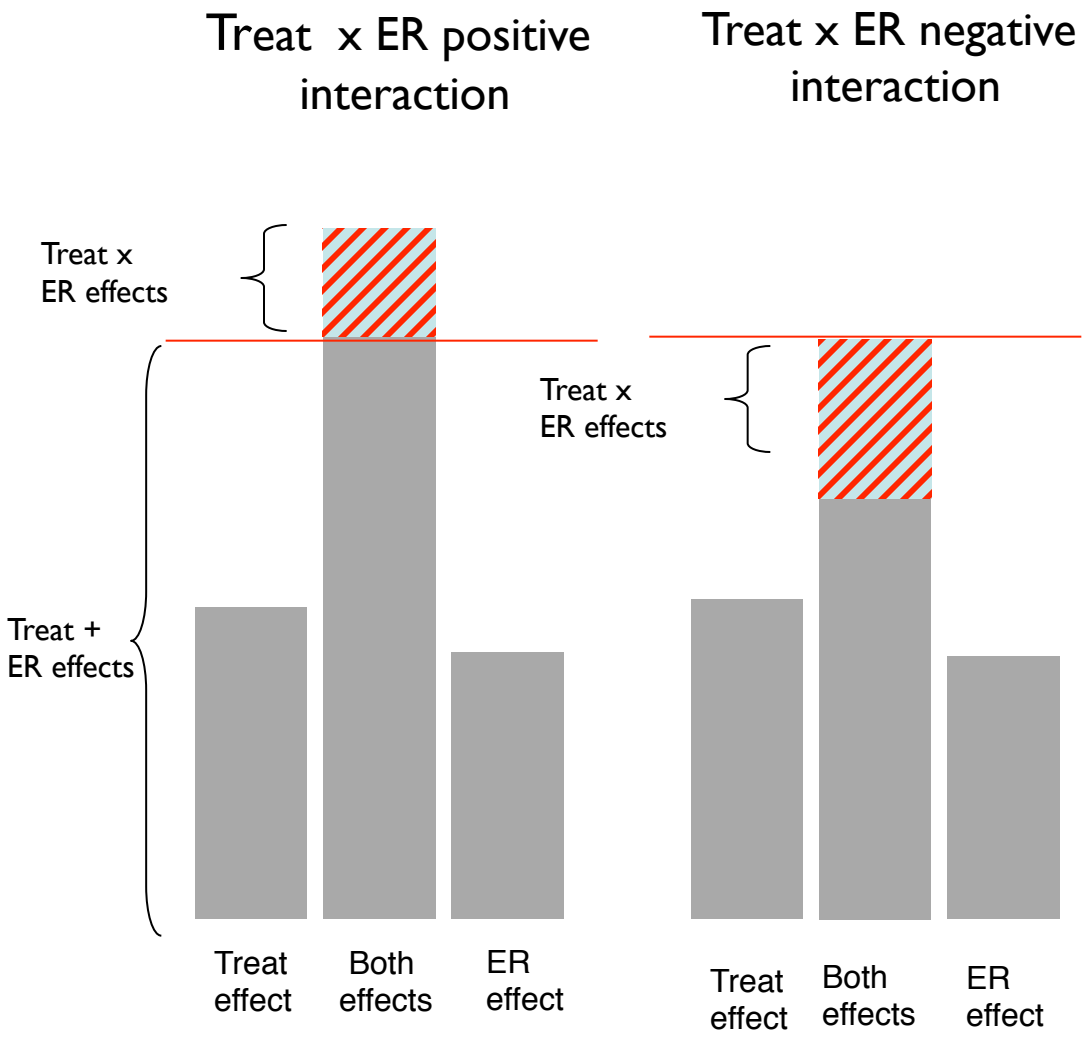
Number of factors: 2

Treatment: Number of levels: 2

ER: Number of levels: 2

Understanding Interactions

	No Treat	Treat A
ER -	S6, S8	S5, S7
ER +	S2, S4	S1, S3



Adapted from Natalie Thorne, Nuno L. Barbosa Morais

Models with 2 factors and no interaction

Model with no interaction: only ***main effects***

Number of coefficients (parameters):

Intercept + (#levels Treat - 1) + (#levels ER - 1) = 3

If we remove the intercept, the additional parameter comes from the missing level in one of the variables, but in models with more than 1 factor it is a good idea to keep the intercept.

Models with 2 factors (no interaction)

R code: `> design.matrix <- model.matrix(~Treatment+ER)` (model with intercept)

$$\begin{bmatrix} S1 \\ S2 \\ S3 \\ S4 \\ S5 \\ S6 \\ S7 \\ S8 \end{bmatrix} = \begin{pmatrix} \end{pmatrix} \begin{bmatrix} \beta_0 \\ a \\ er+ \end{bmatrix}$$

In R, the baseline for each variable is the first level.

	No Treat	Treat A
ER -	S6, S8	S5, S7
ER +	S2, S4	S1, S3

Models with 2 factors (no interaction)

R code: `> design.matrix <- model.matrix(~Treatment+ER)` (model with intercept)

$$\begin{bmatrix} S1 \\ S2 \\ S3 \\ S4 \\ S5 \\ S6 \\ S7 \\ S8 \end{bmatrix} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix} \begin{bmatrix} \beta_0 \\ a \\ er+ \end{bmatrix}$$

	No Treat	Treat A
ER -	S6, S8	S5, S7
ER +	S2, S4	S1, S3

Models with 2 factors and interaction

Model with interaction: ***main effects + interaction***

Number of coefficients (parameters):

Intercept + (#levels Treat - 1) +
(#levels ER - 1) +((#levels Treat - 1) *
(#levels ER - 1)) = 4

Models with 2 factors (interaction)

R code: `> design.matrix <- model.matrix(~Treatment*ER)` (model with intercept)

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \\ Y_6 \\ Y_7 \\ Y_8 \end{bmatrix} = \begin{pmatrix} \end{pmatrix} \begin{bmatrix} \beta_0 \\ a \\ er+ \\ a.er+ \end{bmatrix}$$

"Extra effect" of Treatment A on ER+ samples

	No Treat	Treat A
ER -	S6, S8	S5, S7
ER +	S2, S4	S1, S3

Models with 2 factors (interaction)

R code: `> design.matrix <- model.matrix(~Treatment*ER)` (model with intercept)

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \\ Y_6 \\ Y_7 \\ Y_8 \end{bmatrix} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{pmatrix} \begin{bmatrix} \beta_0 \\ a \\ er+ \\ a.er+ \end{bmatrix}$$

"Extra effect" of Treatment A on ER+ samples

	No Treat	Treat A
ER -	S6, S8	S5, S7
ER +	S2, S4	S1, S3

Paired Designs

Sample	Type
Sample 1	Tumour
Sample 2	Matched Normal
Sample 3	Tumour
Sample 4	Matched Normal
Sample 5	Tumour
Sample 6	Matched Normal
Sample 7	Tumour
Sample 8	Matched Normal

Number of samples: 8
Number of factors: 1
Type: Number of levels: 2

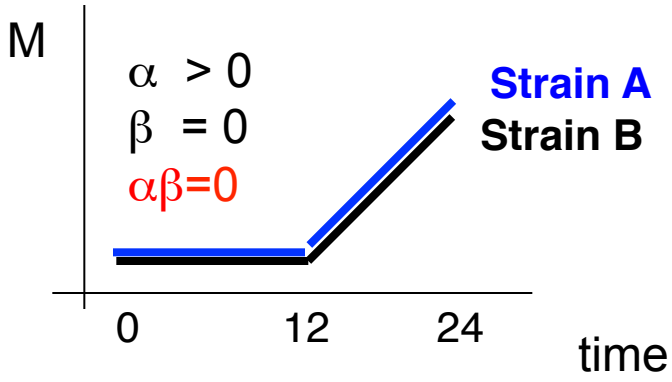
Sample	Type
Sample 1	Tumour
Sample 1	Matched Normal
Sample 2	Tumour
Sample 2	Matched Normal
Sample 3	Tumour
Sample 3	Matched Normal
Sample 4	Tumour
Sample 4	Matched Normal

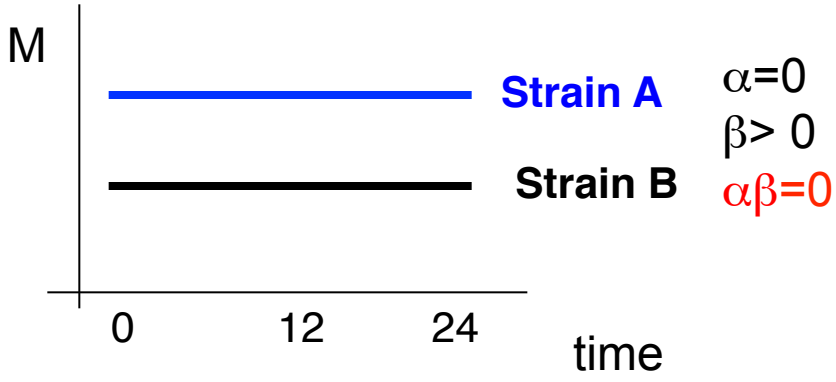
Number of samples: 4
Number of factors: 2
Sample: Number of levels: 4
Type: Number of levels: 2

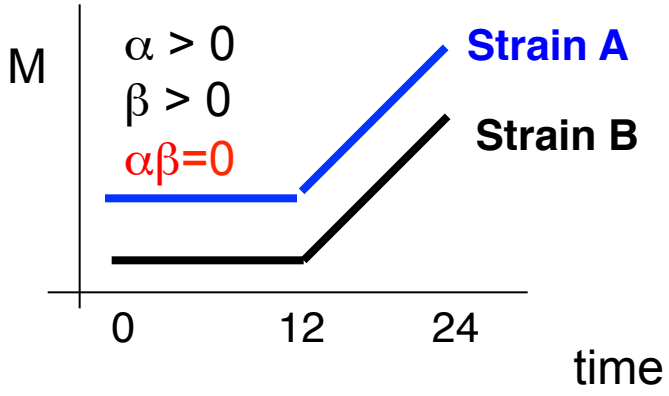
2 by 3 factorial experiment

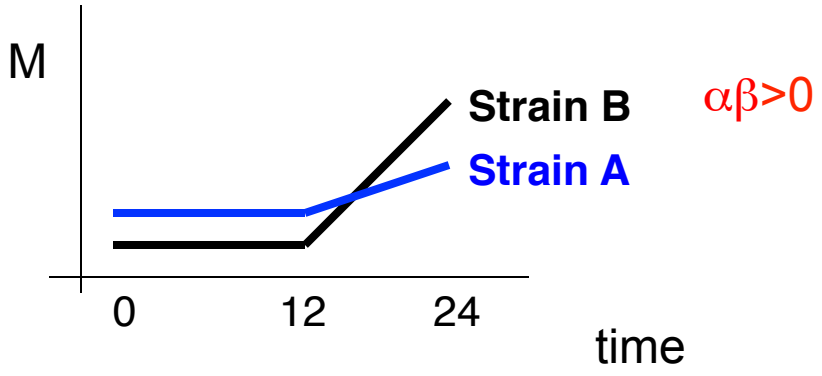
- Identify DE genes that have different time profiles between different mutants.

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









Design matrix for Paired experiments

We can gain precision in our estimates with a paired design, because individual variability is removed when we compare the effect of the treatment within the same sample.

R code: `> design.matrix <- model.matrix(~-1 +Type)` (unpaired; model without intercept)

`> design.matrix <- model.matrix(~-1 +Sample+Type)` (paired; model

Sample	Type
Sample 1	Tumour
Sample 1	Matched Normal
Sample 2	Tumour
Sample 2	Matched Normal
Sample 3	Tumour
Sample 3	Matched Normal
Sample 4	Tumour
Sample 4	Matched Normal

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \\ Y_6 \\ Y_7 \\ Y_8 \end{bmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix} \begin{bmatrix} S1 \\ S2 \\ S3 \\ S4 \\ t \end{bmatrix}$$

These effects only reflect biological differences not related to tumour/ normal effect.

Analysis of covariance (Models with categorical and continuous variables)

Sample	ER	Dose
Sample 1	+	37
Sample 2	-	52
Sample 3	+	65
Sample 4	-	89
Sample 5	+	24
Sample 6	-	19
Sample 7	+	54
Sample 8	-	67

Number of samples: 8
Number of factors: 2
ER: Number of levels: 2
Dose: Continuous

Analysis of covariance (Models with categorical and continuous variables)

R code: `> design.matrix <- model.matrix(~ ER + dose)`

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \\ Y_6 \\ Y_7 \\ Y_8 \end{bmatrix} = \begin{pmatrix} 1 & 1 & 37 \\ 1 & 0 & 52 \\ 1 & 1 & 65 \\ 1 & 0 & 89 \\ 1 & 1 & 24 \\ 1 & 0 & 19 \\ 1 & 1 & 54 \\ 1 & 0 & 67 \end{pmatrix} \begin{bmatrix} \beta_0 \\ er + \\ d \end{bmatrix}$$

If we consider the effect of dose ***linear*** we use 1 coefficient (degree of freedom). We can also model it as non-linear (using splines, for example).

Sample	ER	Dose
Sample 1	+	37
Sample 2	-	52
Sample 3	+	65
Sample 4	-	89
Sample 5	+	24
Sample 6	-	19
Sample 7	+	54
Sample 8	-	67

Analysis of covariance (Models with categorical and continuous variables)

Interaction: ***Is it the effect of dose equal in ER + and ER -?***

R code: `> design.matrix <- model.matrix(~ ER * dose)`

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \\ Y_6 \\ Y_7 \\ Y_8 \end{bmatrix} = \begin{pmatrix} 1 & 1 & 37 & 37 \\ 1 & 0 & 52 & 0 \\ 1 & 1 & 65 & 65 \\ 1 & 0 & 89 & 0 \\ 1 & 1 & 24 & 24 \\ 1 & 0 & 19 & 0 \\ 1 & 1 & 54 & 54 \\ 1 & 0 & 67 & 0 \end{pmatrix} \begin{bmatrix} \beta_0 \\ er+ \\ d \\ er+.d \end{bmatrix}$$

If the interaction is significant, the effect on the dose is different depending on the levels of ER.

Sample	ER	Dose
Sample 1	+	37
Sample 2	-	52
Sample 3	+	65
Sample 4	-	89
Sample 5	+	24
Sample 6	-	19
Sample 7	+	54
Sample 8	-	67

Time Course experiments

Sample	Time
Sample 1	0h
Sample 1	1h
Sample 1	4h
Sample 1	16h
Sample 2	0h
Sample 2	1h
Sample 2	4h
Sample 2	16h

Number of samples: 2
Number of factors: 2

Sample: Number of levels: 2
Time: Continuous or categorical?

Intermediate solution: ***splines***

Main question: how does expression change over time?

If we model time as categorical, we don't make assumptions about its effect, but we use too many degrees of freedom.

If we model time as continuous, we use less degrees of freedom but we have to make assumptions about the type of effect.

Time Course experiments: no assumptions

R code: `> design.matrix <- model.matrix(~Sample + factor(Time))`

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \\ Y_6 \\ Y_7 \\ Y_8 \end{bmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \end{pmatrix} \begin{bmatrix} S_1 \\ S_2 \\ T_1 \\ T_4 \\ T_{16} \end{bmatrix}$$

We can use contrasts to test differences at time points.

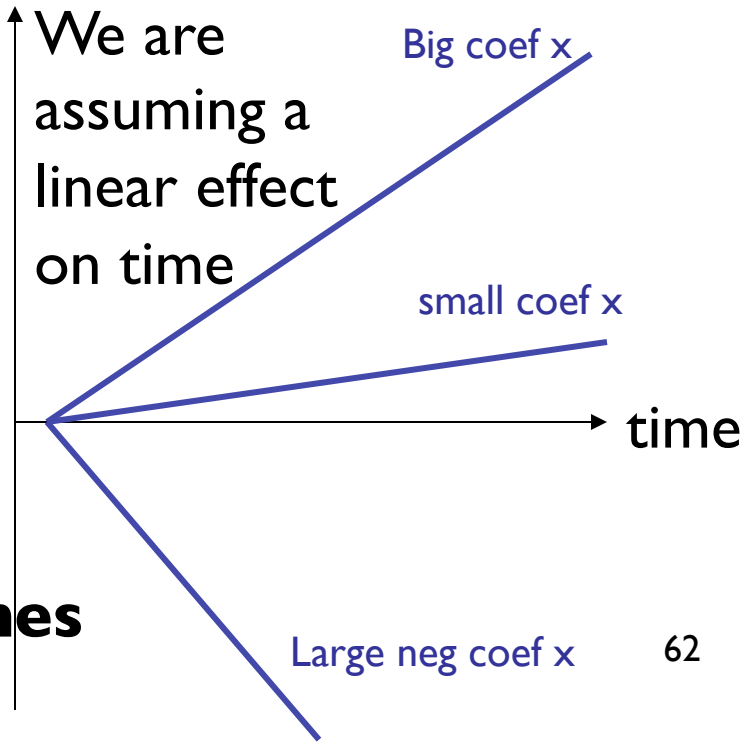
Sample	Time
Sample 1	0h
Sample 1	1h
Sample 1	4h
Sample 1	16h
Sample 2	0h
Sample 2	1h
Sample 2	4h
Sample 2	16h

Time Course experiments

Sample	Time
Sample 1	0h
Sample 1	1h
Sample 1	4h
Sample 1	16h
Sample 2	0h
Sample 2	1h
Sample 2	4h
Sample 2	16h

R code: `> design.matrix <- model.matrix(~Sample + Time)`

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ Y_4 \\ Y_5 \\ Y_6 \\ Y_7 \\ Y_8 \end{bmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 4 \\ 1 & 0 & 16 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 4 \\ 0 & 1 & 16 \end{pmatrix} \begin{bmatrix} S_1 \\ S_2 \\ X \end{bmatrix}$$



Intermediate models are possible: **splines**

Multi factorial models

- We can fit models with many variables
- Sample size must be adequate to the number of factors
- Same rules for building the design matrix must be used:
  - There will be one column in design matrix for the intercept
  - Continuous variables with a linear effect will need one column in the design matrix
  - Categorical variable will need #levels -1 columns
  - Interactions will need (#levels -1) x (#levels -1)
  - It is possible to include interactions of more than 2 variables, but the number of samples needed to accurately estimate those interactions is large.

Parameter estimation and hypothesis testing

Statistical models

– We want to model the expected result of an outcome (dependent variable) under given values of other variables (independent variables)

Arbitrary function (any shape)

Expected value of variable Y

A set of k independent variables (also called factors)

$$E(Y) = f(X)$$

$$Y = f(X) + \varepsilon$$

This is the variability around the expected mean of y

Linear models

– The observed value of Y is a linear combination of the effects of the independent variables

Arbitrary number of independent variables

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

Polynomials are valid

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \ldots + \beta_p X_1^p$$

$$E(Y) = \beta_0 + \beta_1 \log(X_1) + \beta_2 f(X_2) + \ldots + \beta_k X_k$$

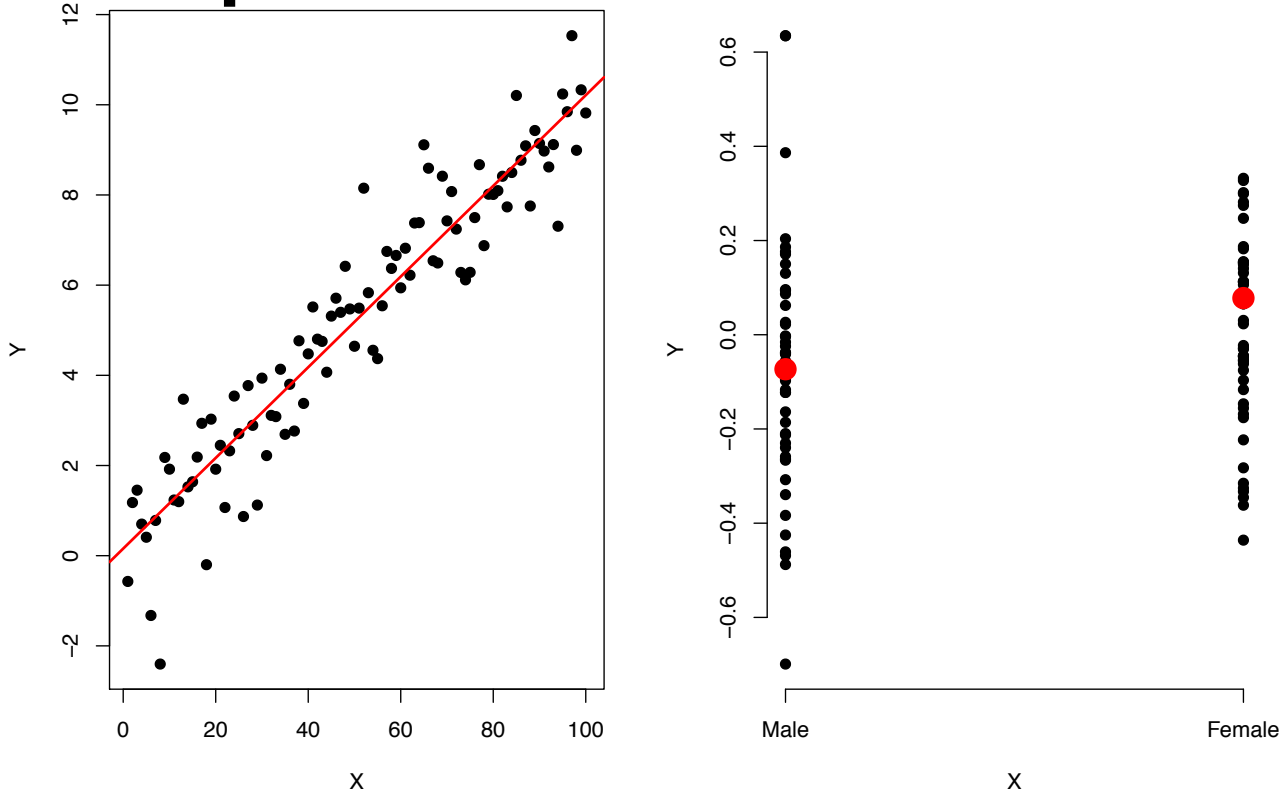
We can use functions of the variables if the effects are linear

Smooth functions: not exactly the same as the so-called ***additive models***

– If we include categorical variables the model is called **General Linear Model**

Model Estimation

We use **least squares estimation**



Given n observations (y_1,..y_n,$\mathbf{x}_1$,..$\mathbf{x}_n$) minimize the differences between the observed and the predicted values

Model Estimation

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

β → Parameter of interest (effect of X on Y)

$\hat{\beta}$ → **Estimator** of the parameter of interest

$se(\hat{\beta})$ → **Standard Error** of the estimator of the parameter of interest

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

$$se(\hat{\beta}_i) = \sigma \sqrt{c_i}$$

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

$\hat{y} = \hat{\beta} x$ → Fitted values (predicted by the model)

$e = y - \hat{y}$ → Residuals (observed errors)

Generalized linear models

- Extension of the linear model to other distributions and non-linearity in the structure (to some degree)

Link function

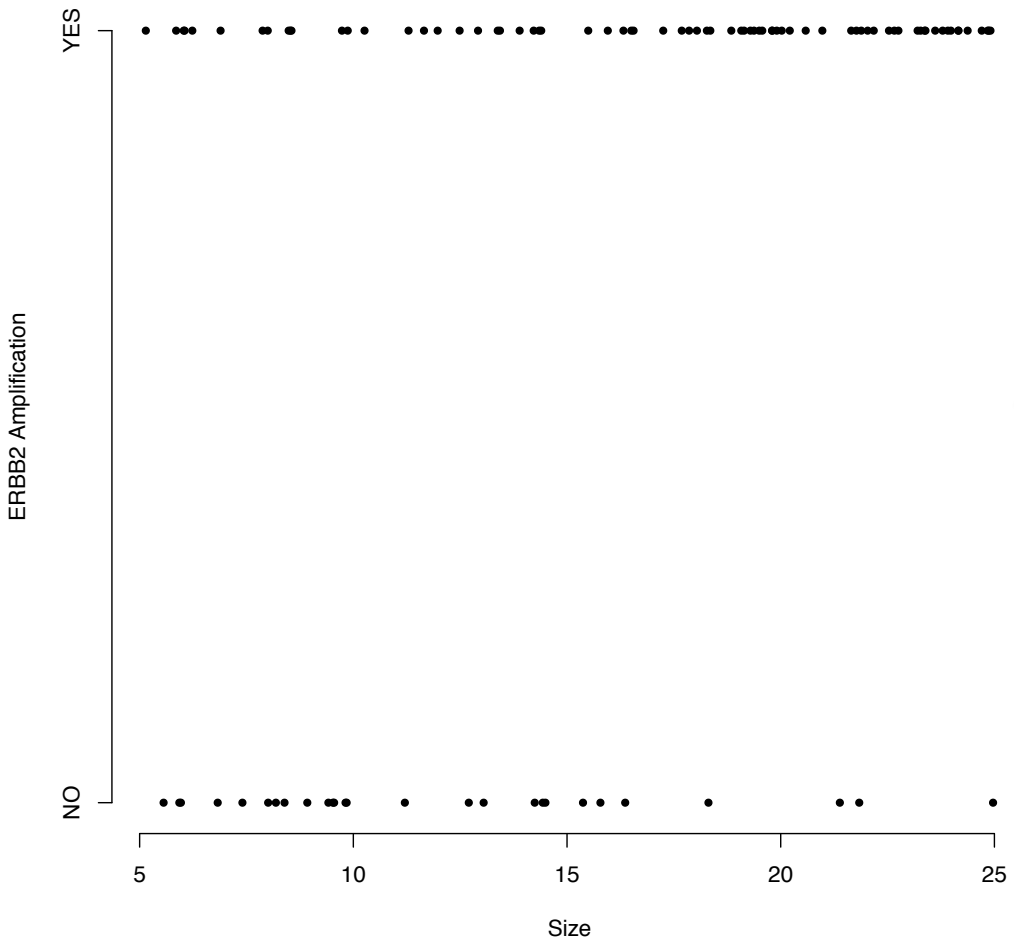
$$g(E(Y)) = X\beta$$

- Y must follow a probability distribution from the exponential family (Bernoulli, Binomial, Poisson, Gamma, Normal,…)
- Parameter estimation must be performed using an iterative method (IWLS).

Example: Logistic Regression

- We want to study the relationship between the presence of an amplification in the ERBB2 gene and the size of the tumour in a specific type of breast cancer.
- Our dependent variable Y, takes two possible values: “AMP”, “NORMAL” (“YES”, “NO”)
- X (size) takes continuous values.

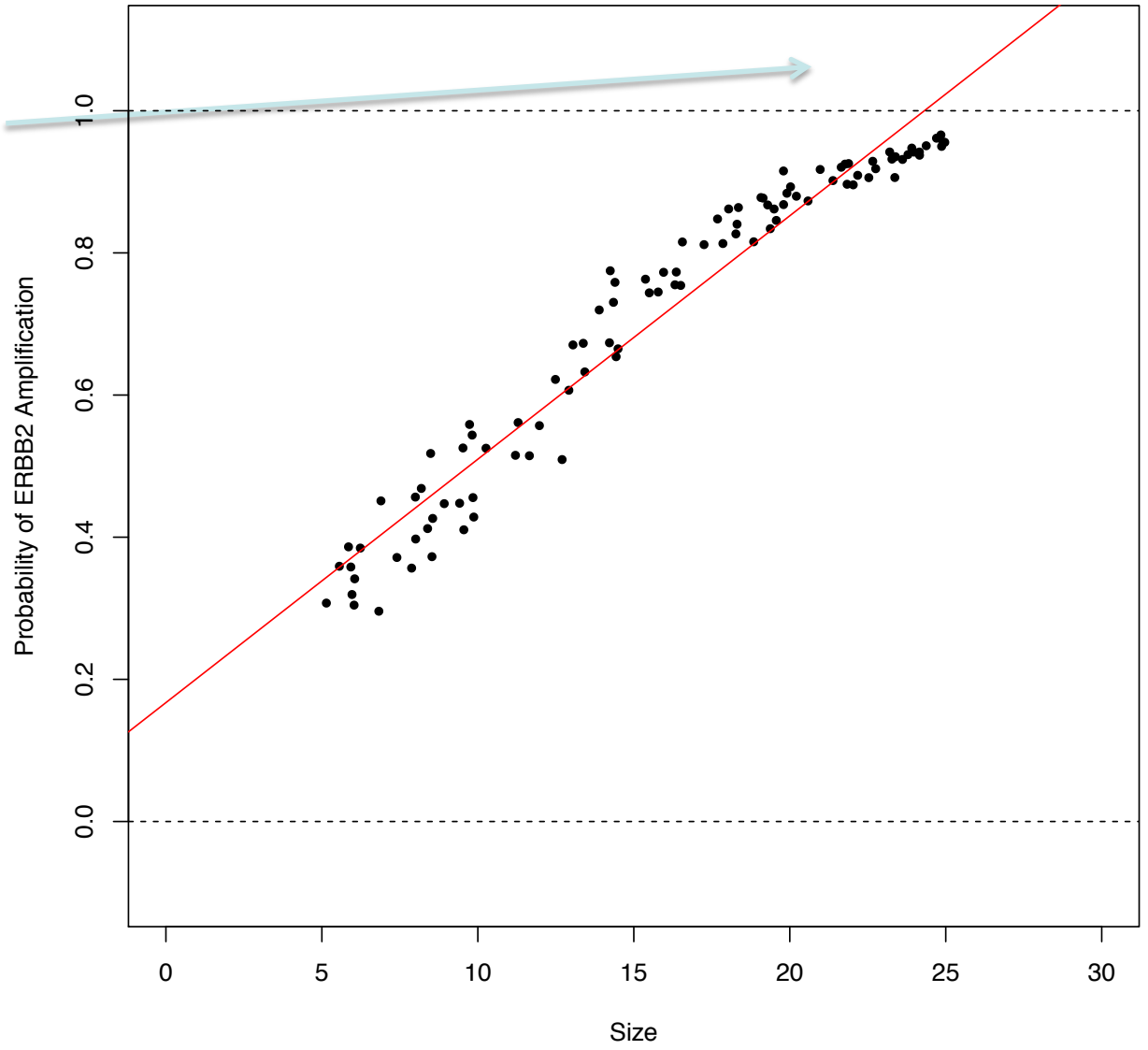
Example: Logistic Regression



It is very difficult to see the relationship. Let's model the **"probability of success"**: in this case, the probability of amplification

Example: Logistic Regression

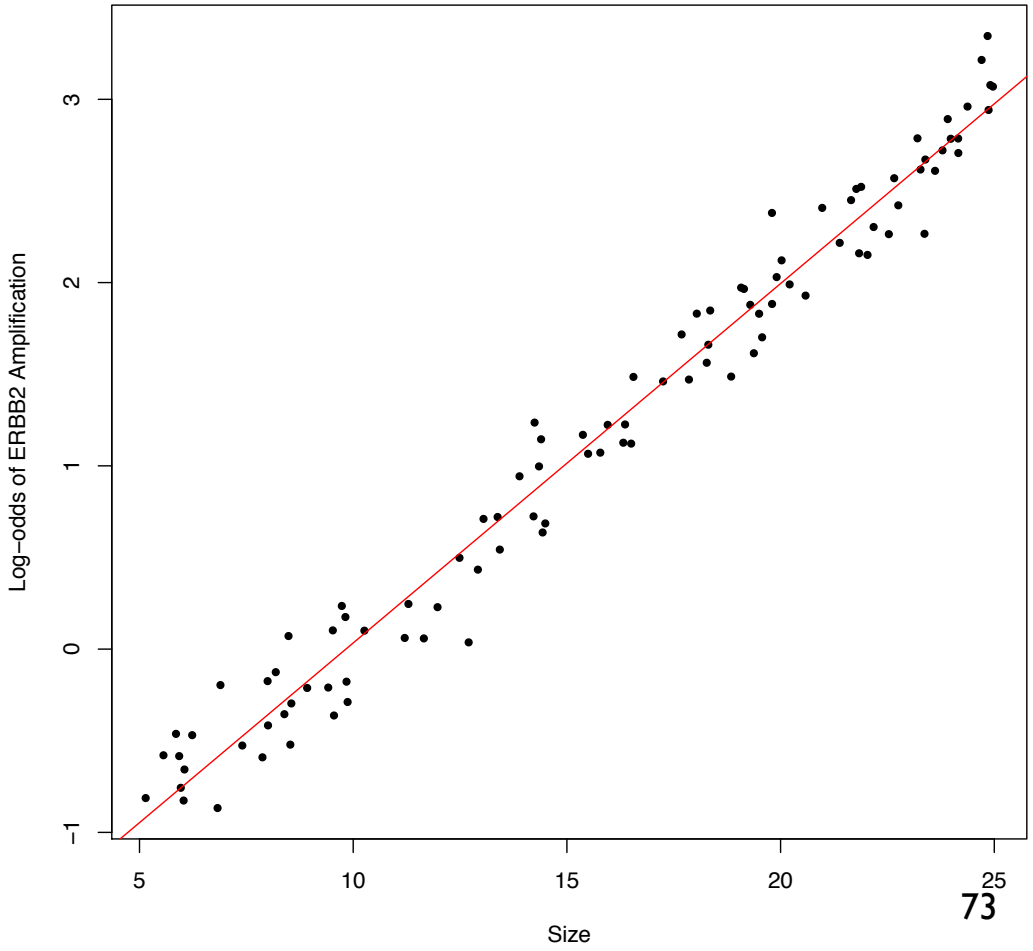
Some predictions are out of the possible range for a probability



Example: Logistic Regression

We can transform the probabilities to a scale that goes from –Inf to Inf using **log odds**

$$\log odds = \log\left(\frac{p}{1-p}\right)$$



Example: Logistic Regression

How does this relate to the generalized linear model?

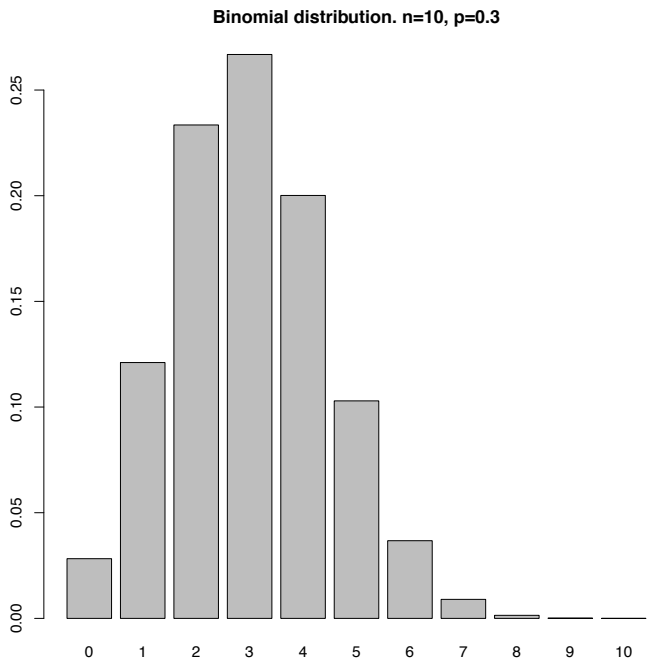
- Y follows a Bernoulli distribution; it can take two values (YES or NO)
- The expectation of Y, p is the probability of YES (EY=p)
- We assume that there is a linear relationship between size and a function of the expected value of Y: the log odds (the **link** function)

$$\log odds(prob.amplif) = \beta_0 + \beta_1 Size$$

$$g(EY) = \beta X$$

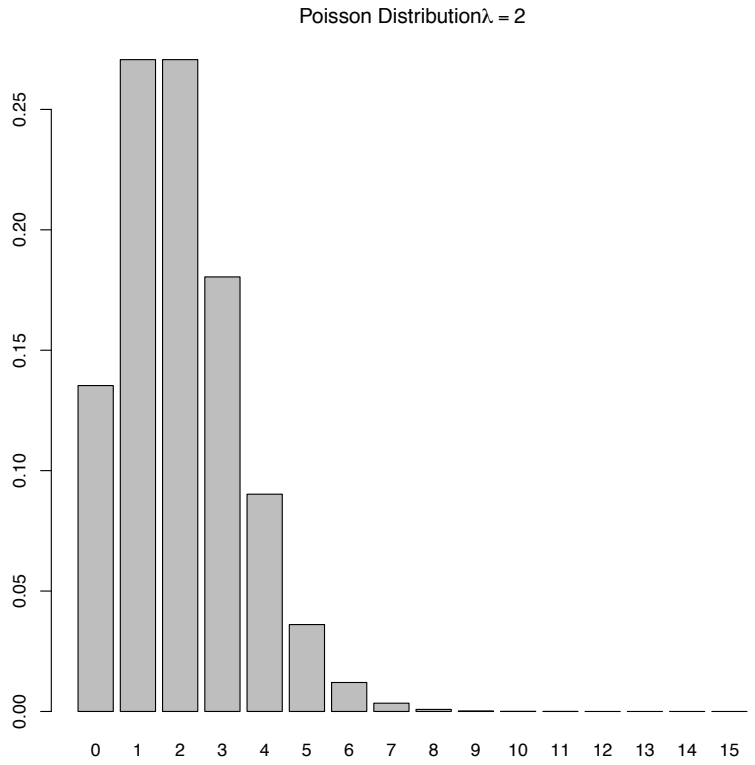
Binomial Distribution

- It is the distribution of the number of events in a series of *n* independent *Bernoulli* experiments, each with a probability of success **p**.
- Y can take integer values from 0 to n
- EY=np
- VarY= np(1-p)



Poisson Distribution

- Let Y ~ B(n,p). If n is large and p is small then Y can be approximated by a Poisson Distribution (*Law of rare events*)
- Y ~ P(λ)
- EY= λ
- VarY= λ

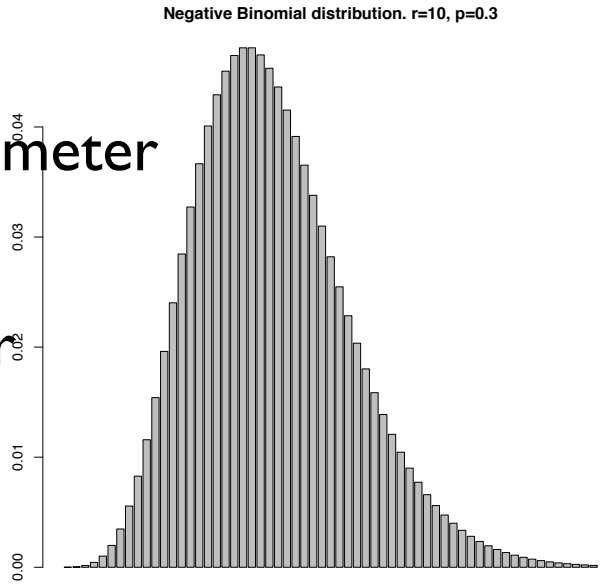


Negative Binomial Distribution

- Let Y ~ NB(r,p)
- Represents the number of successes in a Bernoulli experiment until r failures occur.
- It is also the distribution of a continuous mixture of Poisson distributions where λ follows a Gamma distribution.
- It can be seen as a overdispersed Poisson distribution.

$$p = \frac{\mu}{\sigma^2} \longrightarrow \text{Overdispersion parameter}$$

$$r = \frac{\mu^2}{\sigma^2 - \mu} \longrightarrow \text{Location parameter}$$



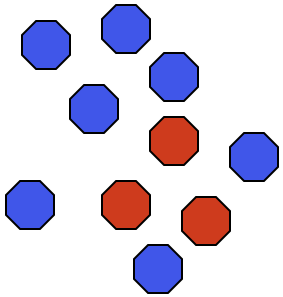
Hypothesis testing

- Everything starts with a biological question to test:
  - **What genes are differentially expressed under one treatment?**
  - **What genes are more commonly amplified in a class of tumours?**
  - **What promoters are methylated more frequently in cancer?**
- We must express this biological question in terms of a parameter in a model.
- We then conduct an experiment, obtain data and estimate the parameter.
- How do we take into account uncertainty in order to answer our question based on our estimate?

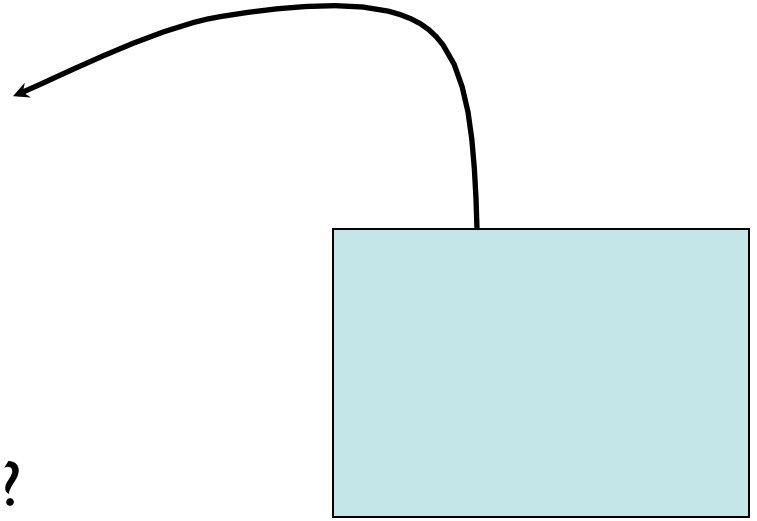
Sampling and testing

Discrete observations

#red = 3



Random sample of 10 balls from the box



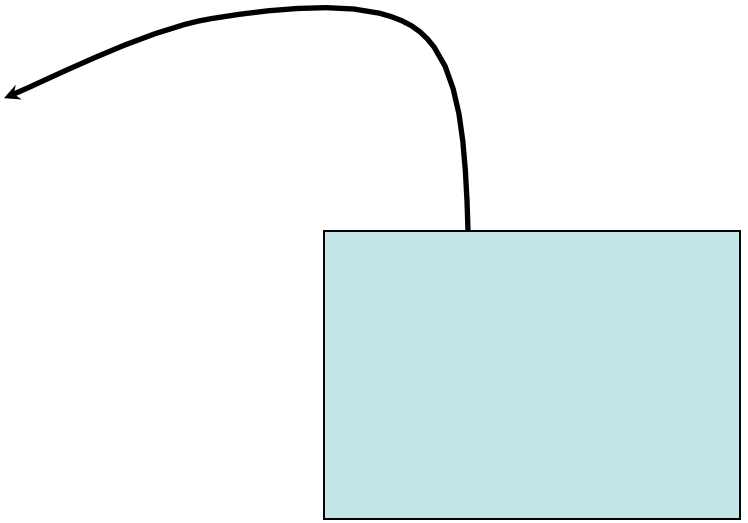
When do I think that I am not sampling from this box anymore?

How many reds could I expect to get <u>just by chance alone!</u>

10% red balls and
90% blue balls

Sample

Random sample of 10 balls from the box



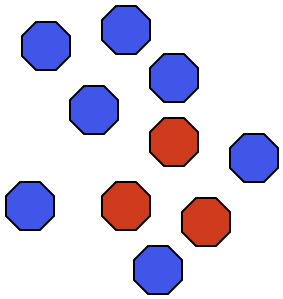
10% red balls and
90% blue balls

Null hypothesis

(about the population that is being sampled)

Discrete observations

#red = 3



Test statistic

Rejection criteria (based on your observed sample, do you have evidence to reject the hypothesis that you sampled from the null population)

Hypothesis testing

- **Null Hypothesis**: Our population follows a (known) distribution defined by a set of parameters: $H_0 : X \sim f(\theta_1,\ldots\theta_k)$
- Take a random sample $(X_1,\ldots X_n) = (x_1,\ldots x_n)$ and observe **test statistic**

  $T(X_1,\ldots X_n)= t(x_1,\ldots x_n)$
- The distribution of T under H_0 is known (g(.))
- **p-value** : probability under H_0 of observing a result as extreme as $t(x_1,\ldots x_n)$

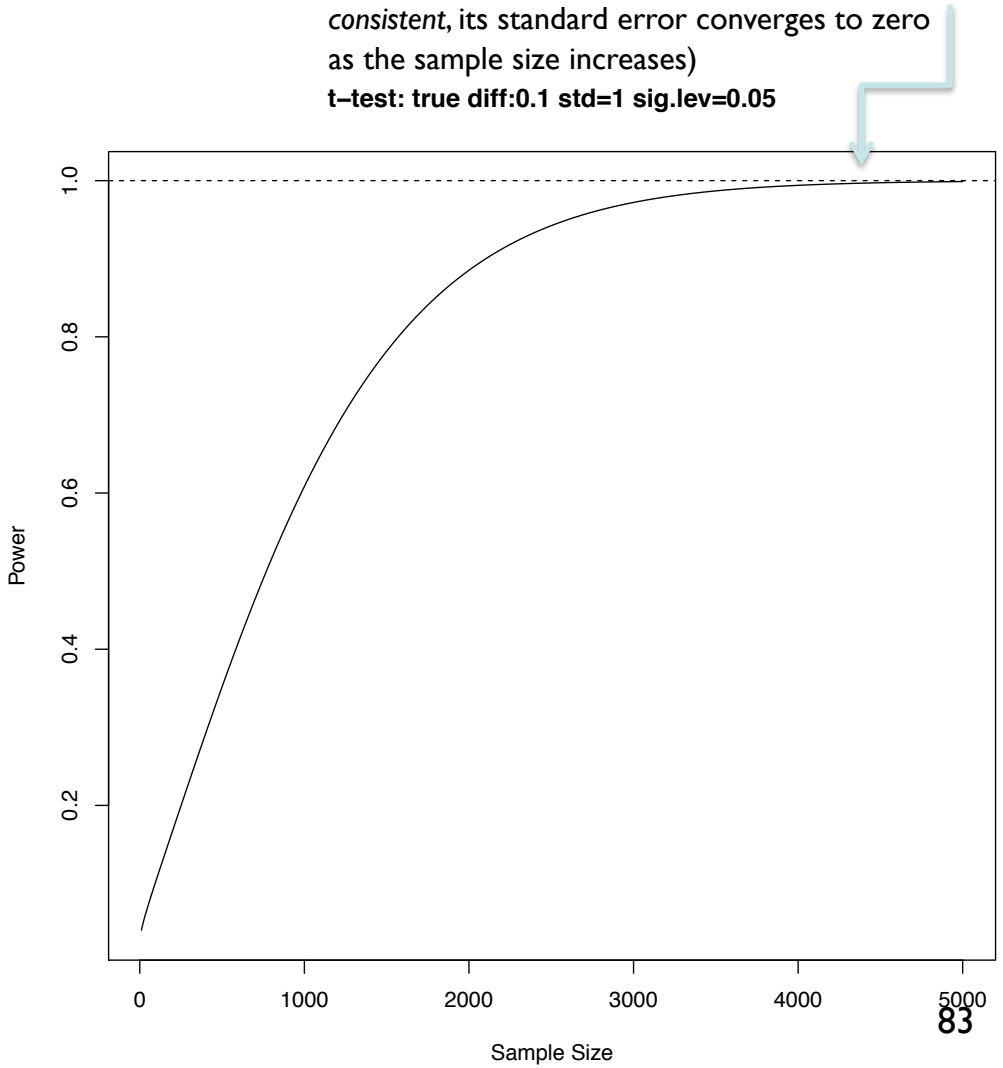
Type I and Type II errors

- Type I error: probability of rejecting the null hypothesis when it is true. Usually, it is the significance level of the test. It is denoted as α
- Type II error: probability of not rejecting the null hypothesis when it is false It is denoted as β
- Decreasing one type of error increases the other, so in practice we fix the type I error and choose the test that minimizes type II error.

The power of a test

- The power of a test is the probability of rejecting the null hypothesis at a given significance level when a specific alternative is true
- For a given significance level and a given alternative hypothesis in a given test, the power is a function of the sample size
- What is the difference between statistical significance and biological significance?

With enough sample size, we can detect ***any*** alternative hypothesis (if the estimator is *consistent*, its standard error converges to zero as the sample size increases)



The Likelihood Ratio Test (LRT)

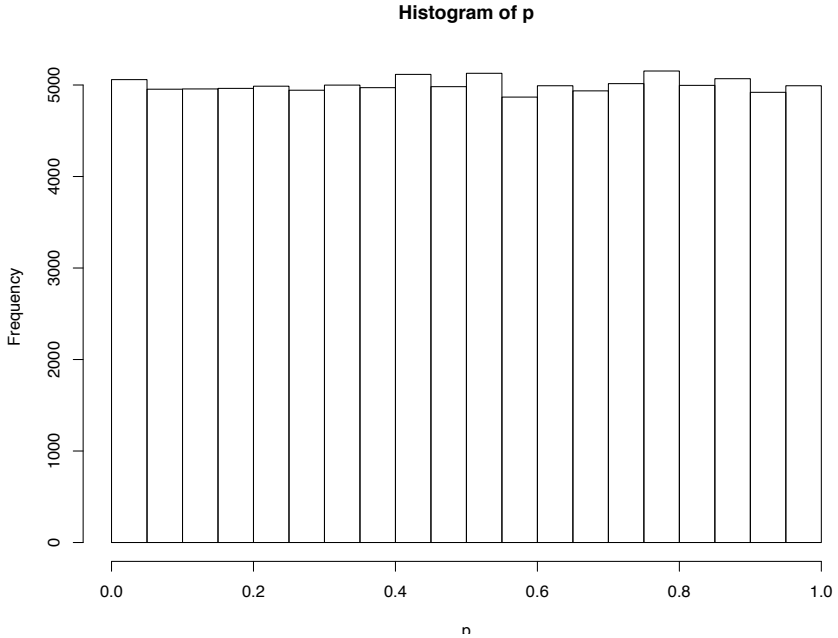
- We are working with models, therefore we would like to do hypothesis tests on coefficients or contrasts of those models
- We fit two models M_1 without the coefficient to test and M_2 with the coefficient.
- We compute the likelihoods of the two models (L_1 and L_2) and obtain $LRT=-2\log(L_1 / L_2)$ that has a known distribution under the null hypothesis that the two models are equivalent. This is also known as ***model selection***.

Multiple testing problem

- In microarray/sequencing experiments we are fitting one model for each probe/gene/ exon/sequence of interest, and therefore performing thousands of tests.
- Type I error is not equal to the significance level of each test.
- Multiple test corrections try to fix this problem (Bonferroni, FDR,…)

Distribution of p-values

If the null hypothesis is true, the p-values from the repeated experiments come from a Uniform(0,1) distribution.



Slide by Alex Lewin, Ernest Turro, Paul O'Reilly

Controlling the number of errors

N = number of hypothesis tested
R = number of rejected hypothesis
n_0 = number of true hypothesis

	Null Hypothesis True	Alternative Hypothesis True	Total
Not Significant (don't reject)	# True Negative	# False Negative (Type II error)	N- # Rejections
Significant (Reject)	# False positive (Type I error)	# True positive	# Total Rejections
Total	n_0	$N-n_0$	N

Bonferroni Correction

If the tests are independent:

P(at least one false positive | all null hypothesis are true) =

P(at least one p-value < α | all null hypothesis are true) = $1 - (1-\alpha)^m$

Usually, we set a threshold at α / n.

Bonferroni correction: reject each hypothesis at $\alpha / \mathbf{N}$ level

It is a very conservative method

False Discovery Rate (FDR)

N = number of hypothesis tested
R = number of rejected hypothesis
n_0 = number of true hypothesis

	Null Hypothesis True	Alternative Hypothesis True	Total
Not Significant (don't reject)	# True Negative	# False Negative (Type II error)	N- # Rejections
Significant (Reject)	V= # False positive (Type I error)	# True positive	R= # Total Rejections
Total	n_0	$N-n_0$	N

Family Wise Error Rate: FWER = $P(V \geq 1)$
False Discovery Rate: FDR = $E(V/R \mid R>0)\ P(R>0)$

FDR aims to control the set of false positives among the rejected null hypothesis.

Benjamini & Hochberg (BH) step-up method to control FDR

Benjamini & Hochberg proposed the idea of controlling FDR, and used a step-wise method for controlling it.

Step 1: compare **largest** p-value to the specified significance level α:
If $p_m^{ord} > \alpha$ then don't reject corresponding null hypothesis

Step 2: compare second largest p-value to a modified threshold:
If $p_{m-1}^{ord} > \alpha * (m-1)/m$ then don't reject corresponding null hypothesis

Step 3:
If $p_{m-2}^{ord} > \alpha * (m-2)/m$ then don't reject corresponding null hypothesis

. . .

Stop when a p-value is lower than the modified threshold:

All other null hypotheses (with smaller p-values) are rejected.

Slide by Alex Lewin, Ernest Turro, Paul O'Reilly

Adjusted p-values for BH FDR

The final threshold on p-values below which all null hypotheses are rejected is $\alpha j^*/m$ where j^* is the index of the largest such p-value.

BH:
compare p_i to $\alpha j^*/m$ $\longleftrightarrow$ compare mp_i/j^* to α

Can define 'adjusted p-values' as mp_i/j^*

But these 'adjusted p-values' tell you the level of FDR which is being controlled (as opposed to the FWER in the Bonferroni and Holm cases).

Slide by Alex Lewin, Ernest Turro, Paul O'Reilly

Multiple power problem

- We have another problem related to the power of each test. Each unit tested has a different test statistic that depends on the variance of the distribution. This variance is usually different for each gene/transcript,…
- This means that the probability of detecting a given difference is different for each gene; if there is low variability in a gene we will reject the null hypothesis under a smaller difference
- Methods that shrinkage variance (like the empirical Bayes in limma for microarrays) deal with this problem.

Differential expression for microarray data

Differential Expression

• Compare gene expression under different conditions (treated vs untreated, wild type vs knockout, normal vs diseased tissue)

• Differential expression (DE) as list-making exercise: rank genes according to likelihood of (evidence for) DE

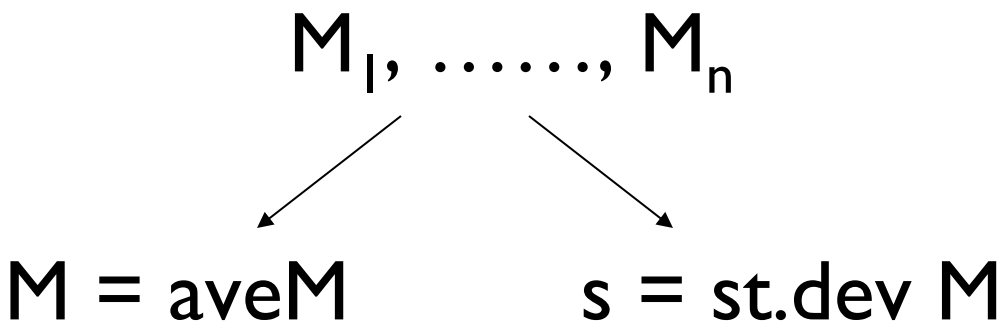
• Trade off: list length vs false positive (type 1 error) and false negative (type 2 error).

• What determines fold-change threshold?

• Some p-value for assessing significance would be nice . . .

Gene-wise summaries

- Each gene give a series of log-ratios
- Summarize log-ratios by the average and standard deviation for each gene



Summarising replicates to determine differential expression

Obvious thing : average M's

avM

But averages can be driven by outliers

Better than that : account for variability

t = avM / SE

But with 10,000 or so genes, some will have very small SE

Better still : use smoothed SE's

t* = avM / SE*

This is a modified t-statistic (also referred to as a moderated t).

SAM: a modified t-statistic

Significance analysis of microarrays applied to the ionizing radiation response

Virginia Goss Tusher*, Robert Tibshirani†, and Gilbert Chu*‡

*Departments of Medicine and Biochemistry, Stanford University, 269 Campus Drive, Center for Clinical Sciences Research 1115, Stanford, CA 94305-5151; and †Department of Health Research and Policy and Department of Statistics, Stanford University, Stanford, CA 94305

5116–5121 | PNAS | April 24, 2001 | vol. 98 | no. 9

Microarrays can measure the expression of thousands of genes to identify changes in expression between different biological states. Methods are needed to determine the significance of these changes while accounting for the enormous number of genes. We describe a method, Significance Analysis of Microarrays (SAM), that assigns a score to each gene on the basis of change in gene expression relative to the standard deviation of repeated measurements. For genes with scores greater than an adjustable threshold, SAM uses permutations of the repeated measurements to estimate the percentage of genes identified by chance, the false discovery rate (FDR). When the transcriptional response of human cells to ionizing radiation was measured by microarrays, SAM identified 34 genes that changed at least 1.5-fold with an estimated FDR of 12%, compared with FDRs of 60 and 84% by using conventional methods of analysis. Of the 34 genes, 19 were involved in cell cycle regulation and 3 in apoptosis. Surprisingly, four nucleotide excision repair genes were induced, suggesting that this repair pathway for UV-damaged DNA might play a previously unrecognized role in repairing DNA damaged by ionizing radiation.

SAM: a modified t-statistic

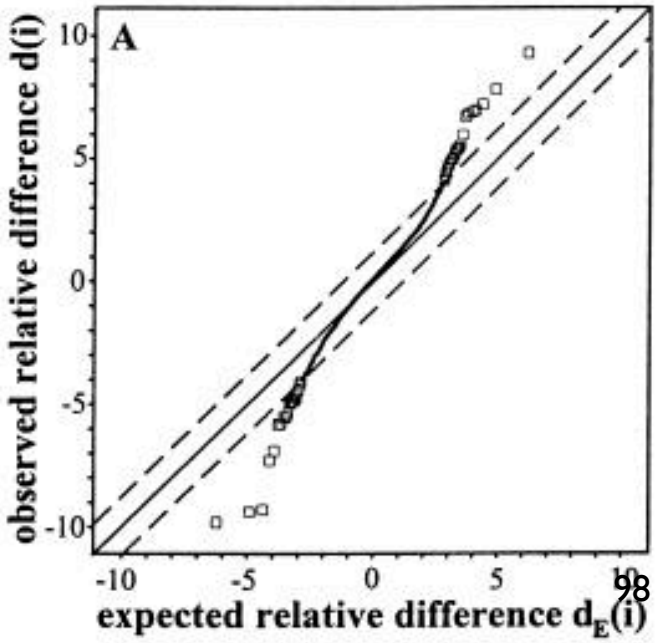
Significance **A**nalysis of **M**icroarrays: very popular method for DE

For each gene *i* , calculate: $$d(i) = \frac{\overline{M}_i}{s_i + s_0}$$

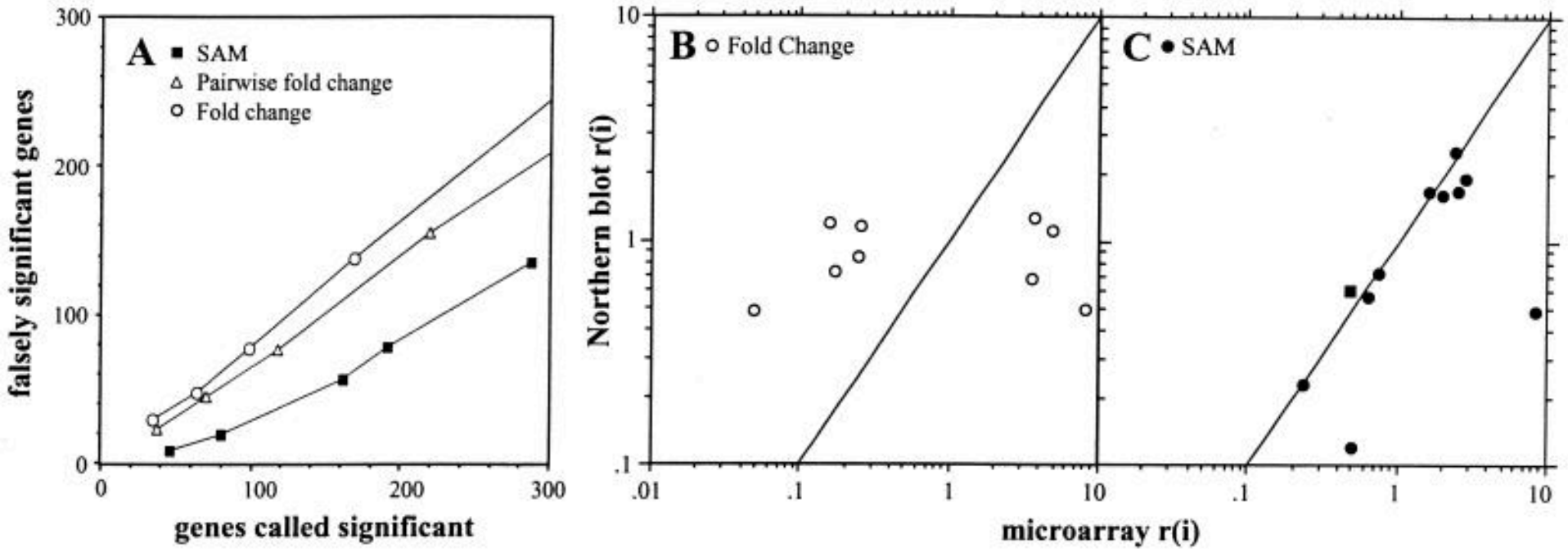
where s_i is the standard deviation of M and s_0 is constant (typically 0-5 %ile of s_i) to minimise variation in *d(i)* with levels of gene expression.

Comparing real *d(i)* with bootstrap samples:

- Rank genes by *d(i)*,so d(1) is largest relative difference.
- For each bootstrap sample (shuffling expression values, *B* bootstraps), calculate $d_p(i)$ and again rank by value, largest first.
- $$d_E(i) = \frac{\sum_p d_p(i)}{B}$$
- Expect $d_E(i) = d(i)$ for most genes.
- $| d(i) - d_E(i) | > \Delta \Rightarrow$ DE candidate.



SAM validation



Tusher VG, Tibshirani R, Chu G. "Significance analysis of microarrays applied to the ionizing radiation response". *Proc Natl Acad Sci U S A*. 2001 Apr 24;98(9):5116-21.

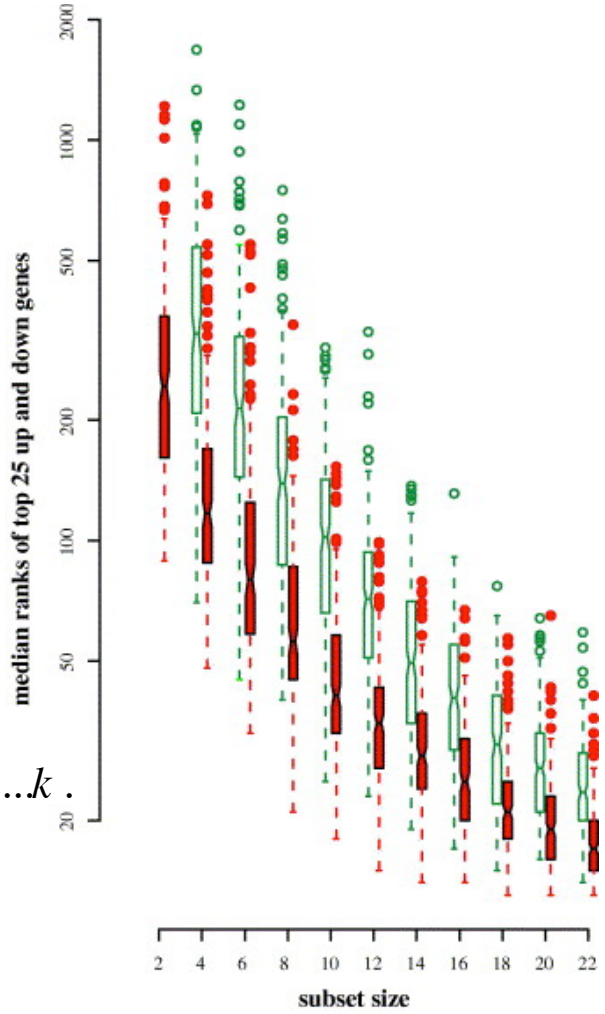
Better than SAM: Rank Products

- Assume 2 colour cDNA chips (single-channels can be accounted for) and compute M of each gene on each chip.
- If gene not DE, very unlikely for gene to be consistently ranked at top of lists sorted by M across replicate chips.
- Looking for up-regulated genes; down-regulated genes handled similarly:

$$RP_g^{up} = \left(\prod_{i=1}^{k} r_{i,g}^{up} \right)^{\frac{1}{k}} \qquad \text{geometric mean}$$

where $r_{i,g}^{up}$ is rank of gene g ($1 = \text{highest}$; $n_i = \text{lowest M}$) in chip $i = 1..k$.

- Significance of rankings assessed using the bootstrap: $q_g = E(RP_g)/rank(g) < FDR$.



Breitling R, Armengaud P, Amtmann A, Herzyk P. "Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments". *FEBS Lett,* 2004 Aug 27;573(1-3):83-92.

Even better: the B-statistic (borrowing information from genes)

Similar to a modified t-statistic (smoothes standard errors)

It is the log odds of differential expression (LODS, LOR)

- When there are thousands of genes we can get a better idea of the variability than from just the individual gene variance estimates
- We can't borrow information when there are only a few genes, but when there are tens of thousands of genes we can.
- We want a compromise between individual gene variance estimates and a single variance estimate for all genes.
- The compromise is achieved by empirical Bayes methods which give a weighted combination…

Smyth GK. "Linear models and empirical bayes methods for assessing differential expression in microarray experiments". *Stat Appl Genet Mol Biol*, 2004;3:Article3

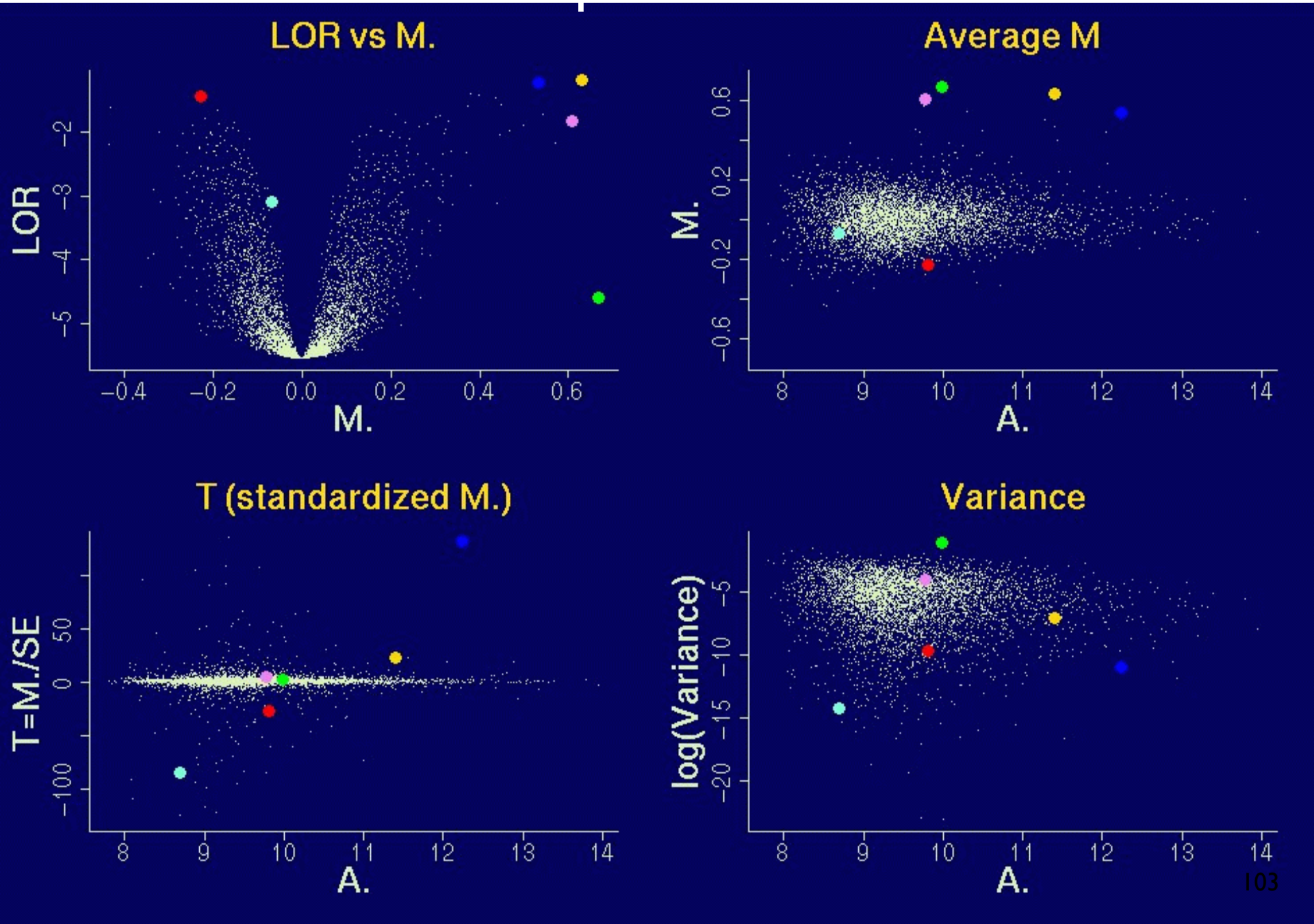
B statistic

$$\tilde{s}_g^2 = \frac{d_0 s_0^2 + d_g s_g^2}{d_0 + d_g}. \qquad \tilde{t}_{gj} = \frac{\hat{\beta}_{gj}}{\tilde{s}_g \sqrt{v_{gj}}}.$$

d_0 and d_g are constants
$V_{g,j} = 1/n$

Posterior odds of differential expression:

$$O_{gj} = \frac{p(\beta_{gj} \neq 0 \,|\, \tilde{t}_{gj}, s_g^2)}{p(\beta_{gj} = 0 \,|\, \tilde{t}_{gj}, s_g^2)}$$


LOR vs M.
LOR
M.
Average M
M.
A.
T (standardized M.)
T=M./SE
A.
Variance
log(Variance)
A.

Summary

- Microarray experiments typically have thousands of genes, but only few (1-10) replicates for each gene.
- Averages can be driven by outliers.
- *t*-statistics can be driven by tiny variances.
- B (or moderated t-statistic)
  - use information from all the genes
  - combine the best of *M*. and *t*
  - avoid the problems of *M*. and *t*

Ranking on B could be helpful.

What we want to do is…

- Analyse data all at once
- Use standard deviances not just fold changes
- Use ensemble information to shrink variances
- Assess differential expression for all comparisons together (because microarray experiments will rarely be just a simple comparison between two samples)

Ranking is easier

- How many genes are differentially expressed?
- If there was only one gene, a t-test would give a reliable P-value for judging whether the true log-ratio was zero
- With so many genes, computing absolute P-values on the basis of probability models is problematic
- Much easier to simply rank the genes in order of evidence for differential expression

Why judging significance is hard

- Log-ratios aren't normally distributed, hard to check log-ratios for different genes are correlated in unknown way
- High level of multiple testing means that very small p-values are required – distributional assumptions must hold in extreme tail

Choosing a cut-off

- Could choose a threshold for differential expression if there were known DE and non-DE genes
- Print artificial genes on microarray, then spike corresponding RNA into target RNA before labelling and hybridization
- Choose a cut-off that seem sensible!! Careful and thorough graphical exploration and the choice of ranking statistic are probably the most important aspect to choosing DE. Follow up experimentation that the biologist intends to perform will also play an important role.

Typical limma commands

```
> ct <- factor(targets$Type)
> design <- model.matrix(~0+ct)
> colnames(design) <- levels(ct)
#Define the design matrix with no intercept
> fit <- lmFit(y,design)
#fit the linear model
> contrasts <- makeContrasts(MS-mL, MS-pL, mL-pL, levels=design)
#Define the contrast matrix
> contrasts.fit <- eBayes(contrasts.fit(fit, contrasts))
#Get empirical Bayes estimates of variance
> topTable(contrasts.fit, coef=1)
#See results
```

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**. *Biostaistics.*

- 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

References

- Anders and Huber. Genome Biology, 2010; 11:R106
- Auer and Doerge. Genetics 2010, 185:405-416
- Harrell. Regression Modeling Strategies
- Robles et al. BMC Genomics 2012, 13:484
- Venables and Ripley. Modern Applied Statistics with S
- Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A*. 2001 Apr 24;98(9): 5116-21
- Breitling R, Armengaud P, Amtmann A, Herzyk P. Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Lett,* 2004 Aug 27;573(1-3):83-92.
- Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*, 2004;3:Article3