

# Analysis of publicly available microarray data

20<sup>th</sup>-21<sup>st</sup> February 2017

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@cancer.org.uk](mailto:Oscar.Rueda@cancer.org.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]

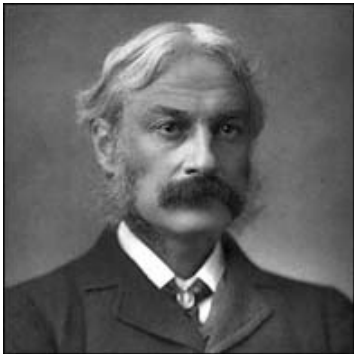
*An approximate answer to the right problem is worth a good deal more than an exact answer to an approximate problem.*



John Tukey (1915-2000)

[Statistician]

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



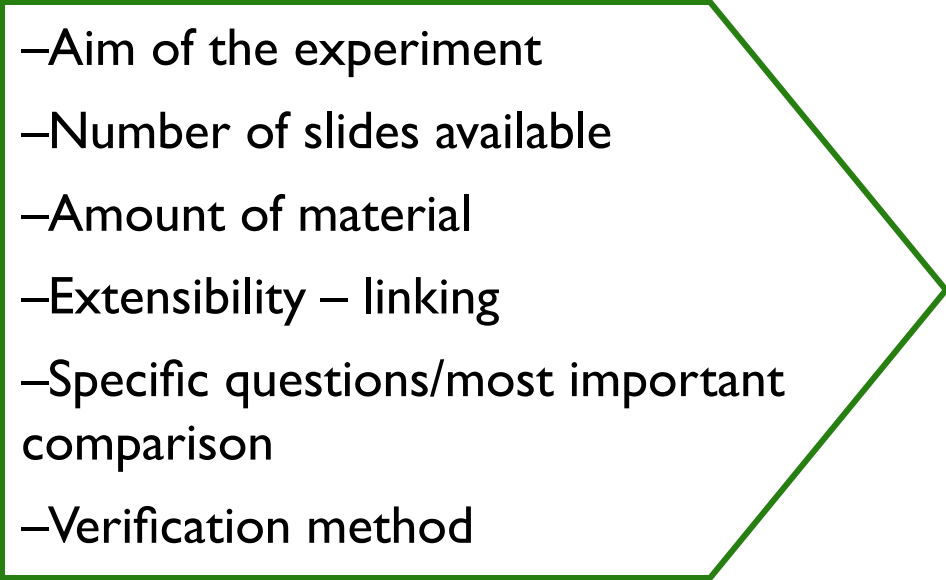
Andrew Lang (1844-1912)

[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

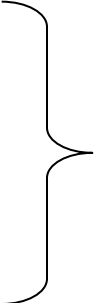
### **– Replication**

### **– 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.



# Design and contrast matrices

# 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

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

The diagram shows the linear model equation  $Y = \beta X + \varepsilon$ . Four arrows point from labels below to the terms in the equation: an arrow from 'Response variable' points to  $Y$ ; an arrow from 'Parameter vector' points to  $\beta$ ; an arrow from 'Design matrix' points to  $X$ ; and an arrow from 'Stochastic error' points to  $\varepsilon$ .

Response variable

Parameter vector

Design matrix

Stochastic error

# Types of designs considered

- 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.

# Strategy

- Define our set of samples
- Define the factors, type of factors (continuous, categorical), number of levels...
- Define the set of parameters: the effects we want to estimate
- Build the design matrix, that relates the information that each sample contains about the parameters.
- Estimate the parameters of the model: testing
- Further estimation (and testing): contrast matrices.

# Models with 1 factor, 2 levels

Sample	Treatment
Sample 1	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
Sample 1	Treatment A
Sample 2	Control
Sample 3	Treatment A
Sample 4	Control
Sample 5	Treatment A
Sample 6	Control

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

Parameters  
(coefficients, levels  
of the variable)

**C** is the mean expression of  
the control

**T** is the mean expression of  
the treatment

Design Matrix


Equivalent to a t-test

# Design matrix for models with 1 factor, 2 levels

Sample	Treatment
Sample 1	Treatment A
Sample 2	Control
Sample 3	Treatment A
Sample 4	Control
Sample 5	Treatment A
Sample 6	Control

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

Parameters  
(coefficients, levels  
of the variable)




Design Matrix

Equivalent to a t-test

# Intercepts

Different parameterization: using intercept

Sample	Treatment
Sample 1	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)

$a = T_A - \text{Control}$  (mean change in expression under treatment)



# Intercept

Different parameterization: using intercept

Sample 1  $\begin{bmatrix} S1 \end{bmatrix}$   
Sample 2  $\begin{bmatrix} S2 \end{bmatrix}$   
Sample 3  $\begin{bmatrix} S3 \end{bmatrix}$   
Sample 4  $\begin{bmatrix} S4 \end{bmatrix}$   
Sample 5  $\begin{bmatrix} S5 \end{bmatrix}$   
Sample 6  $\begin{bmatrix} S6 \end{bmatrix}$

$=$

$\begin{matrix} \text{Intercept} \\ \text{Treatment A} \end{matrix}$   
 $\begin{pmatrix} 1 & 1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 0 \end{pmatrix}$

$\begin{bmatrix} \beta_0 \\ a \end{bmatrix}$

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
Sample 1	Treatment A
Sample 2	Treatment B
Sample 3	Control
Sample 4	Treatment A
Sample 5	Treatment B
Sample 6	Control

ANOVA models

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
Sample 1	Treatment A
Sample 2	Treatment B
Sample 3	Control
Sample 4	Treatment A
Sample 5	Treatment B
Sample 6	Control

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

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

# Design matrix for ANOVA models

Sample	Treatment
Sample 1	Treatment A
Sample 2	Treatment B
Sample 3	Control
Sample 4	Treatment A
Sample 5	Treatment B
Sample 6	Control

Control = Baseline

$T_A = \text{Baseline} + a$

$T_B = \text{Baseline} + b$

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


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

# 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:

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



# Exercise

Build contrast matrices for all pairwise comparisons for these designs:

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

# Exercise

Build contrast matrices for all pairwise comparisons for these designs:

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

# Exercise

Build contrast matrices for all pairwise comparisons for these designs:

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

# Models with 2 factors

Sample	Treatment	ER status
Sample 1	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

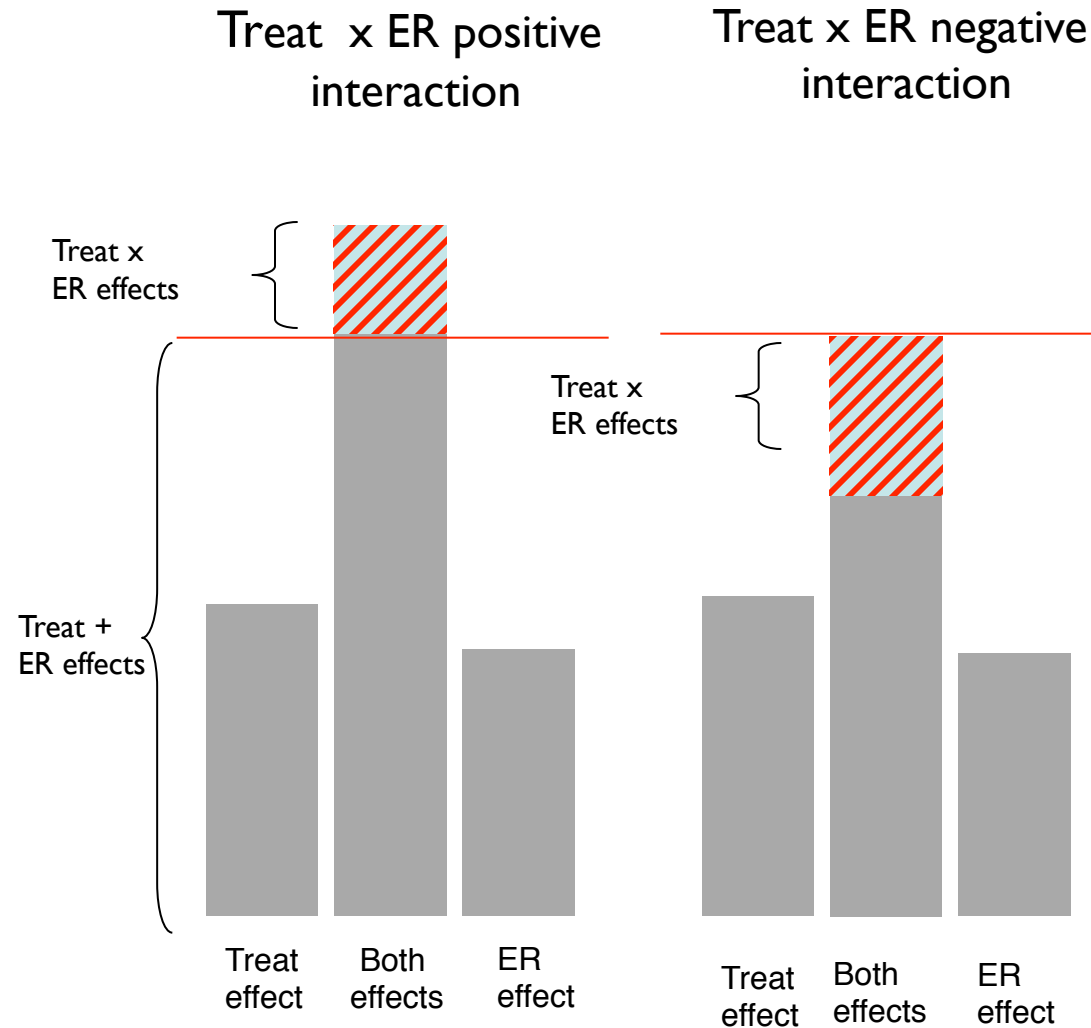
Number of factors: 2

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



# Models with 2 factors and no interaction

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

Number of coefficients (parameters):

$$\text{Intercept} + (\# \text{levels Treat} - 1) + (\# \text{levels ER} - 1) = 3$$

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)

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

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):

$$\text{Intercept} + (\# \text{levels Treat} - 1) + (\# \text{levels ER} - 1) + ((\# \text{levels Treat} - 1) * (\# \text{levels ER} - 1)) = 4$$

# Models with 2 factors (interaction)

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

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

“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)

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

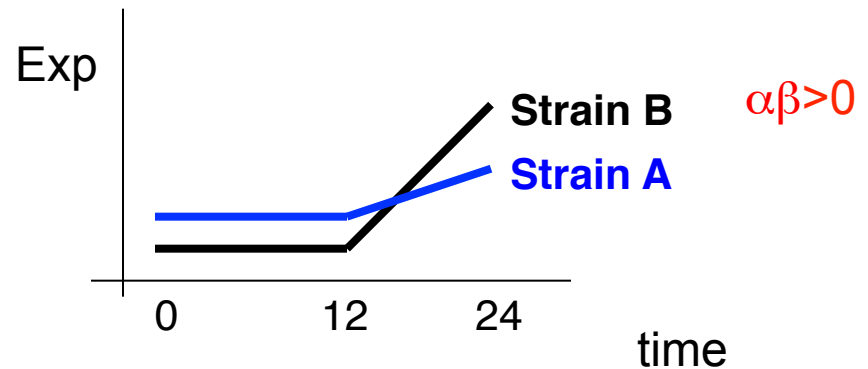
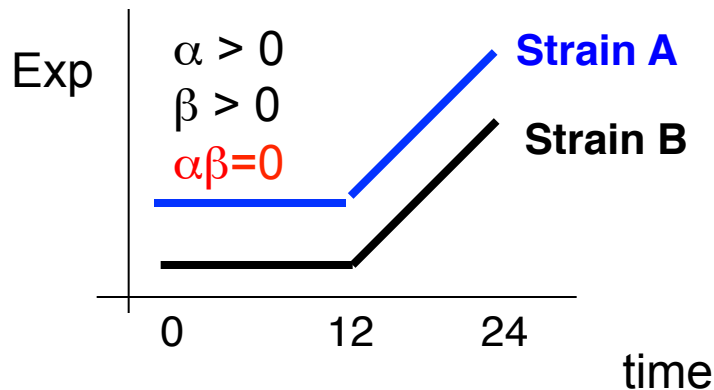
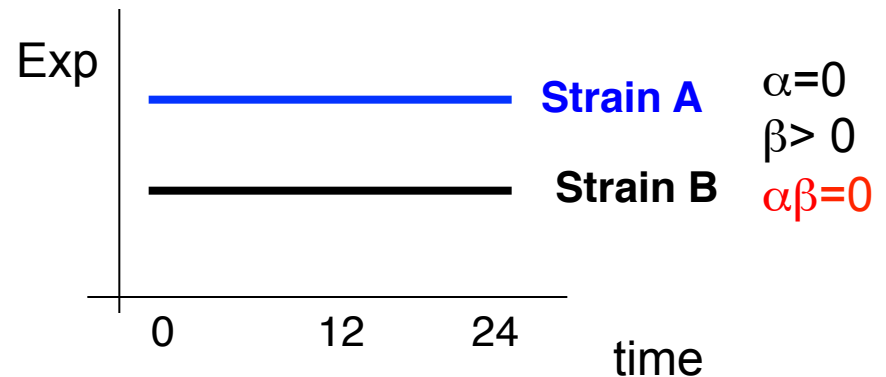
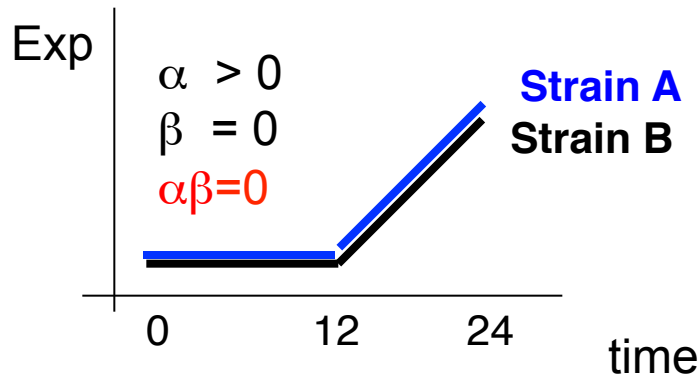
“Extra effect” of Treatment A on ER+ samples

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

# 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



# 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

# 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 without intercept)

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

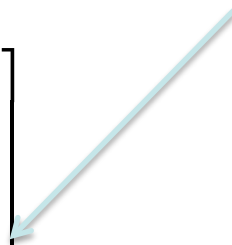
Number of factors: 2

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)`

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

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

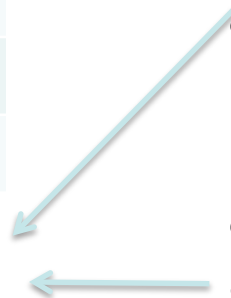
# Time Course experiments

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

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.



Number of samples: 2

Number of factors: 2

Sample: Number of levels: 2

Time: Continuous or categorical?

Intermediate solution: **splines**

# 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.



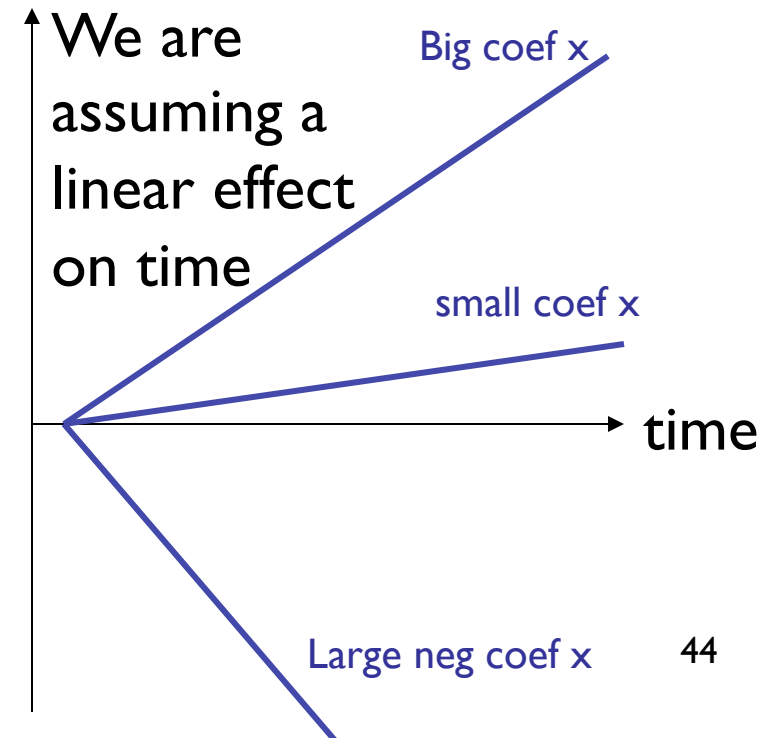
Sample	Time
Sample 1	0h
Sample 1	1h
Sample 1	4h
Sample 1	16h
Sample 2	0h
Sample 2	1h
Sample 2	4h
Sample 2	16h

# Time Course experiments

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

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

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



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  $\# \text{levels} - 1$  columns
  - Interactions will need  $(\# \text{levels} - 1) \times (\# \text{levels} - 1)$
  - It is possible to include interactions of more than 2 variables, but the number of samples needed to accurately estimate those interactions is large.

# Linear models

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

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

Arbitrary number of independent variables

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

Polynomials are valid

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

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

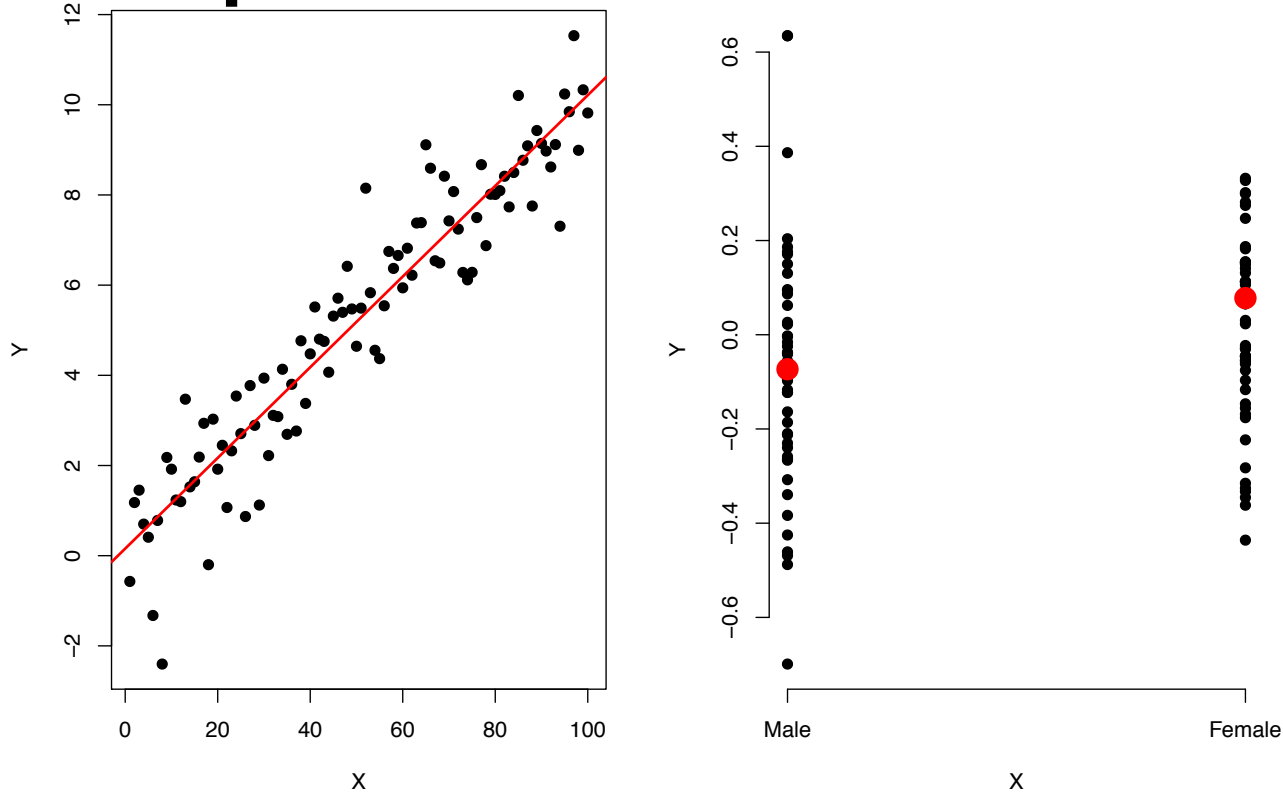
We can use functions of the variables if the effects are linear

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

# Model Estimation

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

$\beta$	→	Parameter of interest (effect of X on Y)
$\hat{\beta}$	→	<b>Estimator</b> of the parameter of interest
$se(\hat{\beta})$	→	<b>Standard Error</b> of the estimator of the parameter of interest

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

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

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

$\hat{y} = \hat{\beta}x$	→	Fitted values (predicted by the model)
$e = y - \hat{y}$	→	Residuals (observed errors)



# Statistical issues in microarray data

# 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?

# Hypothesis testing

- **Null Hypothesis:** Our population follows a (known) distribution defined by a set of parameters:  
 $H_0 : X \sim f(\theta_1, \dots, \theta_k)$

- Take a random sample  $(X_1, \dots, X_n) = (x_1, \dots, x_n)$  and observe **test statistic**

$$T(X_1, \dots, X_n) = t(x_1, \dots, x_n)$$

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

# Type I and Type II errors

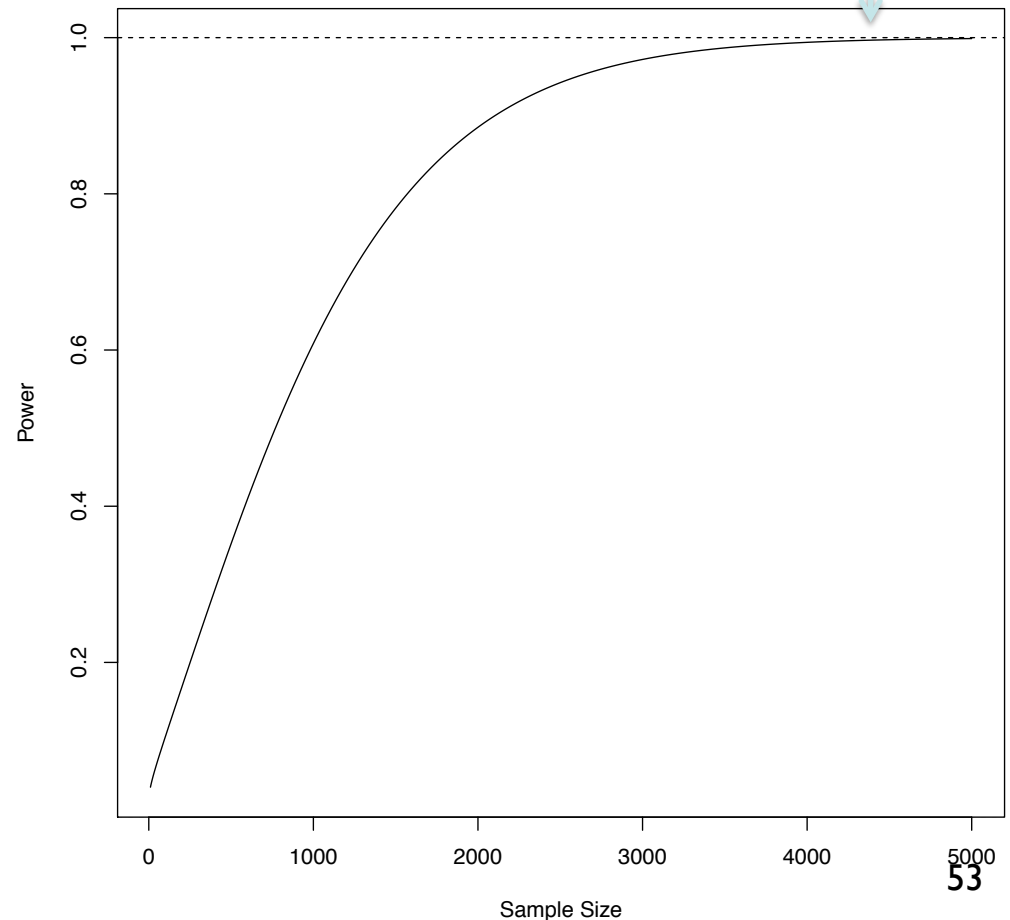
- Type I error: probability of rejecting the null hypothesis when it is true. Usually, it is the significance level of the test. It is denoted as  $\alpha$
- Type II error: probability of not rejecting the null hypothesis when it is false It is denoted as  $\beta$
- Decreasing one type of error increases 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)

t-test: true diff:0.1 std=1 sig.lev=0.05

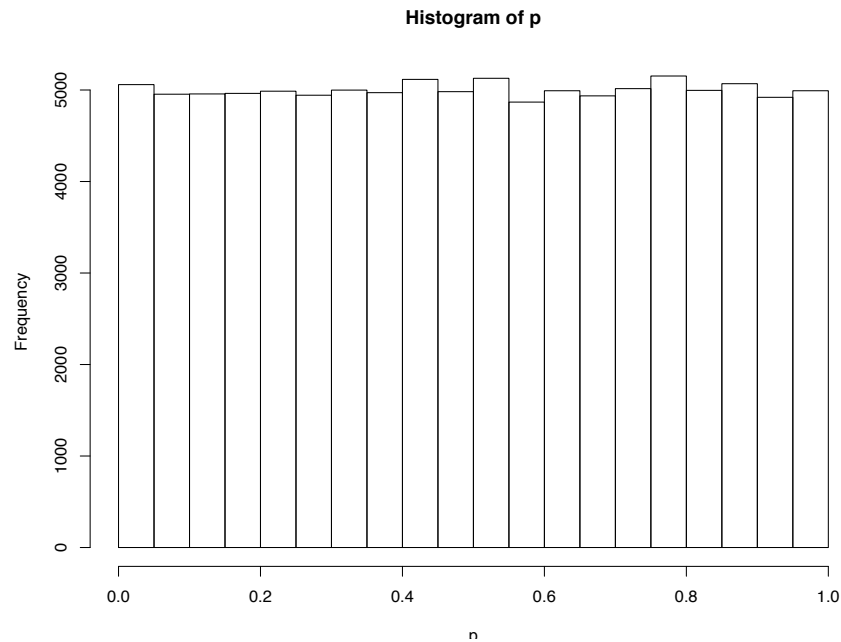


# Multiple testing problem

- 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  $\text{Uniform}(0,1)$  distribution.



# 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)	$N - \# \text{ Rejections}$
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:

$P(\text{at least one false positive} \mid \text{all null hypothesis are true}) =$

$P(\text{at least one p-value} < \alpha \mid \text{all null hypothesis are true}) = 1 - (1 - \alpha)^m$

Usually, we set a threshold at  $\alpha / n$ .

**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)	$N - \# \text{ Rejections}$
Significant (Reject)	$V = \# \text{ False positive (Type I error)}$	# True positive	$R = \# \text{ Total Rejections}$
Total	$n_0$	$N - n_0$	$N$

Family Wise Error Rate:  $\text{FWER} = P(V \geq 1)$

False Discovery Rate:  $\text{FDR} = E(V/R \mid R > 0) P(R > 0)$

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  $\alpha$ :

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.

## 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  $\alpha j^*/m \iff$  compare  $mp_i/j^*$  to  $\alpha$

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

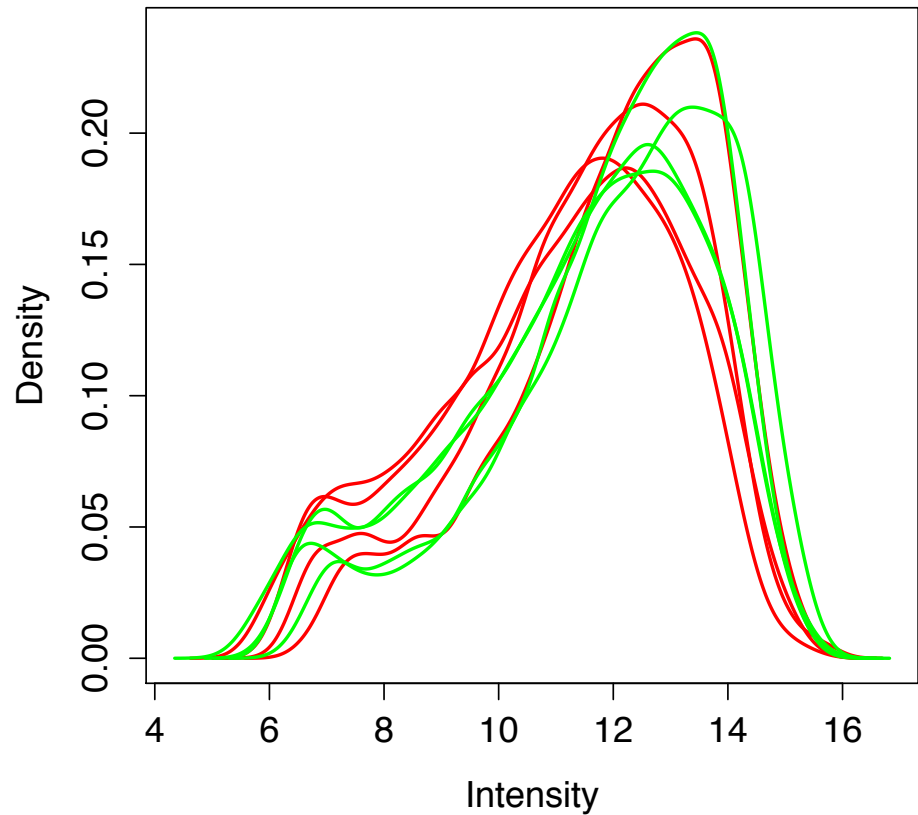
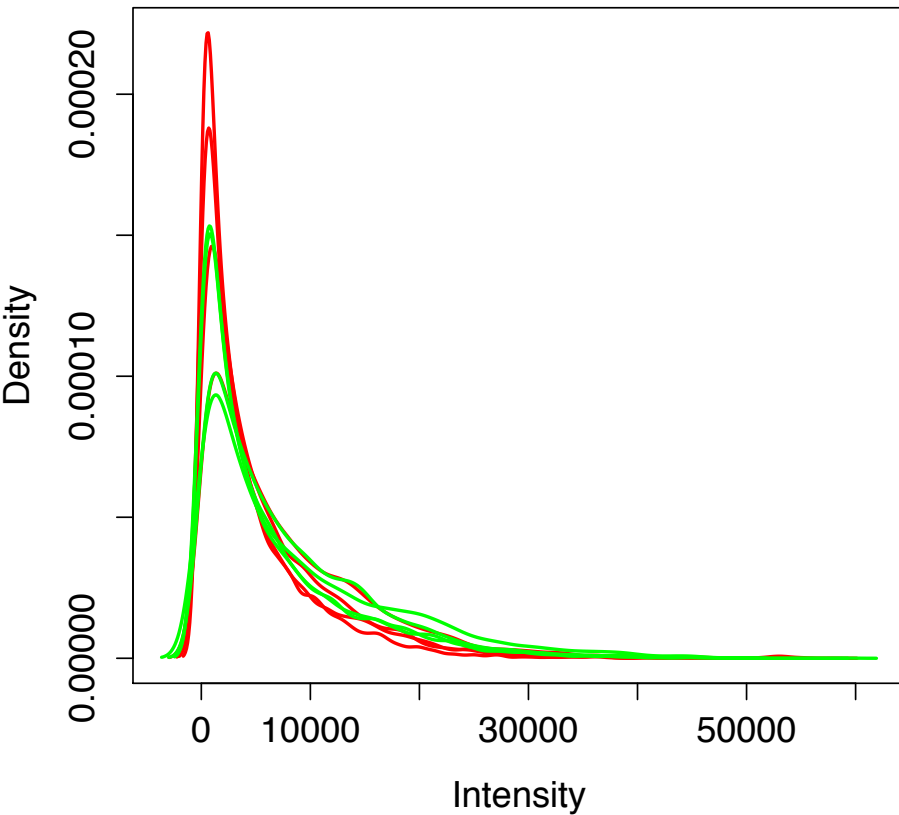
But these 'adjusted p-values' tell you the level of FDR which is being controlled (as opposed to the FWER in the Bonferroni and Holm cases).

# 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.

# Microarray expression data

RG densities Data are color intensities RG densities



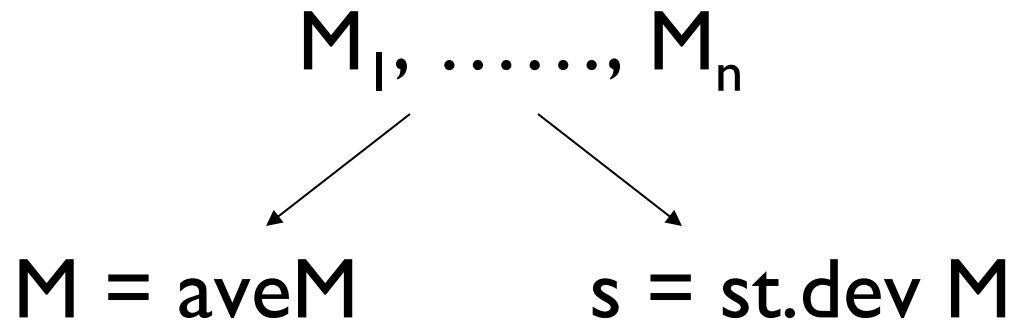
$$\log y_{ij} \sim N(\mu_j, \sigma_j^2)$$

# 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).

# 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...

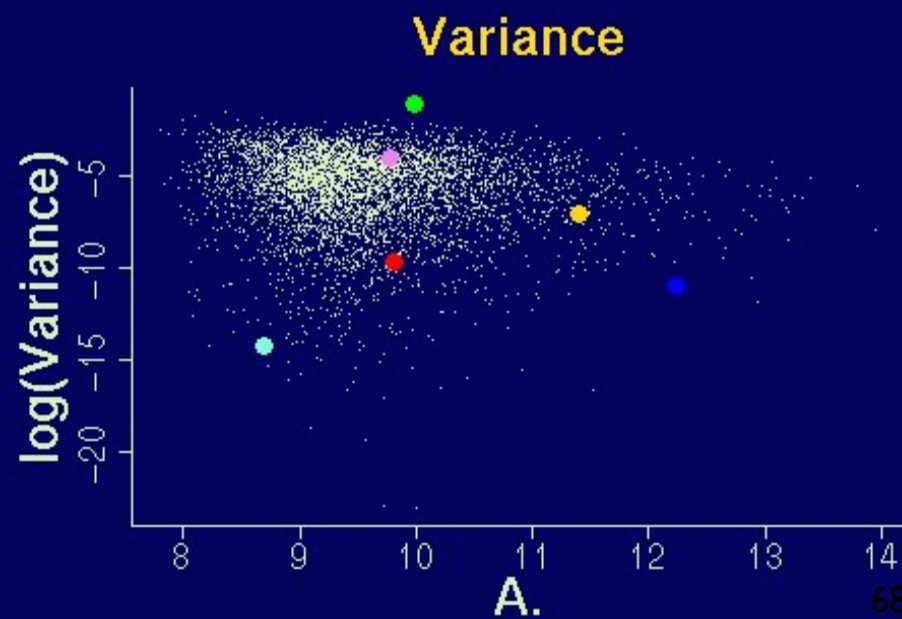
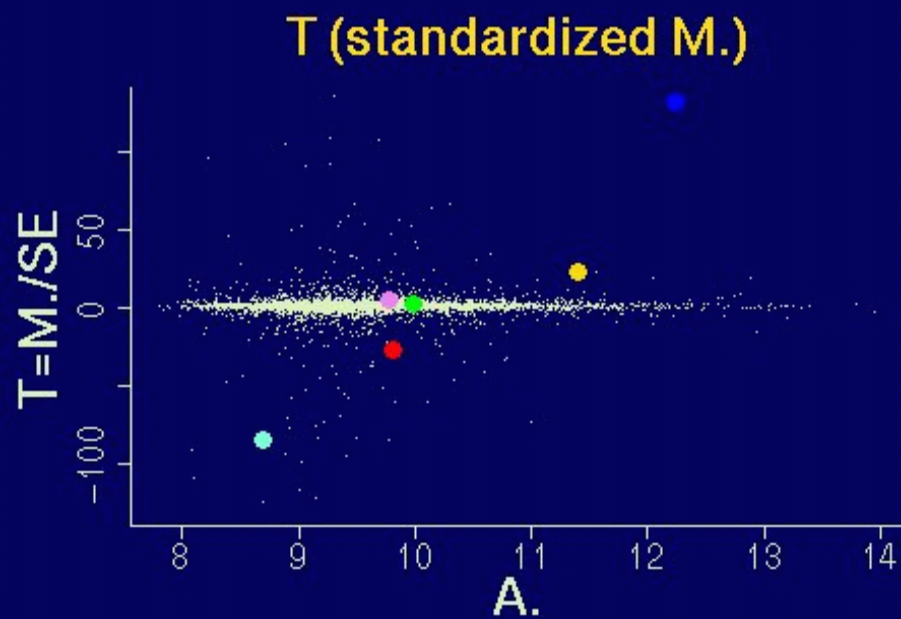
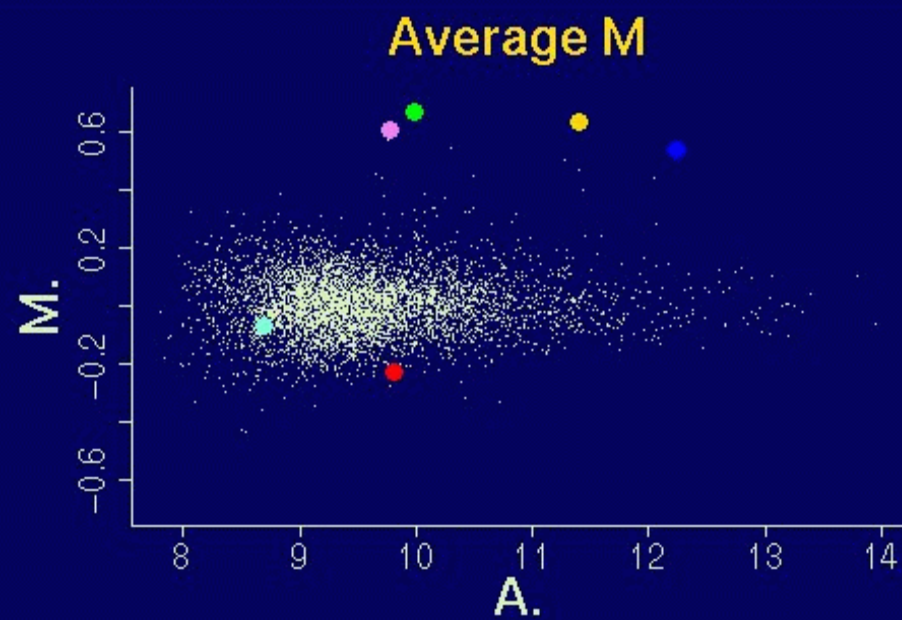
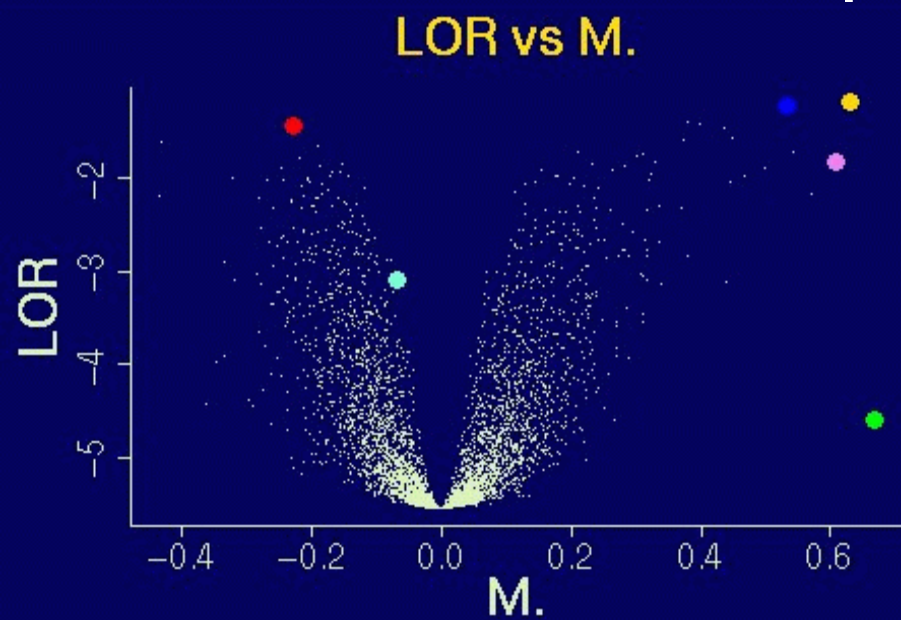
# B statistic

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

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

Posterior odds of differential expression:

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



# 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 <- contrasts.fit(fit, contrasts)
#Estimate the contrasts
> contrasts.fit <- eBayes(contrasts.fit)
#Get empirical Bayes estimates of variance
> topTable(contrasts.fit, coef=1)
#See results
```

# References

- **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