

# Statistical analysis of RNASeq Data

## Introduction to RNA-seq data analysis

dominique-laurent.couturier@cruk.cam.ac.uk [Bioinformatics core]

(Source: O. Rueda, CRUK-CI; G. Marot, INRIA)

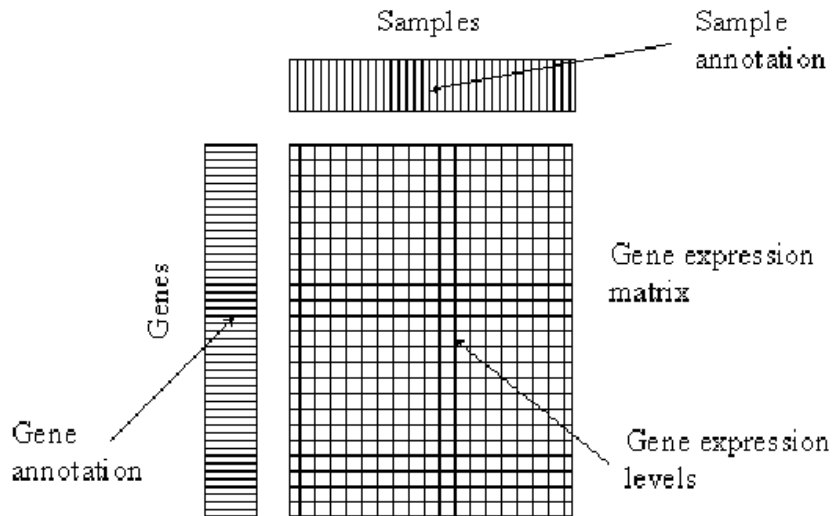
raw count for gene i, sample j

The mean is taken as "normalized counts" scaled by a normalization factor

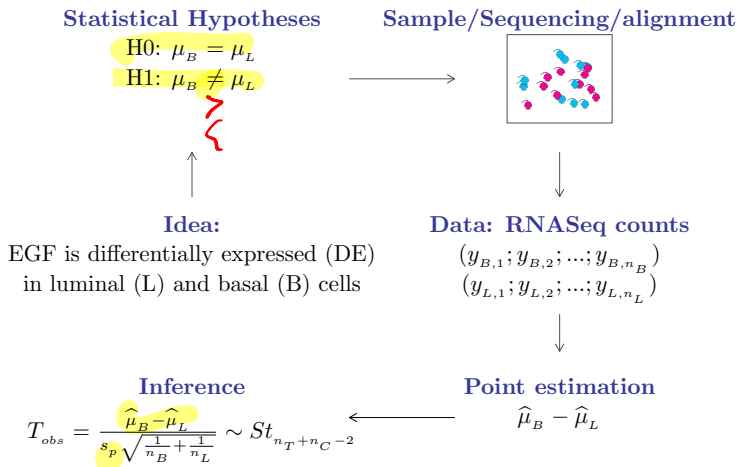
one dispersion per gene

$$K_{ij} \sim \text{NB}(s_{ij}q_{ij}, \alpha_i)$$

# Introduction

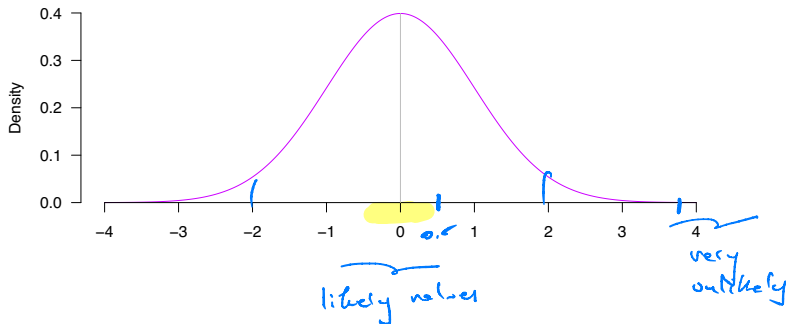


# Grand Picture of Statistics



# Statistical tests

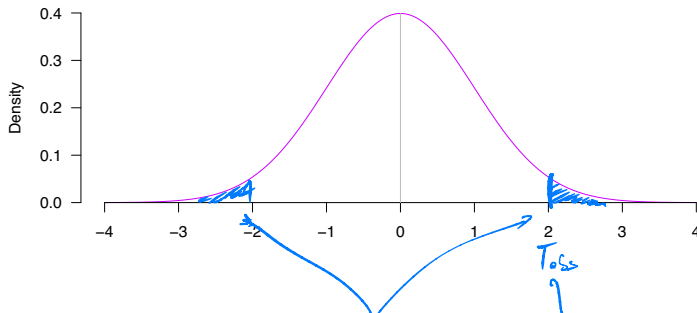
Assess how likely the observed test statistics is compared to the test statistics distribution under  $H_0$ :





# Statistical tests

Assess how likely the observed test statistics is compared to the test statistics distribution under  $H_0$ :



**P-value for a two-sided test:**  $p\text{-value} = P(|T| > T_{obs})$

i.e. the probability of getting a test statistic as extreme or more extreme than the calculated test statistic if  $H_0$  is true

# Statistical tests

## 4 possible outcomes

Conclude:

- ▶ if  $p\text{-value} > \alpha \rightarrow$  do not reject  $H_0$ .
- ▶ if  $p\text{-value} < \alpha \rightarrow$  reject  $H_0$  in favour of  $H_1$ .

		Test Outcome	
		H0 not rejected	H1 accepted
Unknown Truth	H0 true	$1 - \alpha$ [TN]	$\alpha$ [FP]
	H1 true	$\beta$ [FN]	$1 - \beta$ [TP]

where

- ▶  $\alpha$  is the type I error,
- ▶  $\beta$  is the type II error.

# Statistical tests

## 4 possible outcomes

Conclude:

- ▶ if  $p\text{-value} > \alpha \rightarrow$  do not reject  $H_0$ .
- ▶ if  $p\text{-value} < \alpha \rightarrow$  reject  $H_0$  in favour of  $H_1$ .

		Test Outcome	
		H0 not rejected	H1 accepted
Unknown Truth	H0 true	$1 - \alpha$ [TN]	$\alpha$ [FP]
	H1 true	$\beta$ [FN]	$1 - \beta$ [TP]

where

- ▶  $\alpha$  is the type I error,
- ▶  $\beta$  is the type II error.

Want to minimise FP and FN through design

# Experimental design

## 3 fundamental aspects of sounds experiments (Fisher 1935)

### ► Replication

Try to capture all sources of variability  
(Biological versus technical variability)

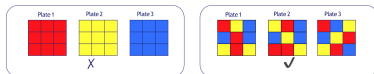
### ► Blocking

Try to remove technical biases/confounding  
(Lane and batch effects)



### ► Randomisation

Try to remove confounding due to other factors



# Experimental design

## Sample size per condition

### Sample size calculation:

Aim is to define the sample size allowing to detect an effect of a given size at the  $\alpha$  level with a given probability (power):

- ▶  $\delta$ , the effect size: function of  $\mu_L$  and  $\mu_B$   
(log fold change, standardised difference),
- ▶  $1 - \beta$ , the power,
- ▶  $\alpha$ , the type I error.
- ▶  $\phi$ , nuisance parameters  
(variability, sequencing depth, multiplicity correction)

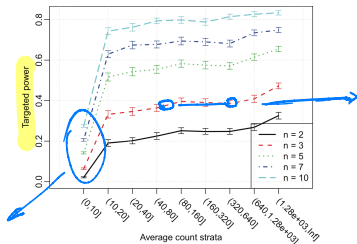
# Experimental design

## Sample size per condition

### Sample size calculation:

Aim is to define the sample size allowing to detect an effect of a given size at the  $\alpha$  level with a given probability (power):

- ▶  $\delta$ , the effect size: function of  $\mu_L$  and  $\mu_B$   
(log fold change, standardised difference),
- ▶  $1 - \beta$ , the power,
- ▶  $\alpha$ , the type I error.
- ▶  $\phi$ , nuisance parameters  
(variability, sequencing depth, multiplicity correction)

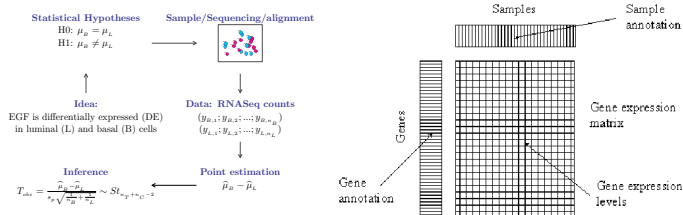


different depth  
but same  
power.

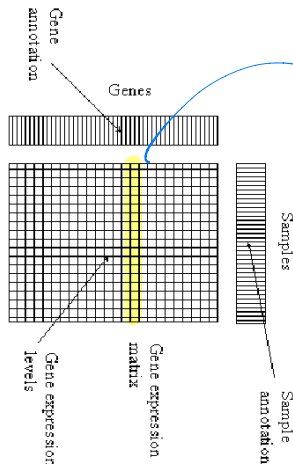
low/no power  
to select DE genes  
with small # reads

(Wu, Wang and Wu (2015))

# Statistical modelling



# Statistical modelling



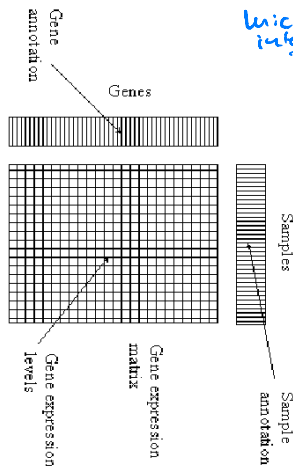
$$y = f(X) + \epsilon$$
$$E[y] = f(X)$$

where

- ▶  $y$  denotes the  $(n \times 1)$  vector of expression intensities of a given gene,
- ▶  $X$  denotes the  $(n \times p)$  design/predictor matrix,
- ▶  $\epsilon$  denotes the  $(n \times 1)$  stochastic error vector,
- ▶  $E[y]$  denotes the expectation of  $y$



# Statistical modelling : Linear regression



Microarrays intensities

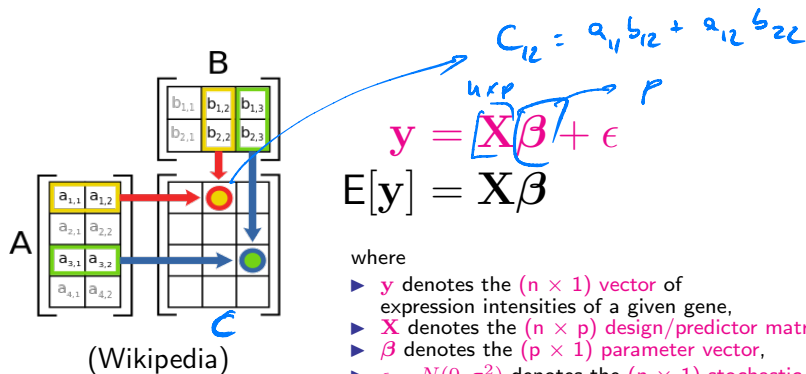
matrix multiplication

$$\log(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$
$$E[\mathbf{y}] = \mathbf{X}\boldsymbol{\beta}$$

where

- ▶  $\mathbf{y}$  denotes the  $(n \times 1)$  vector of expression intensities of a given gene,
- ▶  $\mathbf{X}$  denotes the  $(n \times p)$  design/predictor matrix,
- ▶  $\boldsymbol{\beta}$  denotes the  $(p \times 1)$  parameter vector,
- ▶  $\boldsymbol{\epsilon} \sim N(0, \sigma^2)$  denotes the  $(n \times 1)$  stochastic error vector,
- ▶  $E[\mathbf{y}]$  denotes the expectation of  $\mathbf{y}$

# Statistical modelling : Linear regression



$A$  : (4x2) matrix

$B$  : (2x3) matrix

$C = AB = (4x3)$  matrix

where

- ▶  $y$  denotes the  $(n \times 1)$  vector of expression intensities of a given gene,
- ▶  $X$  denotes the  $(n \times p)$  design/predictor matrix,
- ▶  $\beta$  denotes the  $(p \times 1)$  parameter vector,
- ▶  $\epsilon \sim N(0, \sigma^2)$  denotes the  $(n \times 1)$  stochastic error vector,
- ▶  $E[y]$  denotes the expectation of  $y$

# Statistical modelling : Strategy

- ▶ Collect the information related to each sample for the predictors of interest,
- ▶ define  $\beta$ , the sets of parameters we are interested in,
- ▶ build the X matrix that relates the sample information with the  $\beta$ ,
- ▶ estimate the  $\beta$ ,
- ▶ use statistical inference to assess significance ( $p$ -values).

# Statistical modelling : Contrast matrices

Contrast matrices for models with

- ▶ one factor / categorical predictor,
  - ▷ two experimental conditions (dichotomous predictor),  
t-test
  - ▷ several experimental conditions,  
ANOVA
- ▶ two factors / categorical predictors,
  - ▷ without interaction,
  - ▷ with interaction,
- ▶ Two-way ANOVA  
categorical and continuous ~~factors~~. predictors

## Design matrix for models with a two-level factor

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 with 2 levels (Control and Treatment A)

Possible parameters (What differences are important)?

- Effect of Treatment A
- Effect of Control

## Design matrix for models with a two-level factor

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

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

$\begin{matrix} \text{C is the mean expression of the control} \\ \text{T is the mean expression of the treatment} \end{matrix}$

$X\beta$

$\begin{matrix} T \\ C \end{matrix} \rightarrow = 1T + 0C$   
 $\begin{matrix} C \\ T \end{matrix} \rightarrow = 0T + 1C$

# Design matrix for models with a two-level factor

Different parameterisation: 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:

$\beta$  }  $\begin{cases} C = \text{Control (mean expression of the control)} \\ a = T_A - \text{Control (mean change in expression under treatment)} \end{cases}$

## Design matrix for models with a two-level factor

Different parameterization:  
using an intercept

$$\begin{array}{l} \text{Sample 1} \\ \text{Sample 2} \\ \text{Sample 3} \\ \text{Sample 4} \\ \text{Sample 5} \\ \text{Sample 6} \end{array} \begin{bmatrix} S1 \\ S2 \\ S3 \\ S4 \\ S5 \\ S6 \end{bmatrix} = \begin{array}{c} \text{Intercept} \\ \text{Treatment A} \end{array} \begin{pmatrix} 1 & 1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 0 \end{pmatrix} \begin{bmatrix} \beta_0 \\ a \\ \beta_0 + a \\ \beta_0 \\ \beta_0 + a \\ \beta_0 \end{bmatrix}$$

$\beta$  Parameter vector  
 $\beta_0 + a = T$   
 $\beta_0 = C$

X design Matrix

The Intercept measures the baseline expression and  $a$  measures now the differential expression between Treatment A and Control



## Design matrix for models with a two-level factor

The two parameterizations are equivalent but allows to test different contrasts/parameters

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

 **Contrast matrix**

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

## Design matrix for models with a three-level factor

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 with 3 levels (Control, Treatment A, Treatment B)

Possible parameters (What differences are important)?

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

# Design matrix for models with a three-level factor

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

Control = Baseline

$T_A = \text{Baseline} + a$

$T_B = \text{Baseline} + b$

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


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

*Note: In the second matrix equation, the first column of the design matrix is all ones, and the parameters are  $\beta_0$ ,  $a$ , and  $b$ . Blue arrows indicate the mapping from the 'Control' row in the table to the first column of the second matrix, and from 'Treatment A' and 'Treatment B' to the second and third columns respectively.*

# Design matrix for models with a three-level factor

The model with intercept always take one level as a **reference group**:

The **reference group** here is treatment A, the coefficients are comparisons against it!


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

By default, R uses the first level as baseline

## Design matrix for models with a three-level factor: R code

```
> one3levelfactor = data.frame(condition =  
    rep(c("TreatmentA", "TreatmentB", "Control"), 2))  
  
# model without intercept and default levels:  
> X1 = model.matrix(~ condition - 1, data = one3levelfactor)  
  
# model with intercept and default levels  
> X2 = model.matrix(~ condition, data = one3levelfactor)  
  
# model with intercept and self-defined levels  
> levels(one3levelfactor$condition)  
> levels(one3levelfactor$condition) = c("TreatmentB", "TreatmentA", "Control")  
> X3 = model.matrix(~ condition, data = one3levelfactor)
```

## Design matrix for models with a three-level factor: Exercise

Build contrast matrices for all pairwise comparisons for this design:

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

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

## Design matrix for models with a three-level factor: Exercise

Build contrast matrices for all pairwise comparisons for these designs:

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

$$\begin{bmatrix} \hat{\beta}_0 \\ \hat{a} \\ \hat{b} \end{bmatrix}$$

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

## 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 two-level factors

```
> two2levelfactor = data.frame(treatment = rep(c("TreatA", "NoTreat"), 4),  
                               er = rep(c("+", "-"), each=4))
```



# Models with 2 factors: no interaction

```
X1 = model.matrix(~ treatment + er, data=two2levelfactor)
```

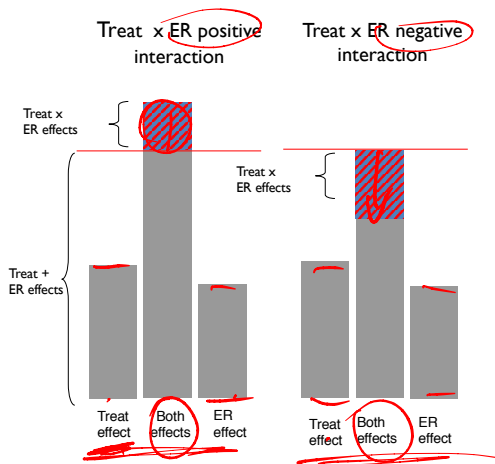
$$\begin{bmatrix} S1 \\ S2 \\ S3 \\ S4 \\ S5 \\ S6 \\ S7 \\ S8 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} \beta_0 \\ a \\ \text{er}+ \end{bmatrix}$$

ER - Treat A

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

# Models with 2 factors: 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: with interaction

```
> X2 = model.matrix(~ treatment * er, data=two2levelfactor)
> X3 = model.matrix(~ treatment + er + treatment:er, data=two2levelfactor)
```

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

*Note: Red handwritten boxes highlight the first two columns of the matrix and the  $a.er+$  term in the coefficient vector.*

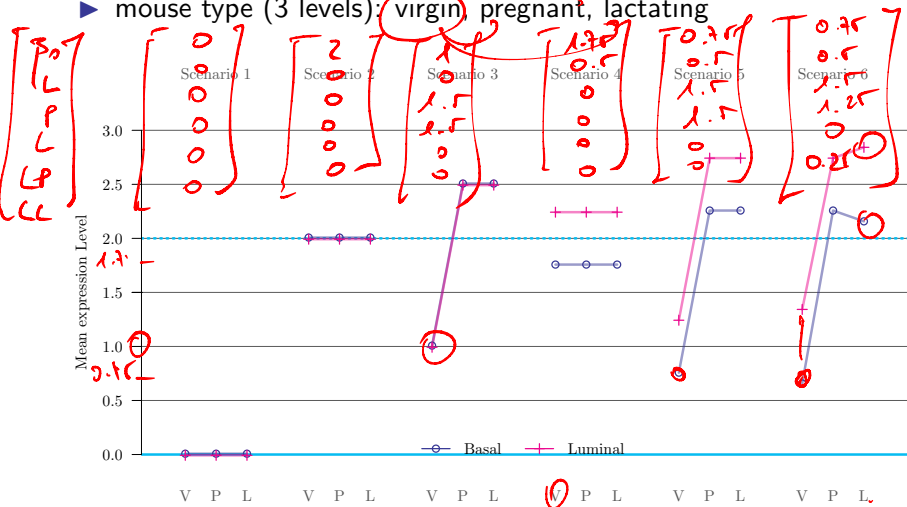
Interaction effect of  
Treatment A on ER+ samples

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

# Models with 2 factors: possible scenarios

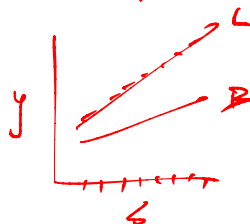
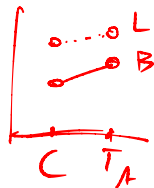
2 factors:

- ▶ cell type (2 levels): luminal versus basal
- ▶ mouse type (3 levels): virgin, pregnant, lactating



## Models with 2 predictors: a factor and a continuous one

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

2 predictors: ER (a two-level factor) and Dose (a continuous predictor)

```
> mixedpredictors = data.frame(er = rep(c("+", "-"), 4),  
                                dose = c(37, 52, 65, 89, 24, 19, 54, 67))
```

## Models with 2 predictors: a factor and a continuous one

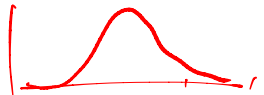
```
X = model.matrix(~ er + dose, data= mixedpredictors)
```

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

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

$$T_{oss} = \frac{\hat{\beta}}{\hat{\sigma}_{\hat{\beta}}} \sim N(0,1) \\ \sim St_{u-p}$$



## Model Estimation and inference

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

$\beta$  → Parameter of interest

$\hat{\beta}$  → Estimate of the parameter of interest

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

$$\hat{\beta} = (X^T X)^{-1} X^T Y \quad MLE : \hat{\beta} = \arg \max \{L(\beta | x)\}$$

$$se(\hat{\beta}_i) = \sigma \sqrt{c_i} \quad \text{where } c_i \text{ is the } i^{\text{th}} \text{ diagonal element of } (X^T X)^{-1}$$

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

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



$\hat{\beta}$

$\varepsilon$

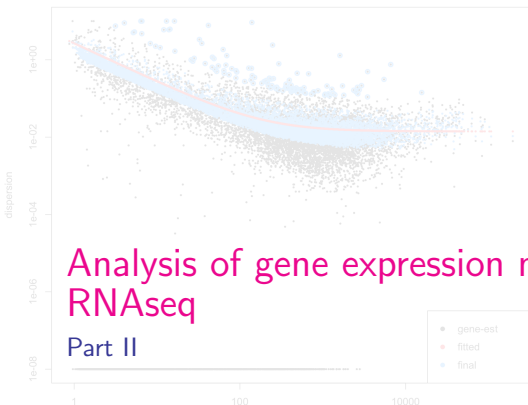


CANCER  
RESEARCH  
UK

CAMBRIDGE  
INSTITUTE



UNIVERSITY OF  
CAMBRIDGE



# Analysis of gene expression measured with RNAseq

## Part II

dominique-laurent.couturier@cruk.cam.ac.uk [Bioinformatics core]

(Source: O. Rueda, CRUK-CI; G. Marot, INRIA)

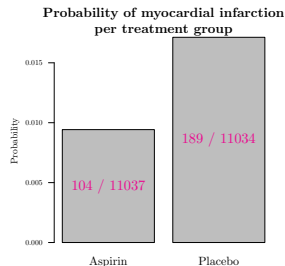
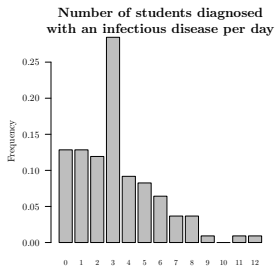
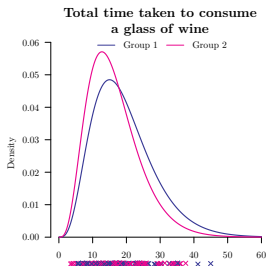
The mean is taken as "normalized  
count" divided by a normalization  
factor

one dispersion per gene

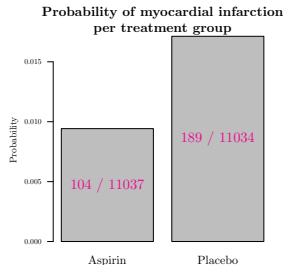
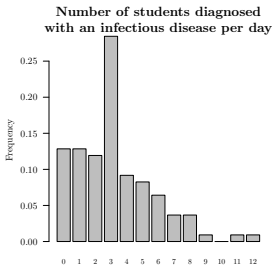
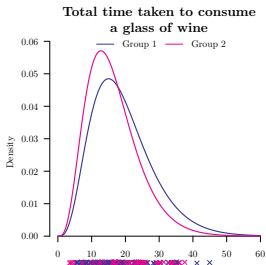
$$K_{ij} \sim \text{NB}(s_{ij}q_{ij}, \alpha_i)$$



# Examples of data with non-normal conditional distributions



# Examples of data with non-normal conditional distributions



Linear model not suitable:

► Assumed model:

$$y = \mathbf{X}\beta + \epsilon \text{ where } \epsilon \sim N(0, \sigma^2),$$

▷ theoretical range of  $\epsilon = [-\infty, +\infty]$ ,

▷  $\mathbf{X}\beta$  not bounded to  $[0, \infty]$  or  $[0, 1]$ ,

▷  $\text{Var}[y]$  independent of  $E[y]$ .

► Solution:

$$y | (\mathbf{X}, \beta, \phi) \sim \text{distribution}(\text{function}(\mathbf{X}\beta), \phi),$$

where *distribution* belongs to the exponential family and *function* is monotonically increasing.

# GLM: conditional distributions

$$\mathbf{y} | (\mathbf{X}, \boldsymbol{\beta}, \phi) \sim \text{distribution}(\text{function}(\mathbf{X}\boldsymbol{\beta}), \phi),$$

- Some possible conditional *distributions* :  
statistical probability mass functions & density functions

- ▷ Within the **exponential family** ['classical' GLM framework]

normal  
exponential  
gamma

chi-squared  
beta  
Dirichlet

Poisson  
Negative Binomial  
Bernoulli

Inverse Wishart  
...

- ▷ **Outside the exponential family** ['extended' GLM framework]

Box-Cox power  
exponential  
exponential Gaussian  
generalized beta  
generalized gamma  
generalized inverse

Gaussian  
inverse Gaussian  
logistic  
power exponential  
reverse Gumbel  
skew power exponential

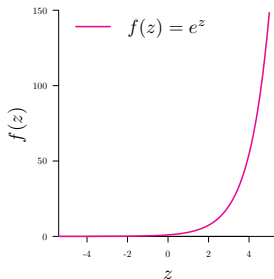
Weibull  
Pareto type I, II, III  
Poisson inverse Gaussian  
...

# GLM: link functions

$$\mathbf{y} | (\mathbf{X}, \boldsymbol{\beta}, \phi) \sim \text{distribution}(\text{function}(\mathbf{X}\boldsymbol{\beta}), \phi),$$

- Most used link *functions* :  
connection between  $\mathbf{y}$  and  $\mathbf{X}\boldsymbol{\beta}$

- ▷ to restrict  $f(\mathbf{X}\boldsymbol{\beta})$  to belong to  $[0, \infty[$ :
  - ▷ log link:  $f(z) = e^z$



# Distribution for count data: Poisson

## Example:

Interest for the number of reads/counts for gene 'X' for a sample basal cells of  $n$  mice

Sample of $n$ mice:	$i = 1$	$i = 2$	$i = 3$	$\dots$	$i = 115$
$y_i$	607	873	1218	$\dots$	2715

If, during a time interval or in a given area,

- ▶ events occur independently,
- ▶ at the same rate,
- ▶ and the probability of an event to occur in a small interval (area) is proportional to the length of the interval (size of the area),

then,

- ▶ a count occurring in a fixed time interval or in a given area,  $Y$ , may be modelled by means of a Poisson distribution with parameter  $\mu$ :

$$Y \sim \text{Poisson}(\mu) \text{ where } \mu = E[Y] = \text{Var}[Y],$$

- ▶ the probability of observing  $x$  events during a fixed time interval or in a given area is given by

$$P(Y = y|\mu) = \frac{\mu^y e^{-\mu}}{y!}.$$

# Distribution for count data: Poisson vs Neg. Bin.

Experimental design

Exploration

Normalization

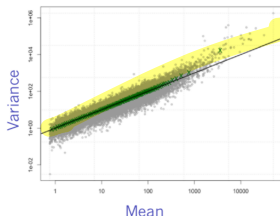
Differential analysis

Multiple testing

## Exploratory data analysis

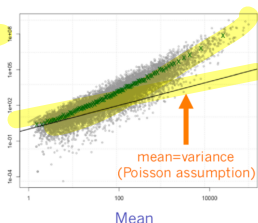
scores between 0 and 1  $\Rightarrow$  underdispersion (variance smaller than mean)

Technical replicates



data from Marioni et al. *Gen Res* 2008

Biological replicates



data from Parikh et al. *Genome Bio* 2010

From D. Robinson and D. McCarthy

scores greater than 1 : overdispersion  $\Rightarrow$  adapted to biological replicates

## 2a/ Negative binomial

- General form:

$$Y_i \sim \text{NB}(\mu_i, \phi)$$

$$f_{Y_i}(y_i | \mu_i, \phi) = \frac{\Gamma(y + \frac{1}{\phi})}{\Gamma(\frac{1}{\phi})\Gamma(y + 1)} \left( \frac{\phi\mu_i}{1 + \phi\mu_i} \right)^y \left( \frac{1}{1 + \phi\mu_i} \right)^{\frac{1}{\phi}}$$

with expectation and variance given by

- $E[Y_i] = \mu_i = \exp(\mathbf{x}_i^T \boldsymbol{\beta})$
- $\text{Var}[Y_i] = \mu_i(1 + \phi\mu_i)$

# Distribution for count data: Poisson vs Neg. Bin.

Experimental design

Exploration

Normalization

Differential analysis

Multiple testing

## Available tests

### Models of count data

- Data transformation and gaussian-based model : limma - voom
- Poisson : TSPM
- Negative Binomial : edgeR, DESeq(2), NBPSeq, baySeq, ShrinkSeq, ...

### Statistical approaches

- Frequentist Approach : edgeR, DESeq(2), NBPSeq, TSPM, ...
- Bayesian Approach : baySeq, ShrinkSeq, EBSeq, ...
- Non-parametric approach : SAMSeq, NOISeq, ...



## 2b/ Negative binomial: Estimation



Experimental design

Exploration

Normalization

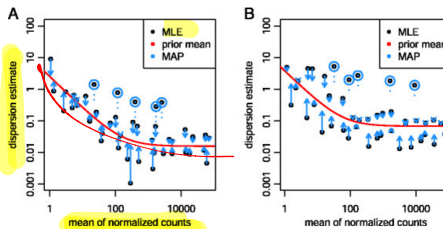
Differential analysis

Multiple testing

### Dispersion estimation with DESeq2

**Hypothesis** : genes of similar average expression strength have similar dispersion

- 1 Estimate **gene-wise dispersion** estimates using maximum likelihood (ML) (black dots)
- 2 Fit a **smooth curve** (red line)
- 3 **Shrink** the gene-wise dispersion estimates (empirical Bayes approach) toward the values predicted by the curve to obtain final dispersion values (blue arrow heads).



## 2b/ Negative binomial: Controlling for library size

$$S_i \mu_i = \underline{e^{x_i^T \beta} S_i}$$

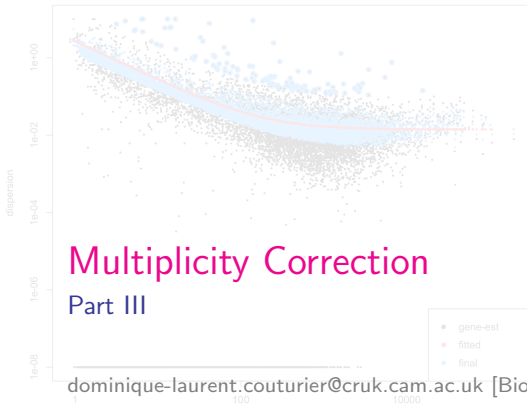
- For a given gene, the variance of the Negative Binomial for the  $i$ th sample is given by

$$\text{Var}(Y_i) = \mu_i (1 + \phi \mu_i)$$

- To control for the library size  $S_i$  of the  $i$ th sample, DESeq2 uses

$$\text{Var}(Y_i) = S_i \mu_i (1 + \phi S_i \mu_i)$$

$$\begin{aligned} \mu_i^* &= S_i e^{\beta_0 + \beta_1 x_i} \\ &= e^{\log(S_i) + \beta_0 + \beta_1 x_i} \\ &= e^{\beta_0 + \beta_1 x_i + \log(S_i)} \end{aligned}$$



## Multiplicity Correction

### Part III

dominique-laurent.couturier@cruk.cam.ac.uk [Bioinformatics core]

(Source: O. Rueda, CRUK-CI; G. Marot, INRIA)

The mean is taken as "normalized counts" scaled by a normalization factor

one dispersion per gene

$$K_{ij} \sim \text{NB}(s_{ij}q_{ij}, \alpha_i)$$

### 3/ Multiplicity correction

Experimental design

Exploration

Normalization

Differential analysis

Multiple testing

## Multiple Testing

False positive (FP) : A non differentially expressed (DE) gene which is declared DE.

For all 'genes', we test  $H_0$  (gene  $i$  is not DE) vs  $H_1$  (the gene is DE) using a statistical test

### Problem

Let assume all the  $G$  genes are not DE. Each test is realized at  $\alpha$  level

Ex :  $G = 10000$  genes and  $\alpha = 0.05 \rightarrow E(FP) = 500$  genes.

### 3/ Multiplicity correction

Experimental design

Exploration

Normalization

Differential analysis

Multiple testing

## The Family Wise Error Rate (FWER)

### Definition

Probability of having at least one Type I error (false positive), of declaring DE at least one non DE gene.

$$FWER = \mathbb{P}(FP \leq 1)$$

### The Bonferroni procedure

Either each test is realized at  $\alpha = \alpha^*/G$  level  
or use of adjusted pvalue  $pBonf_i = \min(1, p_i * G)$  and  $FWER \leq \alpha^*$ .  
For  $G = 2000$ ,  $\leq \alpha^* = 0.05$ ,  $\alpha = 2.510^{-5}$ .

**Easy but conservative and not powerful.**

### 3/ Multiplicity correction

Experimental design

Exploration

Normalization

Differential analysis

Multiple testing

## The False Discovery Rate (FDR)

PROS  $H_0$

Idea : Do not control the ~~error rate~~ but the proportion of error  
 $\Rightarrow$  less conservative than control of the FWER.

### Definition

The false discovery rate of [Benjamini and Hochberg, 1995] is the expected proportion of Type I errors among the rejected hypotheses

$$\text{FDR} = \mathbb{E}(FP/P) \text{ if } P > 0 \text{ and } 0 \text{ if } P = 0$$

### Prop

$$\text{FDR} \leq \text{FWER}$$

### 3/ Multiplicity correction

Experimental design

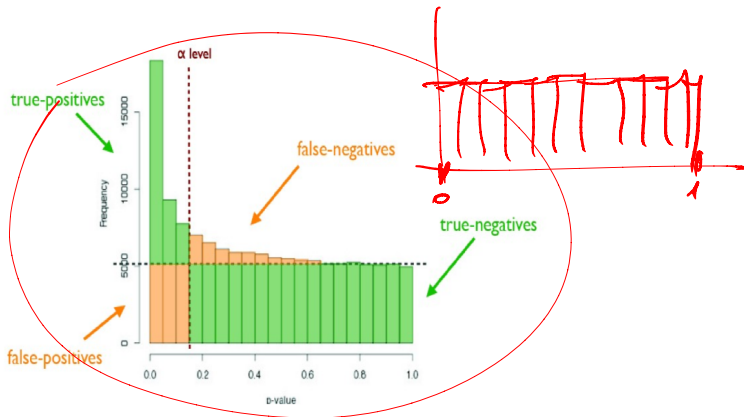
Exploration

Normalization

Differential analysis

Multiple testing

## Standard assumption for p-value distribution



Source : M. Guedj, Pharnext

### 3/ Multiplicity correction

Experimental design

Exploration

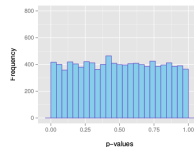
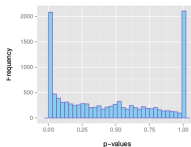
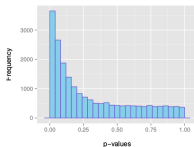
Normalization

Differential analysis

Multiple testing

## p-values histograms for diagnosis

Examples of **expected** overall distribution



- (a) : the most desirable shape
- (b) : very low counts genes usually have large p-values
- (c) : do not expect positive tests after correction



### 3/ Multiplicity correction

Experimental design

Exploration

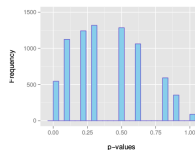
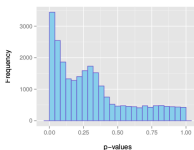
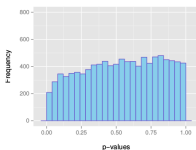
Normalization

Differential analysis

Multiple testing

## p-values histograms for diagnosis

Examples of **not expected** overall distribution



- (a) : indicates a batch effect (confounding hidden variables)
- (b) : the test statistics may be inappropriate (due to strong correlation structure for instance)
- (c) : discrete distribution of p-values : unexpected

# 3/ Multiplicity correction

Experimental design

Exploration

Normalization

Differential analysis

Multiple testing

## Multiple testing : key points

- Important to control for multiple tests
- FDR or FWER depends on the cost associated to FN and FP

### Controlling the FWER :

Having a great confidence on the DE elements (strong control).  
Accepting to not detect some elements (lack of sensitivity  $\Leftrightarrow$  a few DE elements)

### Controlling the FDR :

Accepting a proportion of FP among DE elements. Very interesting in exploratory study.

# 3/ Multiplicity correction

Experimental design

Exploration

Normalization

Differential analysis

Multiple testing

## Multiple testing : key points

- Important to control for multiple tests
- FDR or FWER depends on the cost associated to FN and FP

### Controlling the FWER :

Having a great confidence on the DE elements (strong control).  
Accepting to not detect some elements (lack of sensitivity  $\Leftrightarrow$  a few DE elements)

### Controlling the FDR :

Accepting a proportion of FP among DE elements. Very interesting in exploratory study.