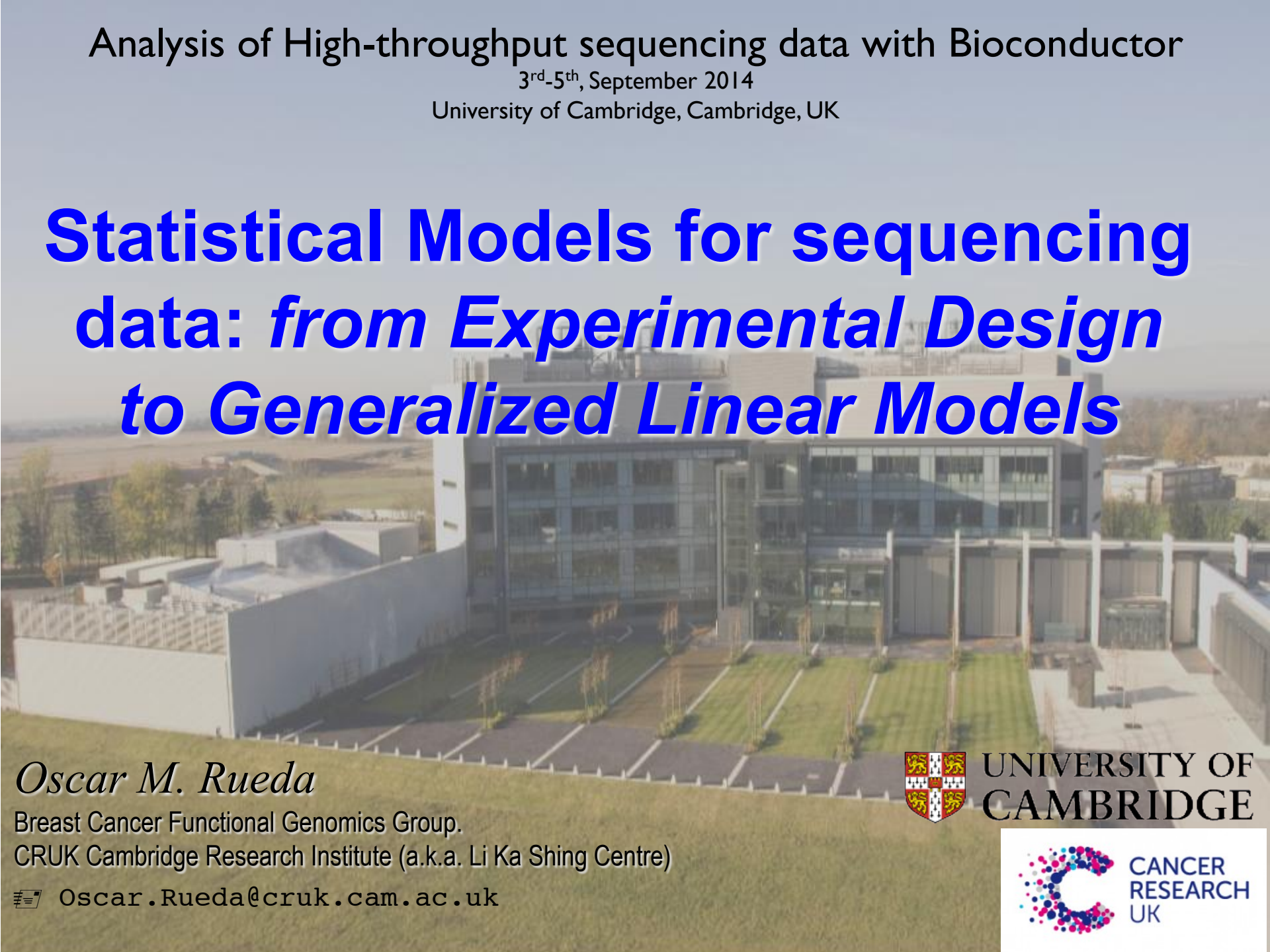
Analysis of High-throughput sequencing data with Bioconductor

3rd-5th, September 2014

University of Cambridge, Cambridge, UK

Statistical Models for sequencing data: *from Experimental Design to Generalized Linear Models*



Oscar M. Rueda

Breast Cancer Functional Genomics Group.

CRUK Cambridge Research Institute (a.k.a. Li Ka Shing Centre)

Oscar.Rueda@cruk.cam.ac.uk

UNIVERSITY OF CAMBRIDGE

CANCER RESEARCH UK

Outline

- Experimental Design
- Design and Contrast matrices
- Generalized linear models
- Models for counting data

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)

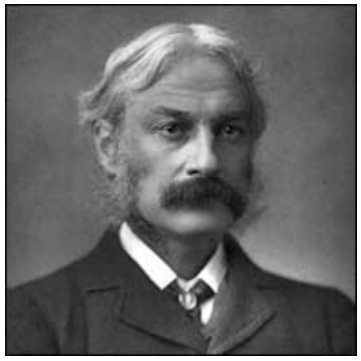
[evolutionary biologist, geneticist and statistician]

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

Design of an experiment

- Select biological questions of interest
- Identify an appropriate measure to answer that question
- Select additional variables or factors that can have an influence in the result of the experiment
- Select a sample size and the sample units
- Assign samples to lanes/flow cells.

Principles of Statistical Design of Experiments

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

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

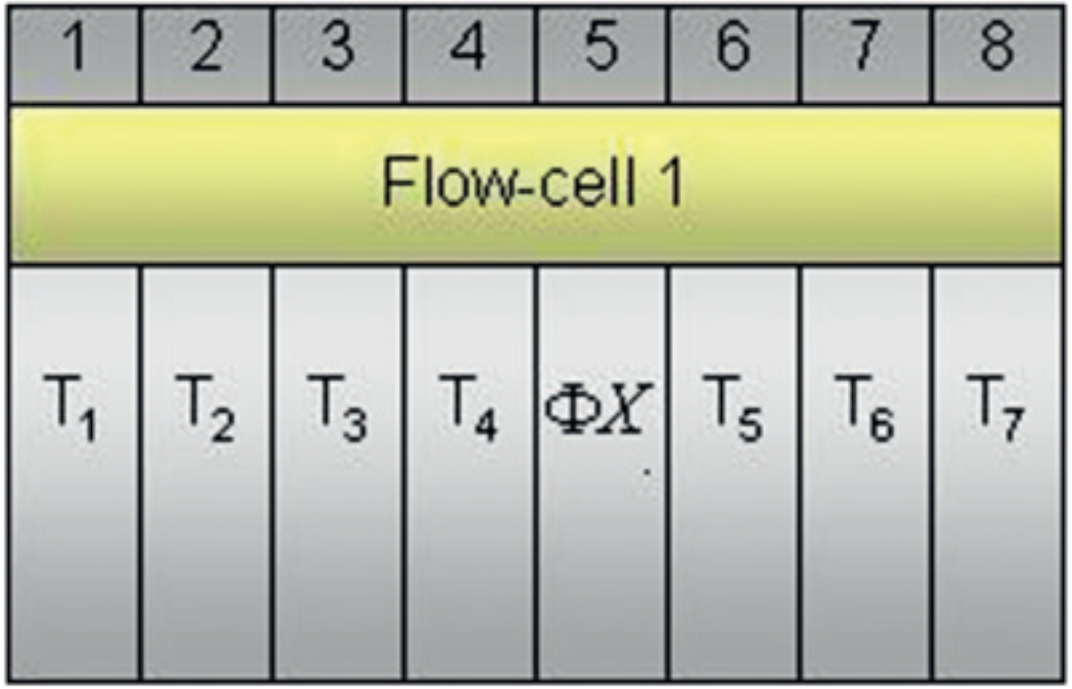
Sampling hierarchy

There are three levels of sampling:

- Subject Sampling
- RNA sampling
- Fragment sampling.

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

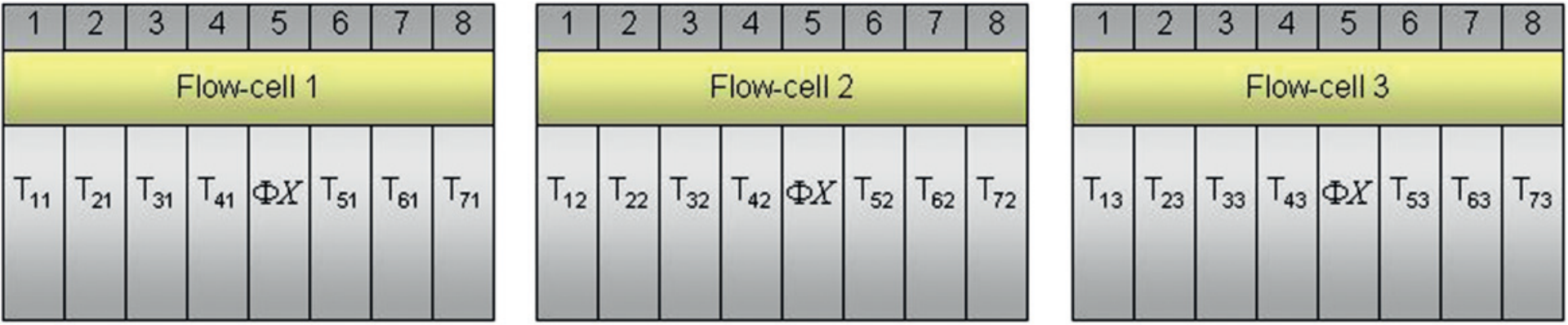
Unreplicated Data



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!

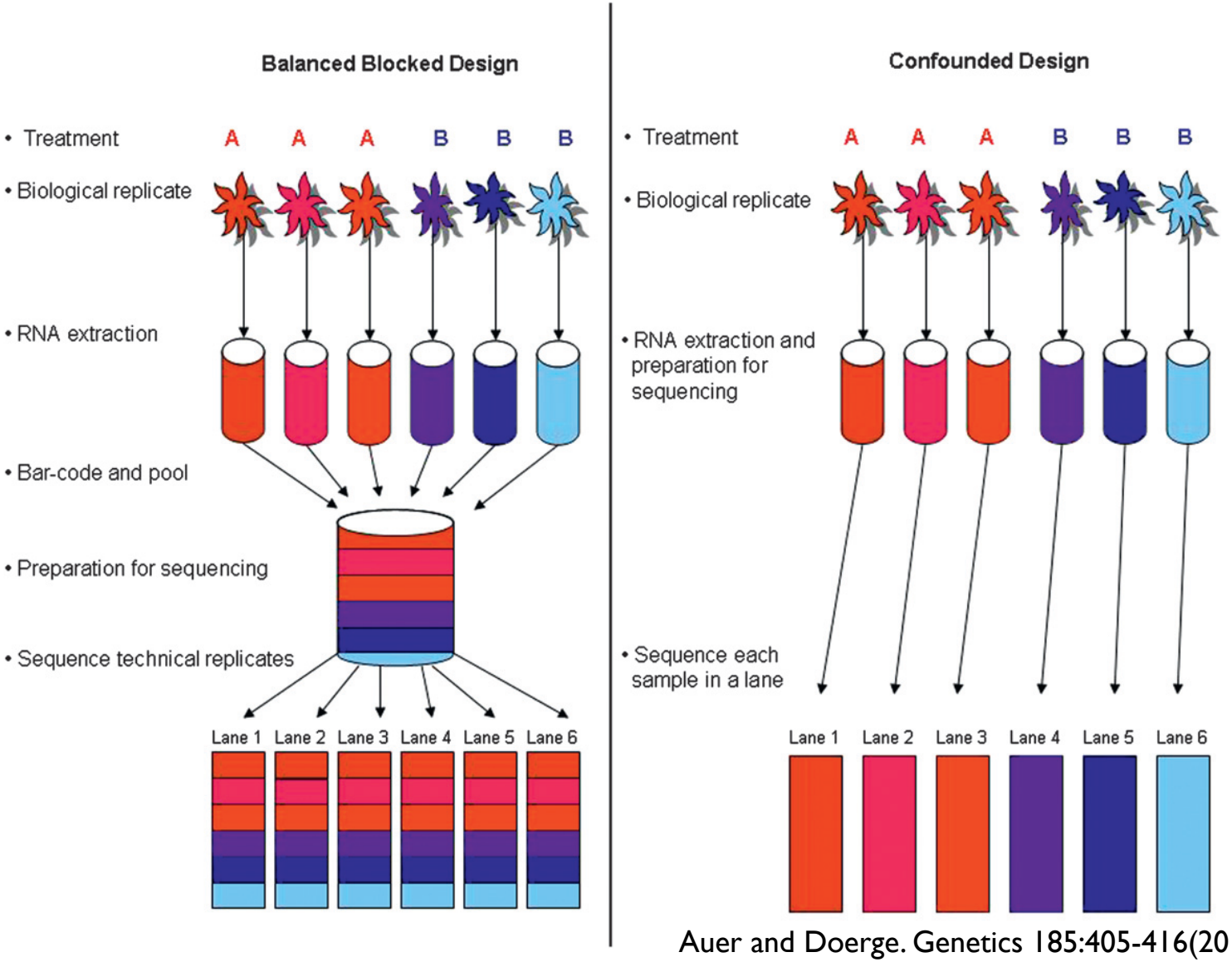
Auer and Doerge. Genetics 185:405-416(2010)

Balanced Block Designs

- Avoids confounding effects:
  - 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…
  - Batch effects (any errors after random fragmentation of the RNA until it is input to the flow cell). Examples: PCR amplification, reverse transcription artifacts…
  - Other effects non related to treatment.

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

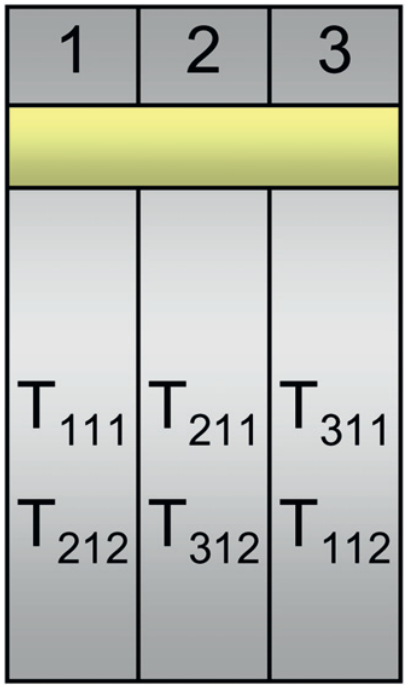
Balanced blocks by multiplexing



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

Balanced incomplete block design and blocking without multiplexing

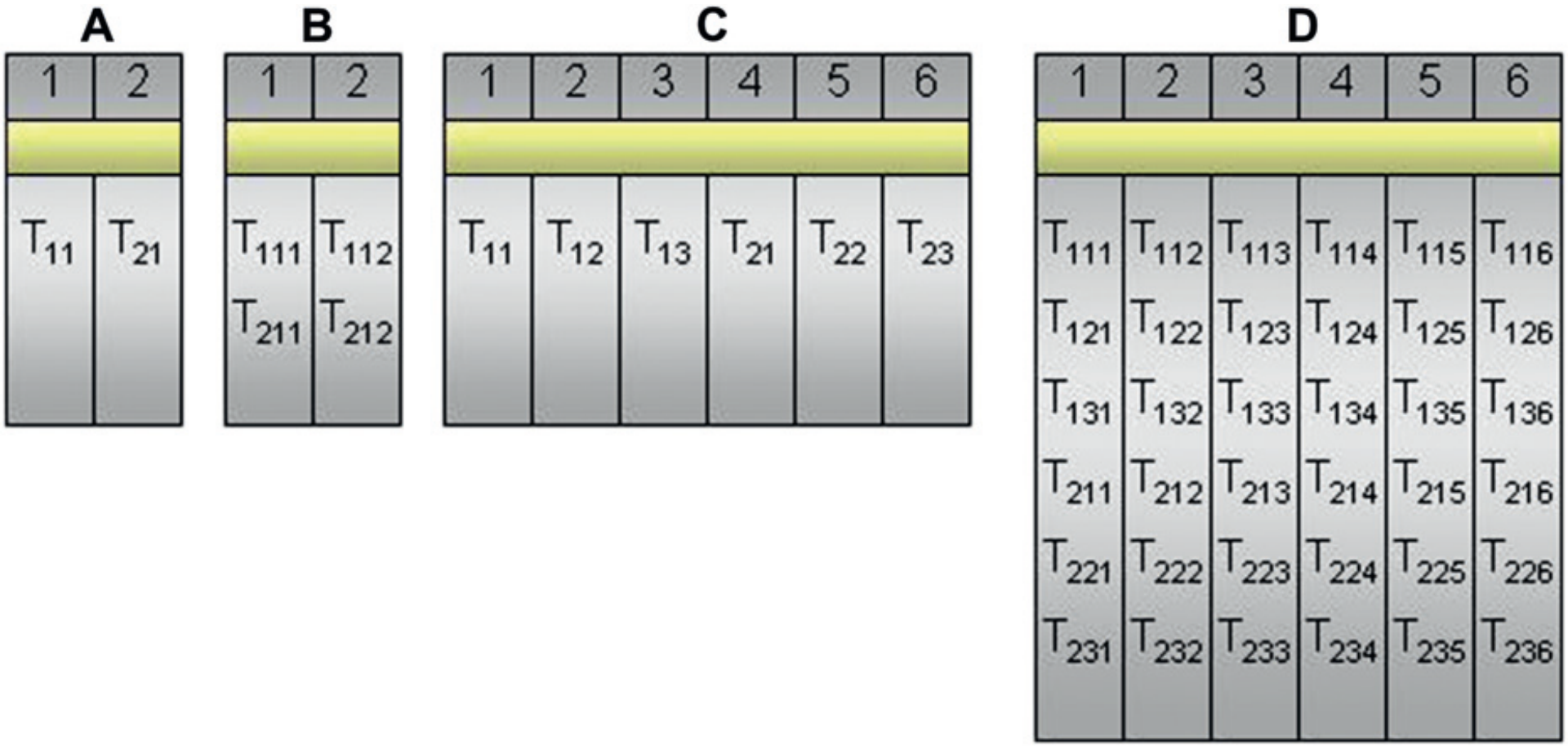
- Usually there are restrictions with the number of treatments, replicates, unique bar codes and available lanes for sequencing.



- 3 treatments (T_1, T_2, T_3)
- 1 subject per treatment (T_{11}, T_{21}, T_{31})
- 2 technical replicates (T_{111}, T_{112}, T_{211}, T_{212}, T_{311}, T_{312})

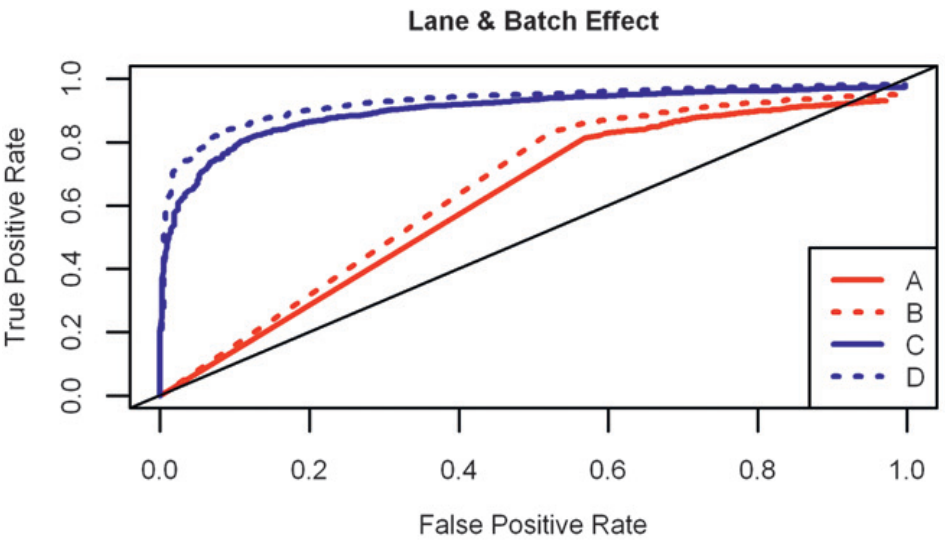
Auer and Doerge. Genetics 185:405-416(2010)

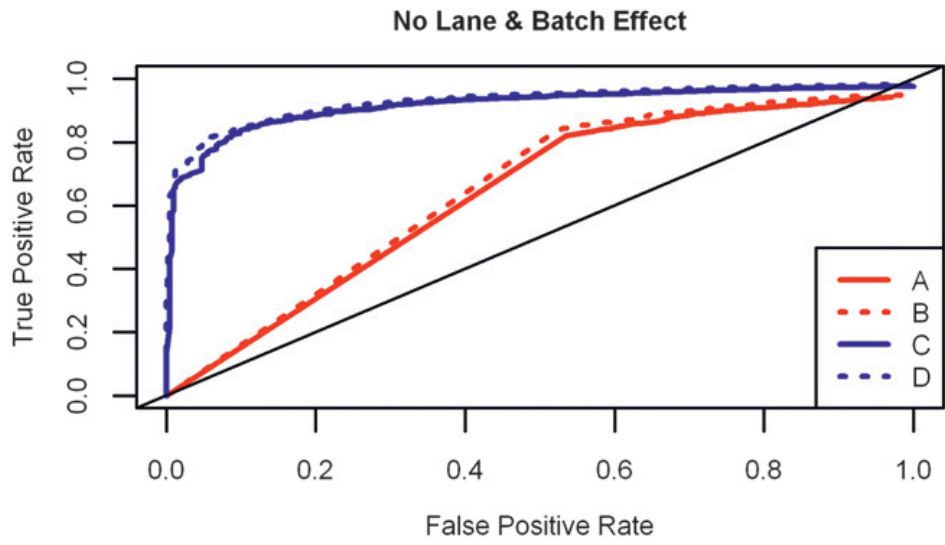
Simulations

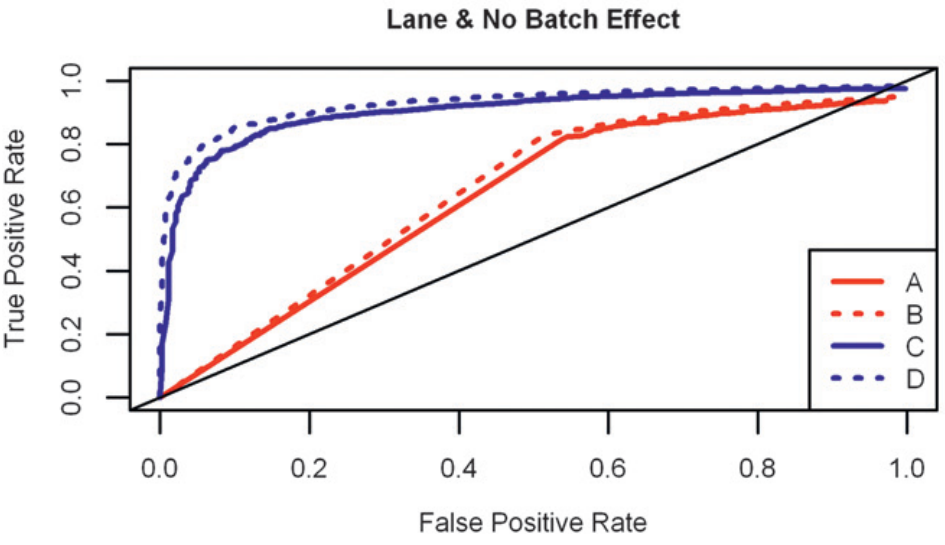


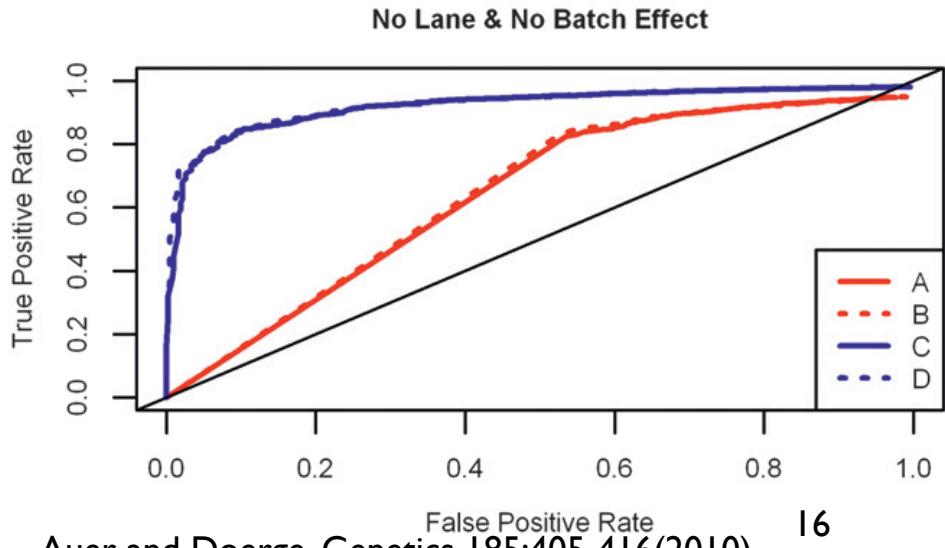
Auer and Doerge. Genetics 185:405-416(2010)

Results









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

Benefits of a proper design

- 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…)

Design and contrast matrices

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)

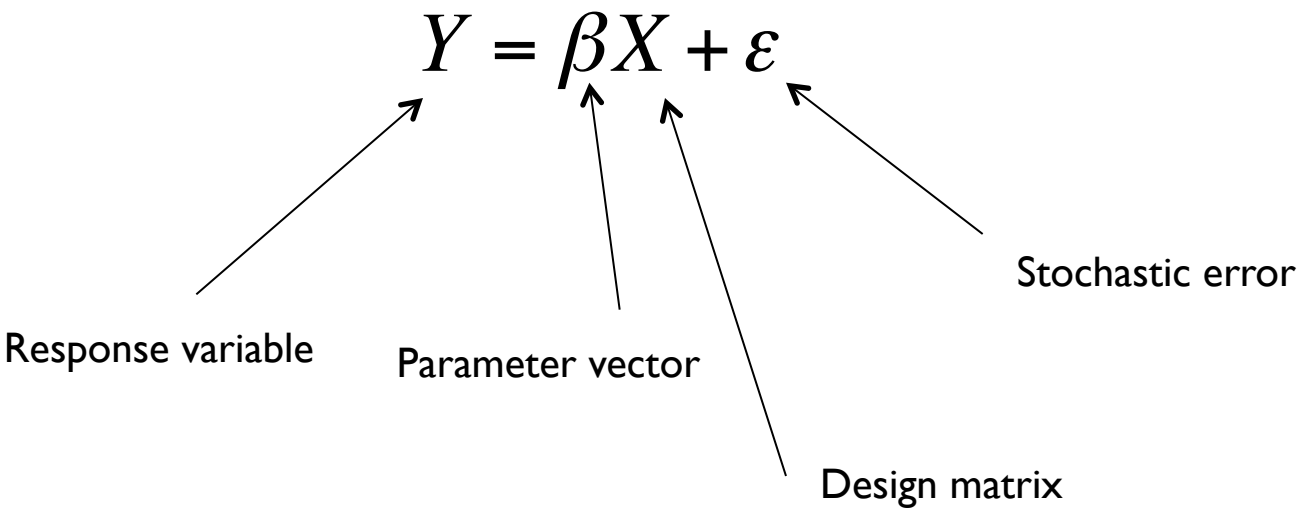
$$E(Y) = f(X)$$

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

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



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

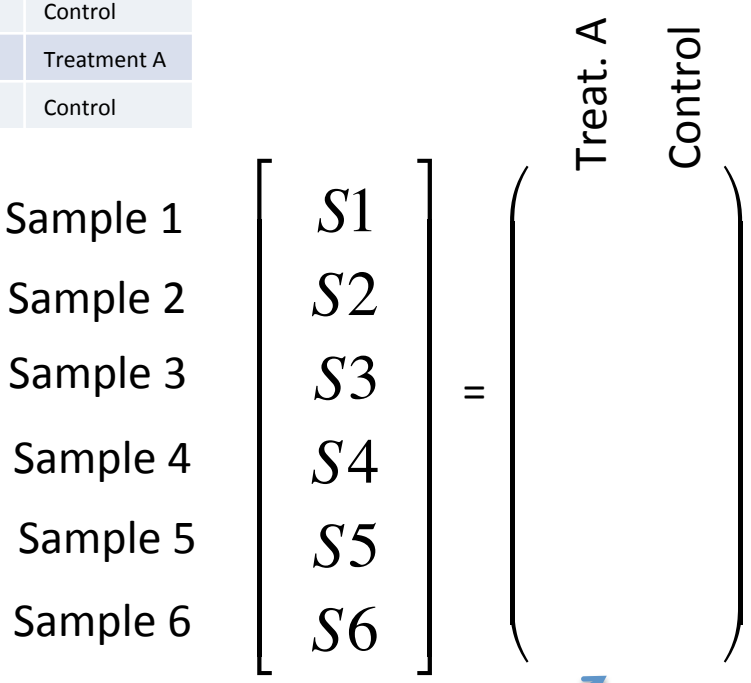
Number of samples: 6
Number of factors: 1
Treatment: Number of levels: 2

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



Parameters (coefficients, levels of the variable)

$$\begin{bmatrix} T \\ C \end{bmatrix}$$

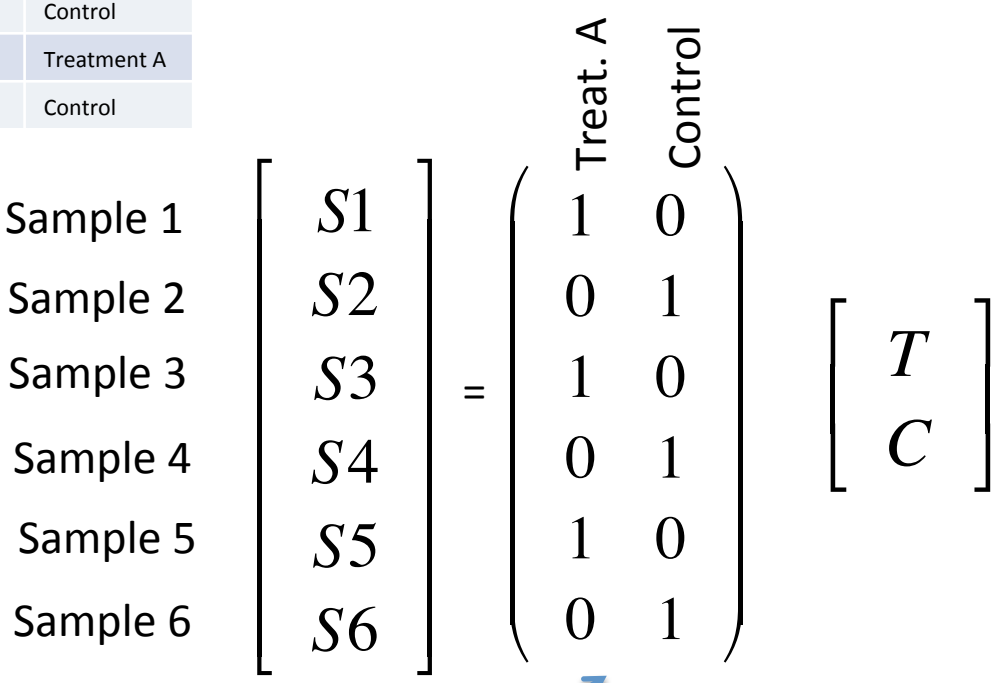
C is the mean expression of the control
T is the mean expression of the treatment

Design Matrix

Equivalent to a t-test

Design matrix for models with 1 factor, 2 levels

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



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:

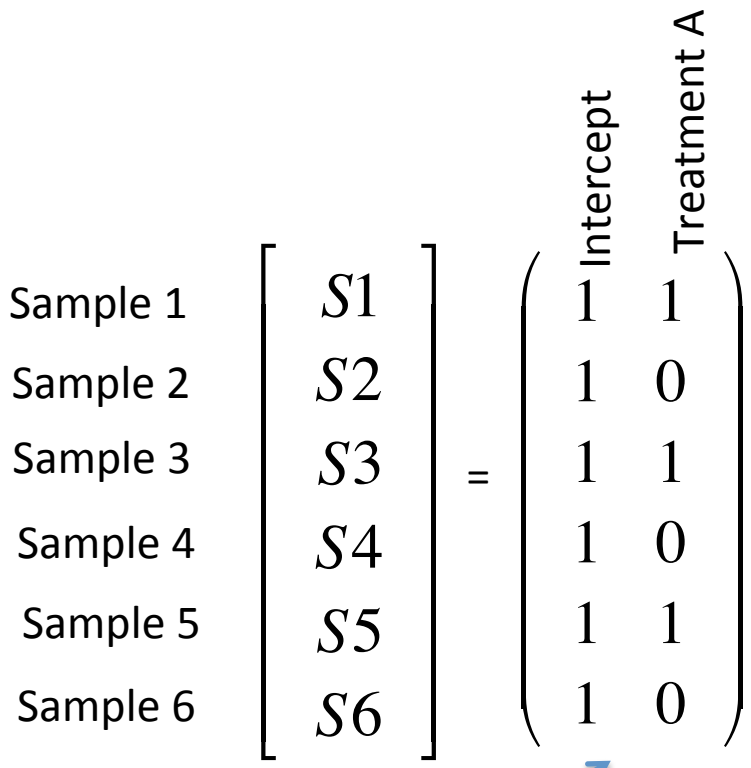
C= Baseline expression
T_A= Baseline expression + effect of treatment

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)

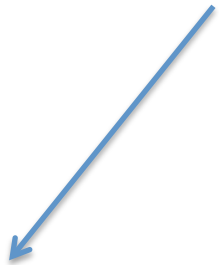
Design Matrix

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

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

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

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

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

Design matrix for ANOVA models

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

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

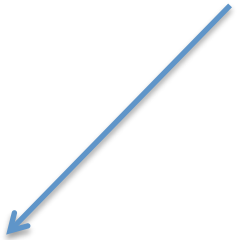
$$\begin{bmatrix} S1 \\ S2 \\ S3 \\ S4 \\ S5 \\ S6 \end{bmatrix} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} \begin{bmatrix} \beta_0 \\ a \\ b \end{bmatrix}$$

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!



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

By default, R uses the first level as baseline

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:

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

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:

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

Exercise

Build contrast matrices for all pairwise comparisons for these designs:

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

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

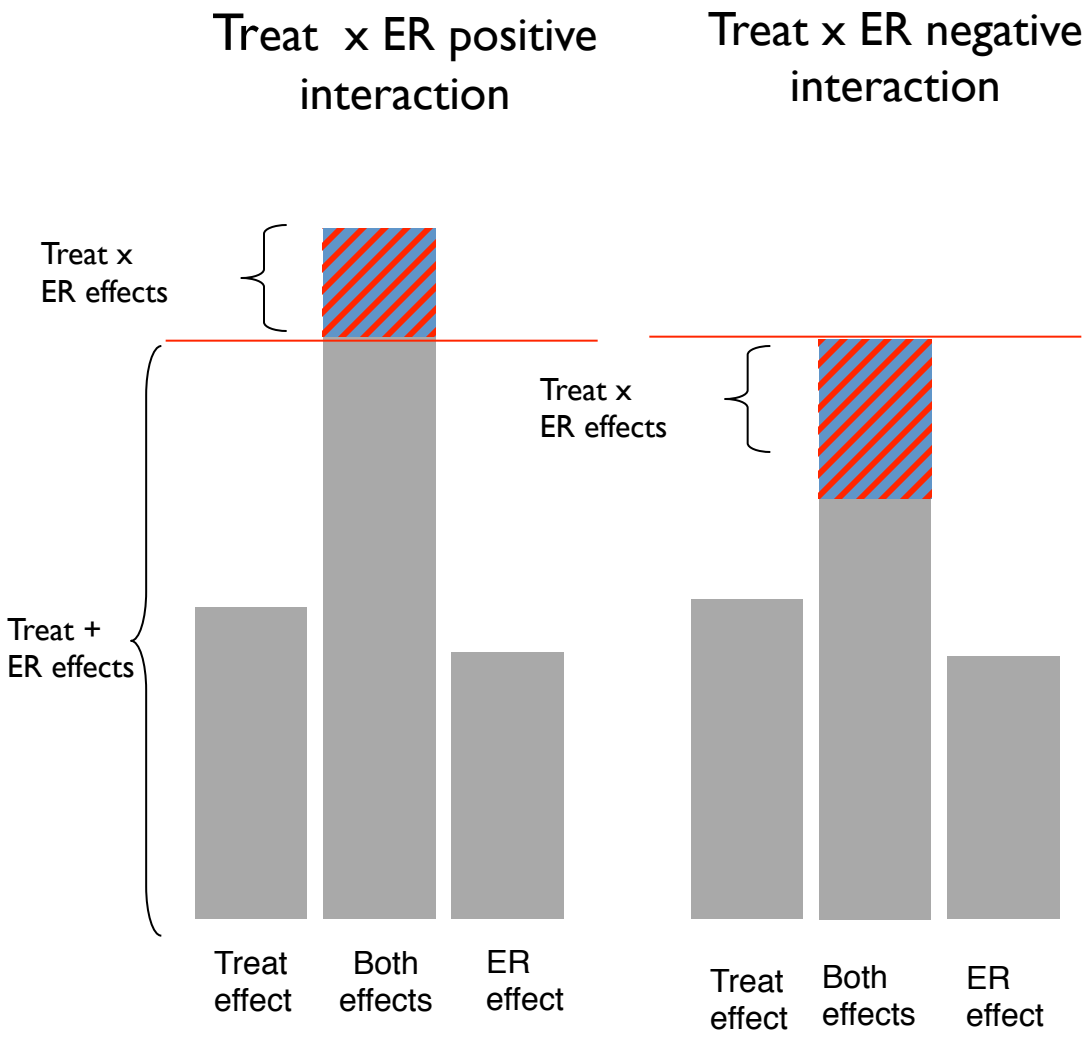
Models with 2 factors

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



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)

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

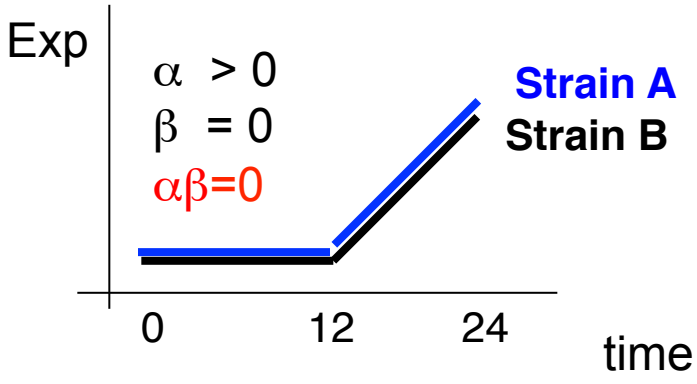
"Extra effect" of Treatment A on ER+ samples

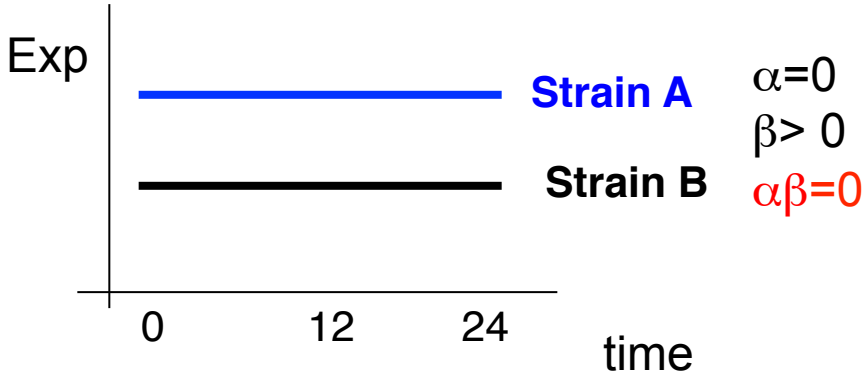
	No Treat	Treat A
ER -	S6, S8	S5, S7
ER +	S2, S4	S1, S3

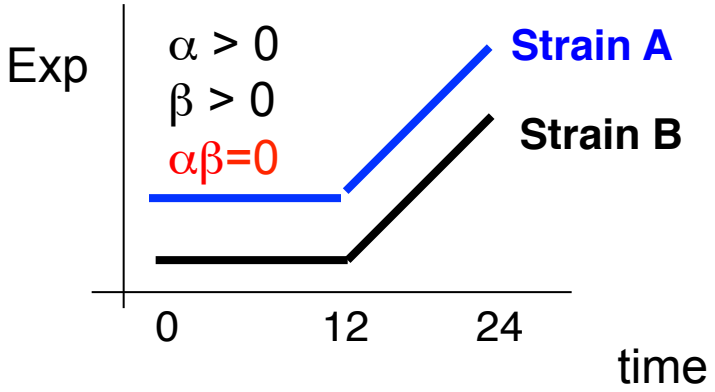
2 by 3 factorial experiment

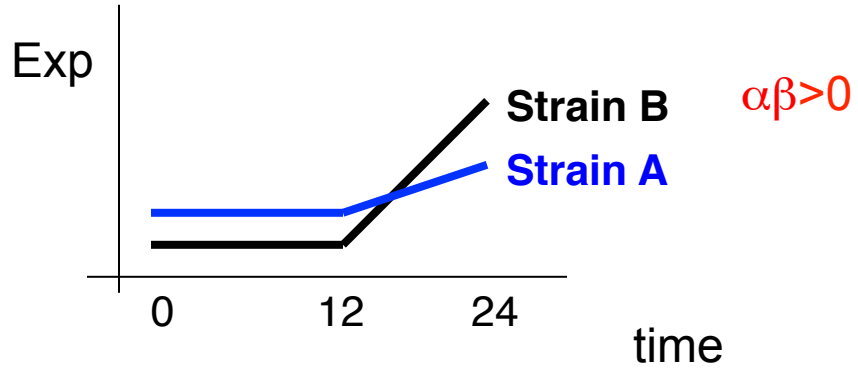
- Identify DE genes that have different time profiles between different mutants.

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









Slide by Natalie Thorne, Nuno L. Barbosa Morais

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)

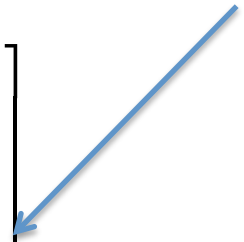
Sample	ER	Dose
Sample 1	+	37
Sample 2	-	52
Sample 3	+	65
Sample 4	-	89
Sample 5	+	24
Sample 6	-	19
Sample 7	+	54
Sample 8	-	67

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)`

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

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

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.

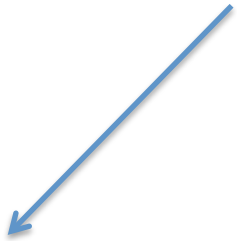
Number of samples: 2
Number of factors: 2
Sample: Number of levels: 2
Time: Continuous or categorical?

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.

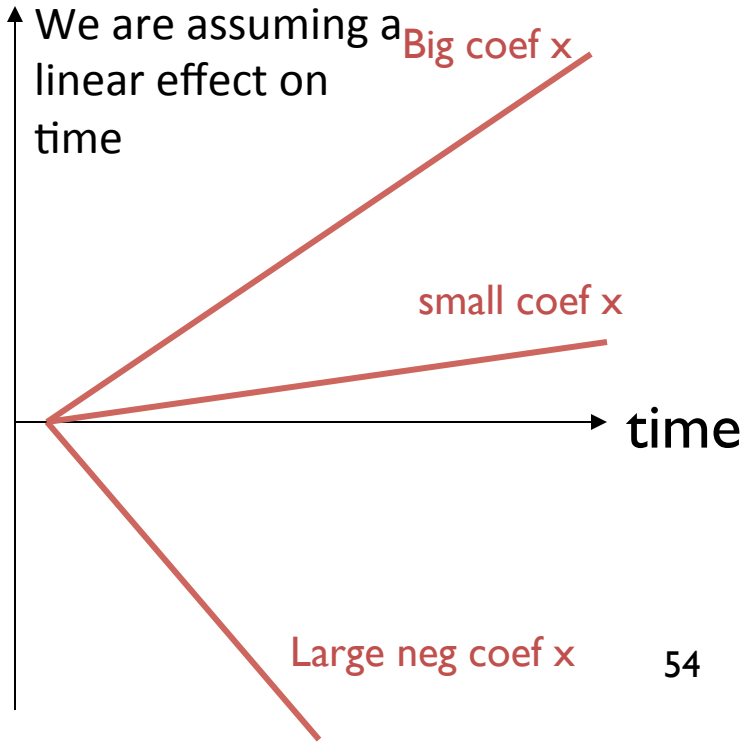
Sample	Time
Sample 1	0h
Sample 1	1h
Sample 1	4h
Sample 1	16h
Sample 2	0h
Sample 2	1h
Sample 2	4h
Sample 2	16h

Time Course experiments

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**

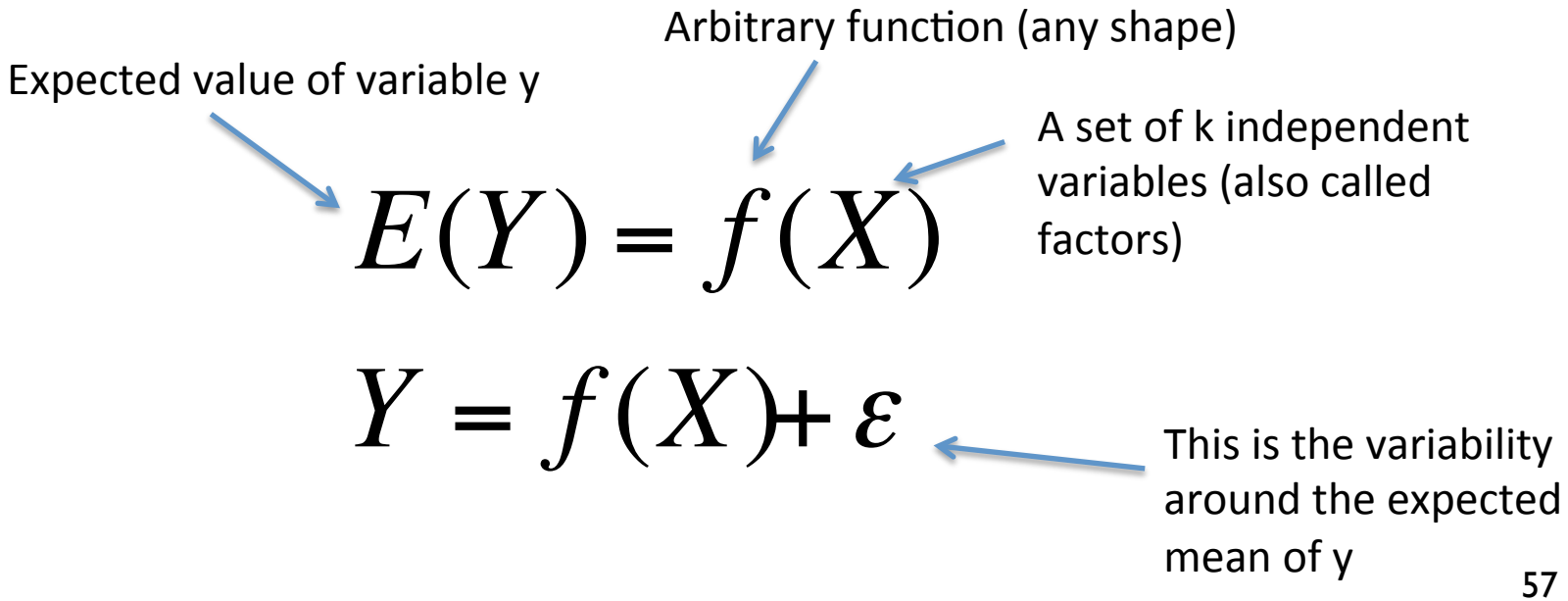
Multi factorial models

- 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.

Generalized linear models

Statistical models

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



Fixed vs Random effects

- If we consider an independent variable X_i as fixed, that is the set of observed values has been fixed by the design, then it is called a fixed factor.
- If we consider an independent variable X_j as random, that is the set of observed values comes from a realization of a random process, it is called a random factor.
- Models that include random effects are called MIXED MODELS.
- In this course we will only deal with fixed factors.

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

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

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

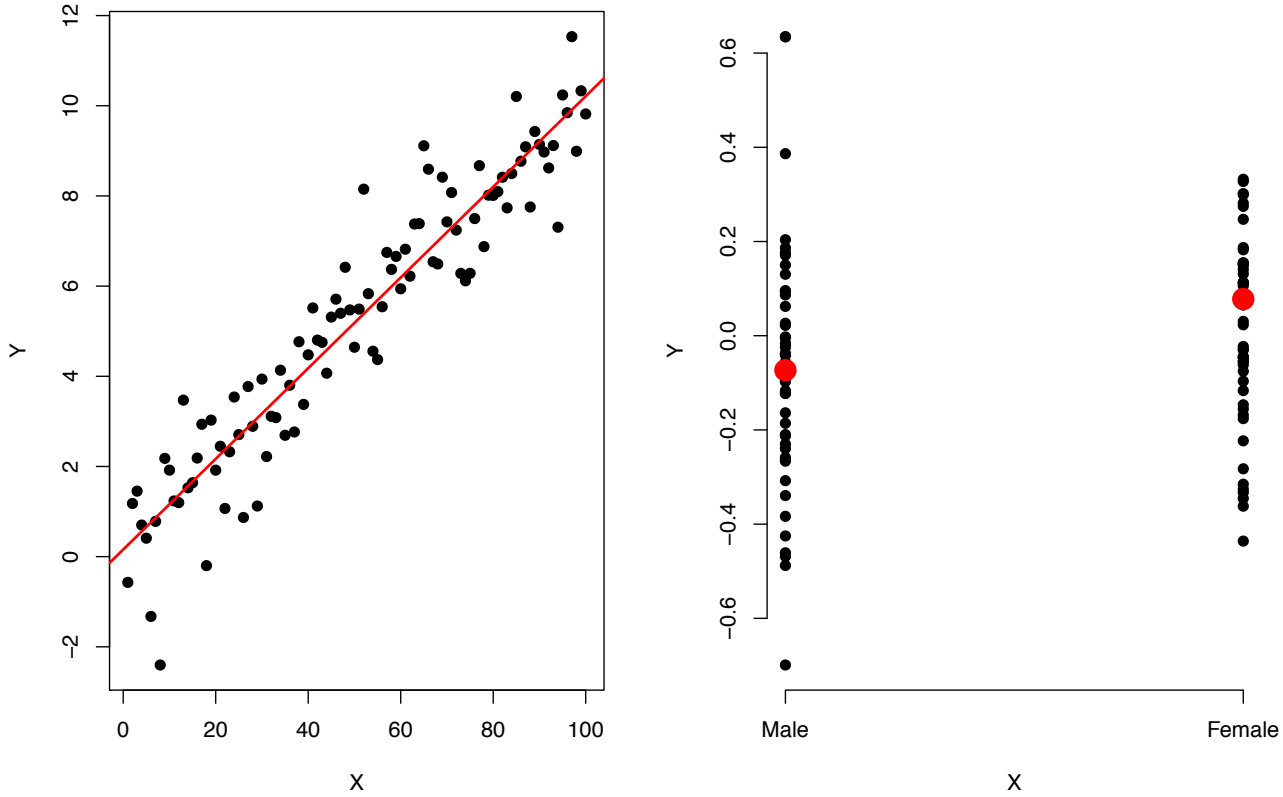
We can use functions of the variables if the effects are linear

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)

Model Assumptions

In order to conduct statistical inferences on the parameters on the model, some assumptions must be made:

- The observations 1,..,n are independent
- Normality of the errors:

$$\varepsilon_i \sim N(0, \sigma^2)$$

- Homoscedasticity: the variance is constant.
- Linearity.

Generalized linear models

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

Link function

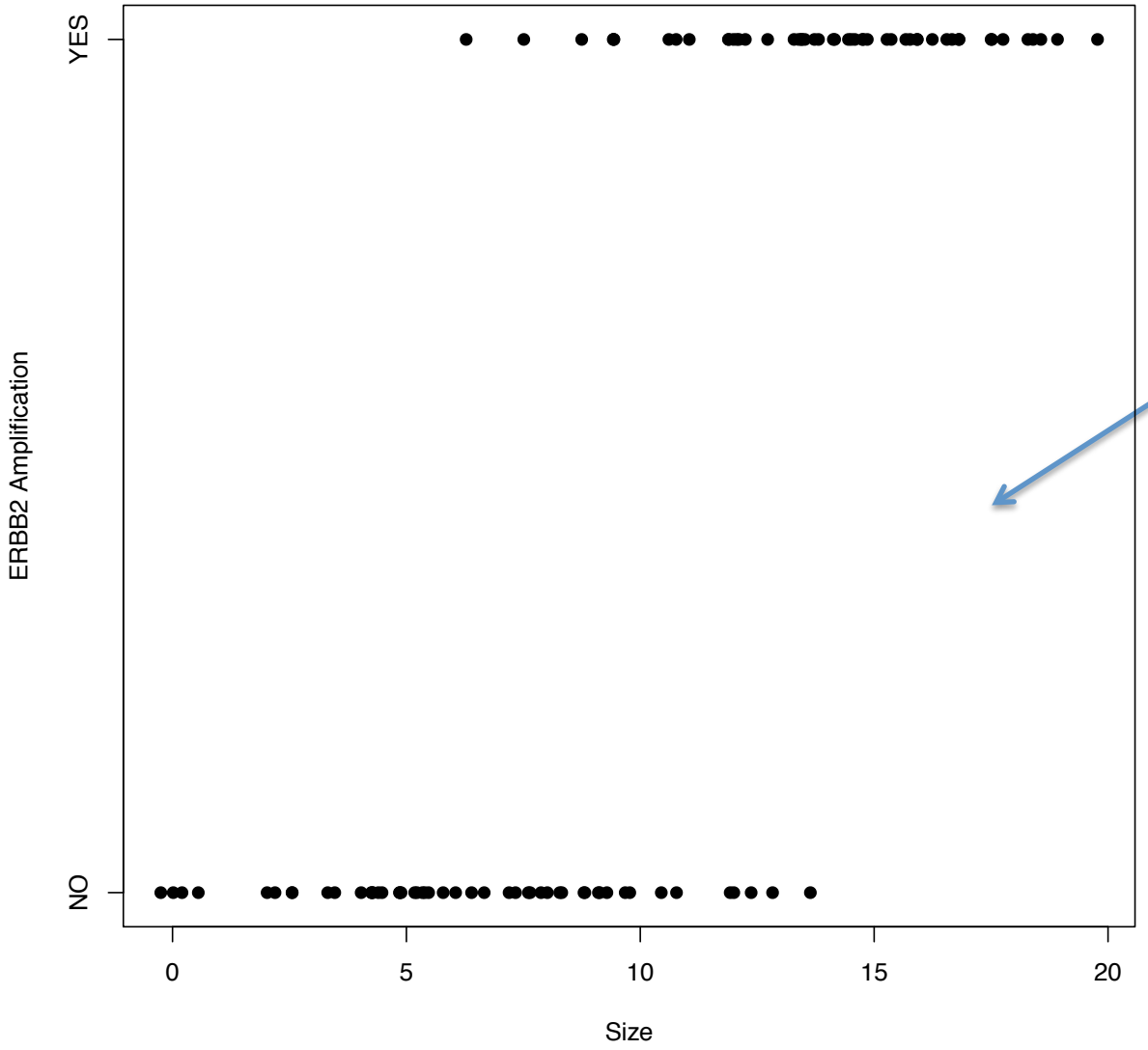
$$g(E(Y)) = X\beta$$

- 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.

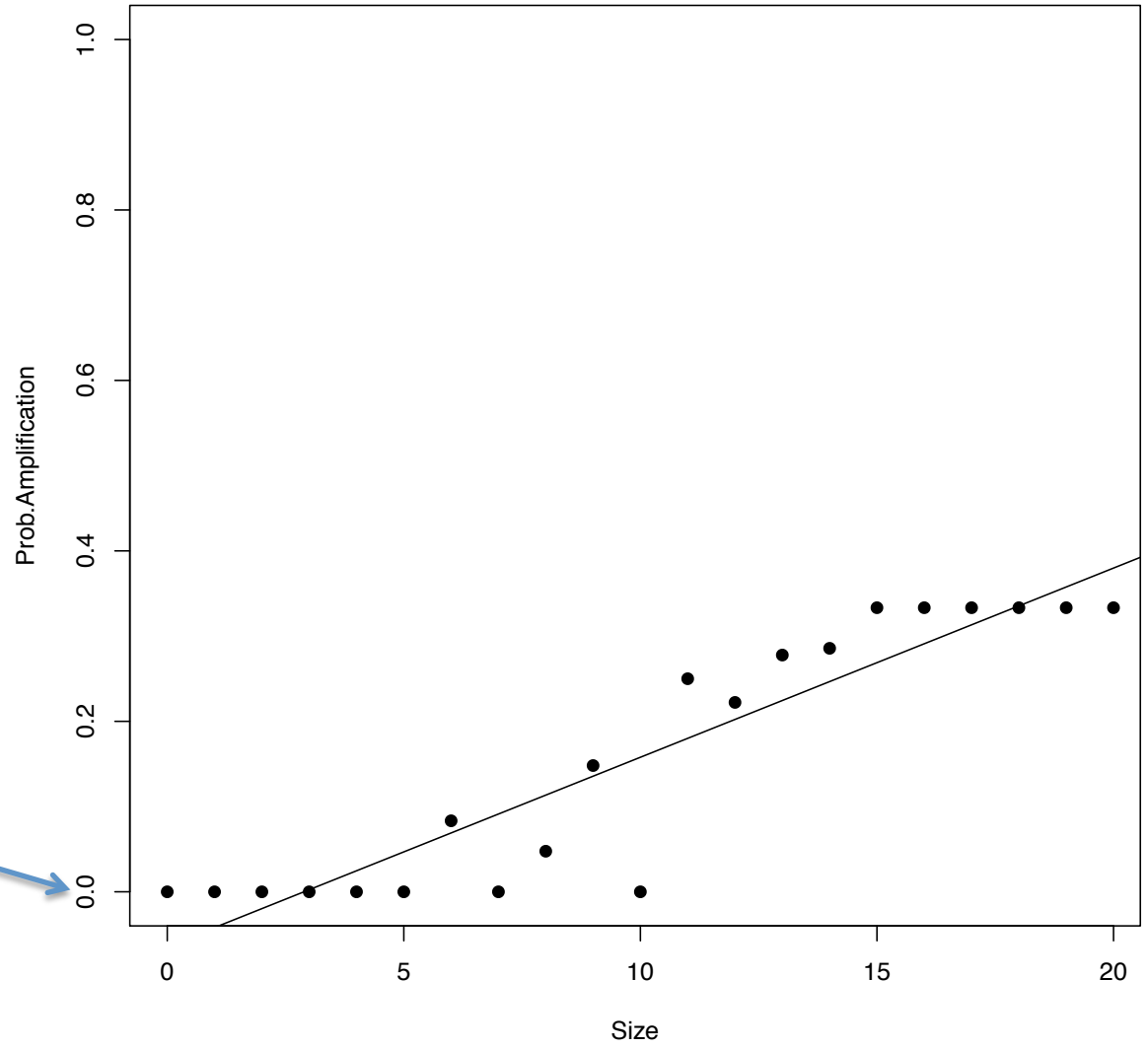
Example: Logistic Regression



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

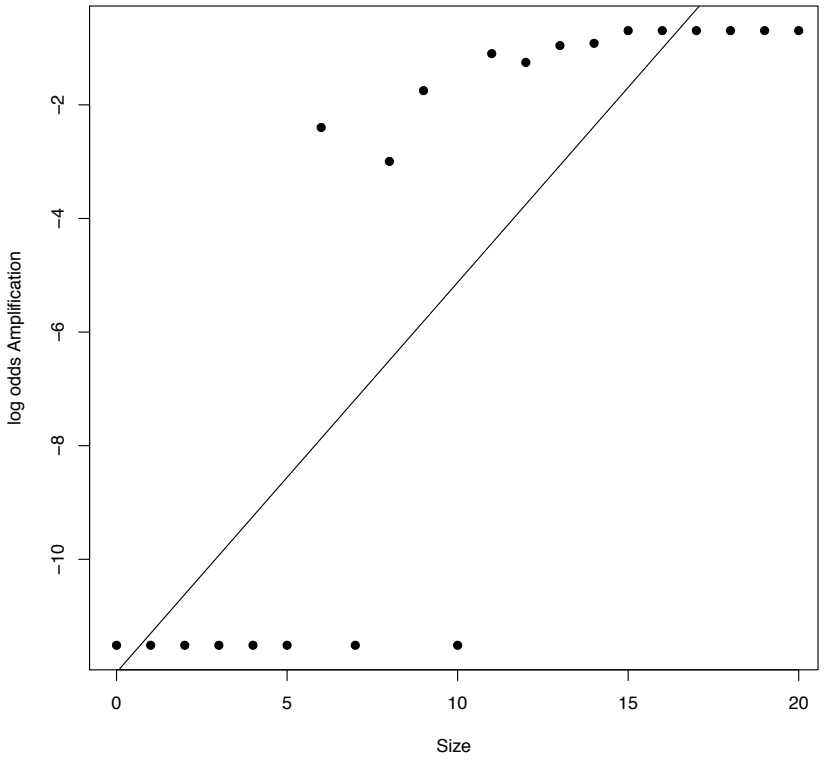
Some predictions are out of the possible range for a probability



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?

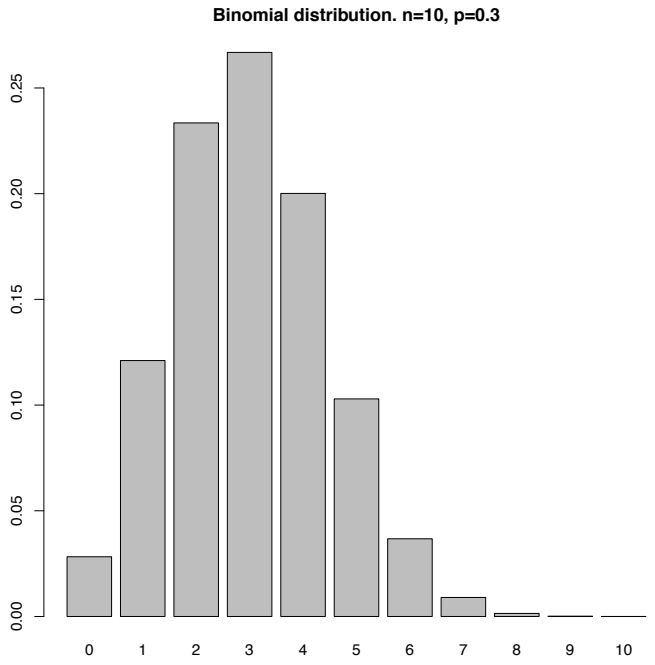
- Y follows a Bernoulli distribution; it can take two values (YES or NO)
- The expectation of Y, p is the probability of YES (EY=p)
- We assume that there is a linear relationship between size and a function of the expected value of Y: the log odds (the **link** function)

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

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

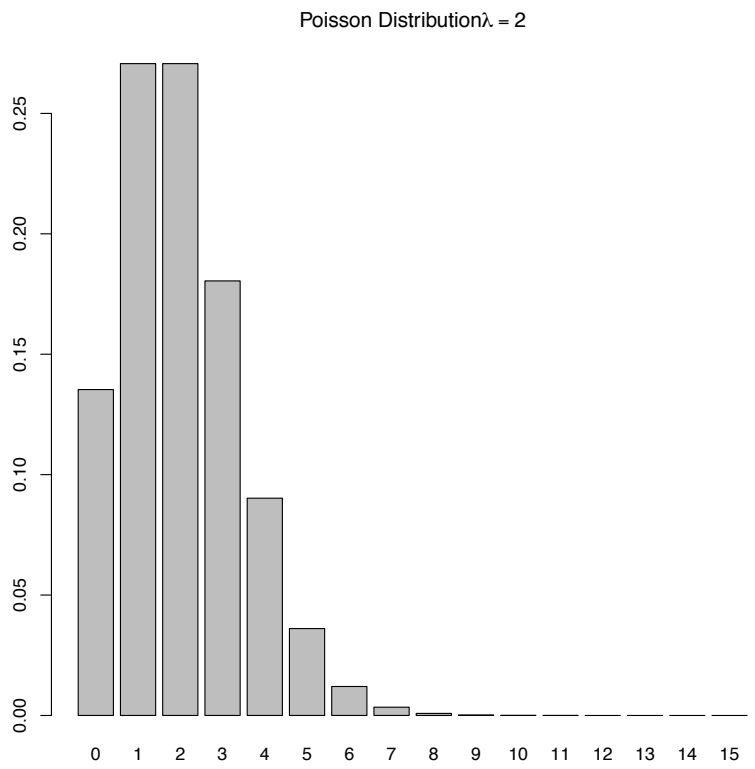
Binomial Distribution

- 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

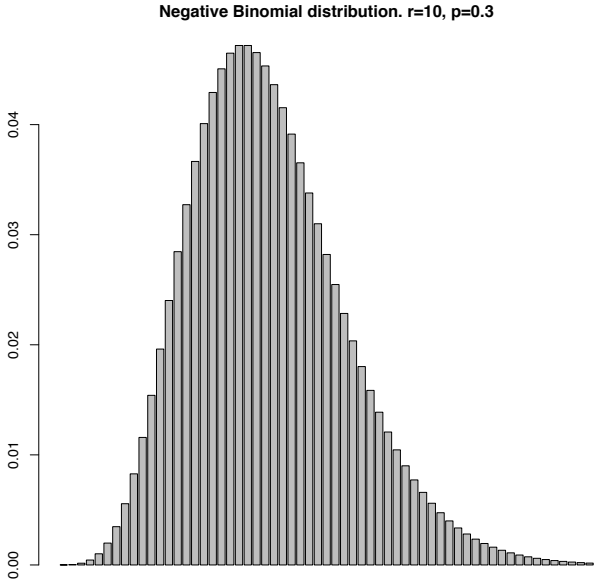


Negative Binomial Distribution

- 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.

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

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



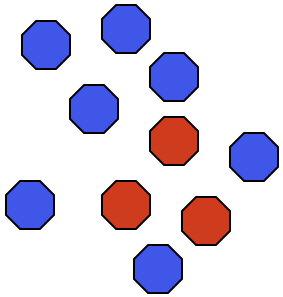
Hypothesis testing

- 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?

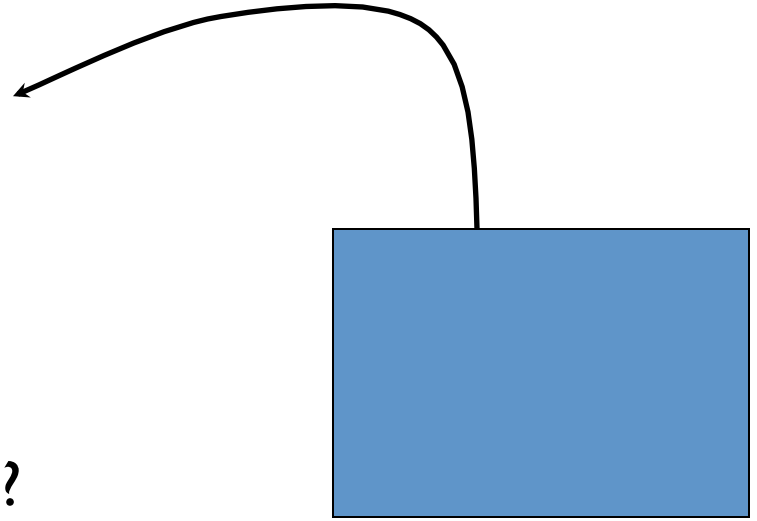
Sampling and testing

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?

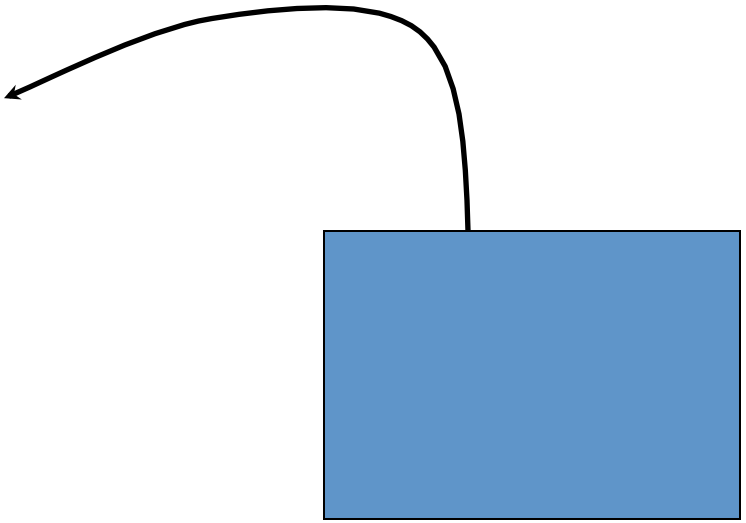
How many reds could I expect to get <u>just by chance alone!</u>

10% red balls and
90% blue balls

Slide by Natalie Thorne, Nuno L. Barbosa Morais

Sample

Random sample of 10 balls from the box



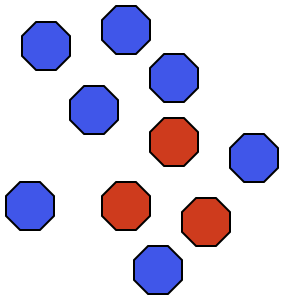
10% red balls and
90% blue balls

Null hypothesis

(about the population that is being sampled)

Discrete observations

#red = 3



Test statistic

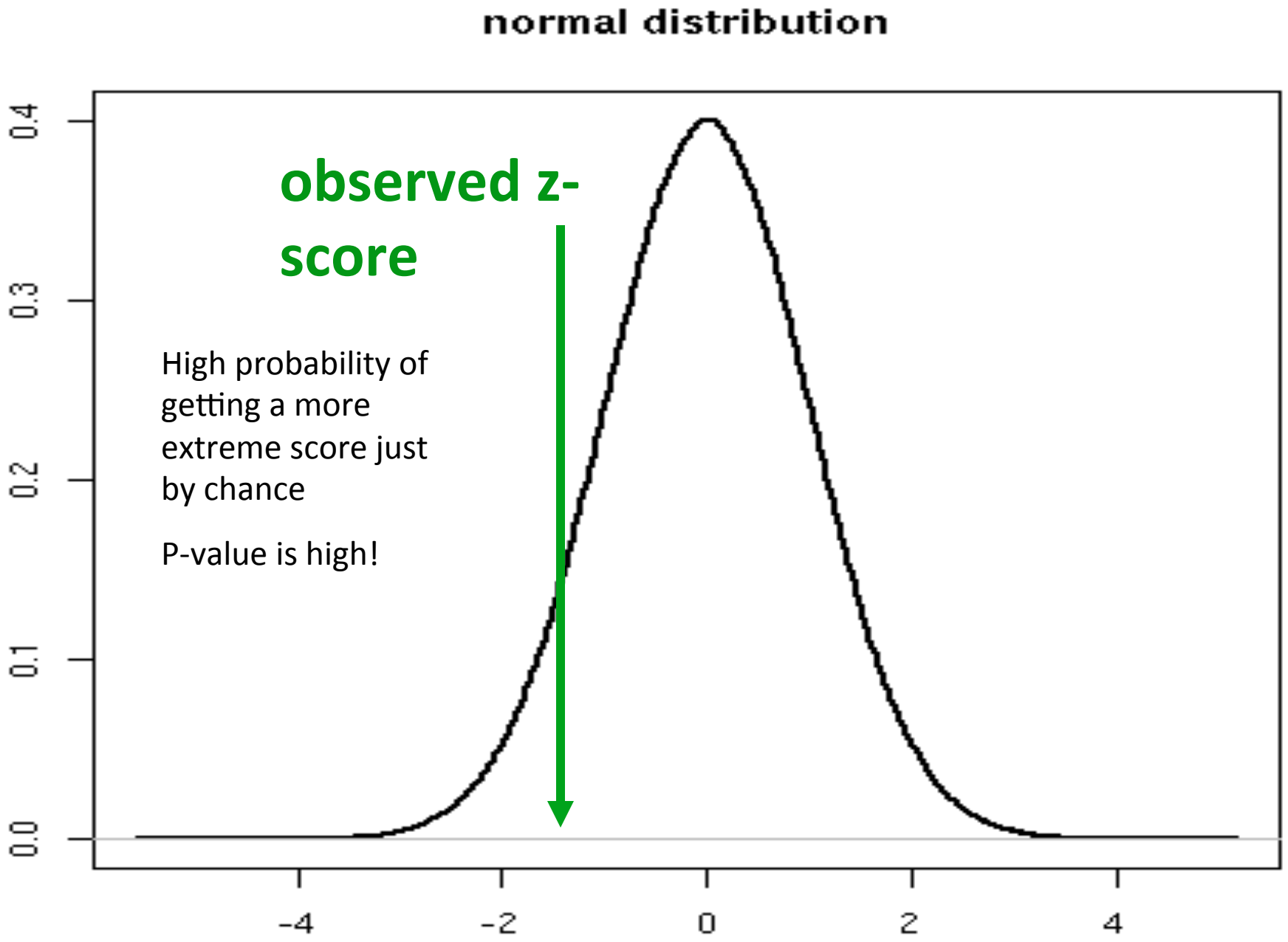
Rejection criteria (based on your observed sample, do you have evidence to reject the hypothesis that you sampled from the null population)

Slide by Natalie Thorne, Nuno L. Barbosa Morais

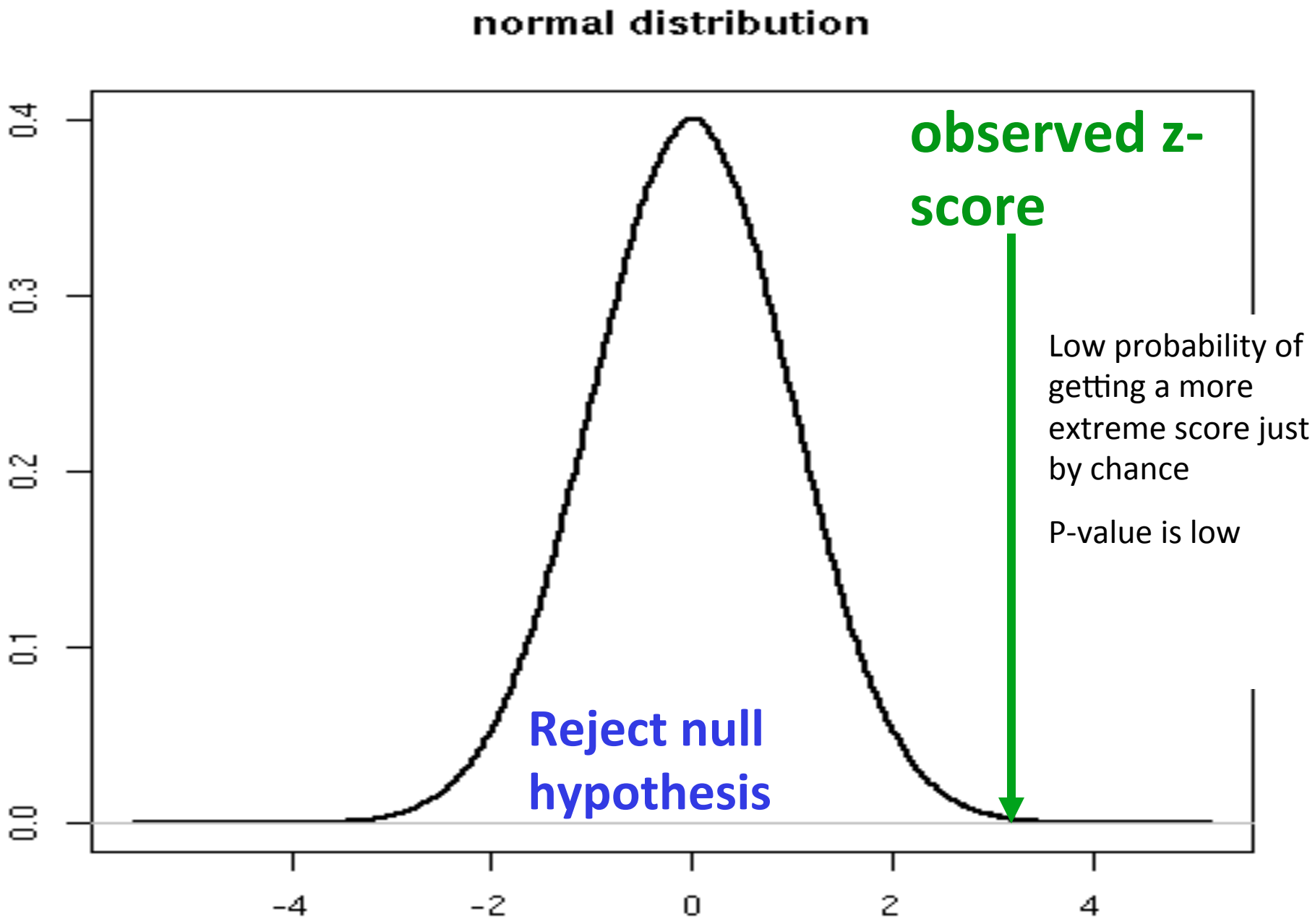
Hypothesis testing

- **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 $t(x_1,\ldots x_n)$



Slide by Natalie Thorne, Nuno L. Barbosa Morais



Slide by Natalie Thorne, Nuno L. Barbosa Morais

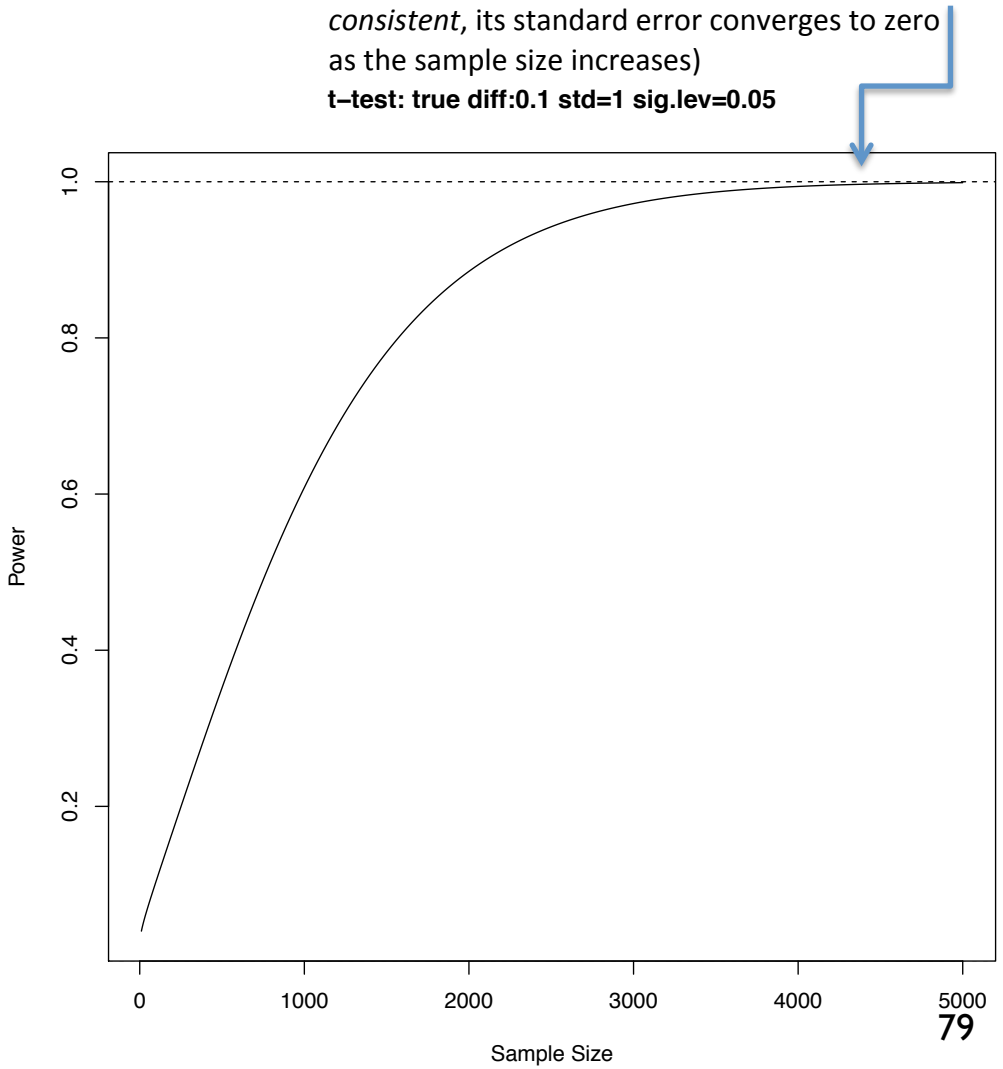
Type I and Type II errors

- 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)

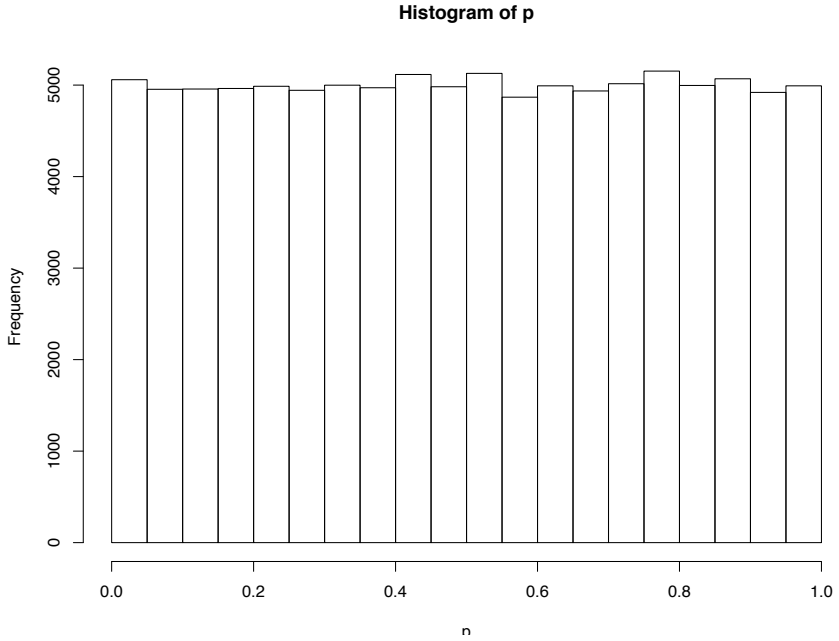
- We are working with models, therefore we would like to do hypothesis tests on coefficients or contrasts of those models
- We fit two models M_1 without the coefficient to test and M_2 with the coefficient.
- We compute the likelihoods of the two models (L_1 and L_2) and obtain LRT=-2log(L_1 /L_2) that 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 High throughput experiments we are fitting one model for each 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 **α/ 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

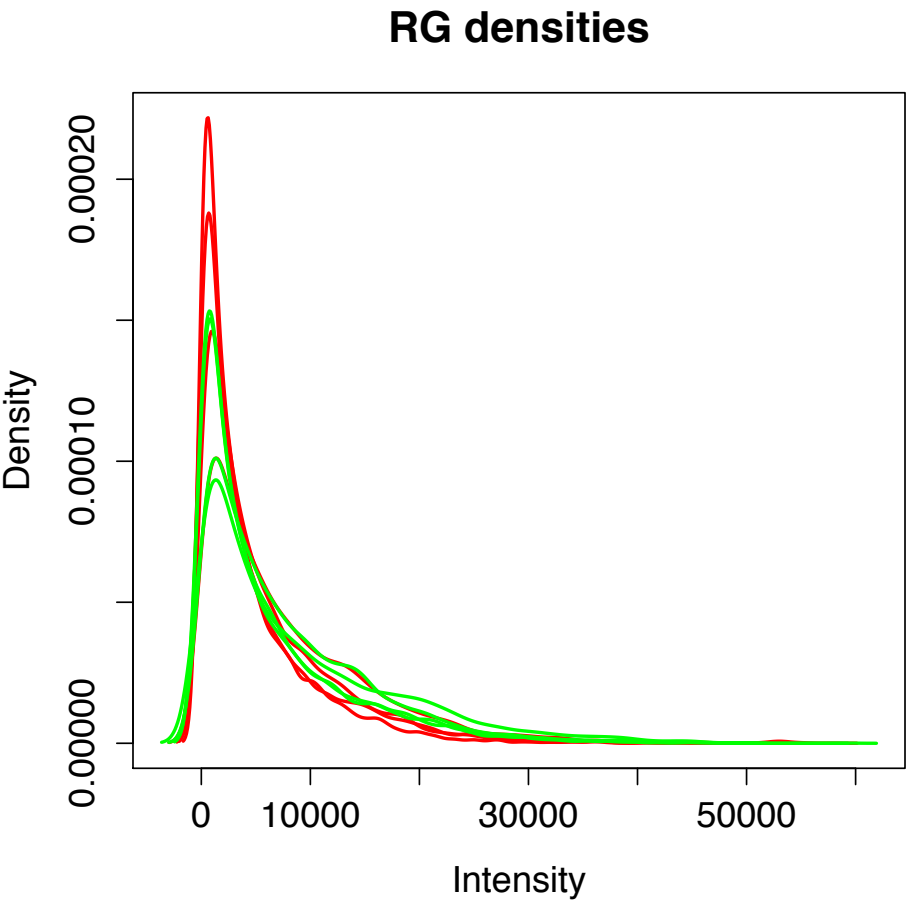
Multiple power problem

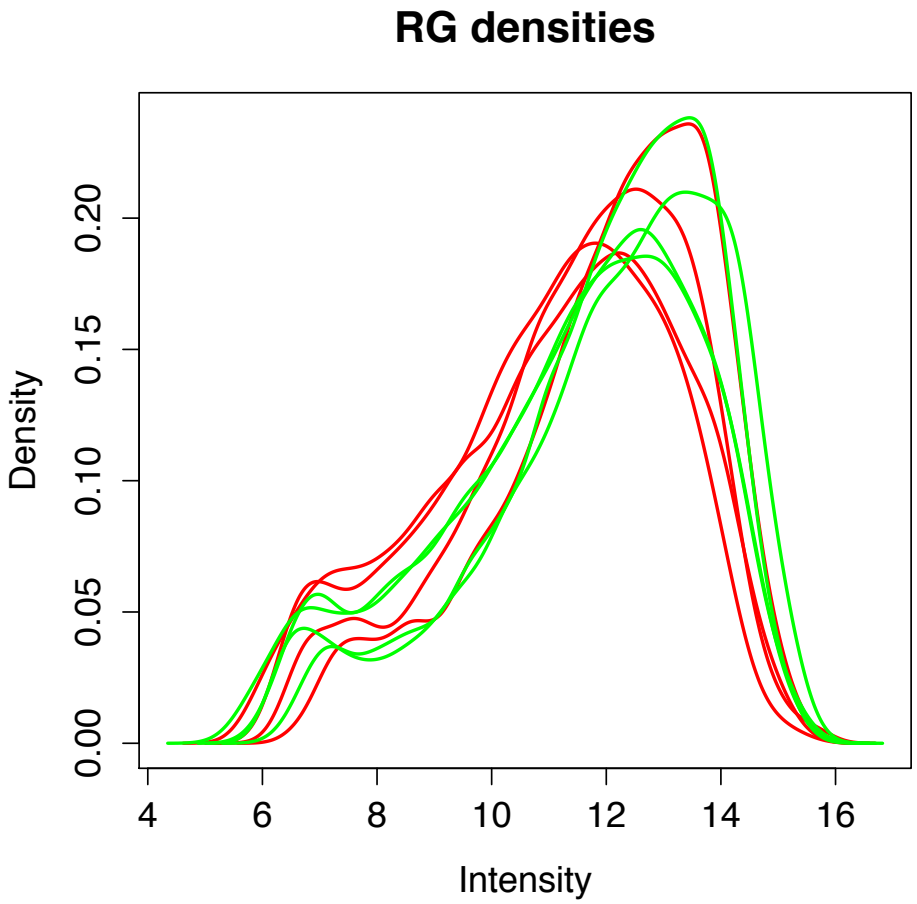
- We have another problem related to the power of each test. Each unit tested has a different test statistic that depends on the variance of the distribution. This variance is usually different for each gene/transcript,…
- This means that the probability of detecting a given difference is different for 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.

Models for counting data

Microarray expression data

Data are color intensities



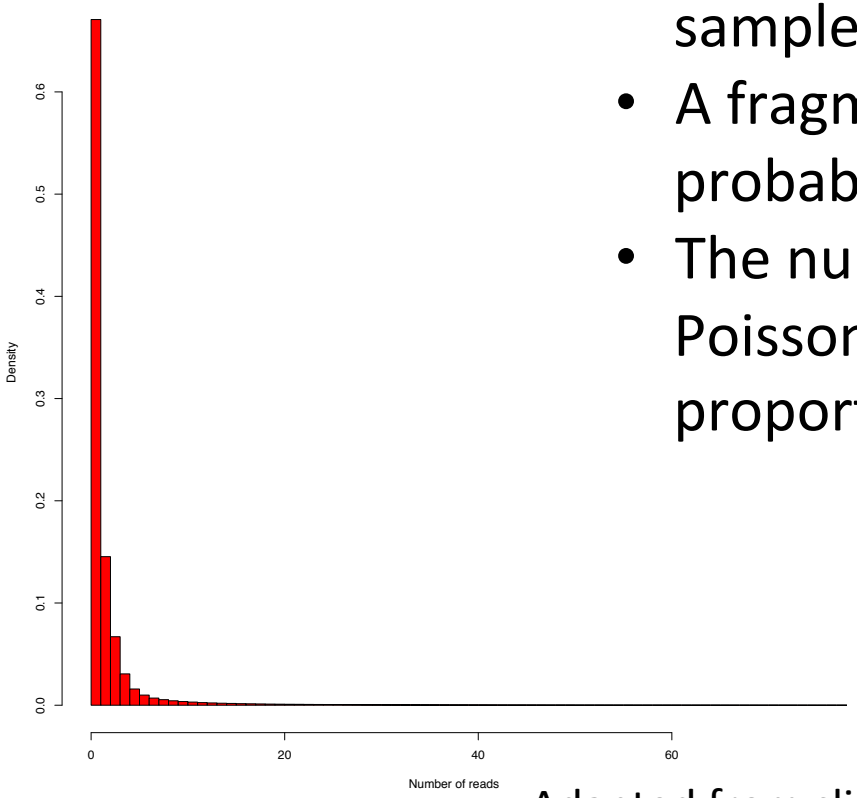


$$\log y_{ij} \sim N(\mu_j, \sigma_j^2)$$

Adapted from slides by Benilton Carvalho

Sequencing data

Gene	Sample 1	Sample 2
ERBB2	0	45
MYC	14	23
ESR1	56	2

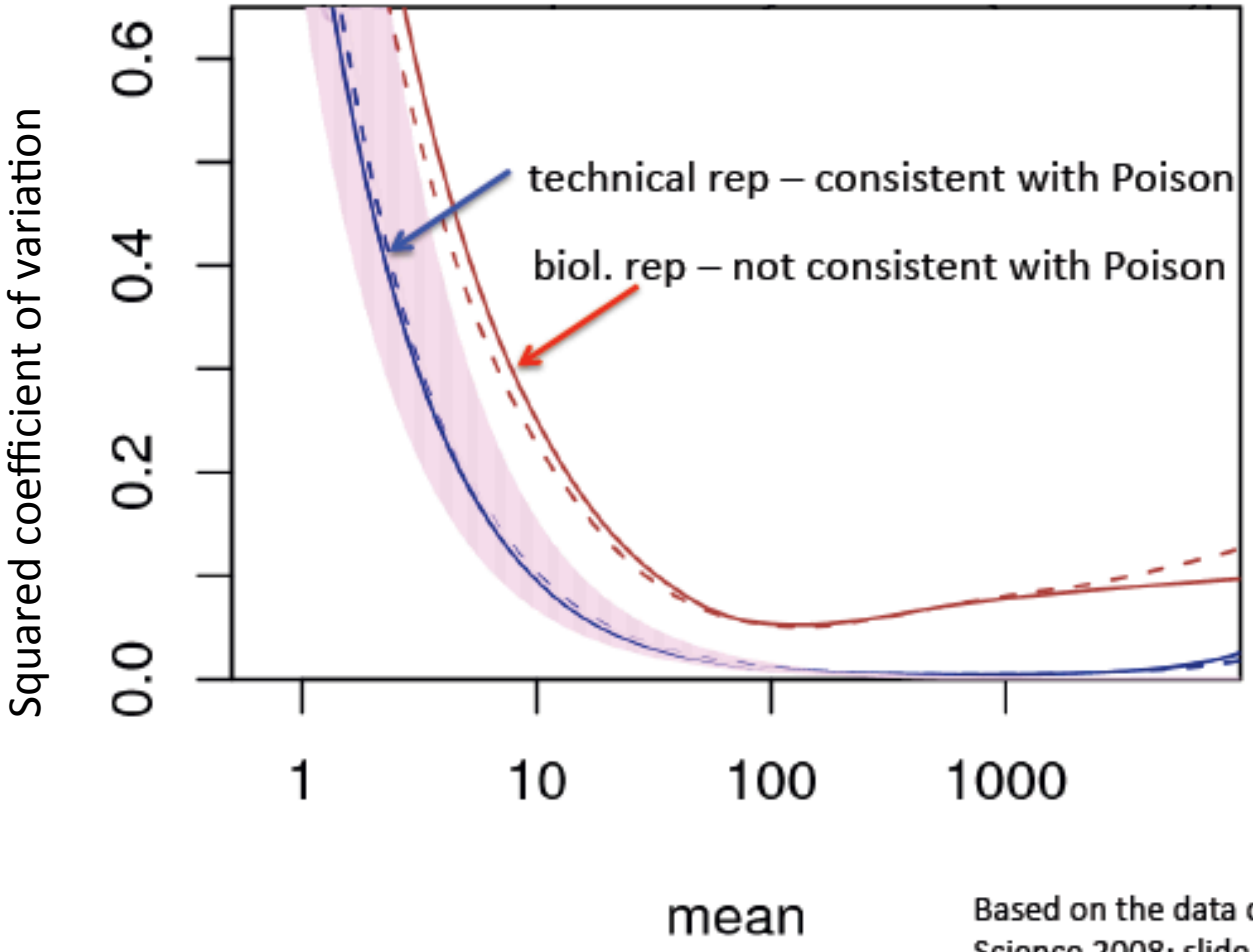


- Transcript (or sequence, or methylation) *i* in sample *j* is generated at a rate λ_{ij}
- A fragment attaches to the flow cell with a probability of p_{ij} (small)
- The number of observed tags y_{ij} follows a Poisson distribution with a rate that is proportional to $\lambda_{ij}p_{ij.}$

The variance in a Poisson distribution is equal to the mean

Adapted from slides by Benilton Carvalho, Tom Hardcastle

Extra variability



Based on the data of Nagalakshmi et al. Science 2008; slide adapted from Huber;

Adapted from slides by Benilton Carvalho

Negative binomial model for sequencing data

- For subject j, on transcript i:

$$Y_{ij}|\lambda_{ij} \sim P(\lambda_{ij})$$

- Different subjects have different rates, which we can model through:

$$\lambda_{ij} \sim \Gamma(\alpha, \beta)$$

- This hierarchy changes the distribution of Y:

$$Y_{ij} \sim \mathrm{NB}\left(\alpha, \frac{1}{1+\beta}\right)$$

$$\begin{aligned} N_{ij}|\eta_{ij} &\sim \mathrm{Poisson}\,(\eta_{ij}) \\ \eta_{ij}|\mu_{ij} &\sim \mathrm{Gamma}\,(\beta_1(\mu_{ij}), \beta_2(\mu_{ij})) \\ N_{ij} &\sim \mathrm{NB}\,(\mu_{ij}, \alpha(\mu_{ij})) \\ \log \mu_{ij} &= s_j + \sum_k \beta_{ik} x_{kj} \end{aligned}$$

smooth dispersion-mean relation α

Adapted from slides by Benilton Carvalho

Estimating Overdispersion with edgeR

- edgeR (Robinson, McCarthy, Chen and Smyth)
- Total CV^2=Technical CV^2 + Biological CV^2

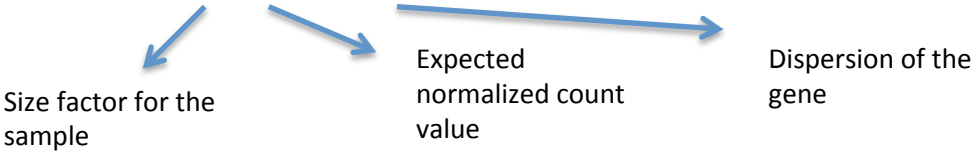
Decreases with sequencing depth

Variability in gene abundance between replicates

- Borrows information from all genes to estimate BCV.
  - Common dispersion for all tags
  - Empirical Bayes to shrink each dispersion to the common dispersion.

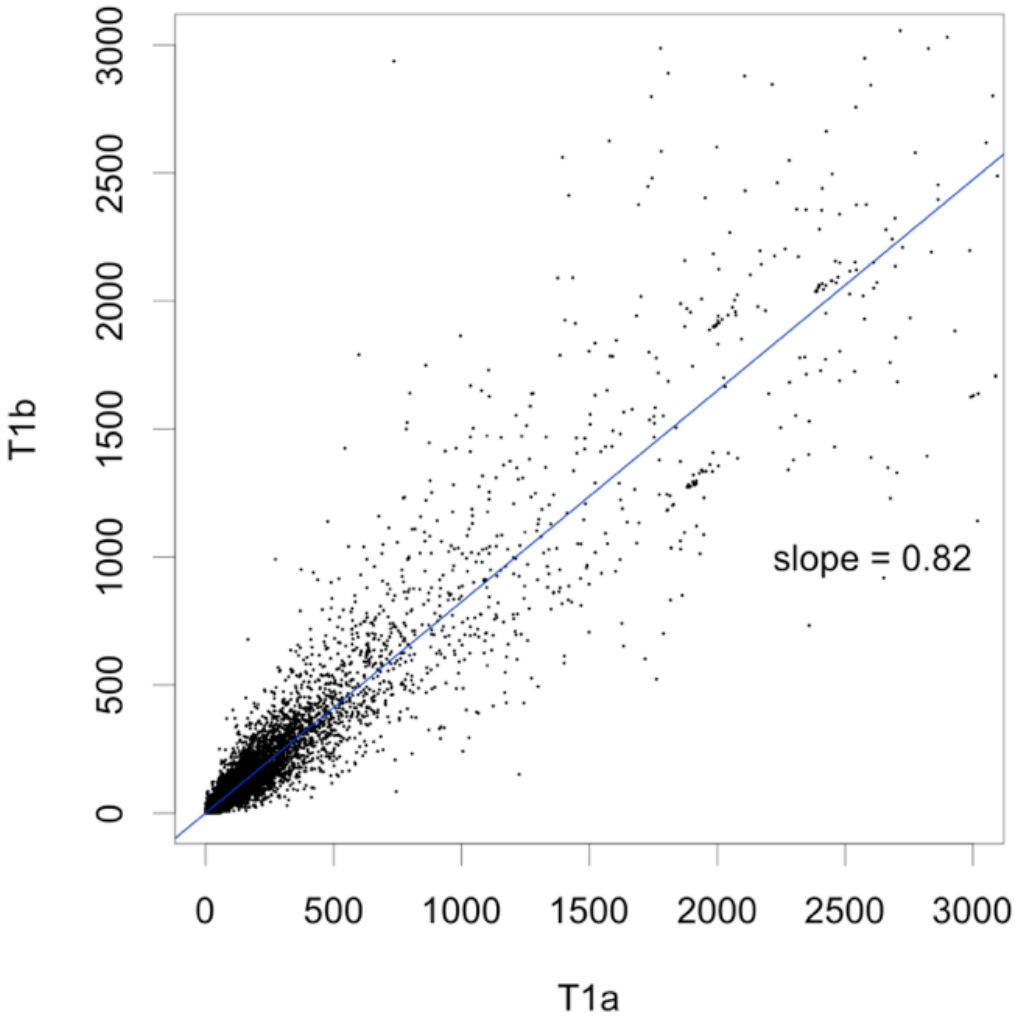
Estimating Overdispersion with DESeq

- DESeq (Anders, Huber)
- $Var = s\mu + \alpha s^2\mu^2$



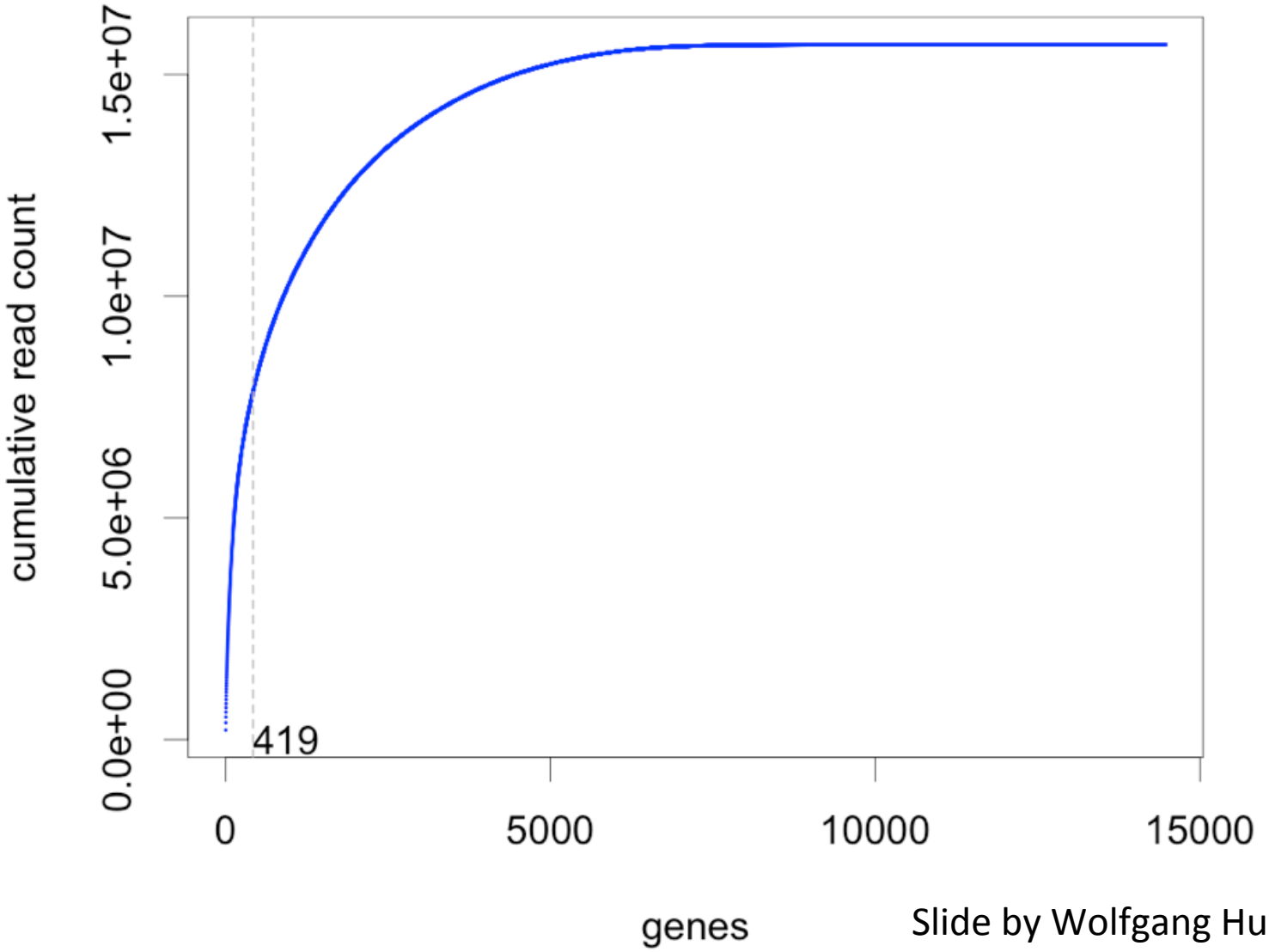
- estimateDispersions()
  1. Dispersion value for each gene
  2. Fits a curve through the estimates
  3. Each gene gets an estimate between (1) and (2).

Reproducibility



Slide by Wolfgang Huber

A few number of genes get most of the reads



Slide by Wolfgang Huber

Effective library sizes

- Also called normalization (although the counts are not changed!!!)
- We must estimate the effective library size of each sample, so our counts are comparable between genes and samples
- Gene lengths?
- This library sizes are included in the model as an ***offset*** (a parameter with a fixed value)

$$\log \mu_{ij} = s_j + \sum_k \beta_{ik} x_{kj}$$

Estimating library size with edgeR

- edgeR (Robinson, McCarthy, Chen and Smyth)
- Adjust for sequencing depth and RNA composition (total RNA output)
- Choose a set of genes with the same RNA composition between samples (with the log fold change of normalised counts) after trimming
- Use the total reads of that set as the estimate.

Estimating library size with DESeq

- DESeq (Anders, Huber)
- Adjust for sequencing depth and RNA composition (total RNA output)
- Compute the ratio between the log counts in each gene and each sample and the log mean for that gene on all samples.
- The median on all genes is the estimated library size.

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