# We are here

1. Basic Statistics and Differential Expression
   - Hypothesis testing
   - Differential expression
   - Practical Session

# Basic Statistics and Differential Expression

# Overview

- Hypotheses
- Test statistics
- Error types
- p-values
- Two-sided versus one-sided
- t-test and other tests
- Multiple tests and *p*-value correction
- Differential expression in RNA-seq data

# Introduction

Question: Which genes have different counts?

Characteristics of RNA-seq data: Small replicate numbers, discreteness, large dynamic range, outliers $\Rightarrow$ a suitable statistical approach is required

Statistical methods try to answer the question:

Which genes are **significantly** differentially expressed?

# Basic terminology

- Population: ensemble of entities
- Sample: subset of population (choice of sample can bias result)
- Random or stochastic variable: variable whose values are affected by chance (result of random experiment)
  If these values are in $\mathbb{R}$ we speak of a *continuous* variable, if they are in $\mathbb{Z}$ we speak of a *discrete* variable.
- Statistic: numerical descriptive measure of a sample

# Example

Population: humans and their heights

Sample: randomly select a small number of humans, measure their height

Random variable $X$: the height of an individual

Statistic: sample mean, used as an estimate for the population's mean height

# Distributions

The probability distribution shows the distribution of the probability of the values of a random variable.

Well-known probability distribution are: Normal or Gaussian distribution, Poisson distribution, Binomial distribution ...

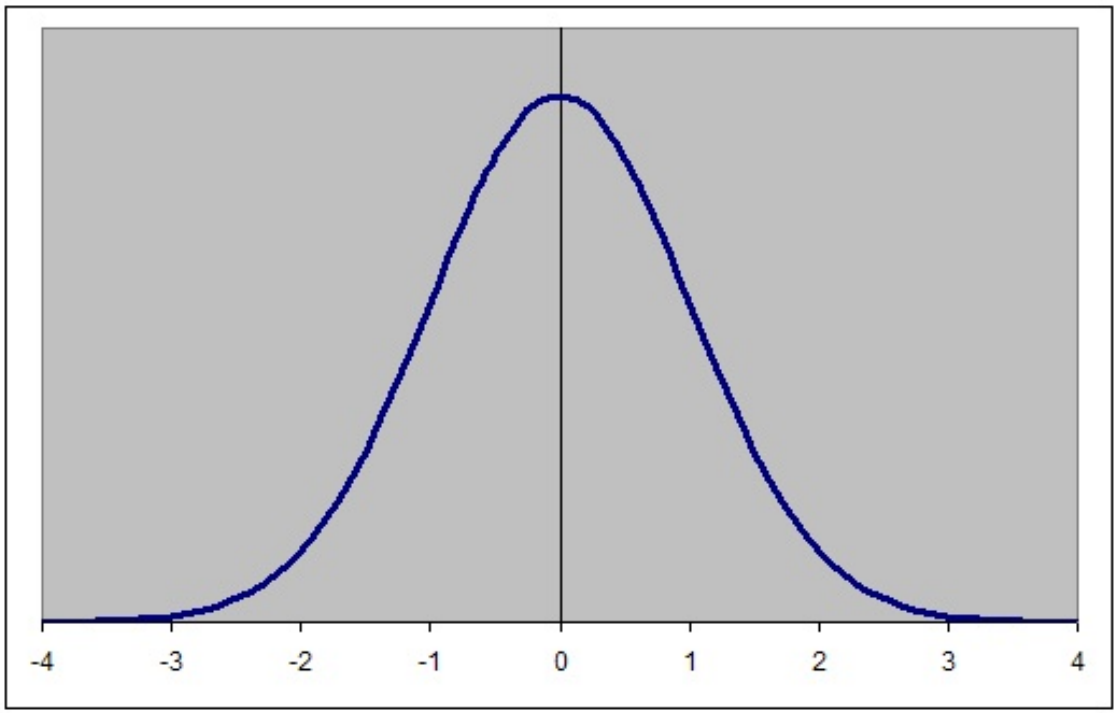
# The normal distribution

*"Everyone believes in the normal law, the experimenters because they imagine it is a mathematical theorem, and the mathematicians because they think it is an experimental fact." (Gabriel Lippman, in Poincaré's Calcul de probabilités, 1896).*

# The normal distribution

The density function of the normal distribution is given by

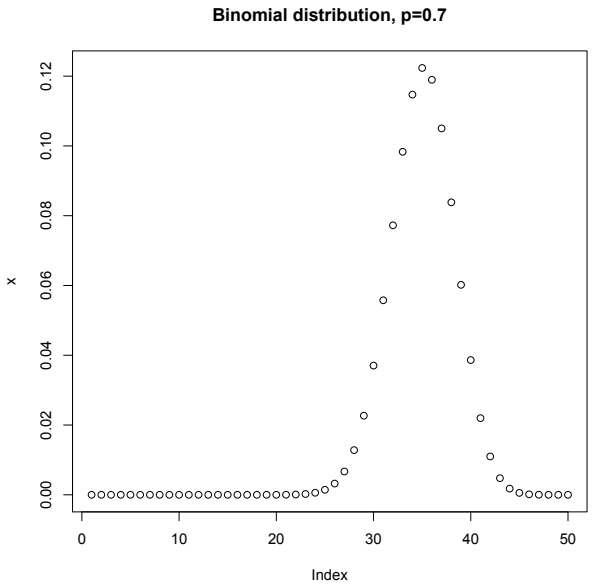
$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \cdot e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



# Distributions

A well-known discrete probability distribution is the Binomial distribution with parameters $n$ (number of trials) and $p$ (prob. of success):

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k},\ 0 \leq k \leq n$$



# Distributions

Distributions are often assessed by summary statistics.

Parameters of summary statistics of a distribution are for example:
mean, median, variance, quantile, quartile.

Gaussian distribution:

Mean: $\mu$

Variance: $\sigma^2$

Binomial distribution:

Mean: $np$

Variance: $np(1-p)$

# Estimator of location

Given a set of data, estimate the true value (location) of the measured quantity.

**Definition**

Estimator of locations are measures of central tendency of a sample.

- Mean:

$$\bar{x} = \frac{1}{n}(\sum_{i=1}^{n} x_i)$$

  Problem: not robust against outliers

- Median: arrange all the observations from lowest value to highest value

$$\text{med} = \begin{cases} x_{\frac{n+1}{2}}; & n \text{ uneven} \\ \frac{1}{2}(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}); & n \text{ even} \end{cases}$$

# Estimator of location

- Quantile: the quantile function returns the value below which random draws from the given distribution would fall, $p \times 100$ % of the time

$$F(x) = Pr(X \leq x) = p$$

P $= 0.3 \Rightarrow Q_{0.3}$ is the value for which 30% of the values of the distribution are below this value.

# Estimator of location

- Quantile: the quantile function returns the value below which random draws from the given distribution would fall, $p \times 100$ % of the time

$$F(x) = Pr(X \leq x) = p$$

P $= 0.3 \Rightarrow Q_{0.3}$ is the value for which 30% of the values of the distribution are below this value.

- Quartile:
first quartile = 0.25-quantile $\rightarrow$ 25% of the data lie below this value
second quartile = median
third quartile = 0.75-quantile
The range between the first and the third quartile is also referred to as the interquartile range (IQB).

# Estimator of deviation

The estimator of deviation describe how much the value of the random variable vary (around the value of central tendency). The most important one are

- Variance:

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- Median absolute deviation (MAD):

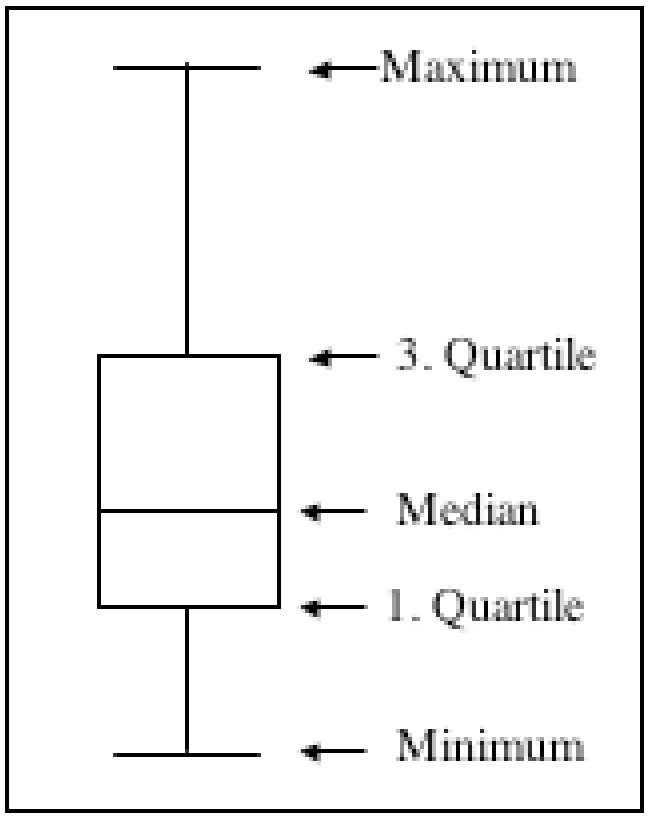
$$\text{MAD} = \text{med}\{|x_i - \text{med}\{x_i\}|)\}$$

# Boxplot

The boxplot visualizes a one-dimensional distribution. It is based on 5 numbers of a distribution:

*minimum, first quartile, median, third quartile and maximum*

Graphically the boxplot looks as follows:



# Hypothesis testing

Generally all hypothesis tests involve comparison of an observation and a value that one would expect by chance for a given test statistic. The following two hypotheses are stated:

Null hypothesis $H_0$: observed value does not differ significantly from the one expected by chance

Alternative hypothesis $H_1$: observed value differs significantly from the one expected by chance

# Example

Which genes are differentially expressed between two different conditions:

*Null hypothesis*: mean of a gene under condition $A$ is not different from the mean of the gene under condition $B$, ie., the gene is not differentially expressed

*Alternative hypothesis*: the gene is differentially expressed

# Statistical hypothesis testing

Statistical hypothesis testing is used to make a decision about whether the data contradicts the null hypothesis: this is called significance testing.

A null hypothesis can either be rejected or not be rejected. It cannot be accepted.

# Error types

All statistical tests can bear errors. We distinguish two error types:

- the *type I error* $\alpha$: the null hypothesis has been falsely rejected - *false alarme*
- the *type II error* $\beta$: the null hypothesis has been falsely not rejected - *missed opprtunity*

# Error types

All statistical tests can bear errors. We distinguish two error types:

- the *type I error* $\alpha$: the null hypothesis has been falsely rejected - *false alarme*
- the *type II error* $\beta$: the null hypothesis has been falsely not rejected - *missed opprtunity*

| | Do not reject $H_0$ | Reject $H_0$ |
|---|---|---|
| $H_0$ is true | true | error type I $\alpha$ |
| $H_0$ is false | error type II $\beta$ | true |

For the error type I one uses a low confidence value $\alpha \leq 1\%$ or $5\%$.

# *p*-value

The *p-value* is the probability that the test statistic $T$ achieved by chance is at least as large as the value $t$ of the observed sample,

$$P(|T| \geq t)|H_0)$$

Thus, the *p*-value is the probability of rejecting the null hypothesis erroneously.

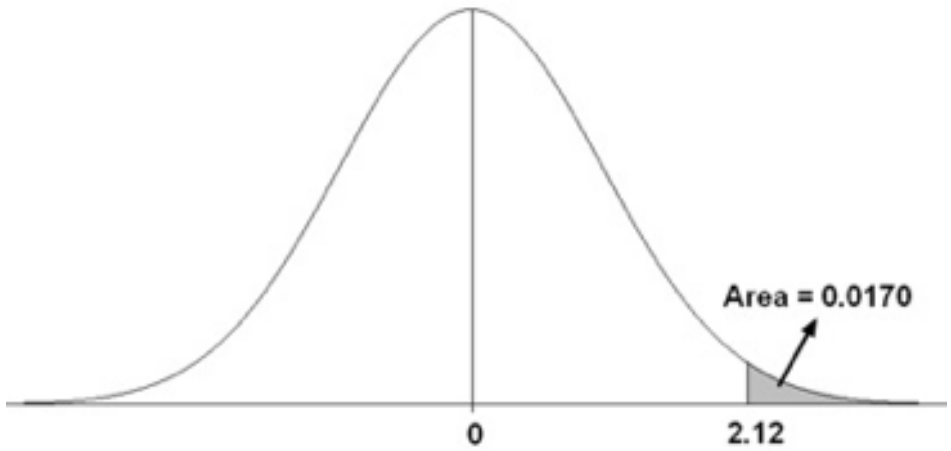
Relationship with type I error $\alpha$: Is the *p*-value smaller than the chosen type I error $\alpha$, the null hypothesis is rejected.

# $p$-value - Example

Assume data is standard normally distributed $\sim \mathcal{N}(0,1)$.
Assume a test statistic computes

$$T = 2.12$$

Then the p-value is




$$p = Prob(T \geq 2.12 | \mathcal{N}(0,1)) = 1 - p(T < 2.12) = 0.017$$

# One-sided and two-sided tests

Example: when testing for differential expression

- two-sided: Does gene from group $A$ have different mean than from group $B$?
- one-sided: Does gene from group $A$ have a larger/smaller mean than from group $B$?

Thus in the case of a two-sided test we are testing whether a gene is differentially regulated, while when applying a one-sided test we are interested whether the gene is up-regulated (larger mean) or down-regulated (smaller mean).

# $t$-test

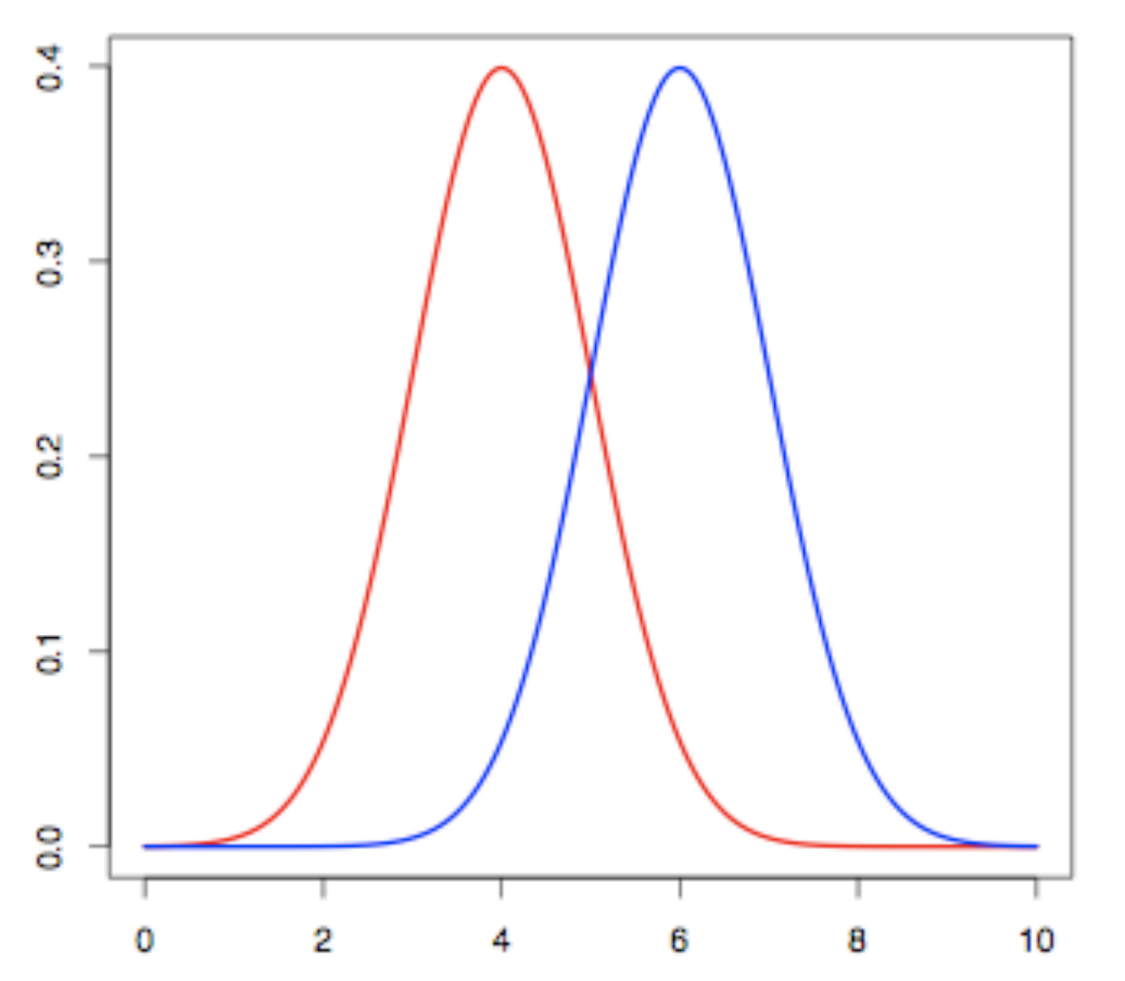
A commonly chosen test to test for equality of means is the $t$-test.

This method (among others) takes the variation of the expression value into account for the computation.

The $t$-test is an example for classical hypothesis testing.

# Testing the (in)equality of means

When we test whether two distributions have equal means, we actually have the goal to compare how separate two distributions are:



# The two-sample test

Given:
expression values $x_1, \ldots, x_{m_1}$ of replicates of group $A$,
expression values $y_1, \ldots, y_{m_2}$ of replicates of group $B$.

Assumptions:

gene expression values in both classes are normally distributed: $x_1, \ldots, x_{m_1} \sim \mathcal{N}(\mu_1, \sigma^2)$ and $y_1, \ldots, y_{m_2} \sim \mathcal{N}(\mu_2, \sigma^2)$ with the same variance $\sigma^2$.

Compute the "pooled" variance estimate $s^2$ as estimate for variance $\sigma^2$ from the data as follows:

$$s^2 = \frac{1}{m_1 + m_2 - 2} \left( \sum_{i=1}^{m_1} (x_i - \overline{x})^2 + \sum_{j=1}^{m_2} (y_j - \overline{y})^2 \right)$$

# The two-sample test

## Definition

The two-sample $t$-statistic is defined by

$$t = \frac{\overline{x} - \overline{y}}{\sqrt{s^2(\frac{1}{m_1} + \frac{1}{m_2})}}$$

Under the null hypothesis the two-sample $t$-statistic follows the $t$-distribution with $(m_1 + m_2 - 2)$ degrees of freedom.

The $t$-test for a gene $g$ is then conducted as follows:

1. Compute $t(g)$
2. Compute a one- or two-sided $t$-statistic for each gene $g$
3. Compute the $p$-value

# The two-sample test

Example:

sample A: $g = (0.45, 0.57, 1.02, 0.97)$
sample B: $g = (1.50, 2.07, 0.51, 1.63)$

Compute means:

$\overline{g(A)} = 0.7525$, $\overline{g(B)} = 1.4275$

# The two-sample test

Example:

sample A: $g = (0.45, 0.57, 1.02, 0.97)$
sample B: $g = (1.50, 2.07, 0.51, 1.63)$

Compute means:

$\overline{g(A)} = 0.7525$, $\overline{g(B)} = 1.4275$

Pooled variance:

$$s^2 = \frac{1}{4+4-2}(0.243675 + 1.300875) = 0.25726$$

# The two-sample test

Example:

sample A: $g = (0.45, 0.57, 1.02, 0.97)$
sample B: $g = (1.50, 2.07, 0.51, 1.63)$

Compute means:

$\overline{g(A)} = 0.7525$, $\overline{g(B)} = 1.4275$

Pooled variance:

$$s^2 = \frac{1}{4+4-2}(0.243675 + 1.300875) = 0.25726$$

t-statistic:

$$t = \frac{|(0.7525 - 1.4275)|}{\sqrt{(0.25726/4 + 0.25726/4)}} = 1.88$$

# The two-sample test

Example:

sample A: $g = (0.45, 0.57, 1.02, 0.97)$
sample B: $g = (1.50, 2.07, 0.51, 1.63)$

Compute means:

$\overline{g(A)} = 0.7525$, $\overline{g(B)} = 1.4275$

Pooled variance:

$$s^2 = \frac{1}{4+4-2}(0.243675 + 1.300875) = 0.25726$$

t-statistic:

$$t = \frac{|(0.7525 - 1.4275)|}{\sqrt{(0.25726/4 + 0.25726/4)}} = 1.88$$

p-value $= 0.1089$.

If we had chosen $\alpha = 0.05$, would we call the gene differentially expressed?

# Non-parametric tests

Further tests

- Make no assumptions about underlying distribution
- Better called distribution free tests
- Examples for testing equality of means:
  - Mann-Whitney test
  - Permutation tests
  - Rank product test: very nice test for microarray data that takes fold changes of all genes into account when testing individual genes for differential expression.

# Multiple testing

- Problem when analyzing a large number of genes: Error $\alpha$ holds only for a single gene
- Given any test procedure, the adjusted p-value corresponding to the test of a single hypothesis $H_j$ can be defined as the level of the entire test procedure at which $H_j$ would just be rejected, given the values of all test statistics involved.

# Multiple testing

Example: let $n = 10,000$ genes be on a chip. Assume not a single gene is differentially regulated. For an individual $\alpha$-value$< 0.01$ one expects to see

$$10,000 \times 0.01 = 100$$

differentially regulated genes just by chance (false positives).

# Correction methods

Type 1 error at experiment level is denoted as $\alpha_E$.

- Bonferroni: $\alpha^* = \alpha_E/n$, very conservative, makes the least false positive calls
- False-discovery date (FDR): less conservative
  - Order all single $p$-values, such that $p_1 \leq p_2 \leq \ldots p_n$
  - For chosen level $\alpha_E$ find largest $k$ such that

$$p_k \leq \frac{k}{n}\alpha_E$$

  - Reject all hypotheses for $p_1, \ldots, p_k$

# Differential expression

As in microarrays, differential expression between two samples $A$, $B$ can be computed as

$$\frac{e_A(G)}{e_B(G)} \qquad \text{for each gene } G$$

This is the fold change.

To simplify interpretation, the logarithm is usually employed:

$$M(G) = \log_2 \left( \frac{e_A(G)}{e_B(G)} \right) = \log_2 e_A(G) - \log_2 e_B(G)$$

Here, $e_X(G)$ is one of the expression measures discussed before.

# Differential expression: Special considerations

RNA-Seq gives a digital measure of expression,
i.e. the number of reads is counted for each gene.
$\Rightarrow$ nonexpressed genes have **zero** reads.

This poses a special problem:

$$\log_2 0 = -\infty$$

What is the fold-change (FC) of a gene with zero reads in one sample?

# Differential expression: Special considerations (II)

What is the fold-change (FC) of a gene with zero reads in one sample?

- The gene is switched on: FC$=+\infty$.
- The gene is switched off: FC$=-\infty$.

$\rightarrow$ A maximum & minimum fold-change can be used.

$\rightarrow$ Pseudo-counts can be used:
Each gene has at least $c$ reads.
Hard to set a value for $c$.

The important question is: **Are such genes interesting?**

| YES | $\leftrightarrow$ | NO |
|---|---|---|
| Major transcriptional event<br>Expression changed *fundamentally*. | | Change can be due to noise:<br>E.g.: Zero vs one read. |

# Differential expression: Statistics

Statistical methods try to answer the question:

> Which genes are **significantly** differentially expressed?

Usual tests (e.g. $t$-test) need several replicates for each condition.

Unfortunately, most RNA-seq experiments so far are performed with **few or no** replicates!

# Differential expression: Statistics (II)

Anders & Huber (2010) proposed the following method, called DESeq:

Idea: Count reads mapping to a gene in two different samples (with or without replicates)

Problem: Which count differences are significant?

Test for differential gene expression between two conditions by

- The number of reads for gene $i$ in sample $j$ is modeled by a negative binomial distribution:

$$K_{ij} \sim NB(\mu_{ij}, \sigma_{ij}^2)$$

- Variance is considered to be a function of the mean which is estimated by a local regression on all genes in all replicates
- P-values are obtained by using a standard conditional test

# Differential expression: Statistics (II)

## Negative binomial distribution

a discrete probability distribution of the number of successes in a sequence of Bernoulli trials before a specified (non-random) number $r$ of failures occurs.

Example:

if one throws a die repeatedly until the third time "1" appears, then the probability distribution of the number of non-"1"s that had appeared will be negative binomial.

# Practical Session

**Questions?**

# Practical Session

**Questions?**

Now off to the fifth practical session

# Learning Objectives of Practical Session

- See how to use DESeq using R / Bioconductor
- Use Mayday to apply t-test, and also other methods if time