University of Essex

RNA-Seq II

Differential Gene Expression

Antonio Marco

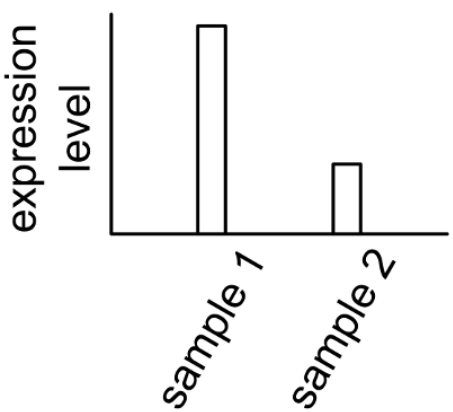
School of Biological Sciences
University of Essex

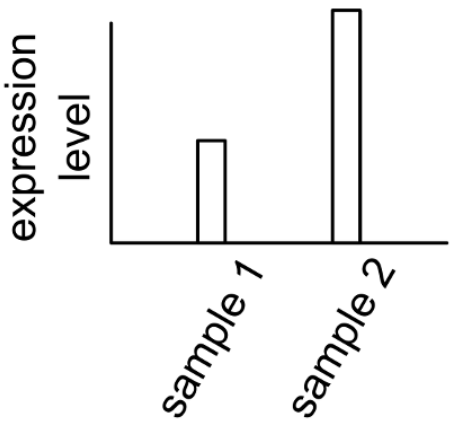
6-Jun-17

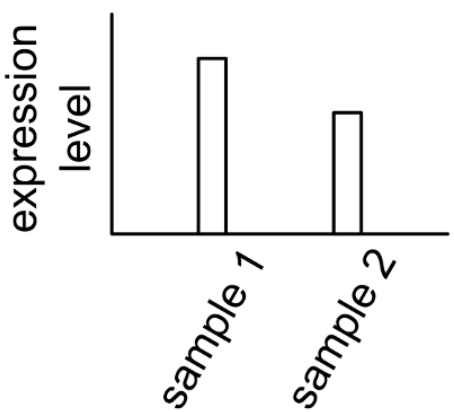
Outline

Comparison of two samples

Which of these pairs are significantly different?

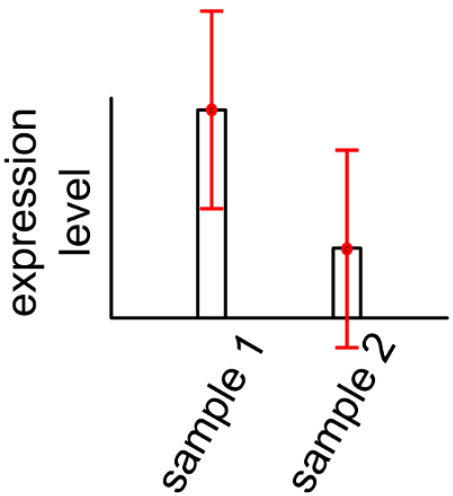


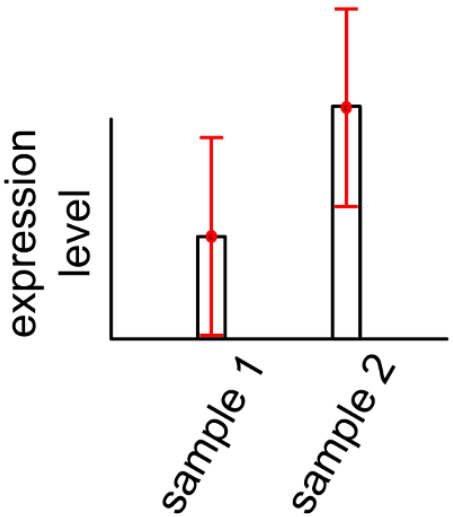


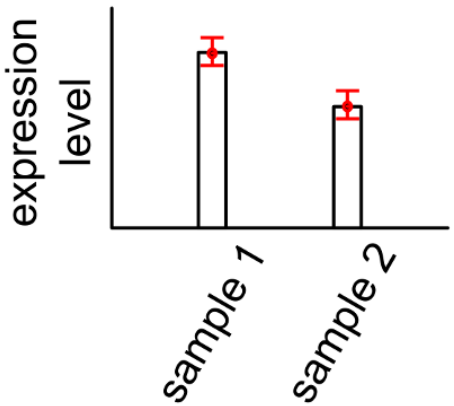


Comparison of two samples

Which of these pairs are significantly different?







- Variability of a set of values is measured as the VARIANCE

$$V = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 \tag{1}$$

- If we know the variance we can estimate the Standard Error of the Mean (SEM)

Biological and technical replicates

To calculate Variances we need multiple measures of each gene:

- Biological replicates: variability among individuals
- Technical replicates: variability due to the method

Assessing differences

- The t-test is also used in differential gene expression analysis

Assessing differences

- The t-test is also used in differential gene expression analysis
- Usually paired t-test, or a non-parametric test (Wilcoxon)

Assessing differences

- The t-test is also used in differential gene expression analysis
- Usually paired t-test, or a non-parametric test (Wilcoxon)
- Use ANOVA for more than 2 groups

Assessing differences

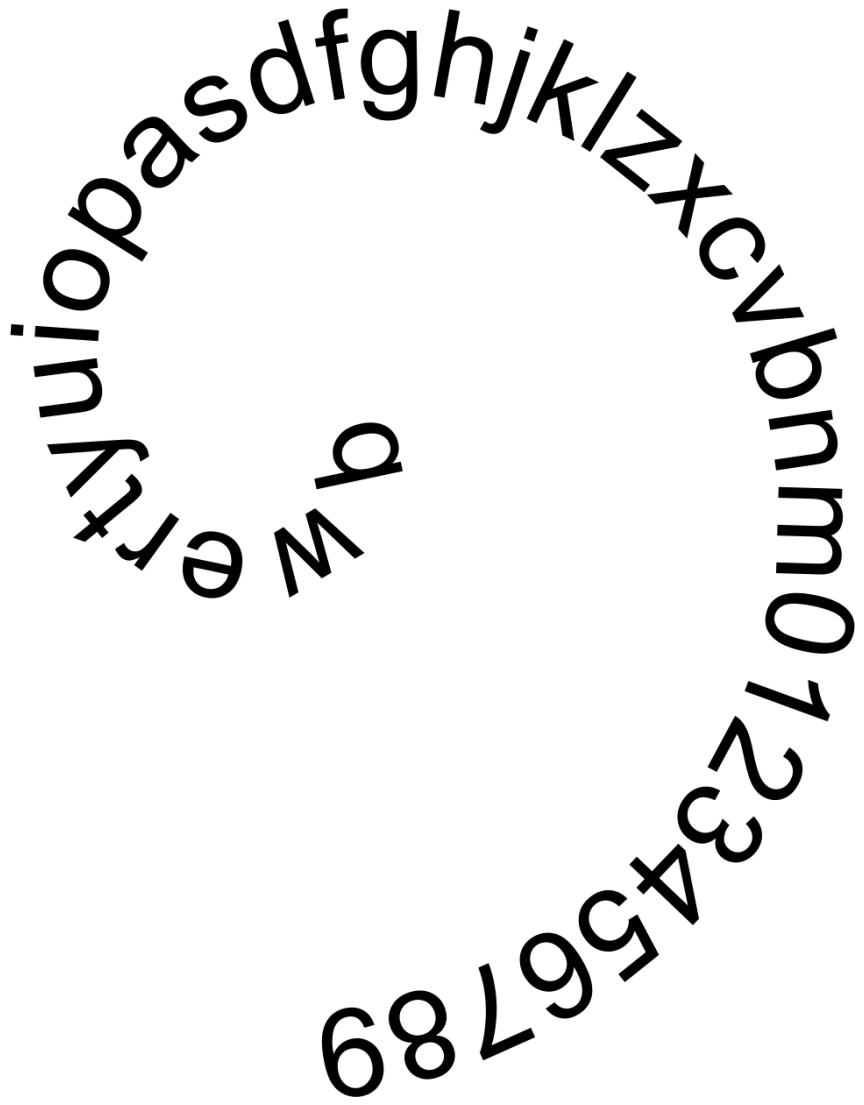
- The t-test is also used in differential gene expression analysis
- Usually paired t-test, or a non-parametric test (Wilcoxon)
- Use ANOVA for more than 2 groups
- State-of-the-art methods use probabilistic models (e.g. Negative Binomial)

Assessing differences

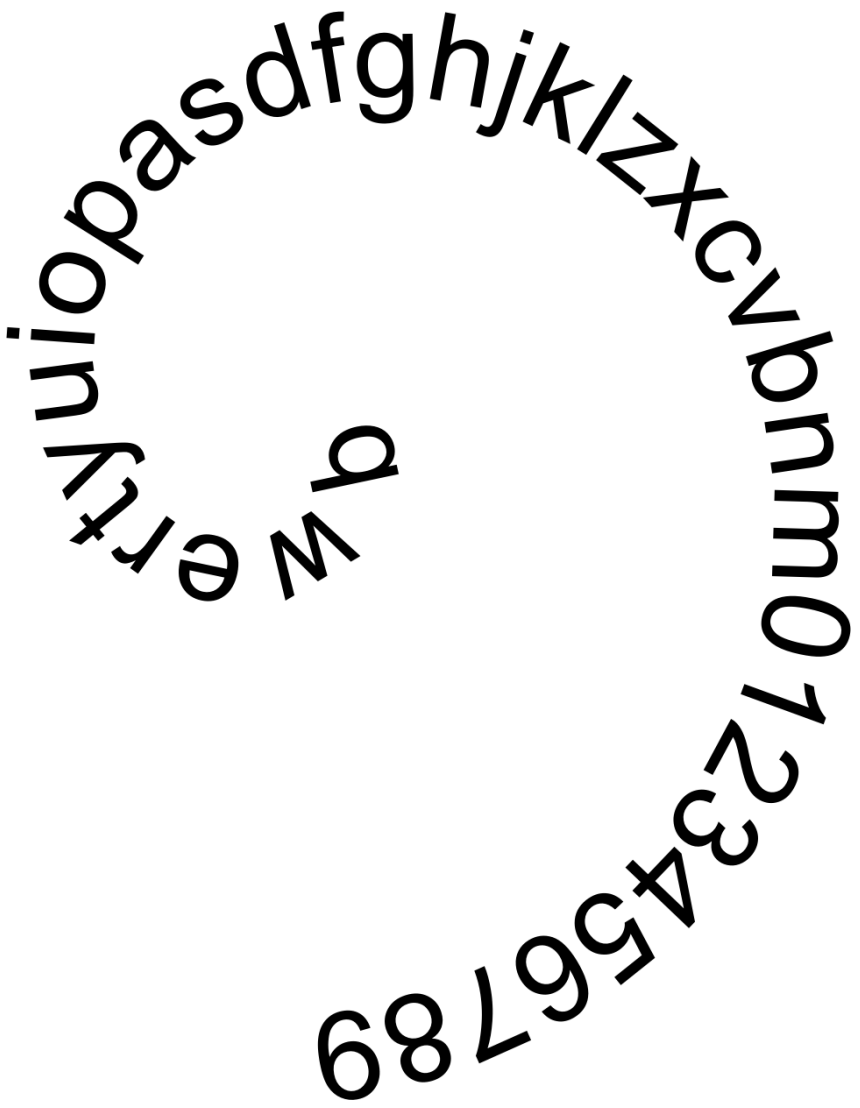
- The t-test is also used in differential gene expression analysis
- Usually paired t-test, or a non-parametric test (Wilcoxon)
- Use ANOVA for more than 2 groups
- State-of-the-art methods use probabilistic models (e.g. Negative Binomial)
- For each gene we will have a p-value: the smaller it is, the more significant the difference

Outline

False positives

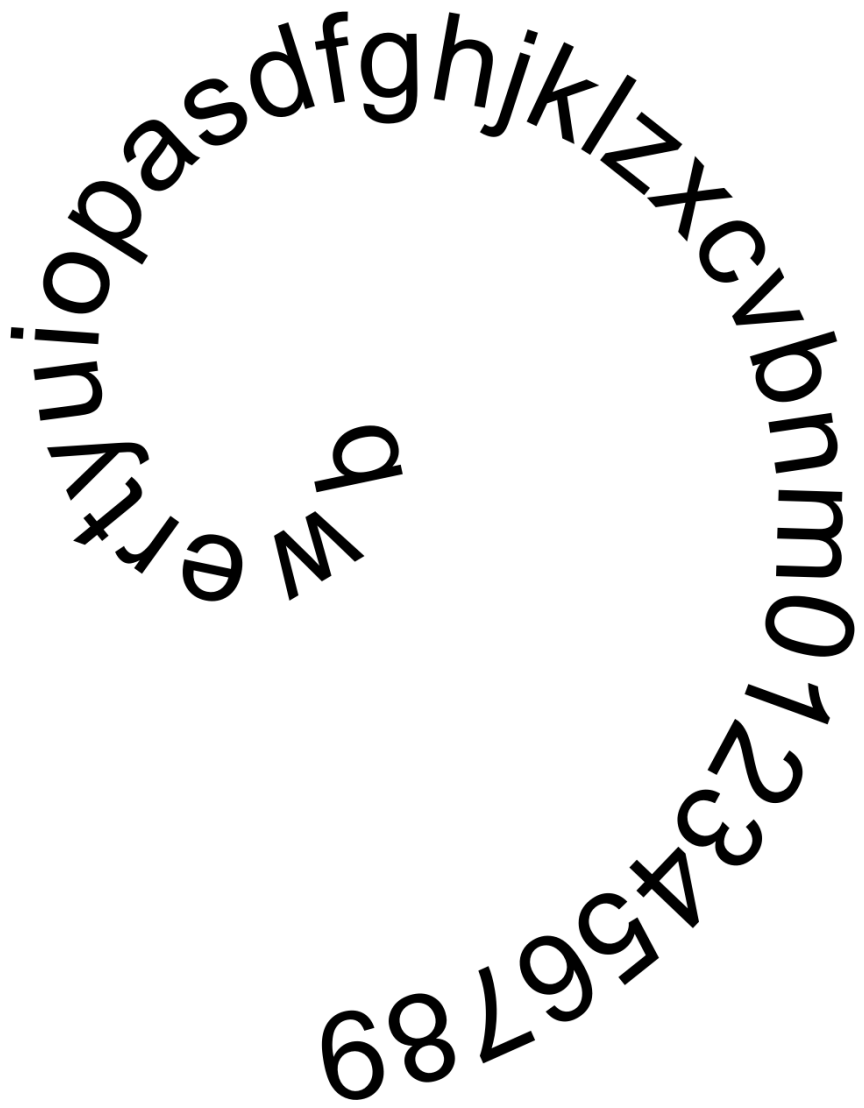


False positives



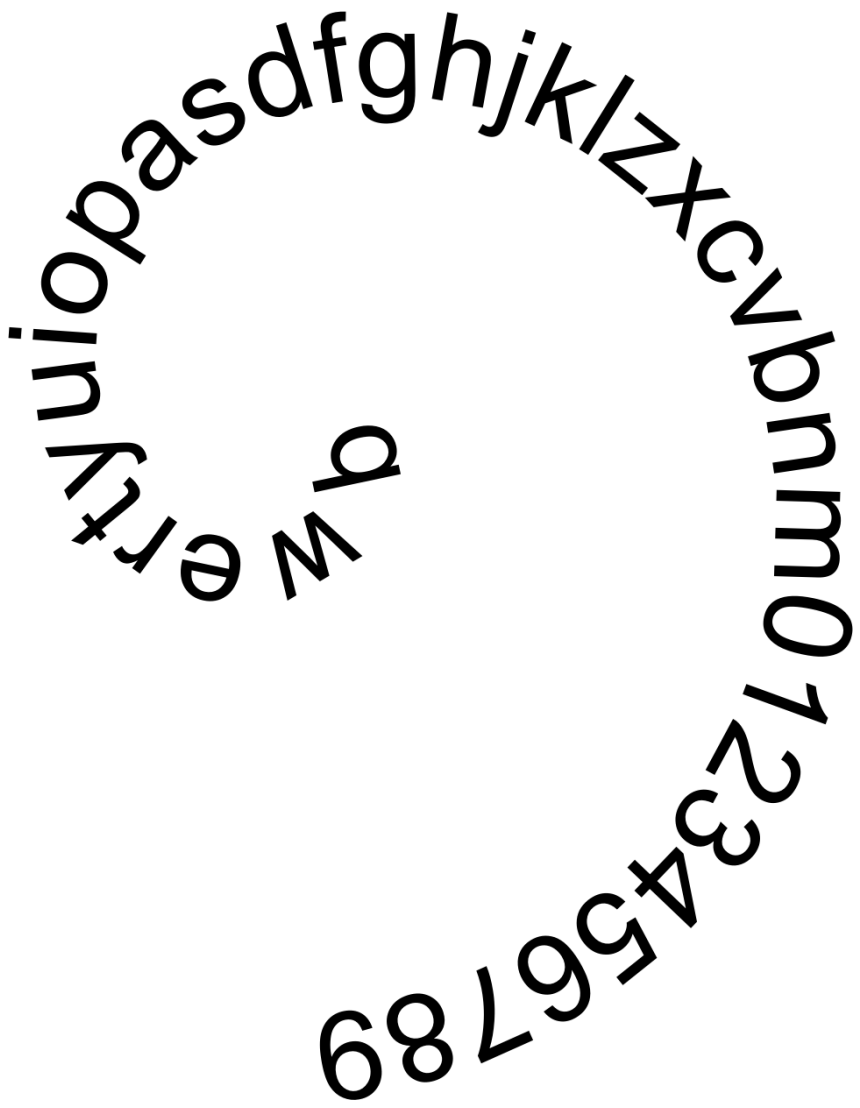
- A p-value is the probability of accepting a hypothesis that is actually false

False positives



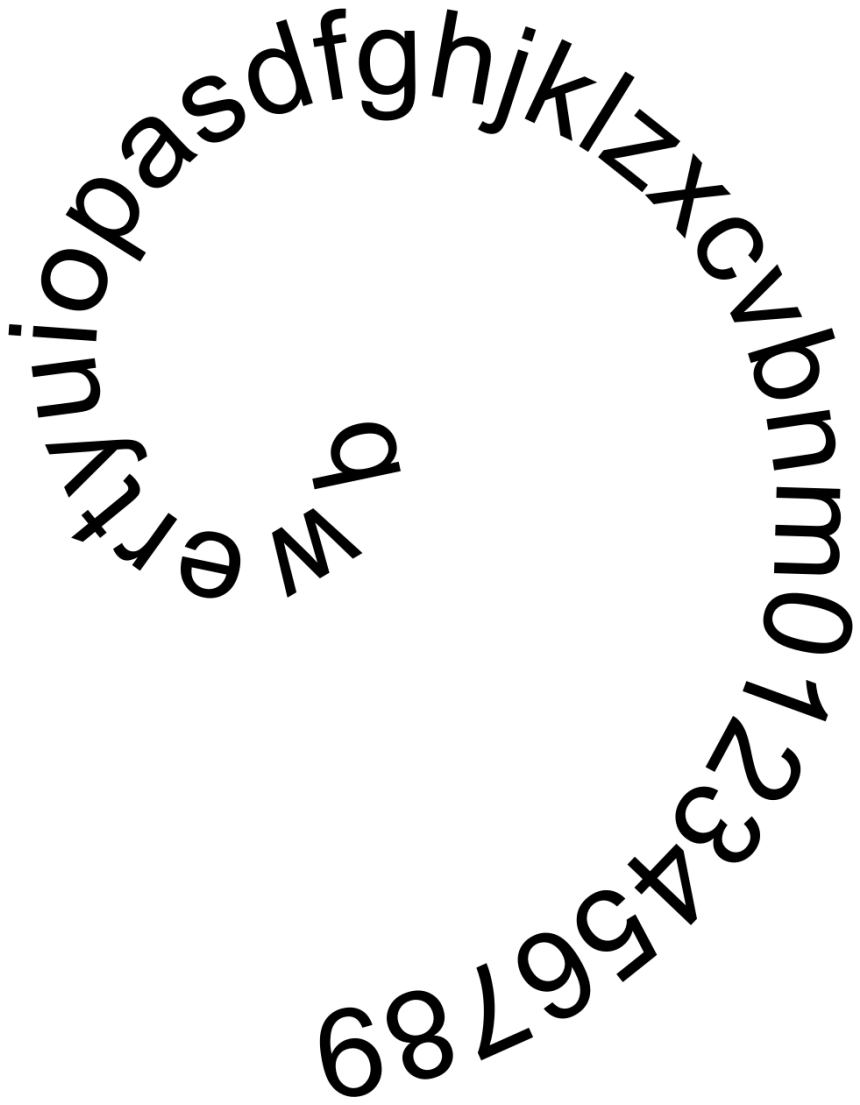
- A p-value is the probability of accepting a hypothesis that is actually false
- In other words, is the probability of a reported difference to be a false positive

False positives



- A p-value is the probability of accepting a hypothesis that is actually false
- In other words, is the probability of a reported difference to be a false positive
- If we perform multiple comparison, a p-value of 0.05 may not be good enough

False positives



- A p-value is the probability of accepting a hypothesis that is actually false
- In other words, is the probability of a reported difference to be a false positive
- If we perform multiple comparison, a p-value of 0.05 may not be good enough
- We have to account for multiple comparisons

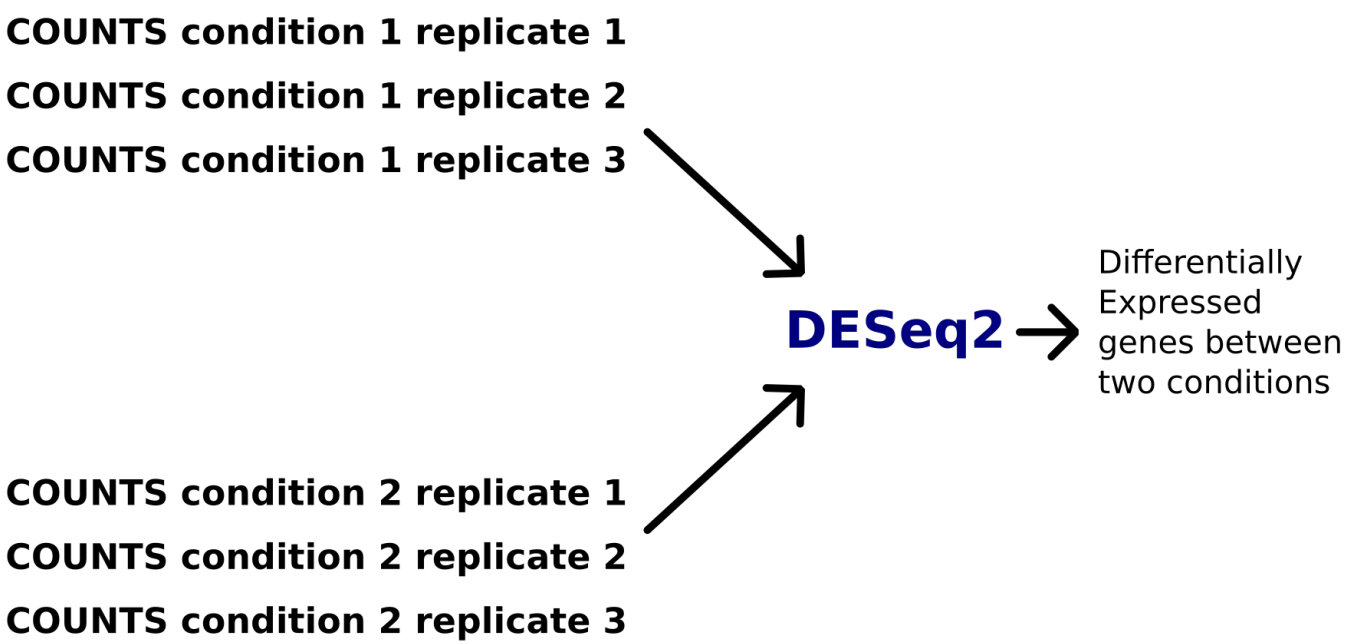
FDR and q-values

- False Discovery Rate (FDR) is a procedure by which we control the maximum number of false positives
- The 'math' bit can be found in Wikipedia

FDR and q-values

- False Discovery Rate (FDR) is a procedure by which we control the maximum number of false positives
- The 'math' bit can be found in Wikipedia
- A modification of the FDR is the q-value, which is the equivalent to a p-value but for a false discovery rate
- The important thing is that you understand that p-values need to be corrected!

Differential Gene Expression Workflow



Output

- Gene
- log Fold Change (IFC)
- q-value