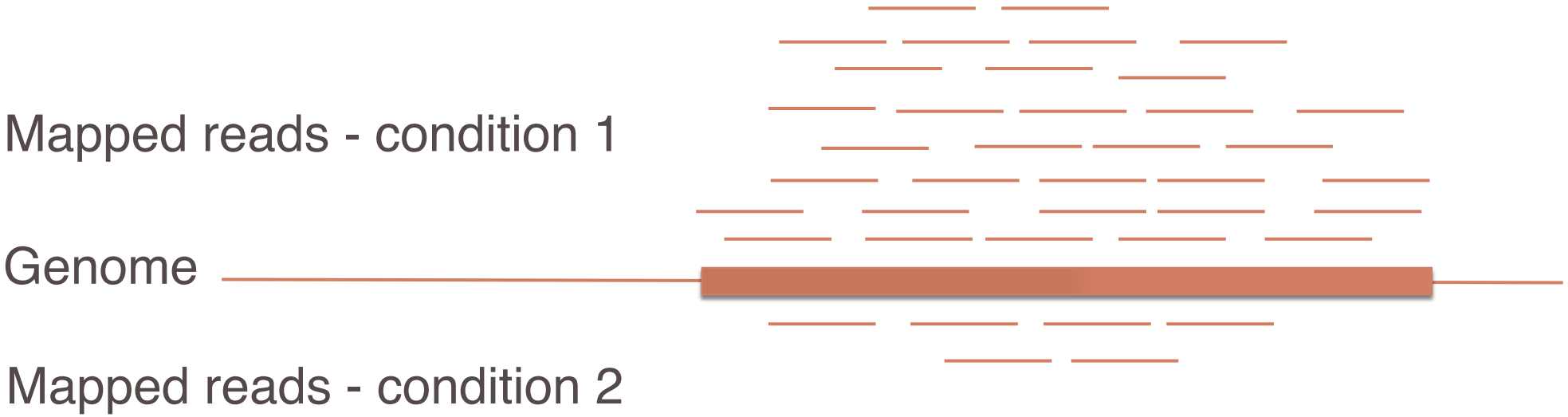
RNA-SEQ DATA ANALYSIS PIPELINE: DIFFERENTIAL EXPRESSION

DIFFERENTIAL EXPRESSION



- The observed difference is statistically significant?
- If so, which are the biological causes and meaning?

DIFFERENTIAL EXPRESSION

A simple way to report expression differences is computing the ration between the expression in one condition and the expression in the other: This is called *fold change*, usually reported as

$$\text{Fold Change} = \log_2 \frac{\text{Expression in condition 1}}{\text{Expression in condition 2}}$$

- Which could be a good threshold for considering a difference in expression as significant?
- This threshold should be the same for all genes or not?
- It is based on expression means or medians, and variance is not considered
- The fold change is not a statistical test

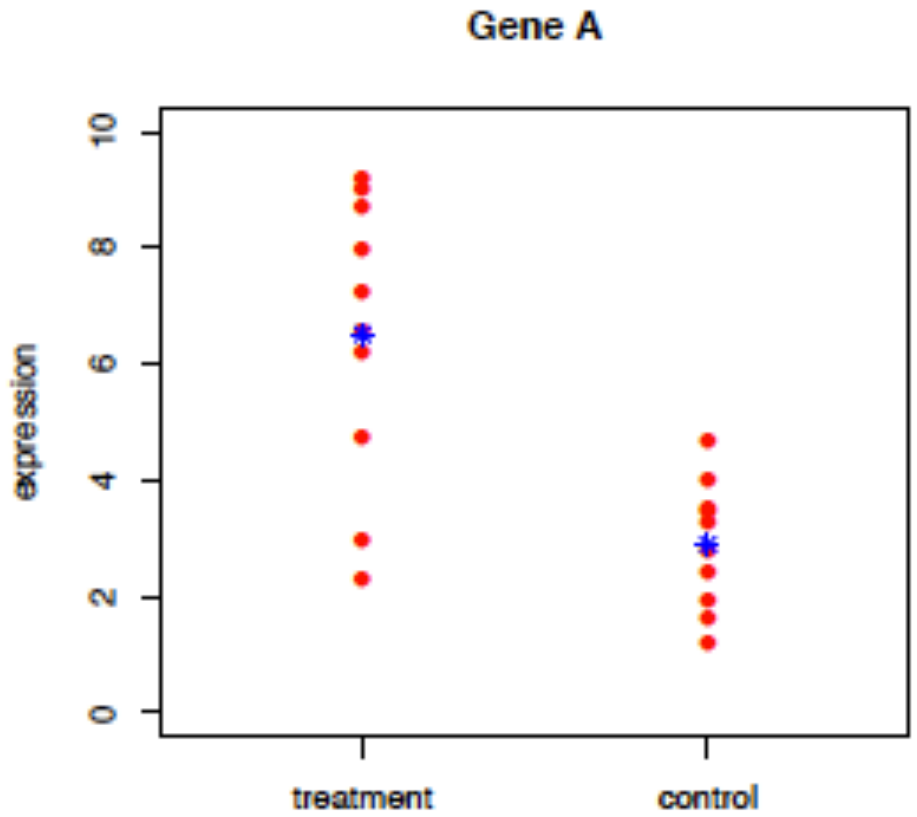
DIFFERENTIAL EXPRESSION

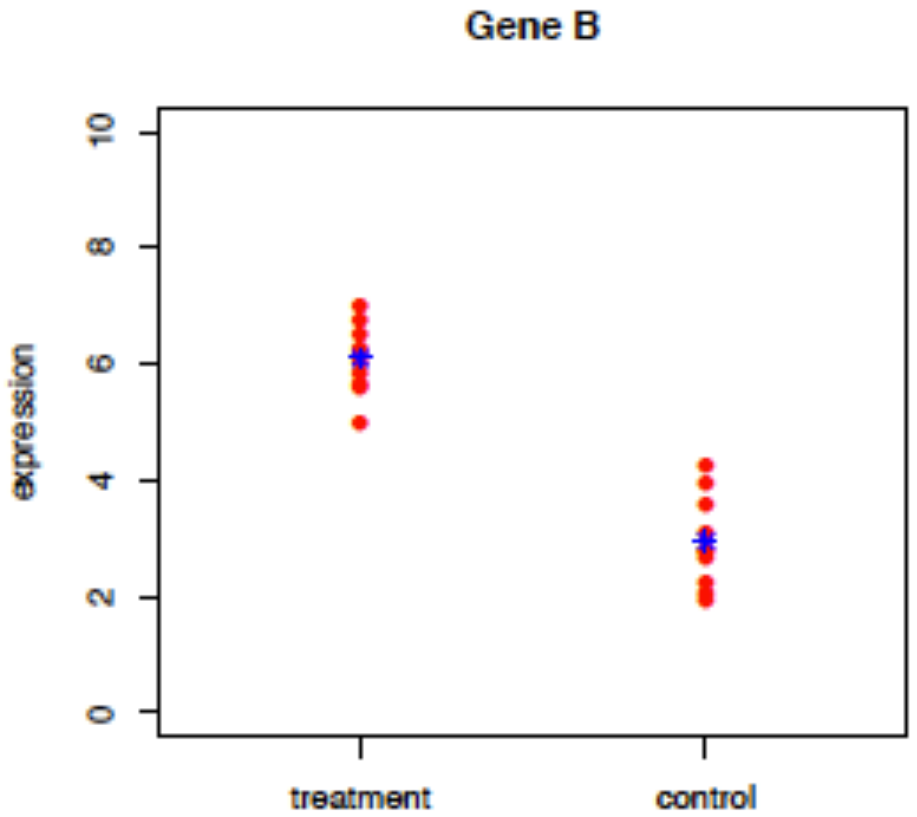
- Differential expression testing methods are based on the estimation of whether the expression variation of a gene in two conditions is higher than the variation that can be due to technical reasons
- For this reason, replicates are essential to have an accurate estimation of the technical and biological variance of gene expression

DIFFERENTIAL EXPRESSION

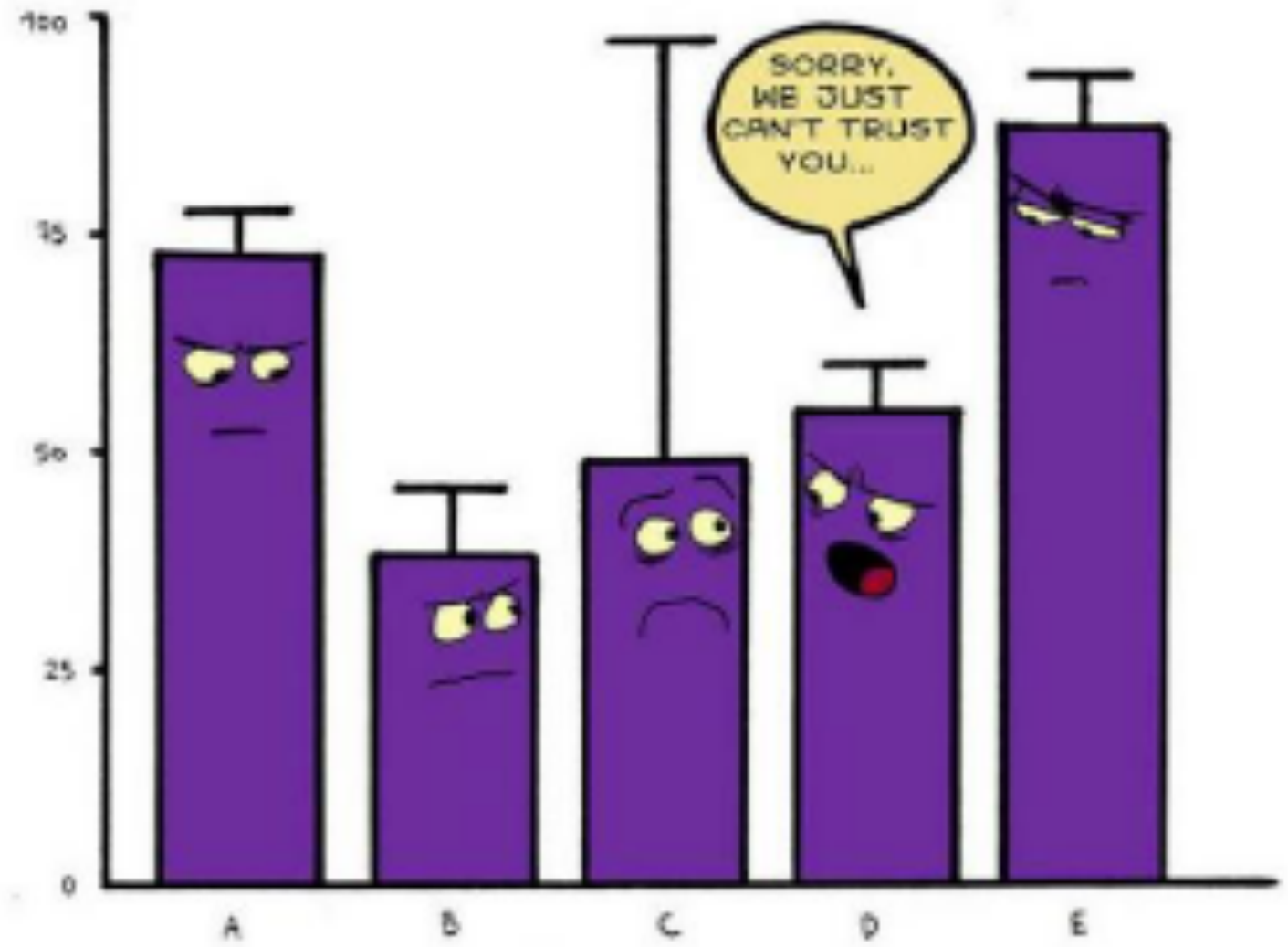
- Some methods employ normalized expression levels expressed in RPKM or TPM, but, more commonly, the read counts (i.e. how many reads can be mapped on a gene) are used
- These counts are always normalized in a way or another, not by the counts of all genes in a condition, but rather on all counts for the same gene in the all analyzed samples
- A common assumption is that gene expression is not globally changing between the two compared conditions, meaning that the number of truly differentially expressed genes is small compared to the total

DIFFERENTIAL EXPRESSION





DIFFERENTIAL EXPRESSION



Sonia Tarazona

DIFFERENTIAL EXPRESSION

A test for differential expression measures the probability that observed expression differences between two conditions are due to a biological difference and not just to random fluctuations.

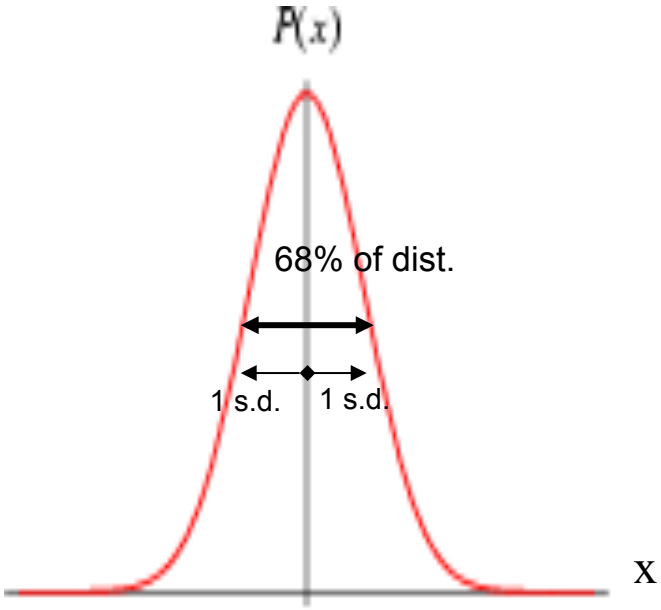
Fluctuations can be due to:

- measure errors
- technical variation due to sample preparation and/or analysis
- biological variation not associated to the differences between the two conditions

Generally, the expected variance is estimated, calibrated on the experimental replicates (when available), and compared to the observed variance based on a probabilistic model of the read mapping. The significance of the difference is evaluated with a statistical test, under the null hypothesis that there is no expression difference between the two conditions.

DIFFERENTIAL EXPRESSION

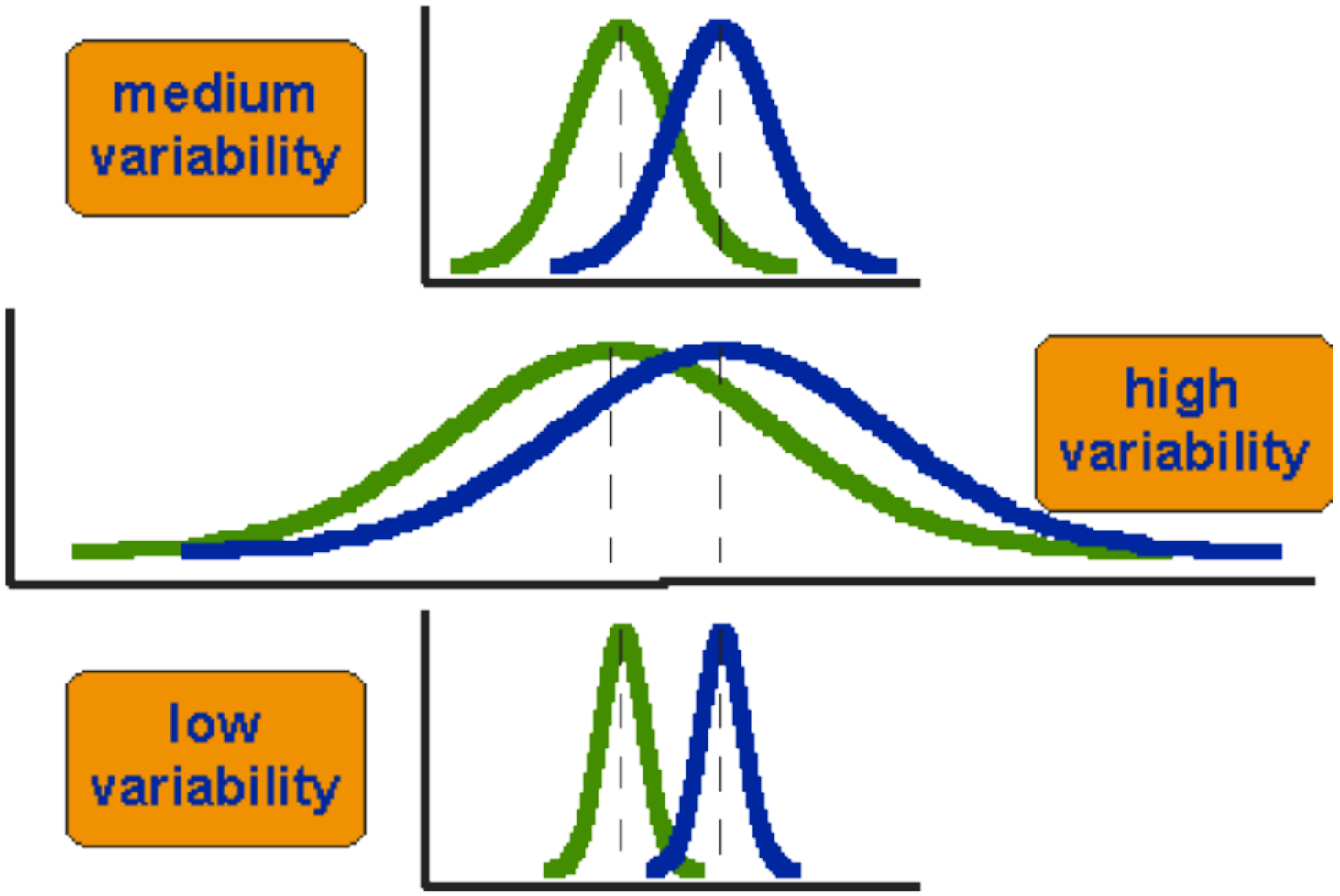
- We want to test:
  - Is the expression of one gene measured in a number of experimental replicates different in condition A and condition B?
  - Is the mean expression of a gene in condition A different from its mean expression B?
  - If the mean is different, with which confidence we can say that this difference is not due by chance only?



- Assumptions:
  - The distribution of expression values of a gene measured in the experimental replicates is normal
  - Replicates are indipendent

$\overline{X}$

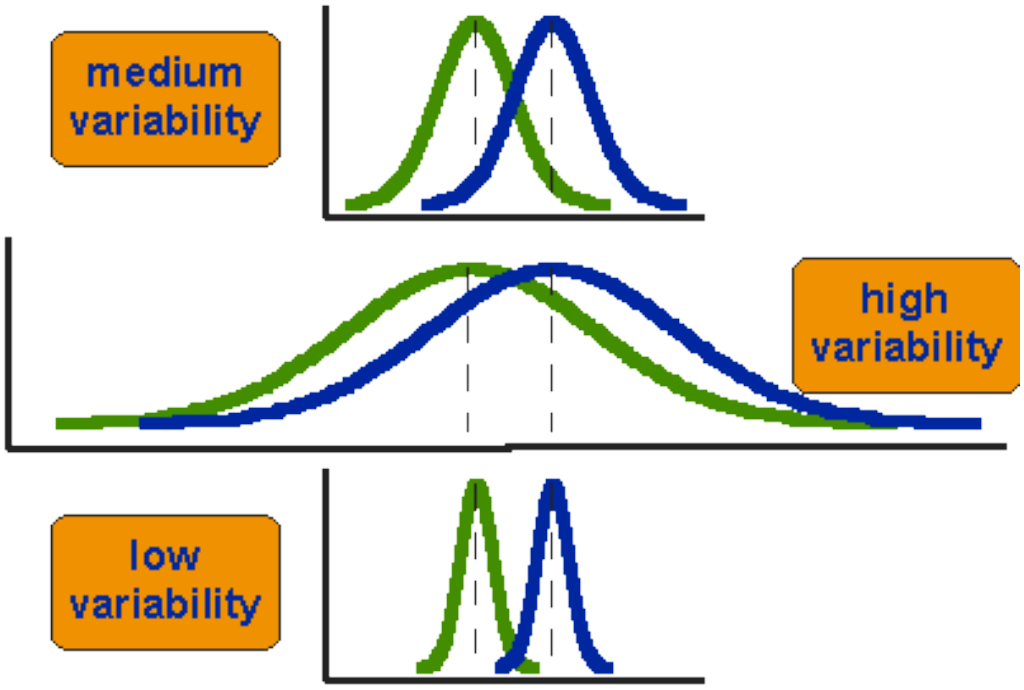
DIFFERENTIAL EXPRESSION



DIFFERENTIAL EXPRESSION

$$t = \frac{\text{Difference of the means}}{\text{Variance}}$$

$$t = \frac{\bar{X}_T - \bar{X}_C}{\sqrt{\frac{var_T}{n_T} + \frac{var_C}{n_C}}}$$



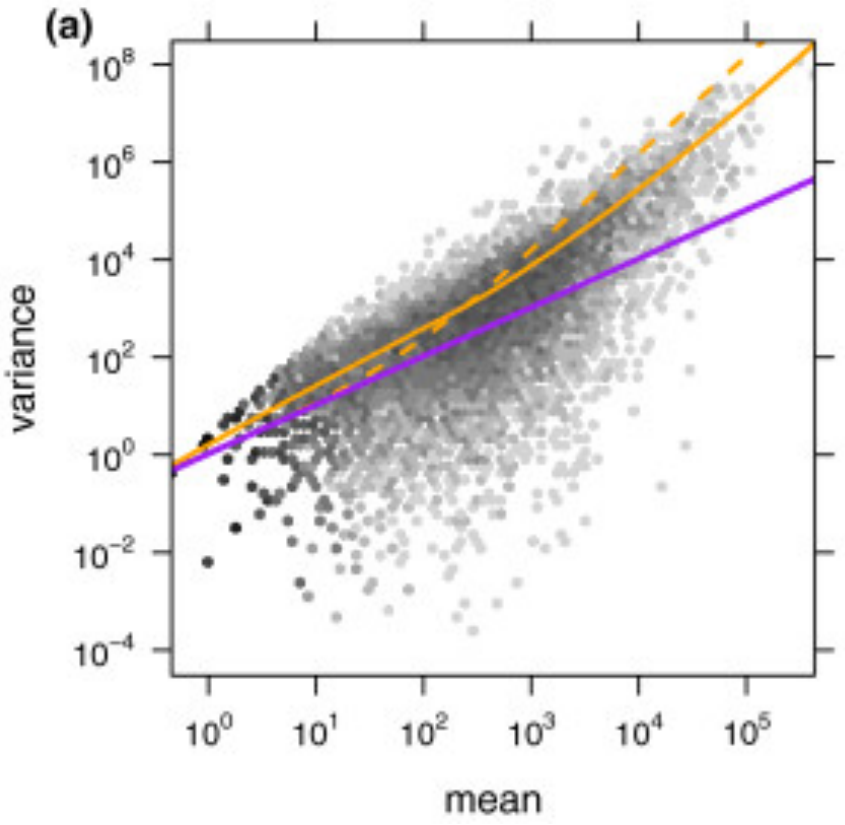
DIFFERENTIAL EXPRESSION

- When differential expression is tested directly on normalized values such as TPM, one can use parametric or non-parametric tests, such as the t-test
- When read counts are used, the distribution that seems to be the better in modeling the experimental variability is the negative binomial (NB), in which the variance can be estimated from the mean using a dispersion parameter Φ
- Dispersion can be estimated from the data, assuming that most genes do not change expression in the two compared conditions
- A gene having mean μ_1 and μ_2 in the two conditions can be evaluated as differentially expressed by computing the probability of having by chance an expression with mean μ_1 when the expected mean is μ_2 and variance is $\sigma^2=\mu_2+\Phi\mu_2^2$

DIFFERENTIAL EXPRESSION

- Negative binomial distribution is often used for RNA-seq data (NB)
- Variance is calibrated on all the genes in the examined samples

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



DIFFERENTIAL EXPRESSION

- There are a number of tests often employed for differential expression testing:
  - Z-test
  - Binomial distribution test
  - Negative binomial distribution test
  - Fisher test
  - Wilcoxon non-parametric

- All these are in some ways based on the comparisons of means and variances to reject a null hypothesis of no difference in expression.

DIFFERENTIAL EXPRESSION

- Hypothesis testing methods:
- For each gene:
  - The null hypothesis (Ho) is that the gene has no expression variation in the two conditions
  - The alternative hypothesis (Ha) is that there is an expression change
  - A test is employed to estimate the probability that the null hypothesis is true (p-value)
  - If the p-value < some threshold, one can reject Ho and accept Ha

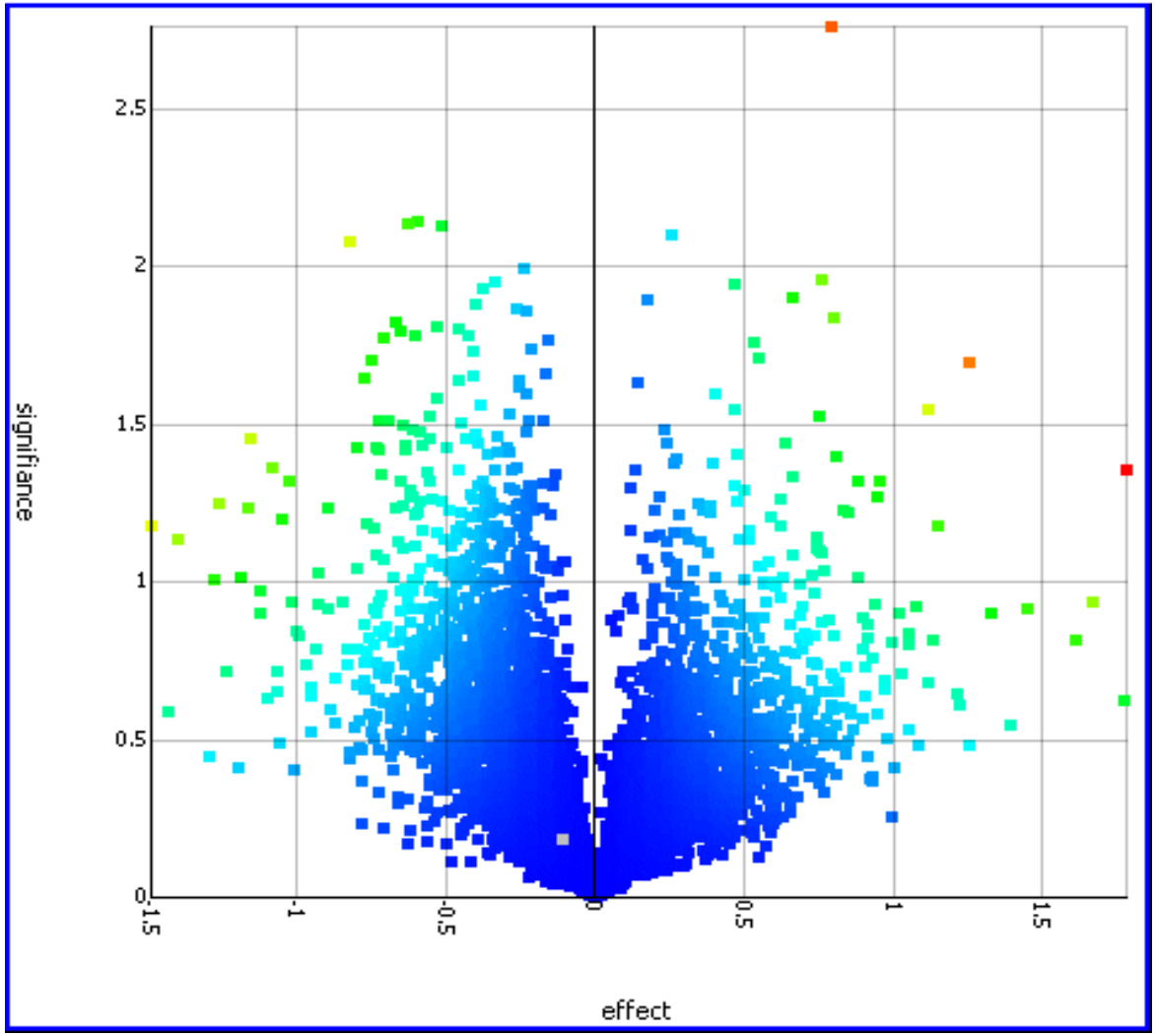
DIFFERENTIAL EXPRESSION

- One must chose the maximum p-value for rejecting the null hypothesis. Usually, genes are considered differentially expressed if p-value < 0.05
- This value corresponds to a risk factor α=5%, which is frequency of rejecting the null hypothesis H_0 even if it is true (i.e. The gene is not differentially expressed)

	Reality: H_0 is false	Reality: H_0 is true
Decision: Reject H_0	H_0 correctly rejected (True Positive)	Type I error (False Positive)
Decision: Do not reject H_0	Type II error (False Negative)	H_0 correctly not rejected (True Negative)

DIFFERENTIAL EXPRESSION

A volcano plot is a scatterplot of -log(p-value) vs. the fold change.



DIFFERENTIAL EXPRESSION

- At the end of the differential expression testing procedure, to each gene a p-value is assigned, that is the probability of rejecting the null hypothesis of having no differences between the two samples for that gene
- In simpler terms, p-value can be considered as the probability that the detected differences in expression values are only due to chance and technical factors
- The smaller is the p-value, the more convincing is that there is a significant expression change between the two conditions
- For each gene, a different test must be performed, raising the issue of multiple testing

DIFFERENTIAL EXPRESSION

Multiple testing problem:

- Each gene requires a different test
- In an RNA-seq experiment, you might perform tens of thousands of tests.
- If the p-value threshold is 0.05, and you are testing 20000 genes (then 20000 tests), you can get in the end with lots of false positives (type I errors), that are genes that you consider differentially expressed even if they are not.
- These false positives can affect the biological characterization of the differences between the two samples.
- There are correction methods (for example the methods of Bonferroni, Benjamini-Hochberg, Sidak, Duncan, Holm)

DIFFERENTIAL EXPRESSION

- If you are testing a null hypothesis that is true, using a threshold of significance α then the probability of having a correct non-significant result is $1-\alpha$
- If you are testing 2 independent hypothesis (e.g. two different genes) the probability that neither test is significant is the product of the two probabilities $(1-\alpha)(1-\alpha)$. In general, for *k* independent tests (for *k* genes), the probability that all tests are not significant is $(1-\alpha)^k$
- Hence, the probability of having at least one significant test by chance (the negation of the previous case, where all were not significant) is $1-(1-\alpha)^k$
- This is called Family Wise Error Rate (FWER)

DIFFERENTIAL EXPRESSION

- For example, if you are testing 20 hypothesis (20 genes) using a significance threshold $\alpha = 0.05$, then the probability that none of them is significant is:

$$0.95^{20} = 0.36$$

- Therefore the probability that at least one test is significant by chance (a false positive) is:

$$1 - 0.36 = 0.64$$

- This probability is quite larger that the threshold value α, hence the need of correction methods

DIFFERENTIAL EXPRESSION

- Most correction methods are based on the assumption that the global error associated to the experiment is function of the p-value of each individual test. If the significance threshold α is lowered, also the global error is reduced
- Methods based on FWER control seek for new thresholds (or, alternatively, they seek corrections of the p-values) in order to reduce false positives
- The most simple and used method is the Bonferroni correction
- Imagine that you are testing 5 hypothesis, and you want FWER to be at most 5%. Then, type I error for each individual test must be set to 0.05/5=0.01, or, alternatively, p-values must be multiplied by 5. These corrected p-values are called adjusted p-values

DIFFERENTIAL EXPRESSION

- Bonferroni correction is very conservative, and when the number of tests increases, thresholds become so small that no test is significant anymore
- This way, false positives are effectively reduced, at the cost of increasing the false negatives, that are tests for which the null hypothesis is truly false but that do not result as significant by the test
- Hence, when testing a large number of genes such as in a RNA-Seq experiment, other methods, such as those limiting the False Discovery Rate (FDR) are preferred

DIFFERENTIAL EXPRESSION

FDR is defined as the expected number of false positives in a set of singificant tests. If the number of tests is *m*, then:

	H_0 true	H_0 false	Total
H_0 rejected (test significant)	V	S	R
H_0 accepted (test not significant)	U	T	m-R
Total	m_0	$m-m_0$	m

R=V+S is the number of discoveries (that can be true or false)

DIFFERENTIAL EXPRESSION

- FDR is defined as:

$$FDR = E\left[\frac{V}{R}\middle| R > 0\right]$$

- Here E is the expected value, V is the number of false positives, R is the number of tests that are significant. So methods must find the threshold that brings to the expected number of false positives
- The difficulty is the estimation of the m0 value. The popular Benjamini-Hochberg method solves this issue by assuming that m_0 is 1, meaning that all tests are true negatives

DIFFERENTIAL EXPRESSION

- Then, it can be demonstrated that p-value correction can be obtained by:

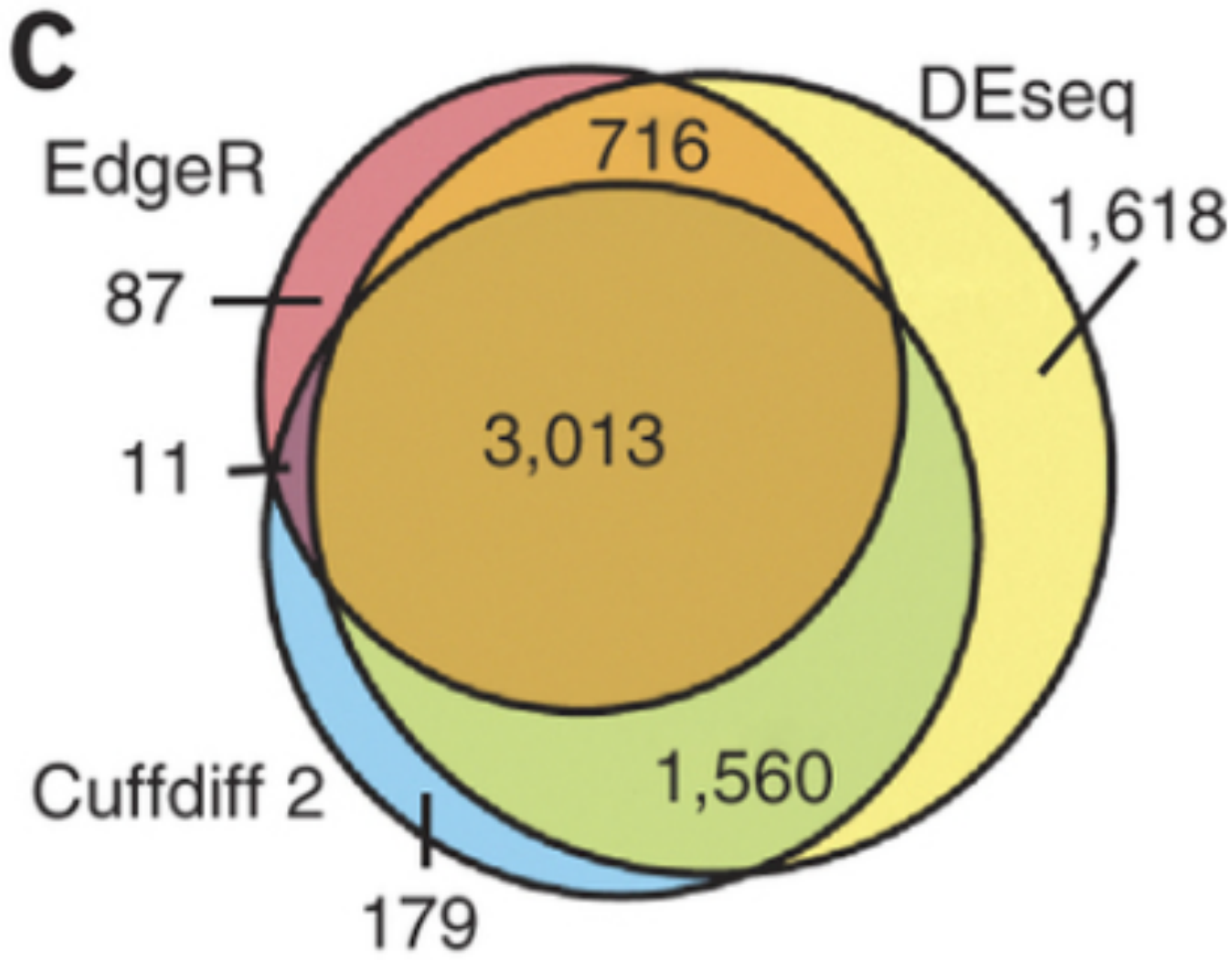
$$q_{(i)} = p_{(i)} \cdot \frac{m}{i}$$

- This adjusted p-value is called q-value
- p(i) and q(i) are the vectors in increasing order of the p-values and q-values, and *i* is the i^{th} position in the ordered vector

DIFFERENTIAL EXPRESSION

- By selecting genes with $q \leq 0.05$, one gets a list of genes with FDR 5%, meaning a list in which the expected rate of false positives is 5%
- Using a p-value threshold of 0.05 means that 5% of all tests will results in false positives. Instead, using a q-value threshold of 0.05 means that 5% of significant tests will results in false positives, which is a smaller amount

DIFFERENTIAL EXPRESSION



Trapnell et al., 2013