RNA-seq

Quantification

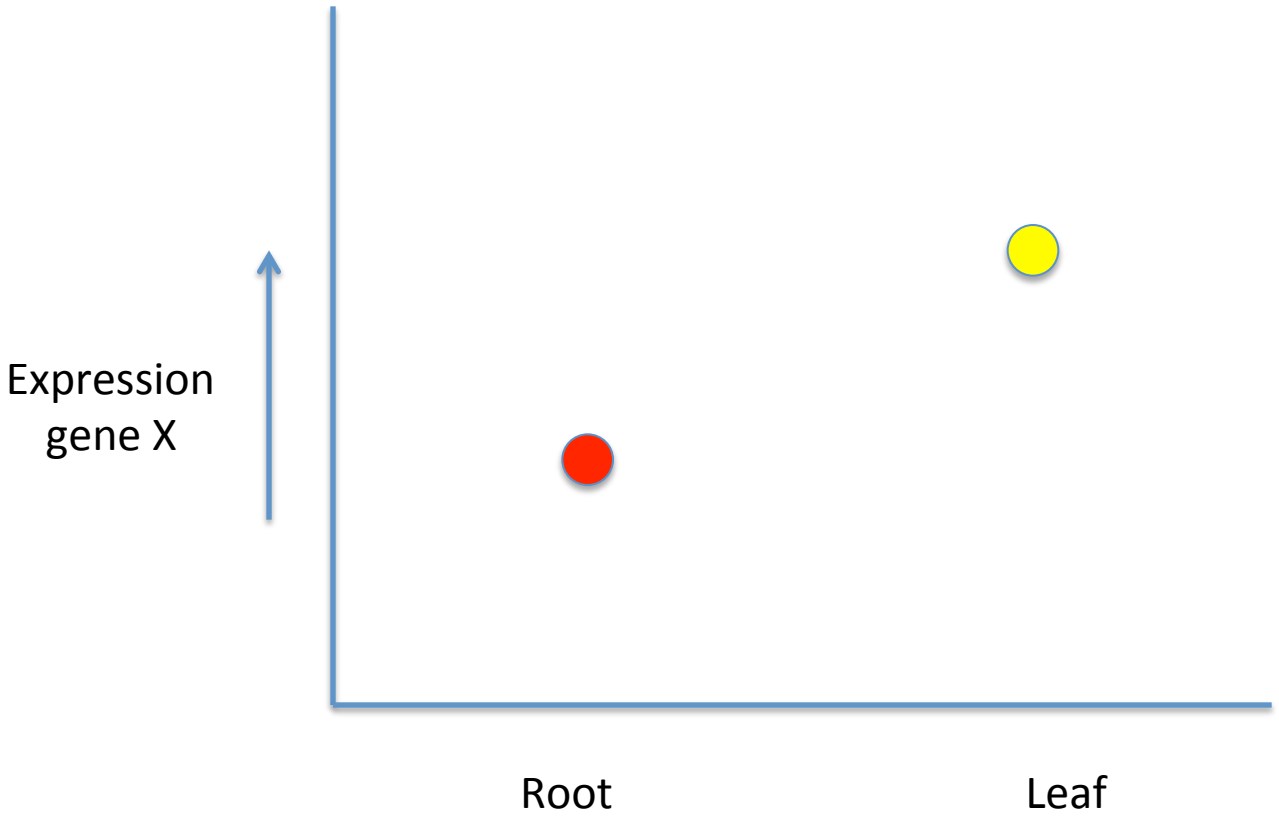
Harm Nijveen

Differential expression

Which genes are **higher/lower** expressed between tissues, after treatment, etc.?

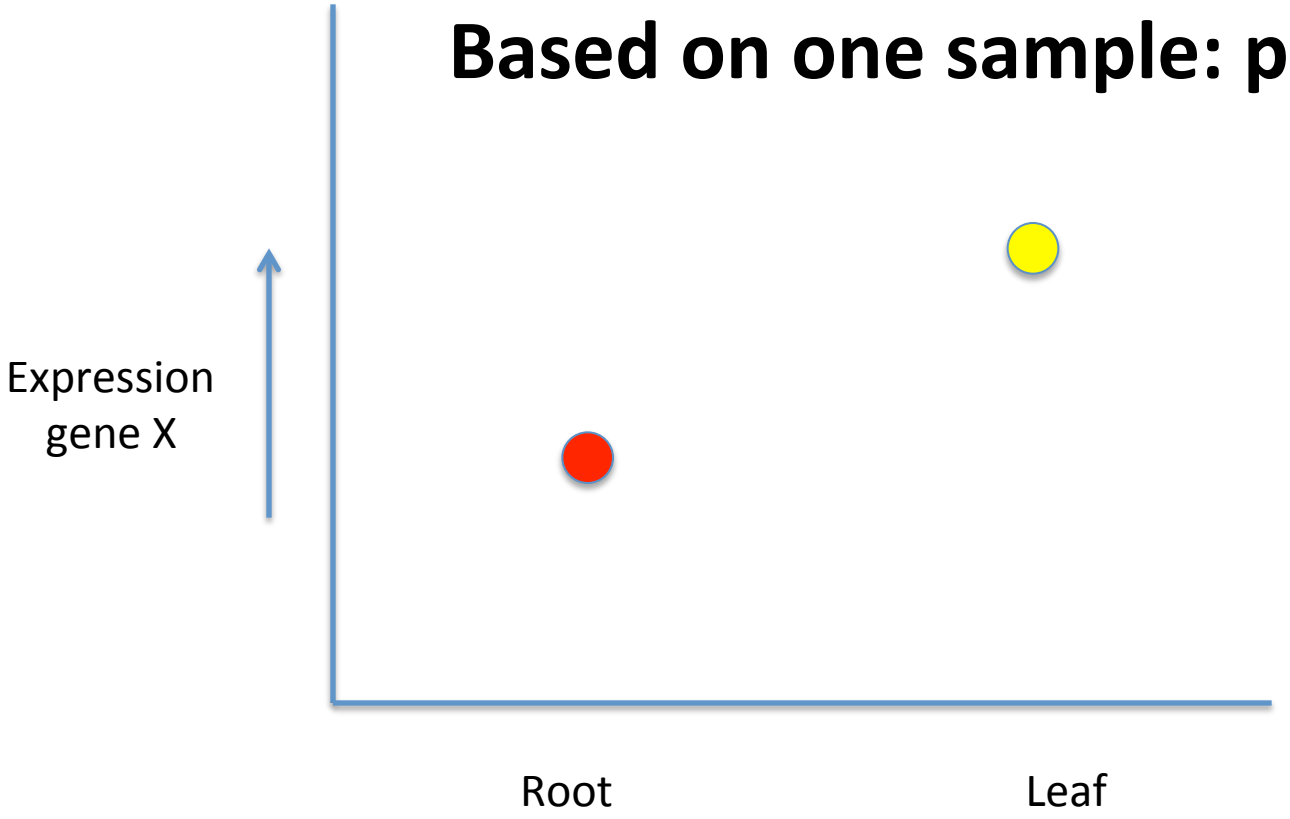
Differentially Expressed genes (DEGs) have an expression level that is **significantly different** between different conditions.

Is the expression of gene X different between root and leaf?

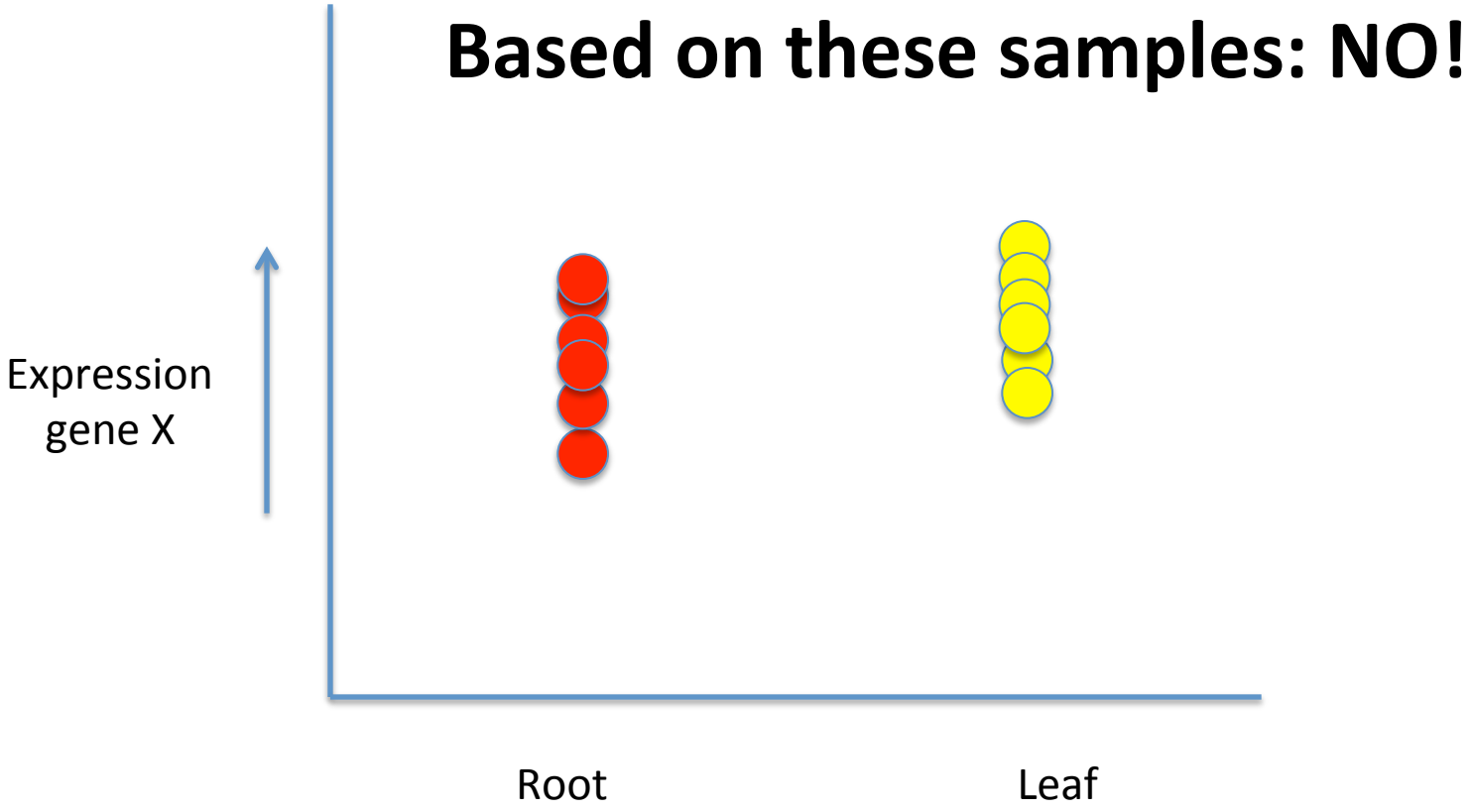


Is the expression of gene X different between root and leaf?

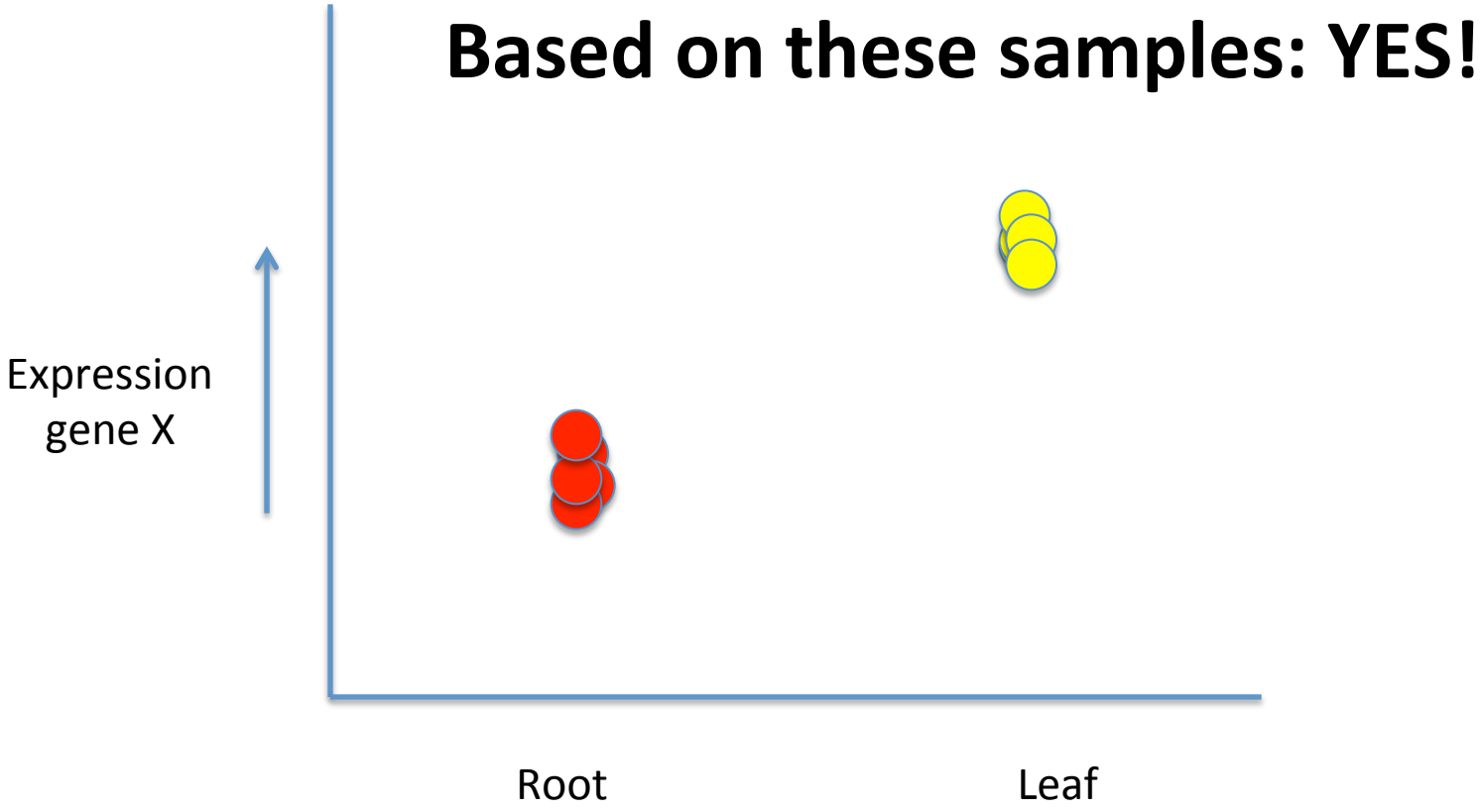
Based on one sample: perhaps…



Is the expression of gene X different between root and leaf?



Is the expression of gene X different between root and leaf?



Is a gene differentially expressed?

With only one measurement : impossible to say

We have to know the within-treatment variation

Determining expression variation

Accurately determining the variation requires many biological samples (replicates)

Unfortunately in most case we only have two or three replicates

Variation has to be estimated

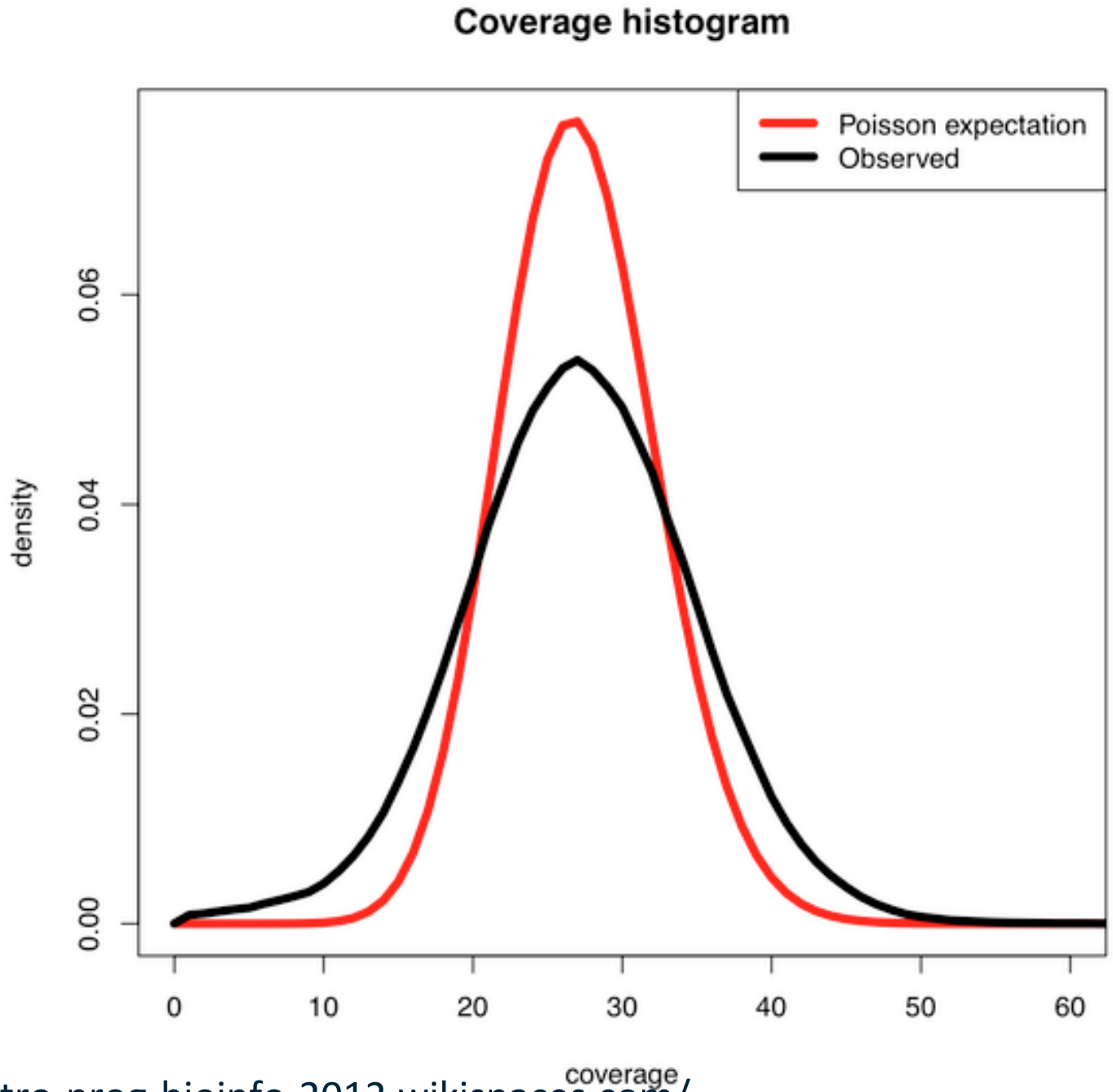
Read count distribution

Poisson distribution: variance = mean

- Holds for technical replicates

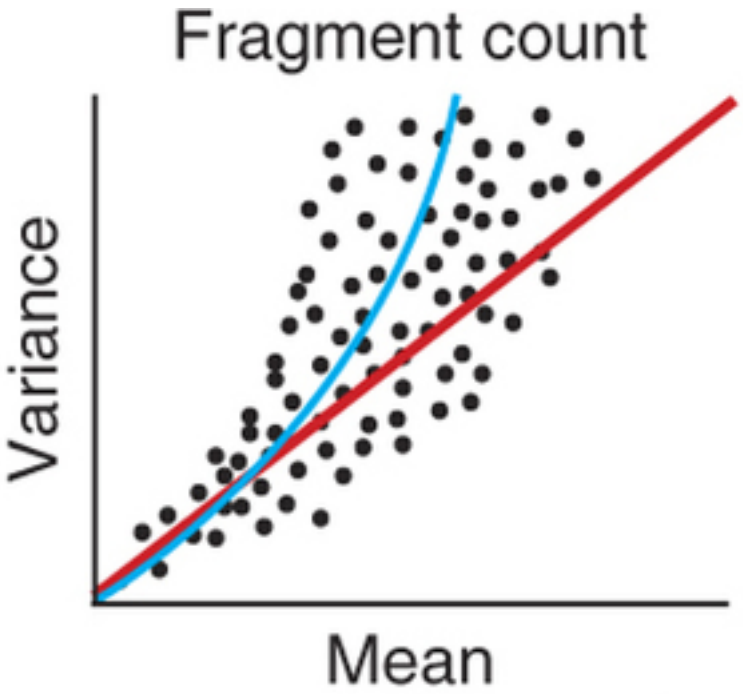
Negative binomial: variance > mean

- Better fit for biological replicates



https://intro-prog-bioinfo-2012.wikispaces.com/

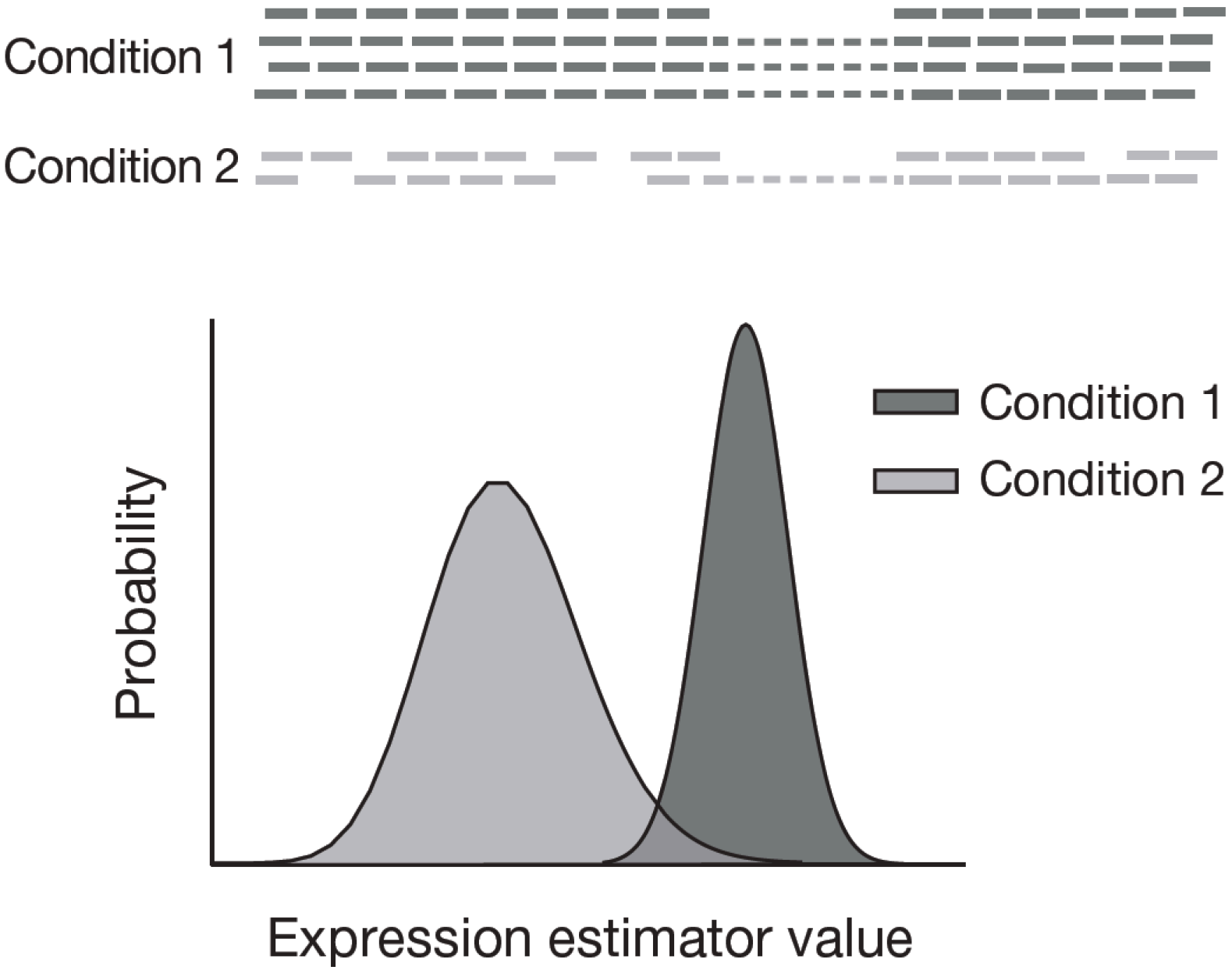
Variance depends on the mean



Trapnell et al. 2012

Variance depends on the mean.

Find a function that best describes the relationship between the mean and variance.


Condition 1
Condition 2
Probability
Condition 1
Condition 2
Expression estimator value

p-value

To find differentially expressed genes we can do a statistical test and determine a *p*-value.

p-value = 0.05 means that there is a 5% chance for a not-differentially expressed gene to show these kind of expression differences

But: with 10,000 genes i.e. 10,000 tests, you can expect 0.05 x 10,000 = 500 false positives!

Multiple testing correction

We need to correct the *p*-value for doing a large number of tests

We can used the False Discovery Rate (FDR) that produces an adjusted *p*-value called *q*-value

q-value = 0.05 means that there is a 5% chance that these expression values are from a not differentially expressed gene

Some tools..

DESeq/DESeq2

EdgeR

Sleuth (kallisto)

HISAT2/StringTie/Ballgown
(can quantify isoforms)

Plotting DEGs

Volcano plot

- x: log2(fold change)
- y: -log10(p-value)

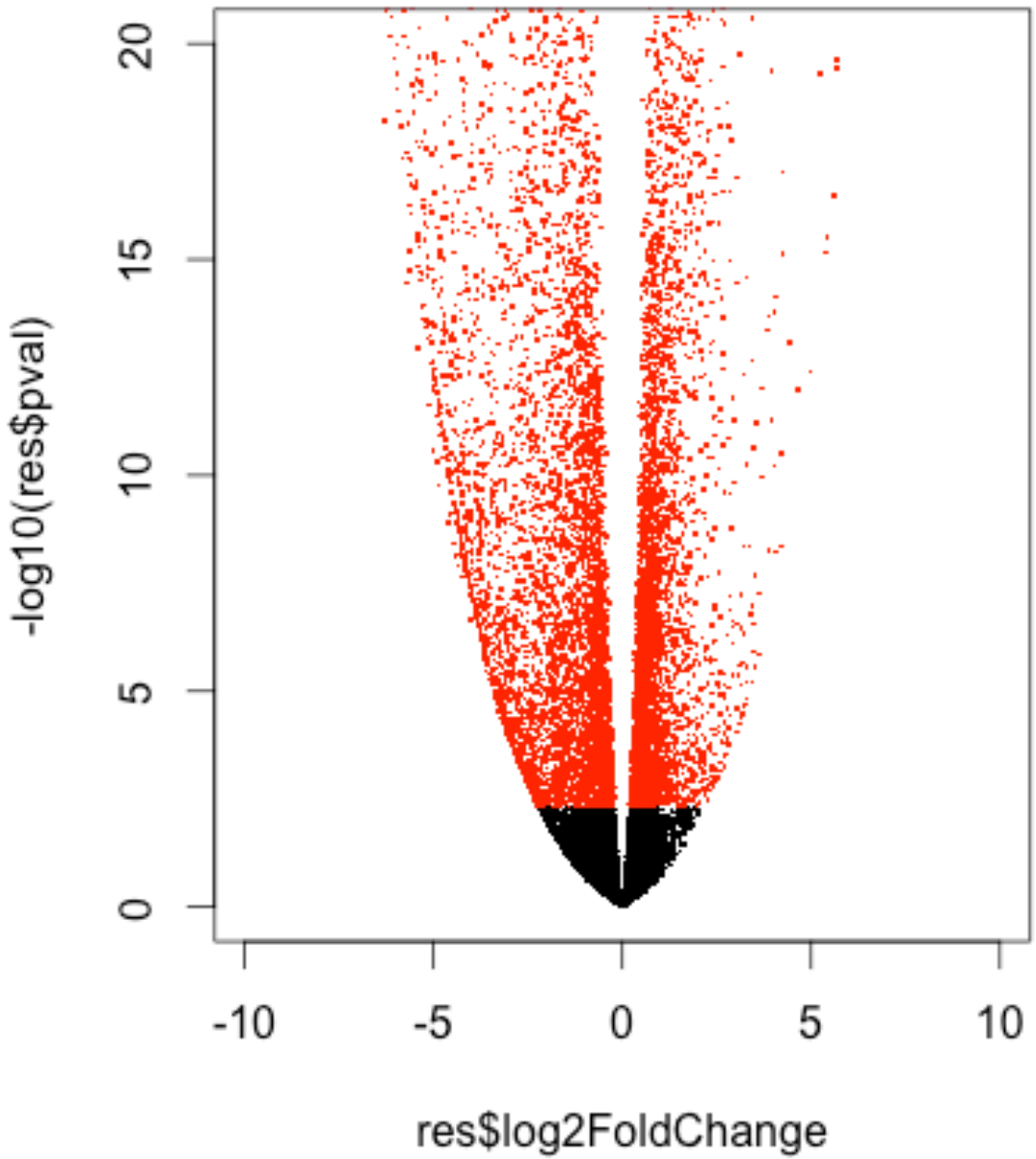
MA plot

- x: mean expression
- y: log2(fold change)

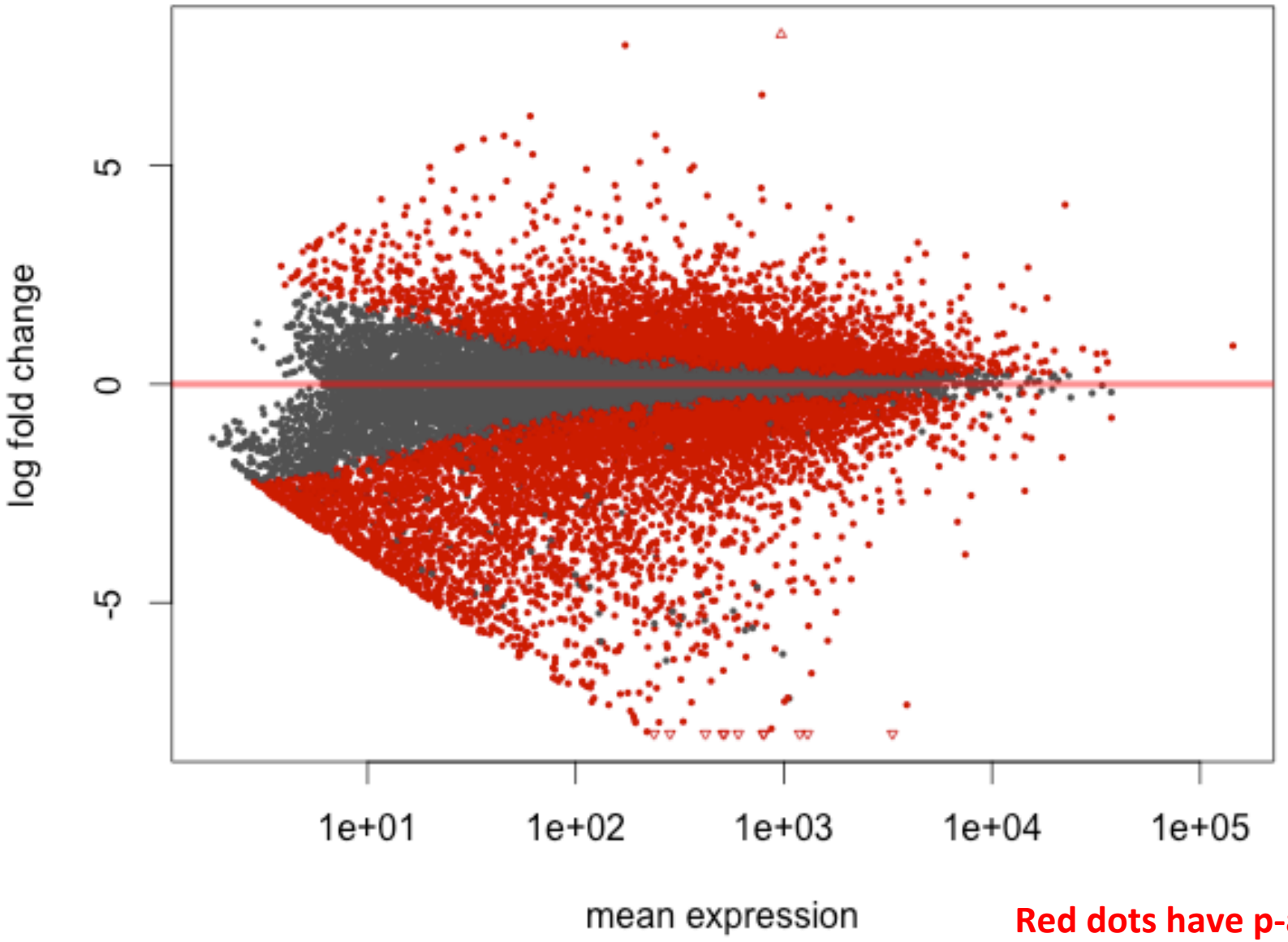
A dot represents one gene

Red dots are significant

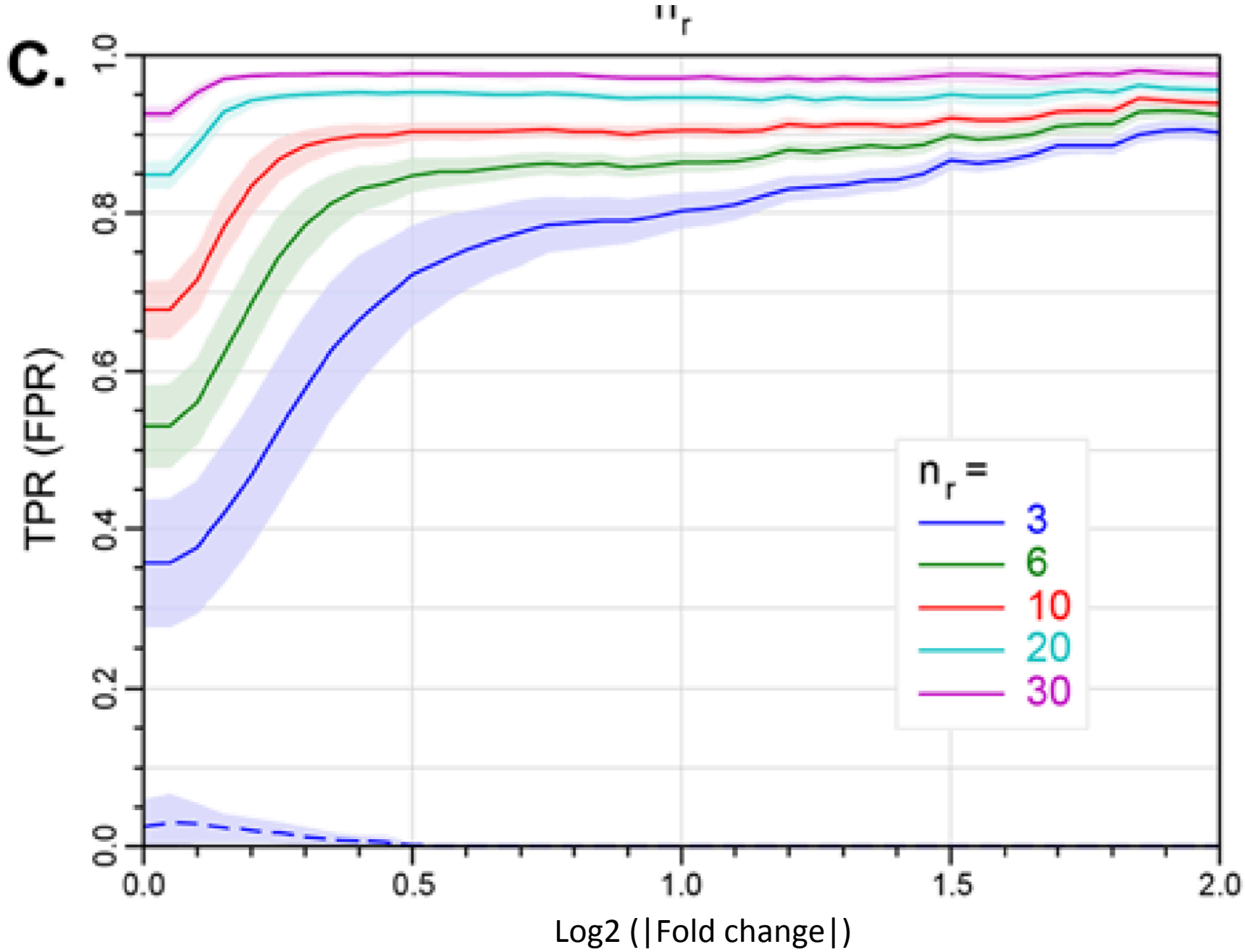
Volcano plot



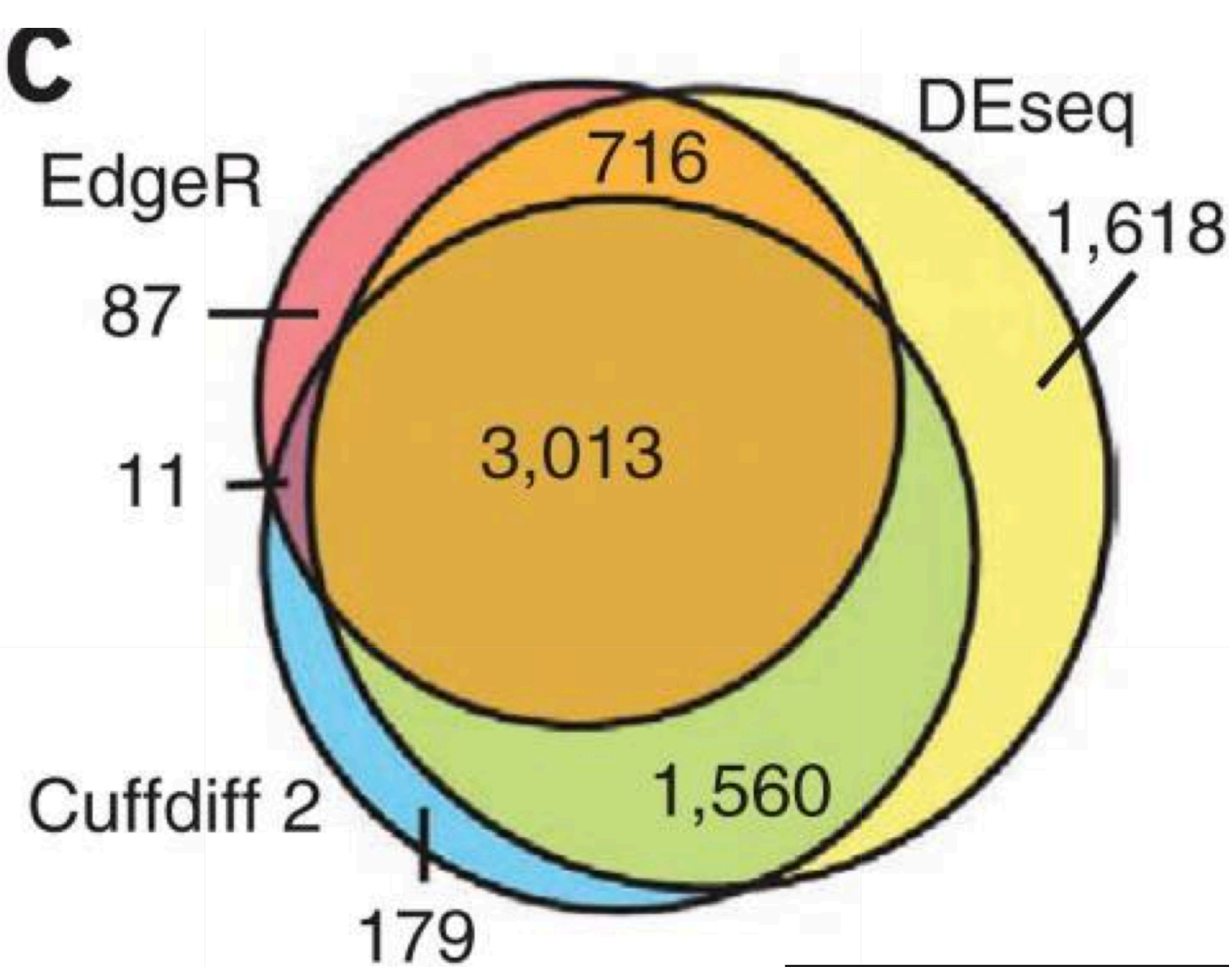
MA plot



Red dots have p-adj < 0.01



Schurch, N. J., P. Schofield, M. Gierliński, C. Cole, A. Sherstnev, V. Singh, N. Wrobel, K. Gharbi, G. G. Simpson and T. Owen-Hughes (2015). "Evaluation of tools for differential gene expression analysis by RNA-seq on a 48 biological replicate experiment." arXiv preprint arXiv:1505.02017.


c
EdgeR
DEseq
716
1,618
87
11
3,013
1,560
Cuffdiff 2
179

Which are the interesting genes?

Highest fold change?

Lowest p-value?

Other?

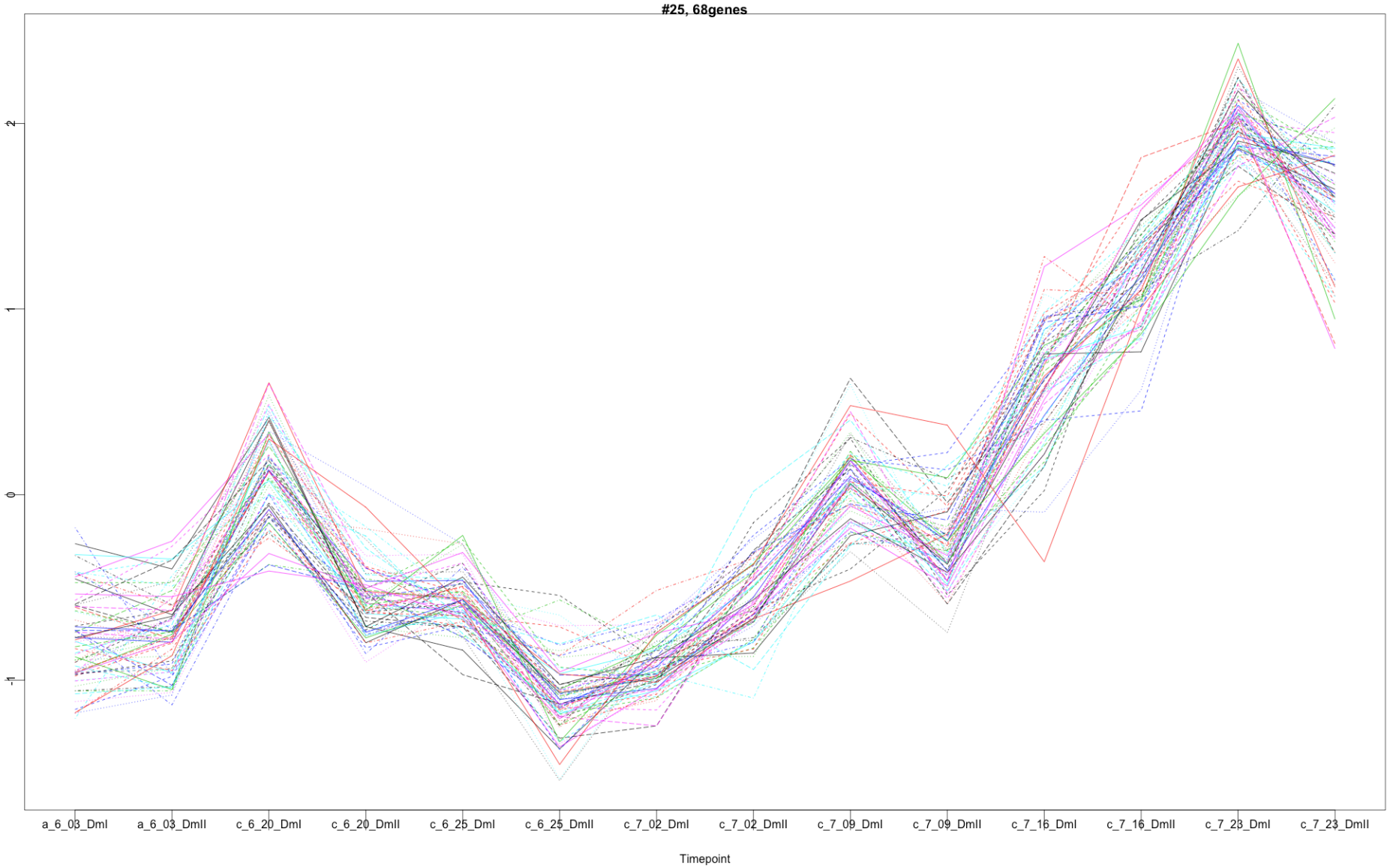
Comparing multiple treatments

Time series, multiple tissues, etc.

Look for genes with a similar expression pattern

Using various kinds of clustering methods

Co-expression



And now?

How do the ‘usual suspects’ behave?

Which biological processes are enriched?

Which pathways are enriched?

Depends on the biological question!

Continued this afternoon…