Testing for Differential Expression with DESeq2

Laura Pikkupeura

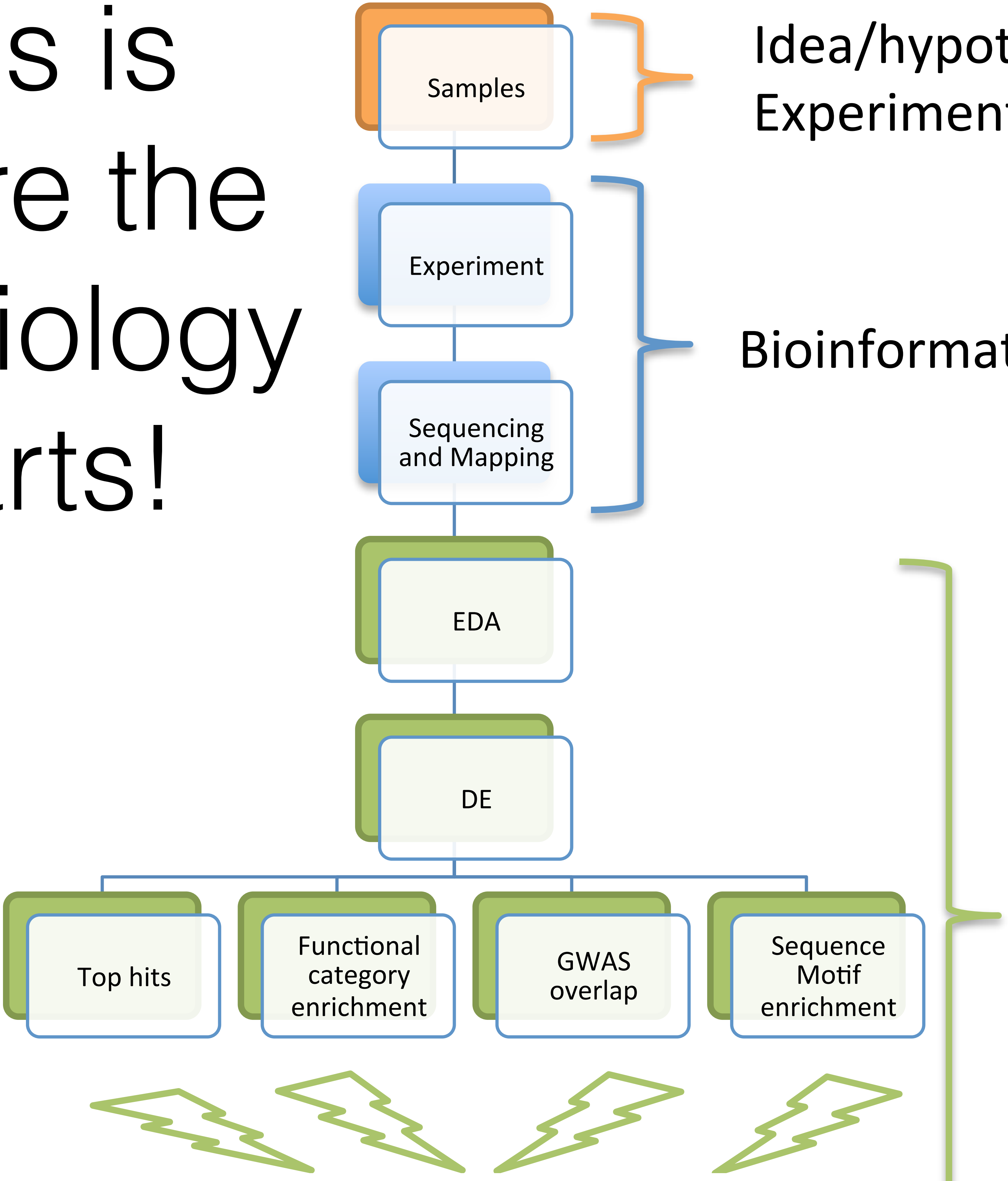
Teacher

- Laura Pikkupeura
- PhD student in Sandelin and Jensen Labs
- Mail: laura.pikkupeura@bric.ku.dk
- Background:
    - MSc Molecular Biomedicine
    - PhD Stem cell biology, genomics and transcriptomics

Lecture outline

- **Theory:**
  - Introduction to testing in transcriptomics
  - Overview of DESeq2
- **Practical:**
  - Gene expression in the fissing yeast stress response
  - Gene expression in Inflammatory bowel disease.
  - DESeq2 applied to pseudo-aligned reads.

This is where the ~~fun~~ biology starts!



Samples

Experiment

Sequencing and Mapping

EDA

DE

Top hits

Functional category enrichment

GWAS overlap

Sequence Motif enrichment

Idea/hypothesis
Experiment Design

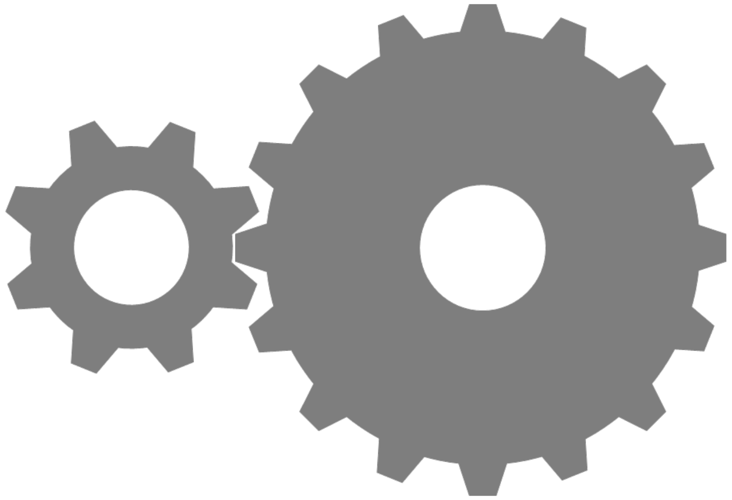
Bioinformatics leg work

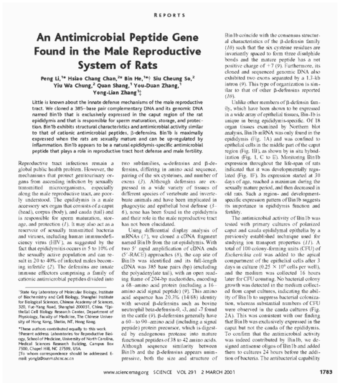
Biological/clinical analysis

Biological interpretation
Candidate validation

Importance of QC and filtering

Cleaned, Trimmed and prepared data





REPORTS

An Antimicrobial Peptide Gene Found in the Male Reproductive System of Rats

Peng Li,[1*] Hsiao Chang Chan,[2*] Bin He,[1*†] Siu Cheung So,[2] Yiu Wa Chung,[2] Quan Shang,[1] You-Duan Zhang,[1] Yong-Lian Zhang[1‡]

Little is known about the innate defense mechanisms of the male reproductive tract. We cloned a 385–base pair complementary DNA and its genomic DNA named Bin1b that is exclusively expressed in the caput region of the rat epididymis and that is responsible for sperm maturation, storage, and protection. Bin1b exhibits structural characteristics and antimicrobial activity similar to that of cationic antimicrobial peptides, β-defensins. Bin1b is maximally expressed when the rats are sexually mature and can be up-regulated by inflammation. Bin1b appears to be a natural epididymis-specific antimicrobial peptide that plays a role in reproductive tract host defense and male fertility.

Reproductive tract infections remain a global public health problem. However, the mechanisms that protect genitourinary organs from ascending infection by sexually transmitted microorganisms, especially along the male reproductive tract, are poorly understood. The epididymis is a male accessory sex organ that consists of a caput (head), corpus (body), and cauda (tail) and is responsible for sperm maturation, storage, and protection (*1*). It may also act as a reservoir of sexually transmitted bacteria and viruses, including human immunodeficiency virus (HIV), as suggested by the fact that epididymitis occurs in 5 to 10% of the sexually active population and can result in 20 to 40% of infected males becoming infertile (*2*). The defensins are innate immune effectors comprising a family of cationic antimicrobial peptides divided into two subfamilies, α-defensins and β-defensins, differing in amino acid sequence, pairing of the six cysteines, and number of exons (*3*). Although defensins are expressed in a wide variety of tissues of different species of vertebrate and invertebrate animals and have been implicated in phagocytic and epithelial host defense (*3*–*6*), none has been found in the epididymis and their role in the male reproductive tract has not been elucidated.

Using differential display analysis of mRNAs (*7*), we cloned a cDNA fragment named Bin1b from the rat epididymis. With two 5′ rapid amplification of cDNA ends (5′-RACE) approaches (*8*), the cap site of Bin1b was identified and its full-length cDNA was 385 base pairs (bp) (excluding the polyadenylate tail), with an open reading frame of 204-bp nucleotides, encoding a 68–amino acid protein (including a 16–amino acid signal peptide) (*9*). This amino acid sequence has 20.3% (14/68) identity with several β-defensins such as bovine neutrophil beta-defensin-9, -3, and -7 found in the cattle (*9*). β-defensins generally have a 60– to 90–amino acid (including a signal peptide) protein precursor, which is digested by endogenous protease into mature functional peptides of 38 to 42 amino acids. Although sequence similarity between Bin1b and the β-defensins appears unimpressive, both the size and structure of Bin1b coincide with the consensus structural characteristics of the β-defensin family (*10*) such that the six cysteine residues are invariantly spaced to form three disulphide bonds and the mature peptide has a net positive charge of +7 (*9*). Furthermore, its cloned and sequenced genomic DNA also exhibited two exons separated by a 1.3-kb intron (*9*). This type of organization is similar to that of other β-defensins reported (*10*).

Unlike other members of β-defensin family, which have been shown to be expressed in a wide array of epithelial tissues, Bin-Ib is unique in being epididymis-specific. Of 18 organ tissues examined by Northern blot analysis, Bin1b mRNA was only found in the epididymis (Fig. 1A) and was confined to epithelial cells in the middle part of the caput region (Fig. 1B), as shown by in situ hybridization (Fig. 1, C to E). Monitoring Bin1b expression throughout the life-span of rats indicated that it was developmentally regulated (Fig. 1F). Its expression started at 30 days of age, reached a maximum during the sexually mature period, and then decreased in old rats. Such a region- and development-specific expression pattern of Bin1b suggests its importance in epididymis function and fertility.

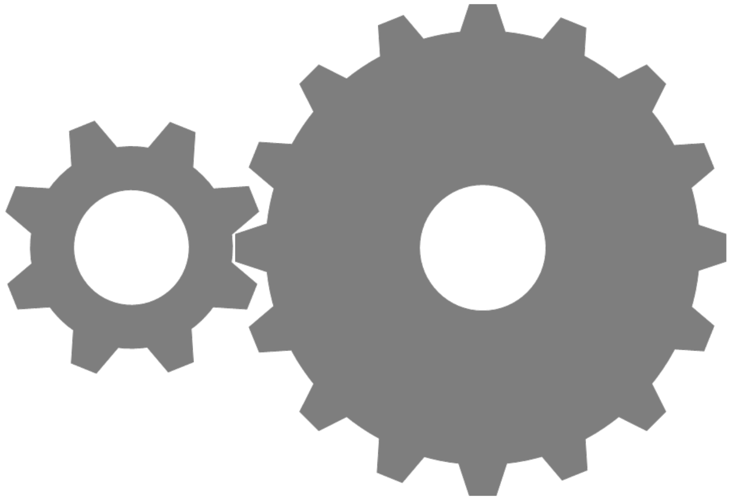
The antimicrobial activity of Bin1b was tested with primary cultures of polarized caput and cauda epididymal epithelia by a previously established technique used for studying ion transport properties (*11*). A total of 100 colony-forming units (CFU) of *Escherichia coli* was added to the apical compartment of the epithelial cells after 3 days in culture (0.25 × 10^6 cells per well), and the medium was collected 16 hours later for CFU counting. No bacterial colony growth was detected in the medium collected from caput cultures, indicating the ability of Bin1b to suppress bacterial colonization, whereas substantial numbers of CFU were observed in the cauda cultures (Fig. 2A). This was consistent with our finding that Bin1b was exclusively expressed in the caput but not the cauda of the epididymis. To confirm that the antimicrobial activity was indeed contributed by Bin1b, we designed antisense oligos of Bin1b and added them to cultures 24 hours before the addition of bacteria. The antibacterial capability

[1] State Key Laboratory of Molecular Biology, Institute of Biochemistry and Cell Biology, Shanghai Institute for Biological Sciences, Chinese Academy of Sciences, 320, Yue-Yang Road, Shanghai 200031, China. [2] Epithelial Cell Biology Research Center, Department of Physiology, Faculty of Medicine, The Chinese University of Hong Kong, Shatin, NT, Hong Kong.

*These authors contributed equally to this work.
†Present address: Laboratories for Reproductive Biology, School of Medicine, University of North Carolina, Medical Sciences Research Building, Campus Box 7500, Chapel Hill, NC 27599, USA.
‡To whom correspondence should be addressed. E-mail: yonglz@sunm.shcnc.ac.cn

Nature publication

Shitty data

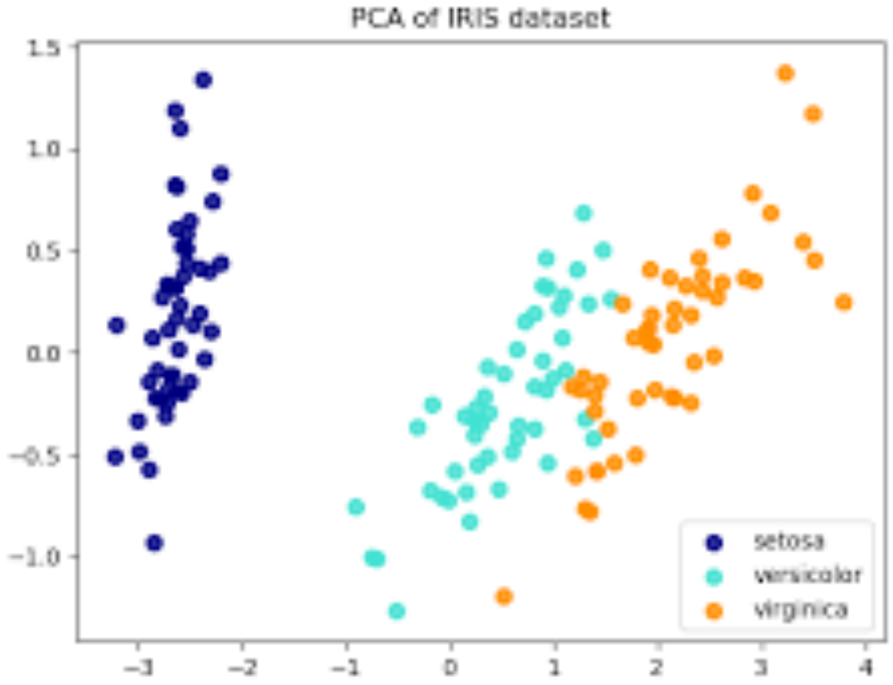


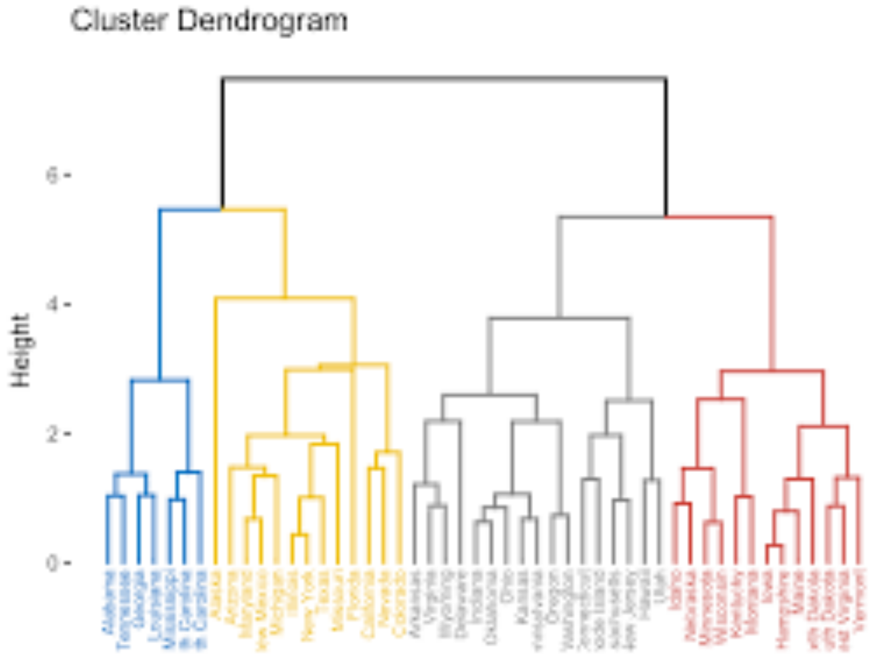
Even more shitty data

Testing in transcriptomics

EDA

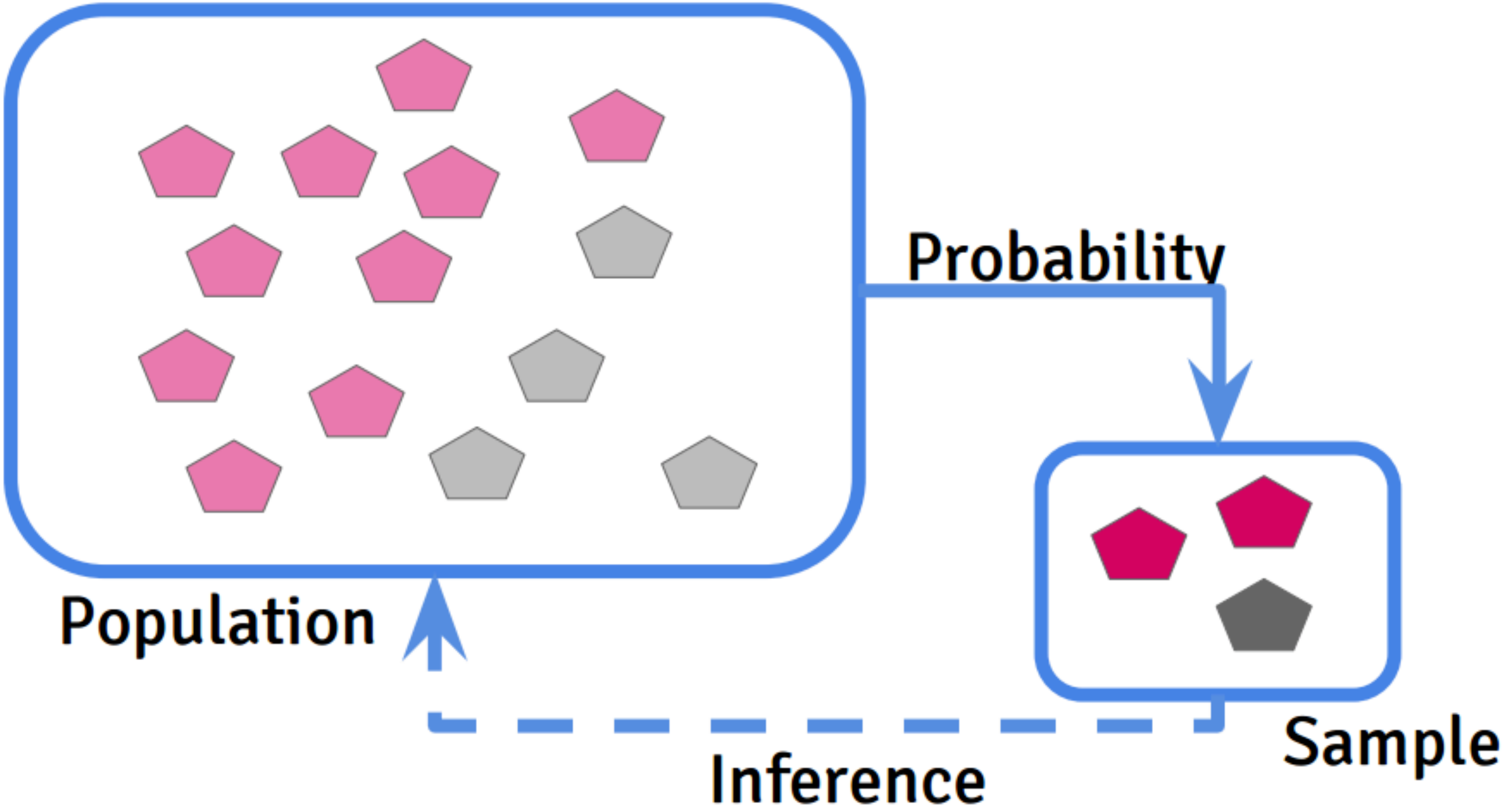
Exploratory data analysis deals with *global differences*



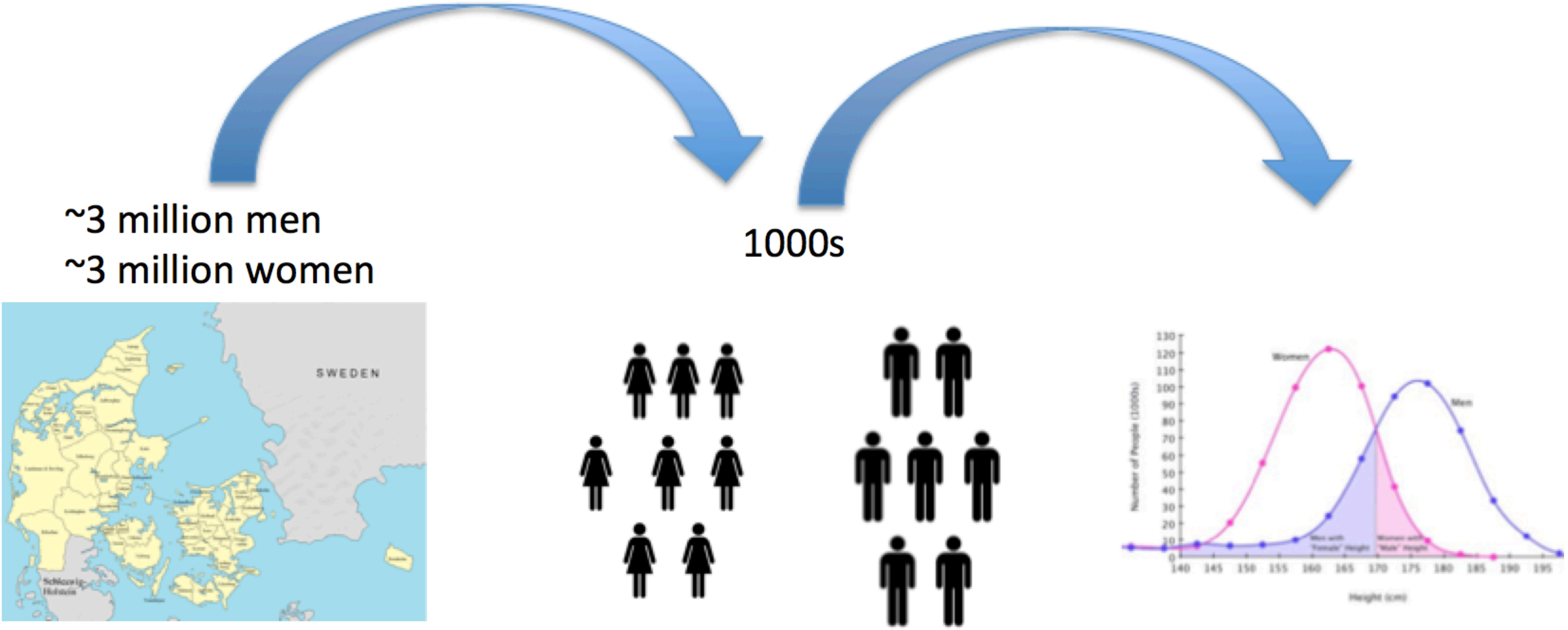


But what if we want to say something about whether *a single gene* is changing?

The basic idea of testing



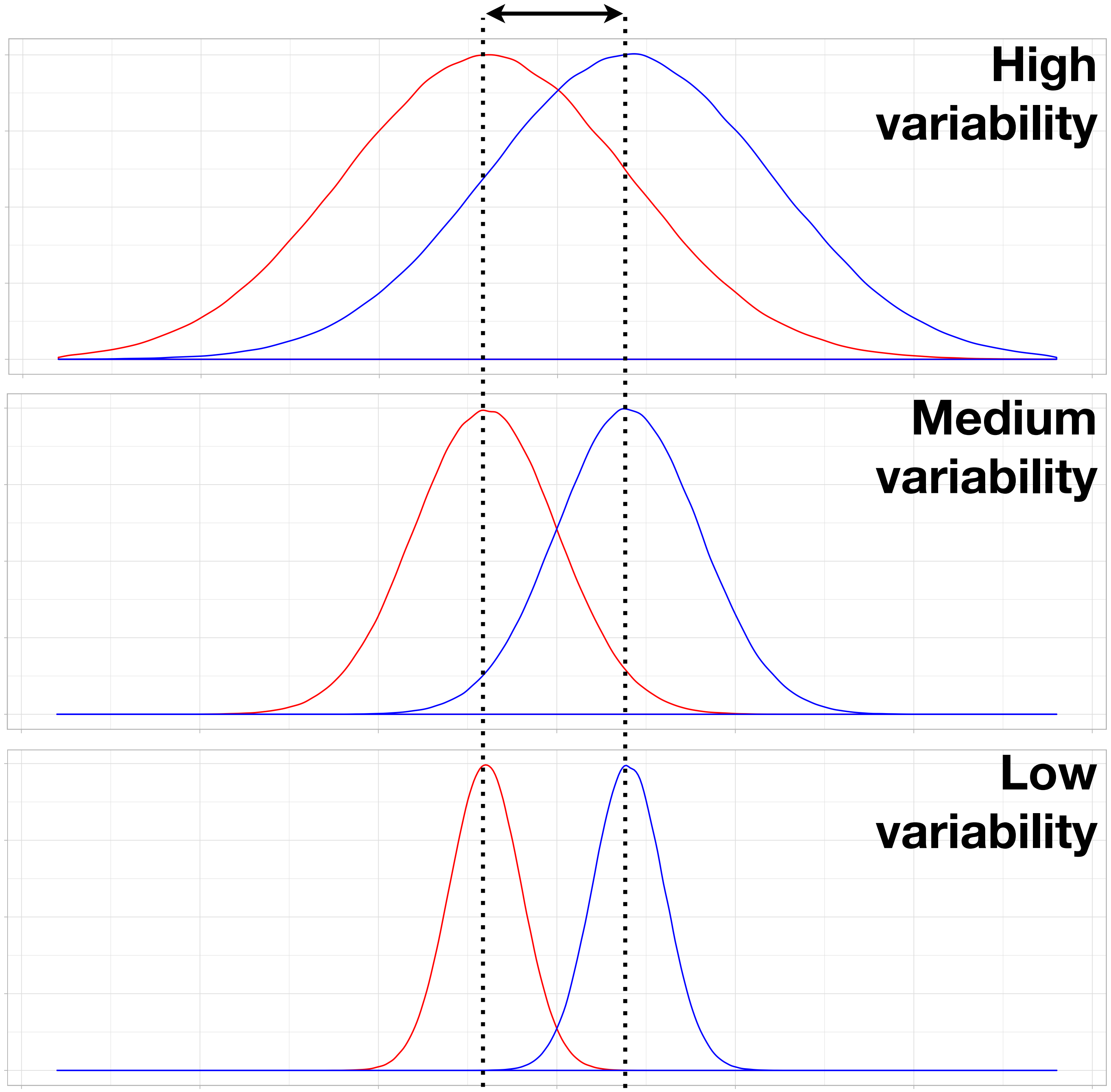
The basic idea of testing



~3 million men
~3 million women

1000s

Effect size vs significance

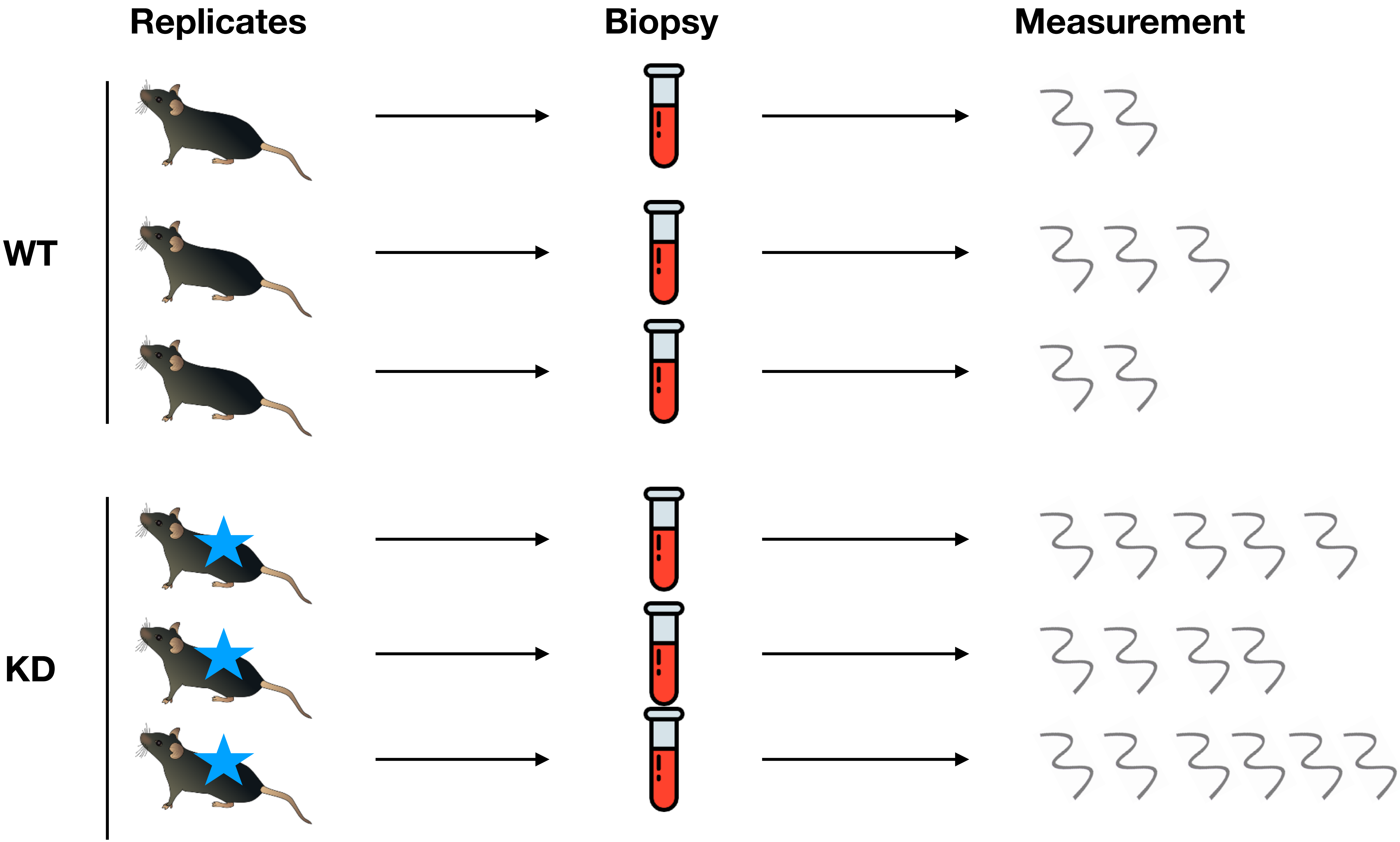


Effect Size:
Magnitude of the effect

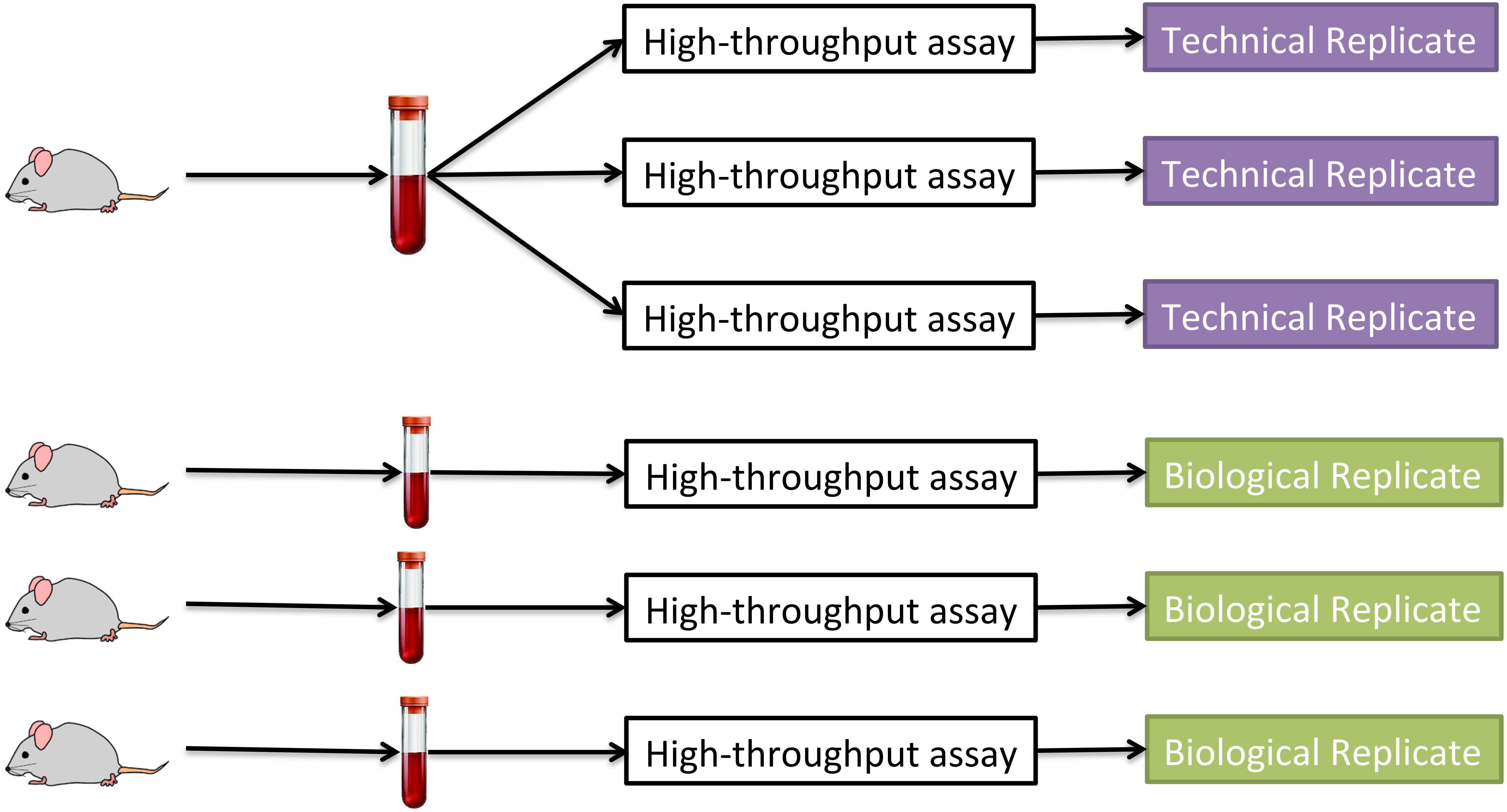
Significance (p-value):
How much we believe the effect

Large sample size + low variance > even small effect sizes can be significant!

The basic idea of testing



Different types of replication



High-throughput assay

Technical Replicate

High-throughput assay

Technical Replicate

High-throughput assay

Technical Replicate

High-throughput assay

Biological Replicate

High-throughput assay

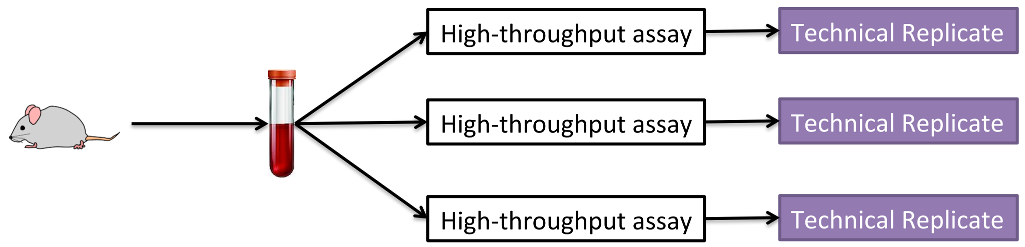
Biological Replicate

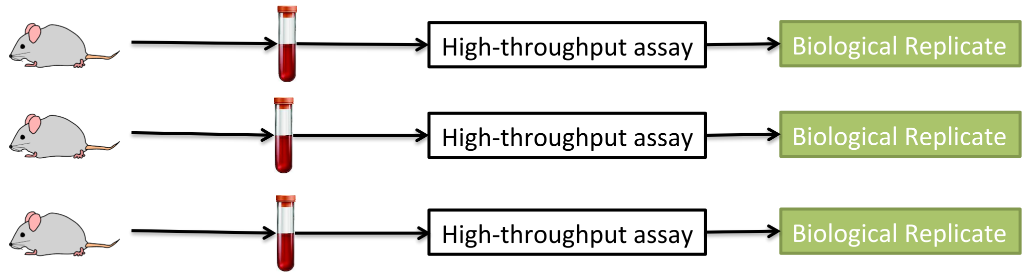
High-throughput assay

Biological Replicate

The use of technical replicates

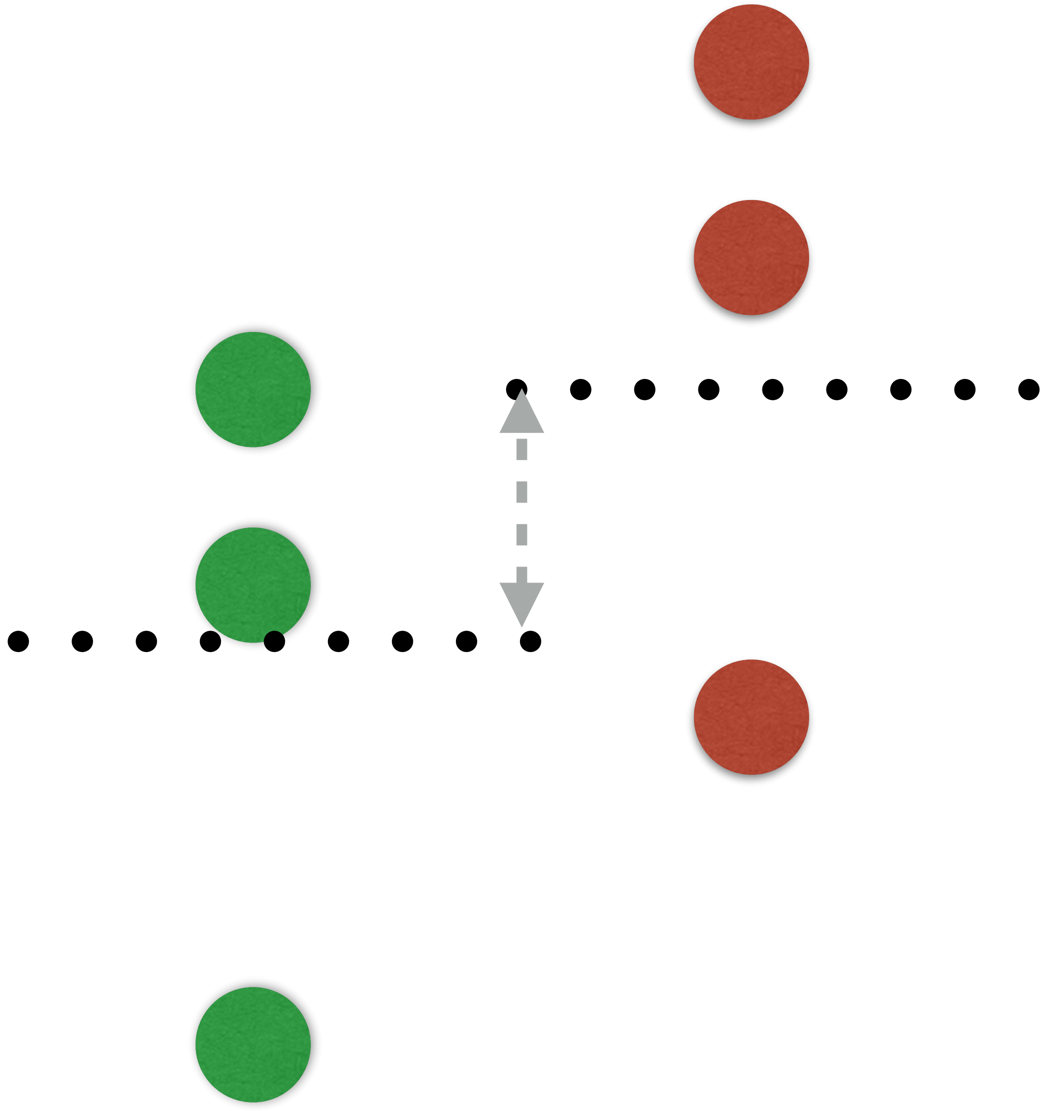
- **Technical replicates:**
  - Useful for EDA.
  - Measure technical noise.
  - Determine whether the sample pipeline is stable - technical replicates should be most similar.
  - Samples are not independent: Should not be (directly) included in the DE analysis.
- **Biological replicates:**
  - Useful for DE analysis.
  - Determine the biological variation in gene expression.





The inherent problem of testing in high-throughput data

- Very small sample size (usually 3 replicates per condition).
- Extremely high number of tests – problems with multiple testing.

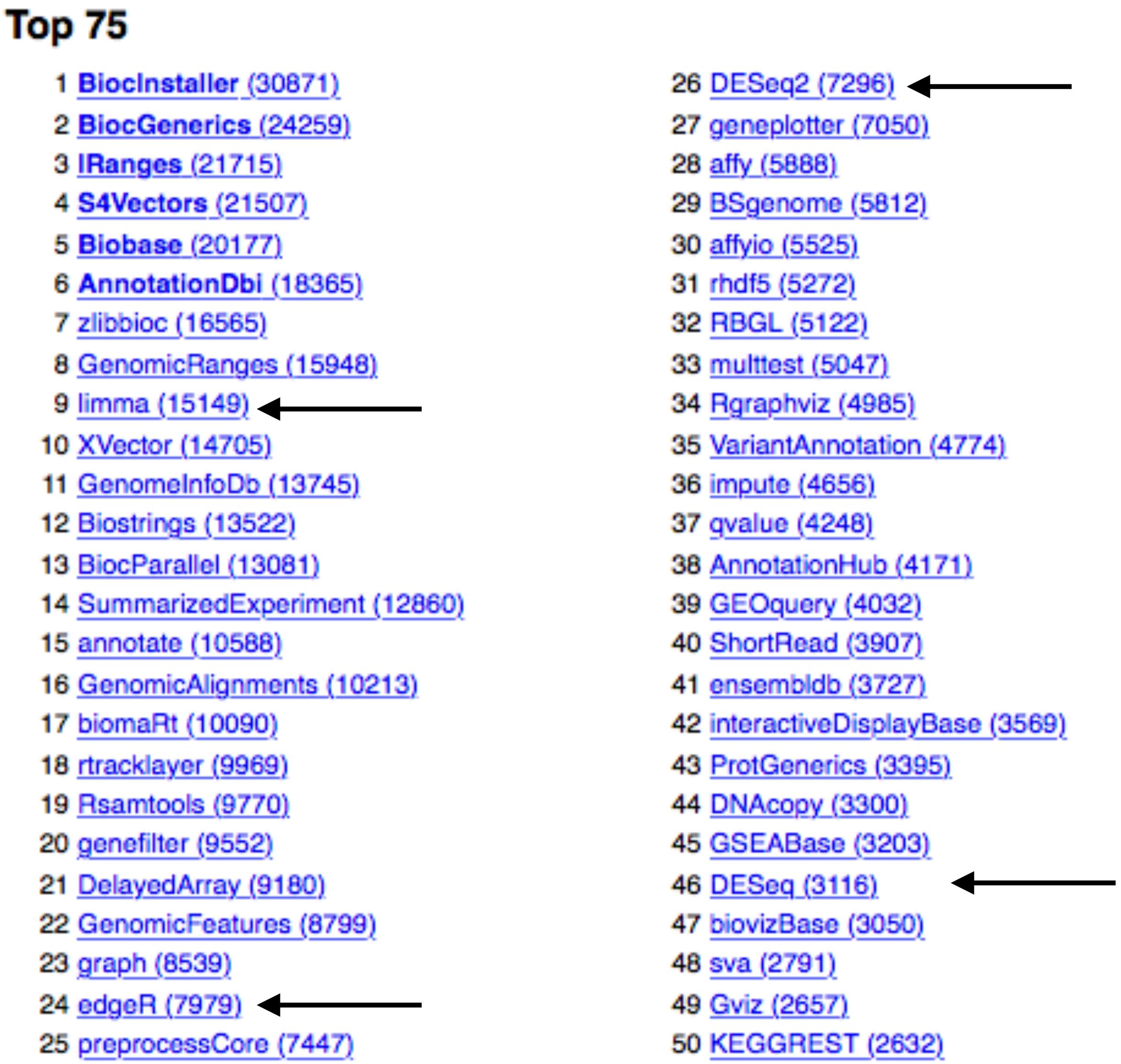


x 20.000

DESeq2

Why DESeq2?

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS



Top 75

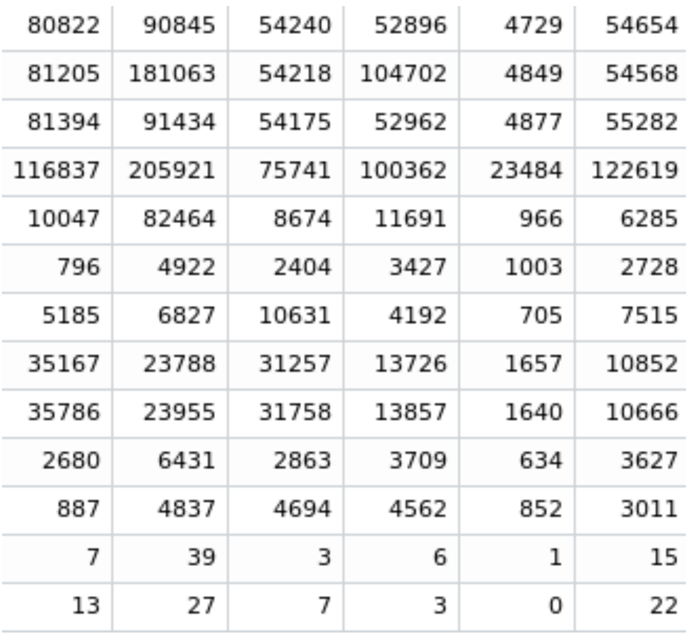
1	**BiocInstaller** (30871)	26	DESeq2 (7296)
2	**BiocGenerics** (24259)	27	geneplotter (7050)
3	**IRanges** (21715)	28	affy (5888)
4	**S4Vectors** (21507)	29	BSgenome (5812)
5	**Biobase** (20177)	30	affyio (5525)
6	**AnnotationDbi** (18365)	31	rhdf5 (5272)
7	zlibbioc (16565)	32	RBGL (5122)
8	GenomicRanges (15948)	33	multtest (5047)
9	limma (15149)	34	Rgraphviz (4985)
10	XVector (14705)	35	VariantAnnotation (4774)
11	GenomeInfoDb (13745)	36	impute (4656)
12	Biostrings (13522)	37	qvalue (4248)
13	BiocParallel (13081)	38	AnnotationHub (4171)
14	SummarizedExperiment (12860)	39	GEOquery (4032)
15	annotate (10588)	40	ShortRead (3907)
16	GenomicAlignments (10213)	41	ensembldb (3727)
17	biomaRt (10090)	42	interactiveDisplayBase (3569)
18	rtracklayer (9969)	43	ProtGenerics (3395)
19	Rsamtools (9770)	44	DNAcopy (3300)
20	genefilter (9552)	45	GSEABase (3203)
21	DelayedArray (9180)	46	DESeq (3116)
22	GenomicFeatures (8799)	47	biovizBase (3050)
23	graph (8539)	48	sva (2791)
24	edgeR (7979)	49	Gviz (2657)
25	preprocessCore (7447)	50	KEGGREST (2632)

- Third most downloaded R-package for DE
- Benchmarks show good performance in most cases
- Well made vignettes and guide available
- Defaults work well out of the box
- Same developers as Salmon, resulting in a highly optimised Salmon > tximport > DESeq2 pipeline.

What DESeq2 does (tl;dr)

Samples

Genes



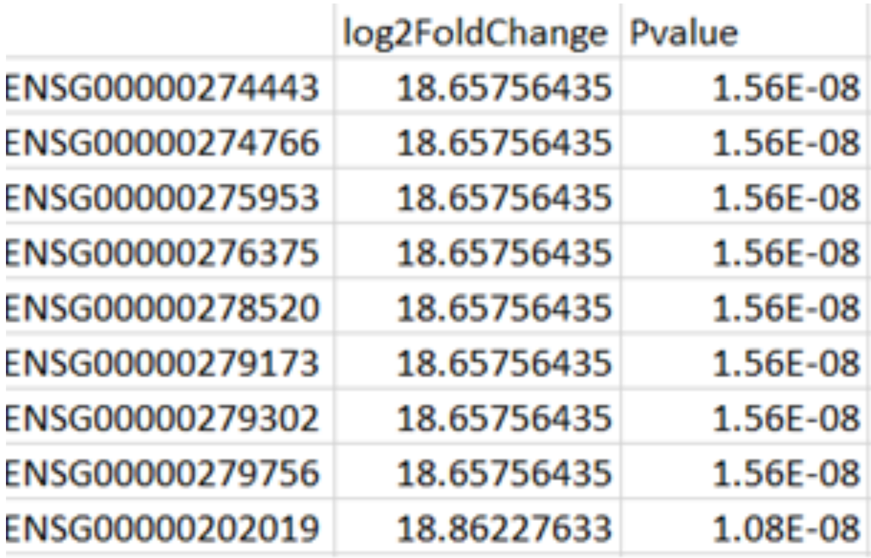
80822	90845	54240	52896	4729	54654
81205	181063	54218	104702	4849	54568
81394	91434	54175	52962	4877	55282
116837	205921	75741	100362	23484	122619
10047	82464	8674	11691	966	6285
796	4922	2404	3427	1003	2728
5185	6827	10631	4192	705	7515
35167	23788	31257	13726	1657	10852
35786	23955	31758	13857	1640	10666
2680	6431	2863	3709	634	3627
887	4837	4694	4562	852	3011
7	39	3	6	1	15
13	27	7	3	0	22

+

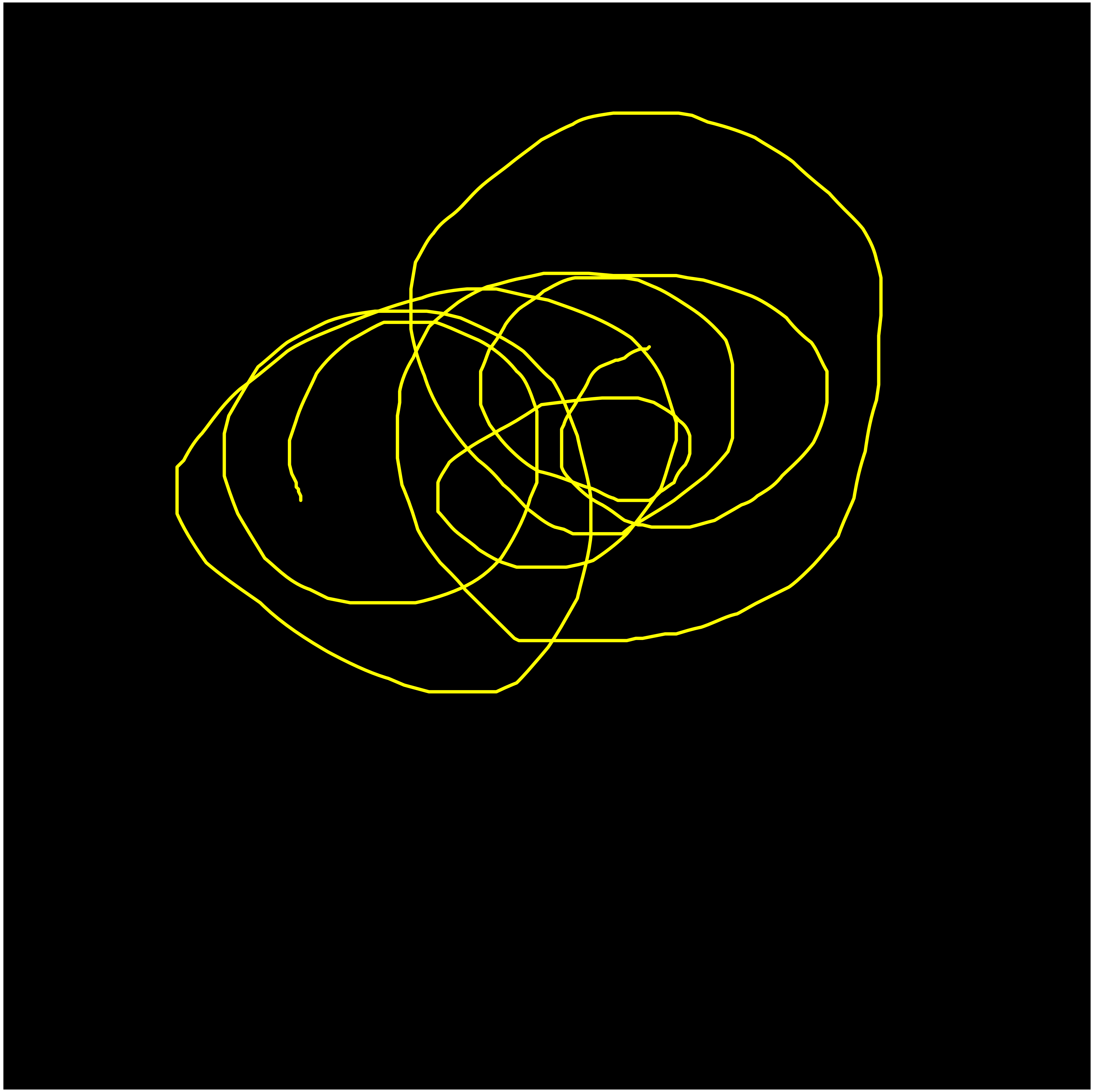
~ Experimental groups

Effect Size

Significance



	log2FoldChange	Pvalue
ENSG00000274443	18.65756435	1.56E-08
ENSG00000274766	18.65756435	1.56E-08
ENSG00000275953	18.65756435	1.56E-08
ENSG00000276375	18.65756435	1.56E-08
ENSG00000278520	18.65756435	1.56E-08
ENSG00000279173	18.65756435	1.56E-08
ENSG00000279302	18.65756435	1.56E-08
ENSG00000279756	18.65756435	1.56E-08
ENSG00000202019	18.86227633	1.08E-08



Components of DESeq2

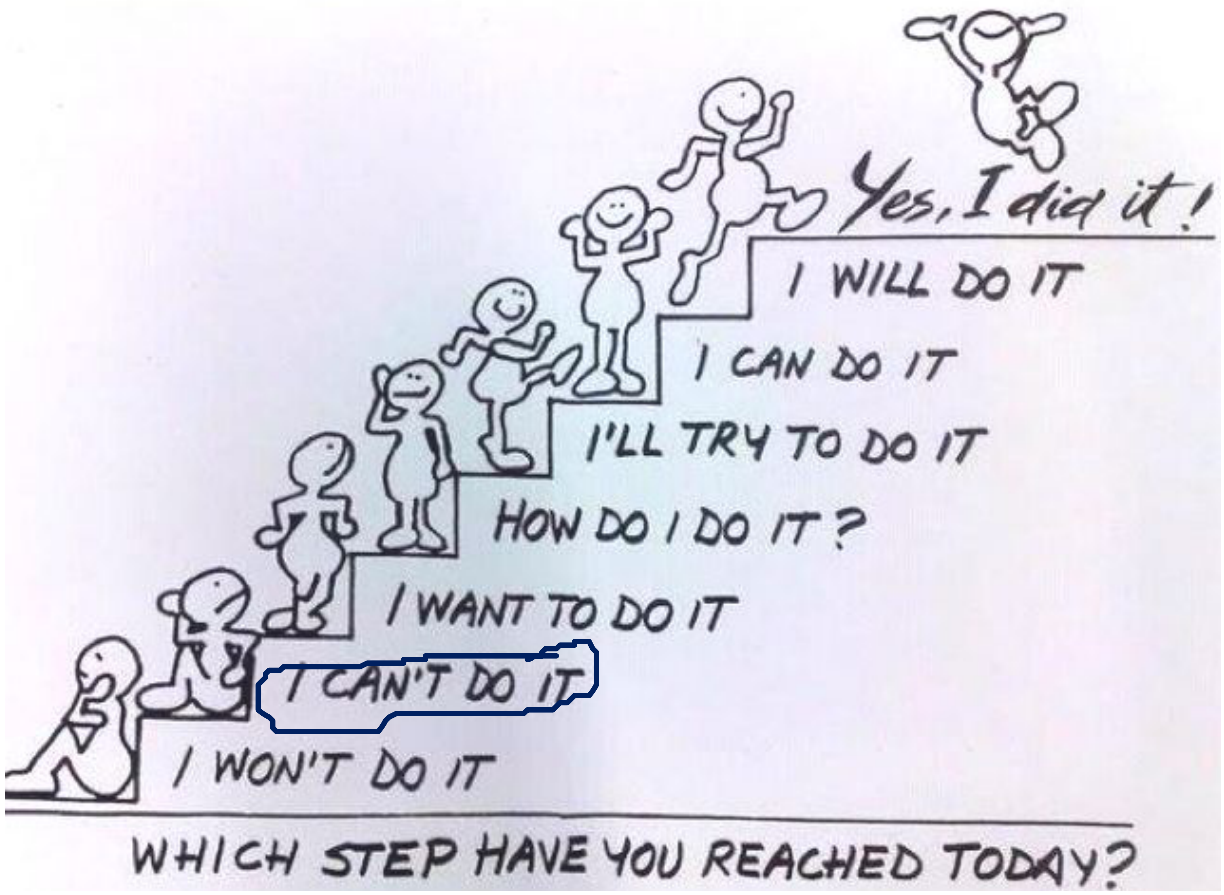
1. Normalisation.

2. Estimating dispersion.

3. Sharing information across genes.

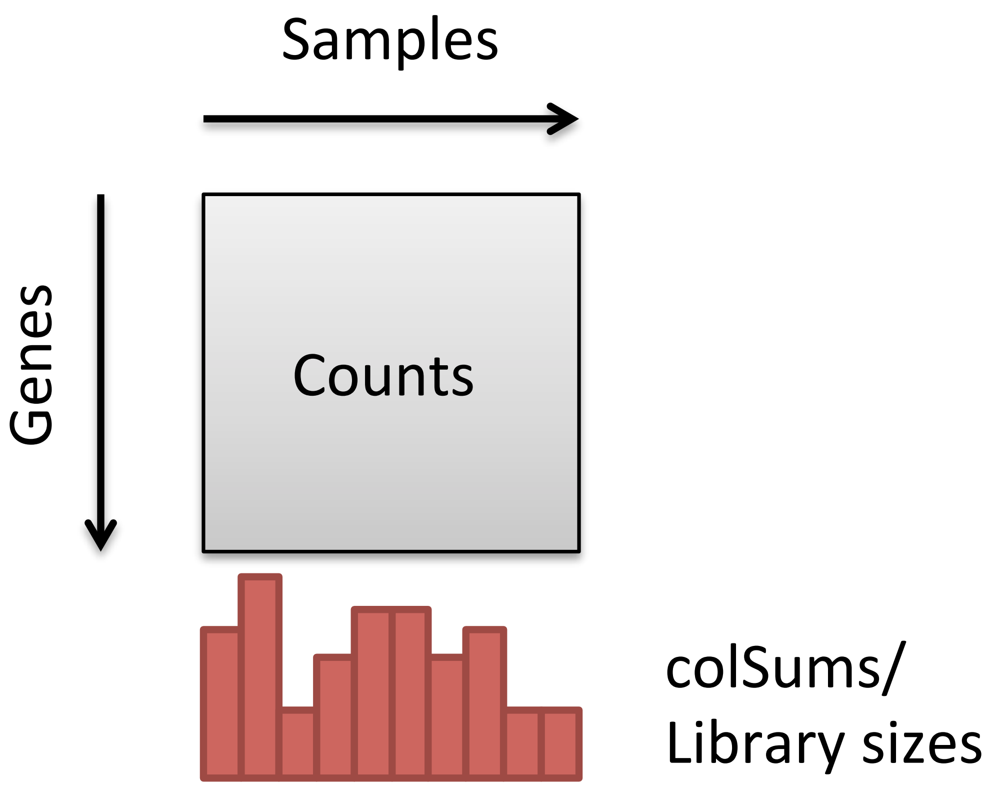
4. Testing (Using GLMs and Wald Tests).

5. Correction for multiple testing with independent filtering.



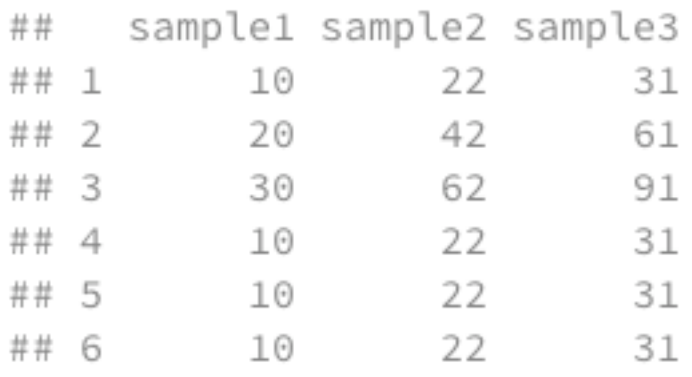
Counts **P**er **M**illion (**CPM**)

- As you already know, the individual samples in the EM cannot be compared before they have been **normalised**.
- The main reason for this is differences in **library size** or **sequencing depth** between samples in the same experiments.
- The basic way of dealing with this problem is to consider counts as fractions of total, usually scaled to a human scale unit, such as counts-per-million (cpm) or tags-per-million (tpm)

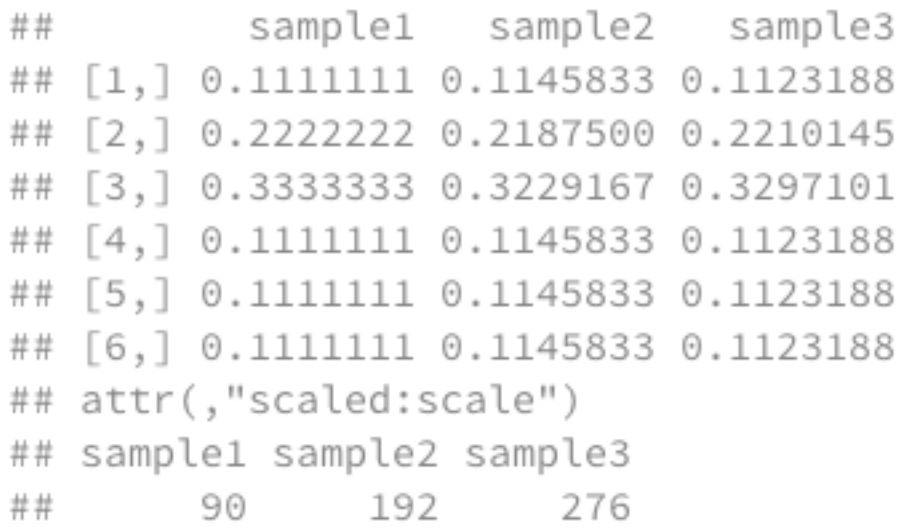


RNA composition

Without DE



```
##   sample1 sample2 sample3
## 1      10      22      31
## 2      20      42      61
## 3      30      62      91
## 4      10      22      31
## 5      10      22      31
## 6      10      22      31
```

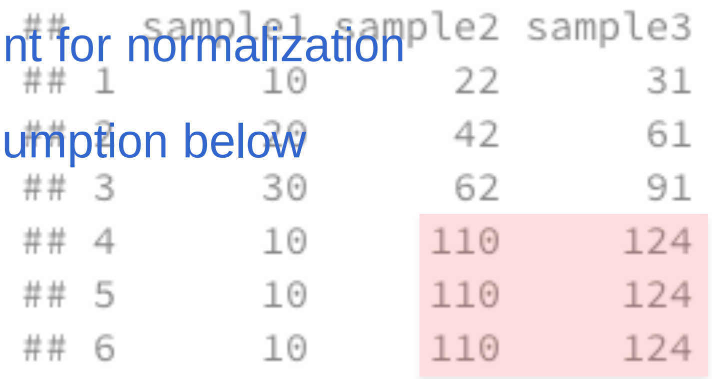


```
##        sample1   sample2   sample3
## [1,] 0.1111111 0.1145833 0.1123188
## [2,] 0.2222222 0.2187500 0.2210145
## [3,] 0.3333333 0.3229167 0.3297101
## [4,] 0.1111111 0.1145833 0.1123188
## [5,] 0.1111111 0.1145833 0.1123188
## [6,] 0.1111111 0.1145833 0.1123188
## attr(,"scaled:scale")
## sample1 sample2 sample3
##      90     192     276
```

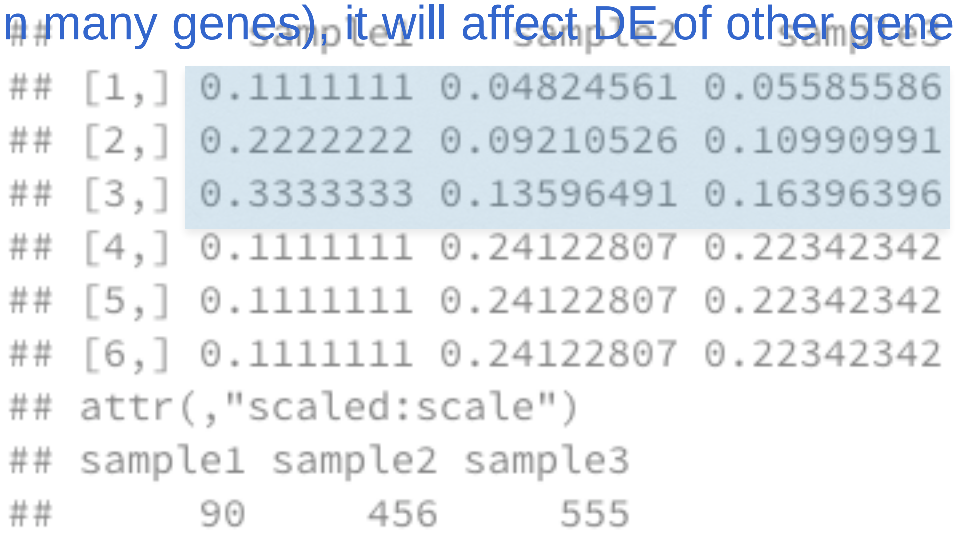
With DE

problems --> add a lot of count to many gene (eq to exp chnage in many genes), it will affect DE of other gene when we use total count for normalization

we need assumption below



```
##   sample1 sample2 sample3
## 1      10      22      31
## 2      20      42      61
## 3      30      62      91
## 4      10     110     124
## 5      10     110     124
## 6      10     110     124
```



```
##        sample1    sample2    sample3
## [1,] 0.1111111 0.04824561 0.05585586
## [2,] 0.2222222 0.09210526 0.10990991
## [3,] 0.3333333 0.13596491 0.16396396
## [4,] 0.1111111 0.24122807 0.22342342
## [5,] 0.1111111 0.24122807 0.22342342
## [6,] 0.1111111 0.24122807 0.22342342
## attr(,"scaled:scale")
## sample1 sample2 sample3
##      90     456     555
```

DESeq2 solves this by estimating normalisation factors.
Importantly, this **assumes that most genes are not DE!**

2. Dispersion

Y_{gj} is the count for gene *g* in sample *j*

$$Y_{gj} \sim NB\left(\mu_{gj}, \mu_{gj} + \mu_{gj}^{2}\phi_{g}\right)$$

negative binomial dist

Where u_{gj} is the mean expression:

$$\mu_{gj} = M_{j} p_{pj}$$

Where *M* is the library size and p_{pj} the fraction of the gene within M.

$$BCV = \sqrt{\phi_{g}}$$

Intuitively:

- Gene expression varies between samples due to sequencing noise (Modelled with a poisson distribution)
- Gene expression varies between samples due to biological variation
- DESeq2 models gene expression as a negative binomial distribution, where the biological coefficient of variation (BCV) measures excess variance compared to poisson

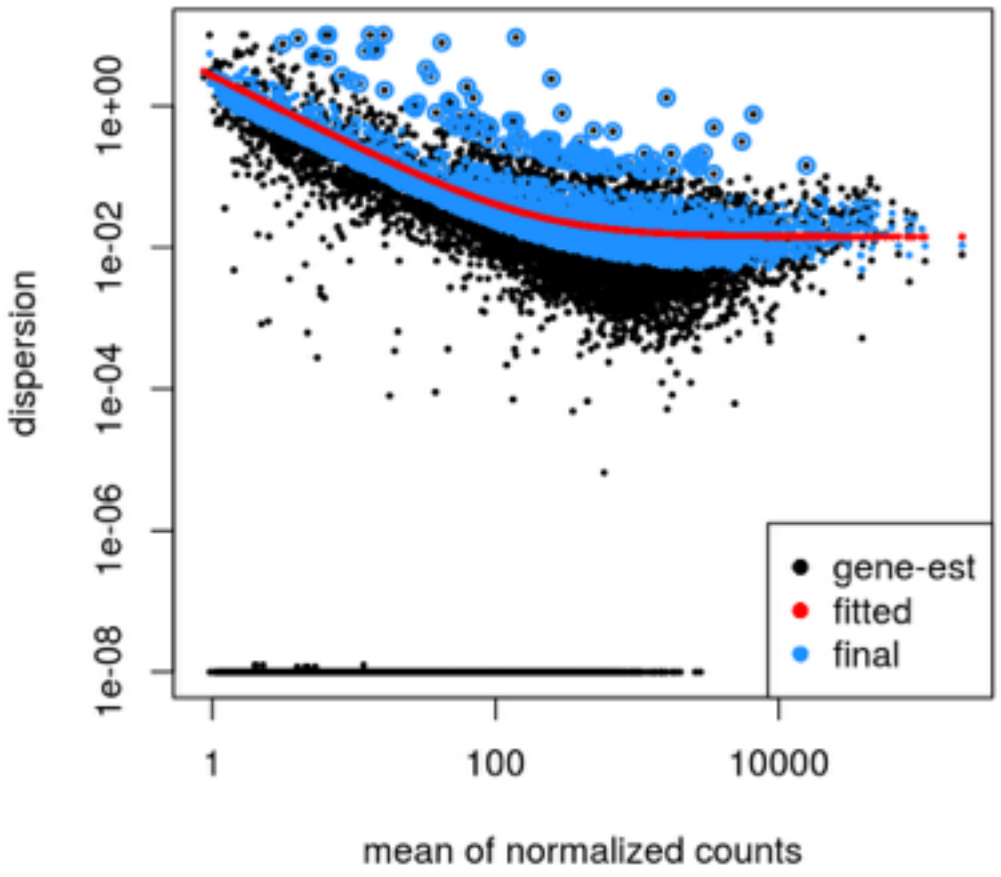
$$\text{Total CV}^2 = \text{Technical CV}^2 + \text{Biological CV}^2$$

3. Information sharing

1. It is difficult to estimate the variance of the expression of a single gene with very few replicates

2. DESeq2 solves this issue by assuming genes with the same expression tend to have the same variance (BCV)

3. By “shrinking” the individual estimates of variance towards the global trend DESeq2 obtains better downstream results.



** Additionally, DESeq2 also filters out outliers using Cook’s distance*

4. Testing

Black box warning!

DESeq2 uses generalised linear models (GLMs) to test for differential expression, allowing for highly complex experimental setups.

Importantly, this allows for correction of *batch effects* as we discussed in the PCA lecture.

Here, we will only deal with simple group-wise comparisons in abstract terms

If you want to know more, take the Advanced Bioinformatic Course in block 1!

4. Testing

- Once shrunken dispersions estimates have been obtained, DESeq2 uses the Wald test to test for differences between groups

- Hypotheses:

  - H_0: $\log_2$FC between two groups is zero

  - H_1: $\log_2$FC between two groups is not zero

- Alternatively, the threshold can also be set to something other than zero:

  - H_0: $\log_2$FC between two groups is less than 1 (in either direction)

  - H_1: $\log_2$FC between two groups is more than 1 (in either direction)

- The results is an estimate of the **logFC** and a **p-value** for each gene!

Summary on Dispersion and Testing

Var(Expression) = Across group variability + Biological Variability + Measurement Error

Effects of conditions, this is what we want to find.

This is what we test for!

Variation in expression between individuals.

Shared across genes.

Dominant for highly expressed genes

Variation in expression due to sequencing.

Sample independent (hopefully!).

Dominant for lowly expressed genes.

Correction for multiple testing

- As we are testing thousands of genes, we must correct for multiple testing.

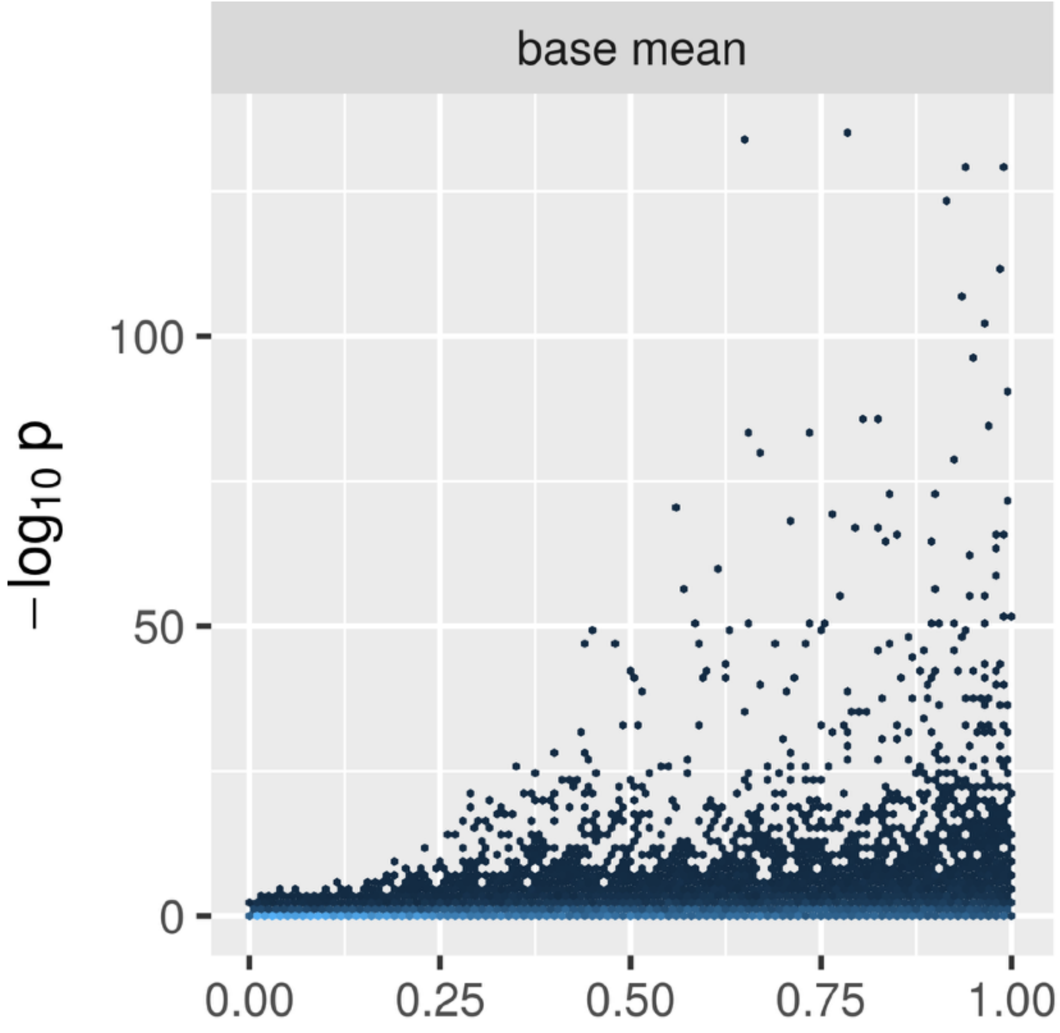
- DESeq2 be defaults uses an extension of the FDR-correction procedure based on *independent filtering:*

  - Due to sequencing noise, lowly expressed features will often vary too much to detect DE even very high logFCs.

  - Including many extra lowly expressed genes therefore decreases the number of DE genes due to more strict correction for multiple testing.

  - DESeq2 removes lowly expressed genes before FDR-correction to maximise the number of genes detected as DE.

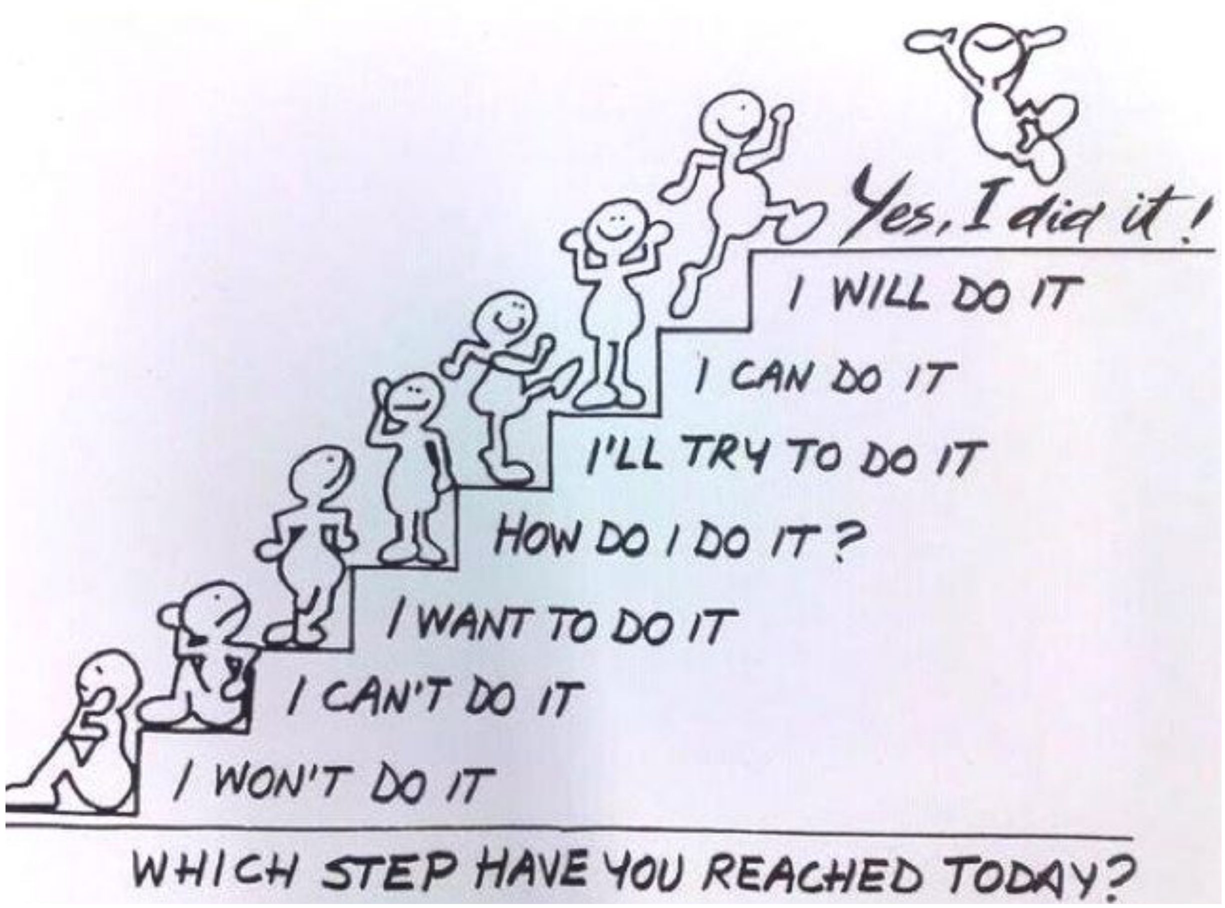
- **Statistical note:** This is only valid if the filtering statistic (mean expression in this case) is *independent under the null hypothesis* (there is no relation between mean expression and tendency to be DE)



Genes ranked by mean expression

Components of DESeq2

1. Normalisation.
2. Estimating dispersion.
3. Sharing information across genes.
4. Testing (Using GLMs and Wald Tests).
5. Correction for multiple testing with independent filtering.



Other functionalities

- Variance-stabilizing transformations: An alternative to the standard log2-transformation (vst and rlog).
- Shrink *both* dispersion and logFCs
- Make PCA-plots
- Use a more advanced independent filtering approach called independent hypothesis weighting (IHW)
- Test complex multifactorial designs using Likehood ratio tests.
- New: Tissue decomposition (unmix)

DESeq2 practical

- Example of complete analysis starting from a count matrix:
  - Setting up the data.
  - Running the DESeq2 functions.
  - Producing essential diagnostic plots.
  - Inspecting the results.
- You will redo the analysis on another dataset
- Demonstration of how to use tximport to import Salmon quantifications into DESeq2.

Warning!

- DESeq2 is based on Bioconductor
- Bioconductor uses a quite rare R-programming system called S4.
- **Bioconductor IS NOT the tidyverse!**
- **DO NOT** expect everything to be a tibble!
- You might to read help files for the functions to know what to do!

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Next lecture

- Now you have used DESeq2 to find genes that show statistically significant changes between groups.

- Can you say something about these sets of genes?

- You can look at whether the genes as a group are associated to some known biological pathway or process.

- Monday you will look at ***GO-term enrichment.***