CEITEC

Central European Institute of Technology

BRNO | CZECH REPUBLIC

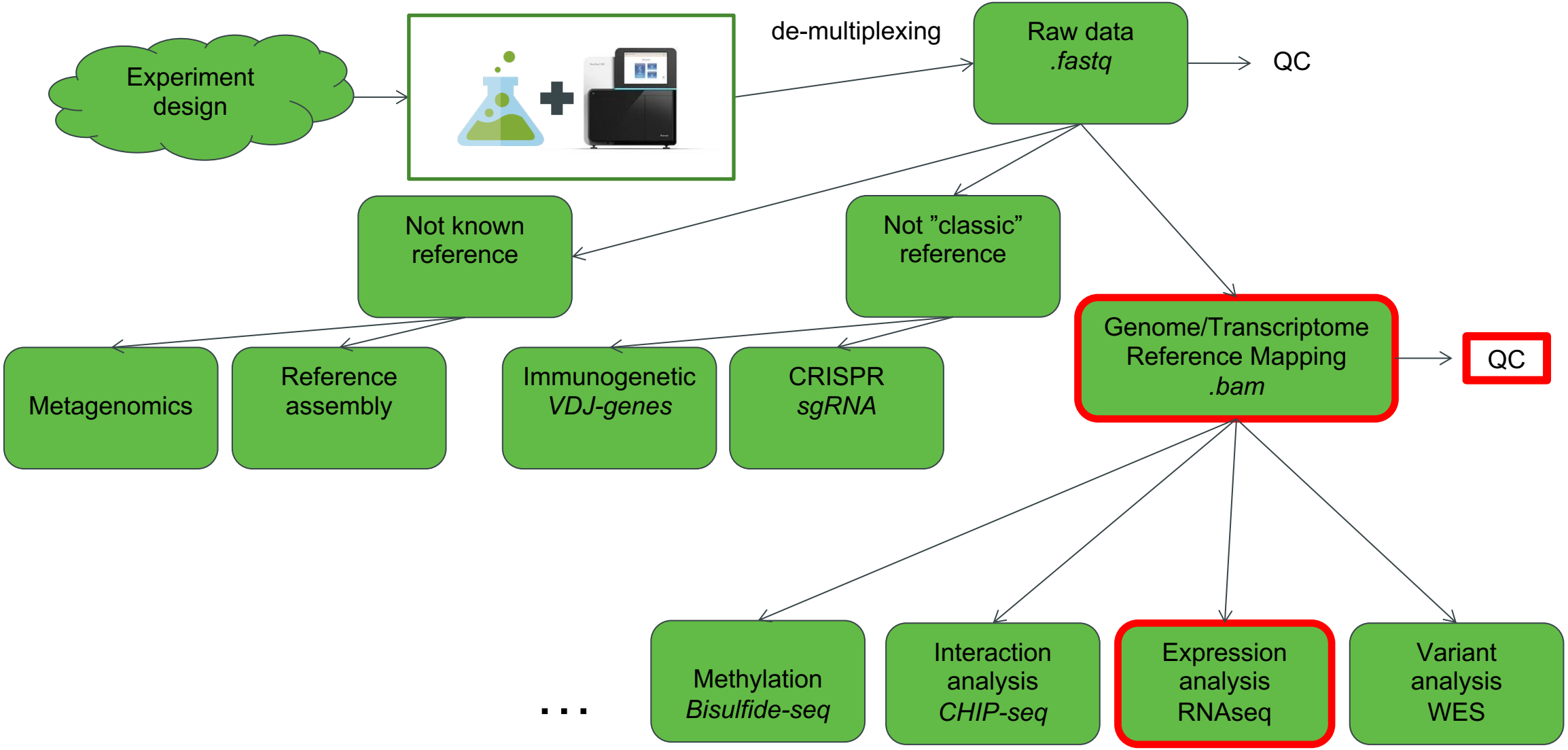
UNIVERSITAS MASARYKIANA BRUNENSIS

Modern methods for genome analysis (PřF:Bi7420)

Lecture 6 : RNA-seq differential expression

Vojta Bystry
vojtech.bystry@ceitec.muni.cz

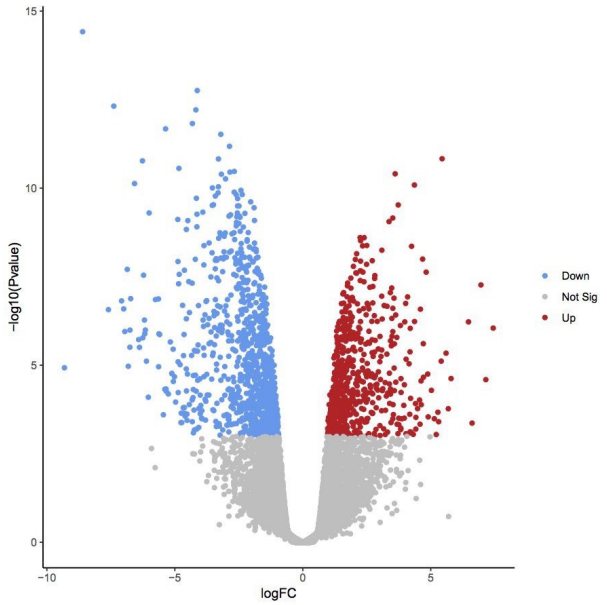
NGS data analysis

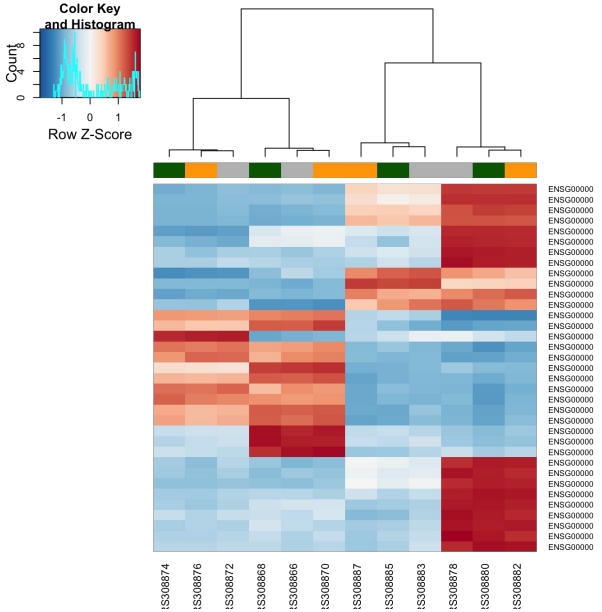


Why RNA-seq?

The goal of RNA-seq is often to perform differential expression testing to determine which genes are expressed at different levels between conditions. These genes can offer biological insight into the processes affected by the condition(s) of interest.

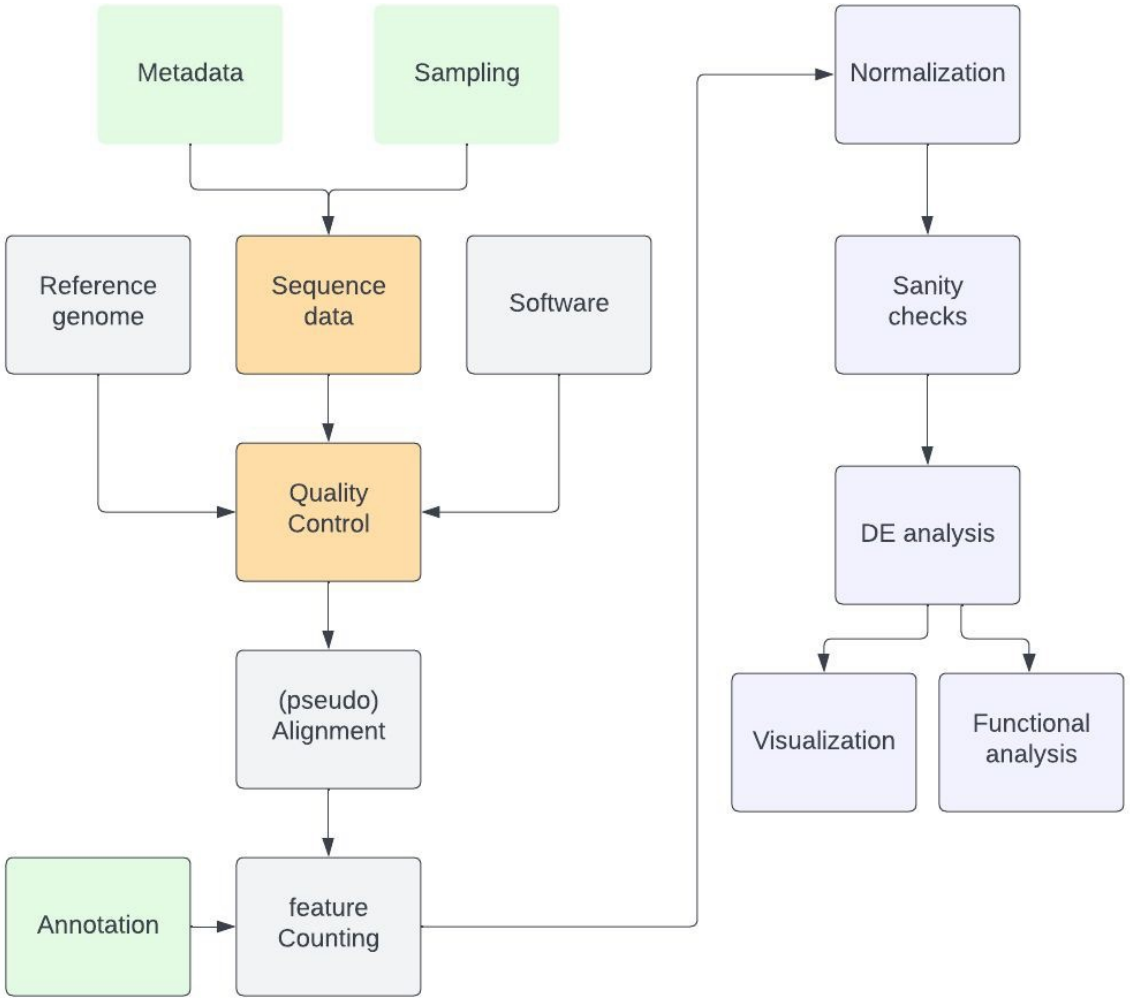
Great resource that has made the bulk of this talk:
https://github.com/hbctraining/DGE_workshop/blob/master/lessons/01_DGE_setup_and_overview.md
Lean tutorial for mostly wetlab humans:
https://git.embl.de/provazni/rna-seq-tutorial/-/tree/EMBLVM?ref_type=heads



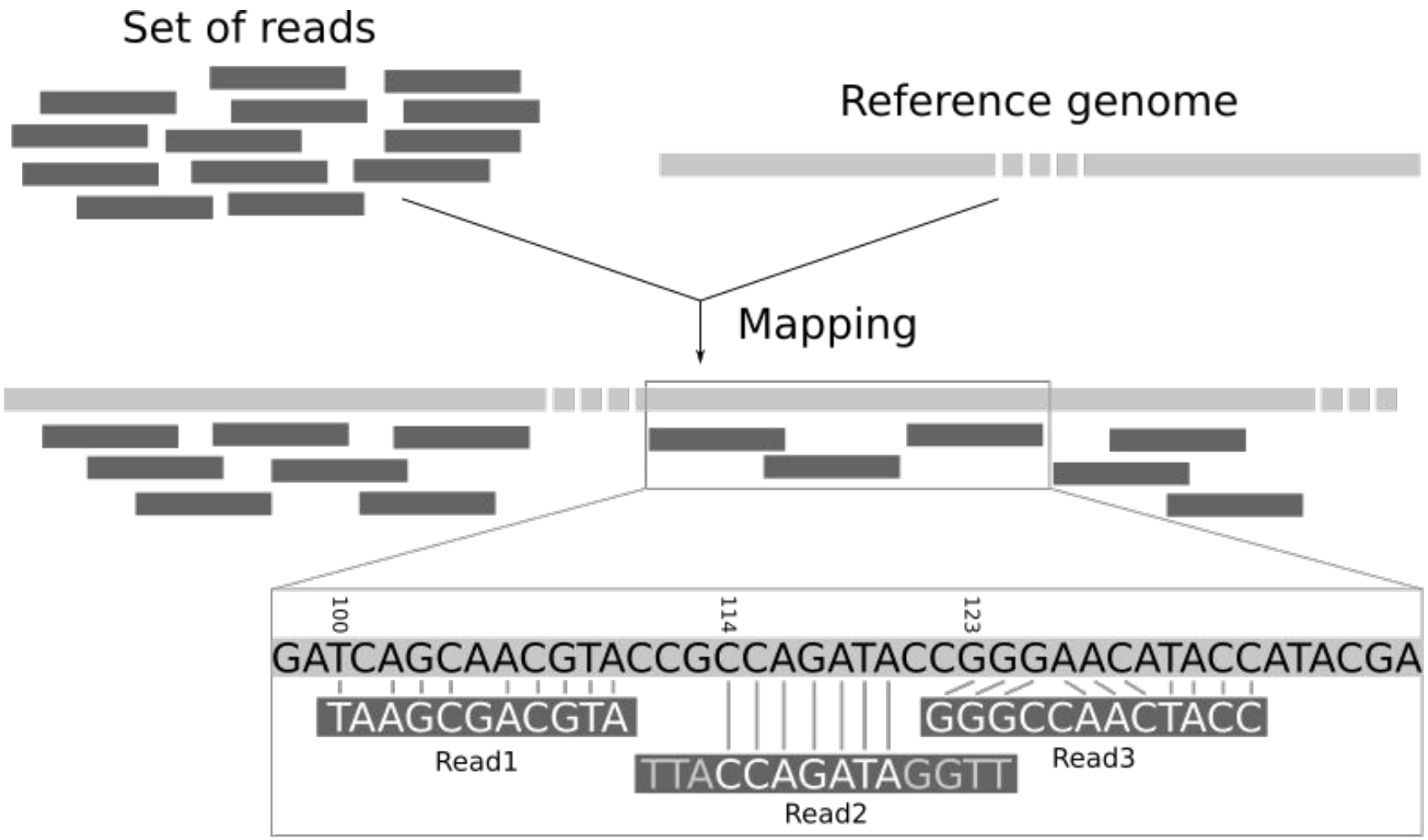


CEITEC

RNA-seq workflow



Read Alignment



https://training.galaxyproject.org/training-material/topics/sequence-analysis/tutorials/mapping/tutorial.html

Read Alignment

Many different aligners/pseudo-aligners:

- Eland, Maq, Bowtie, Bowtie2, BWA, SOAP, SSAHA, TopHat, SpliceMap, Novoalign, STAR, GSNAP …
- Kallisto, Salmon

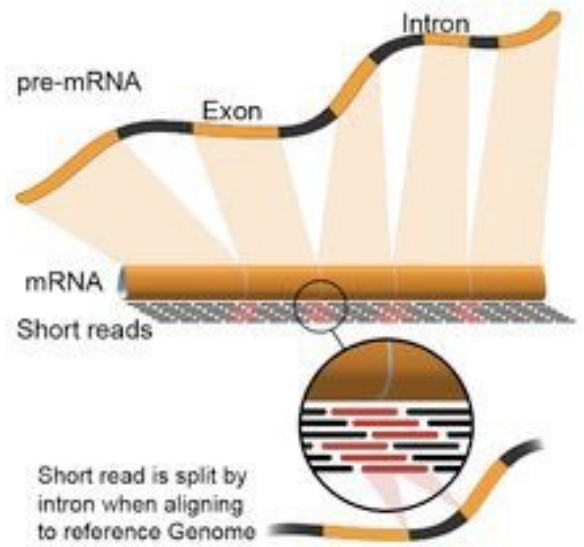
Main differences:

- Publication year, maturity, development after publication, popularity
- usage of base-call qualities, calculation of mapping qualities
- speed-vs-sensitivity trade-off
- suitability for RNA-Seq (“spliced alignment”)
- suitability for special tasks

Expected alignment percentage:

- 70% to 90% on genome
- slightly lower on transcriptome

Genome or annotated transcriptome



https://en.wikipedia.org/wiki/RNA-Seq

Alignment

- Mapping to genome or transcriptome?
- Genome
  - Requires spliced alignment
  - Can find novel genes/isoforms/exons
  - Information about whole genome/transcriptome
- Transcriptome
  - No spliced alignments necessary
  - Many reads will map to multiple transcripts (shared exons)
  - Cannot find anything new
  - Difficult to determine origin of reads (multiple copies of transcripts)

Duplication removal - UMI

- PCR duplicates
- Optical duplicates

- How the tools recognize duplicates
  - Maps to the exact same place
- Problem is it could be identical fragment not PCR duplicate
- UMI helps
  - Maps to the exact same place
  - AND have identical UMI sequence

Post-alignment QC

- Number of mapped reads - unique + multi mapped
- Mapped locations – intron, exon, intergenic
- Duplication rates
- Library strand specificity
- Captured biotypes
- Contamination (rRNA, non-self)
- 5’ to 3’ end coverage bias

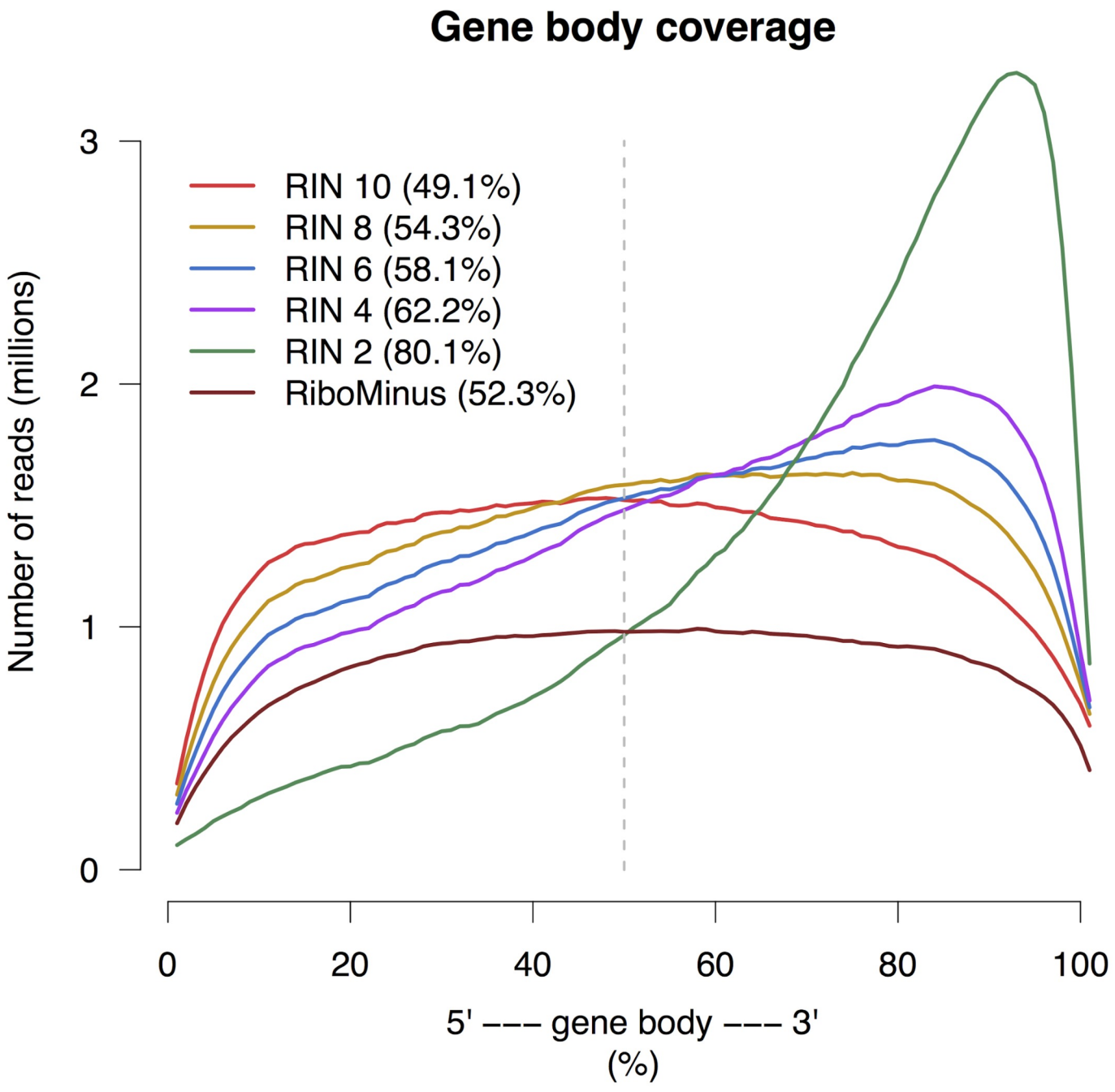
Post-alignment QC - Tools

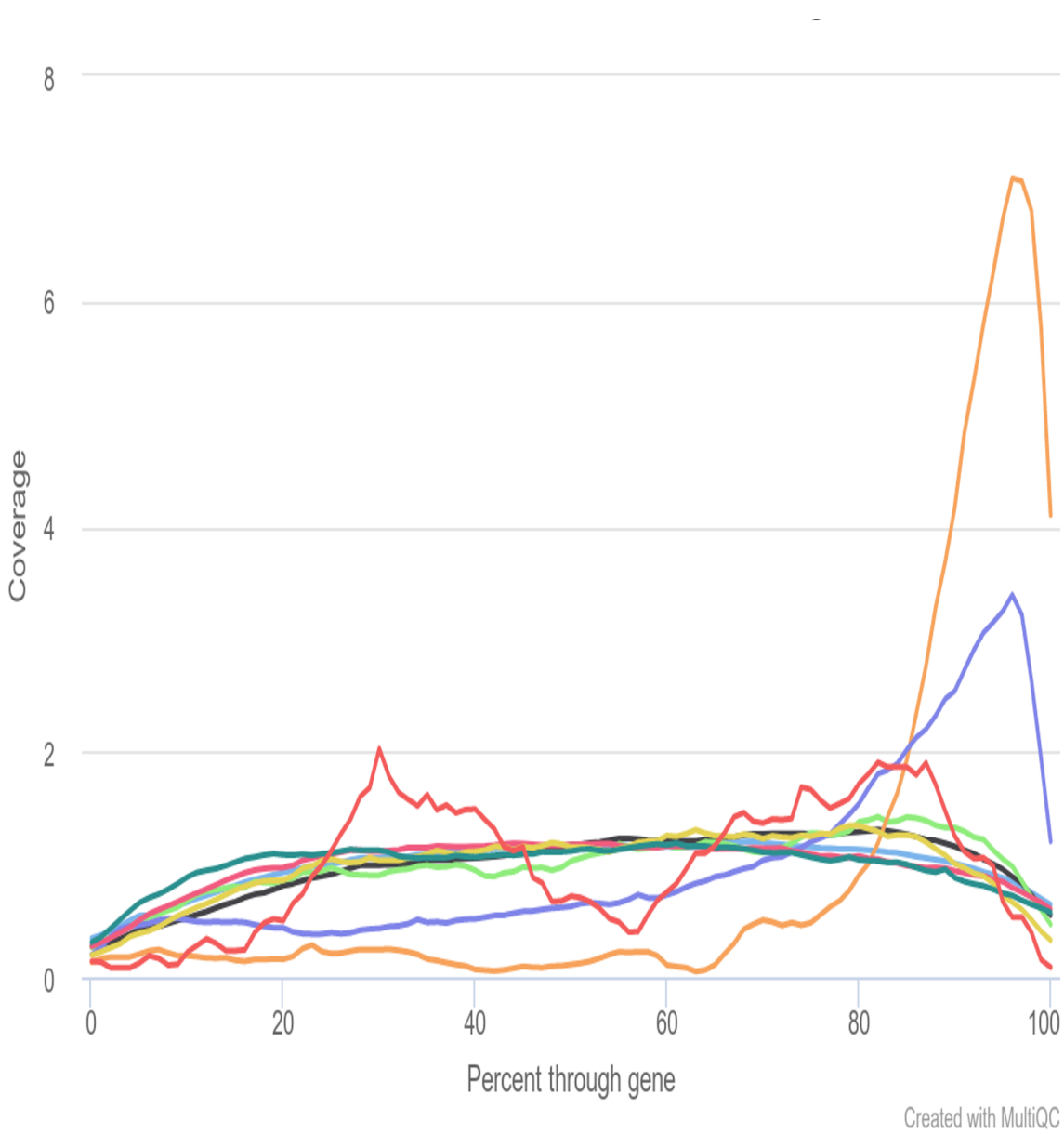
- Aligner report
  - `STAR` – most direct assesment
- General QC tools
  - `RSeQC`
  - `Picard`
  - `Qualimap`
- Feature counting tools
  - `featureCounts`
  - `RSEM`
- Non-aligment tools
  - `FastQ screen`
  - `Biobloom`

Note: Gene body coverage

- Often, libraries with high fragmentation (and low RIN numbers) combined with polyA selection might have strong 3’ end bias
  - This is a result of polyA “pulled” fragments
- Some kits, however, target only the polyA tail or sequences close to it
  - An example is Lexogen QuantSeq which sequences only one read per mRNA molecule close to polyA tail

Source: Sigurgeirsson et al. PLoS ONE 2014





Feature counting

- Now, when we know our alignments are solid we need to get the number of reads mapped to a gene (or other feature)
  - From there, we can calculate the differential expression
- The question is, how do we summarize the counts
  - Do we want only uniquely mapped reads
  - Do we want also multi mapped? And how do we assign them? All? One random? Somehow else?
  - And what if we have multiple genes which overlap each other?

Strand specific library

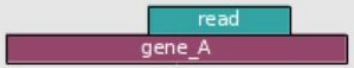
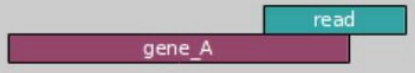
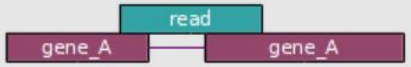
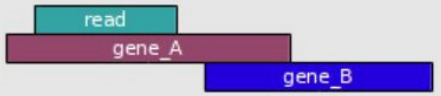
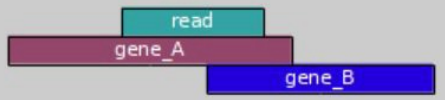
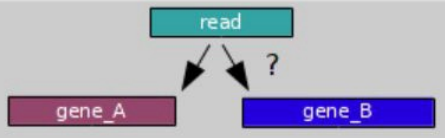
- We can basically have three strand specificities
  - **Non stranded/Unstranded** - not very common anymore
    - Direction of the read mapping is completely random (50/50)
  - **Forward (sense) stranded** - common for target kits and “bacterial kits”
    - Direction of the read mapping is the **same** as the gene it originates from
  - **Reverse (antisense) stranded** - “default” for Illumina and NEB kits
    - Direction of the read mapping is the **opposite** as the gene it originates from
- In case of paired-end sequencing it’s measure by the first (R1) read orientation (FR, RF)

Gene Counts

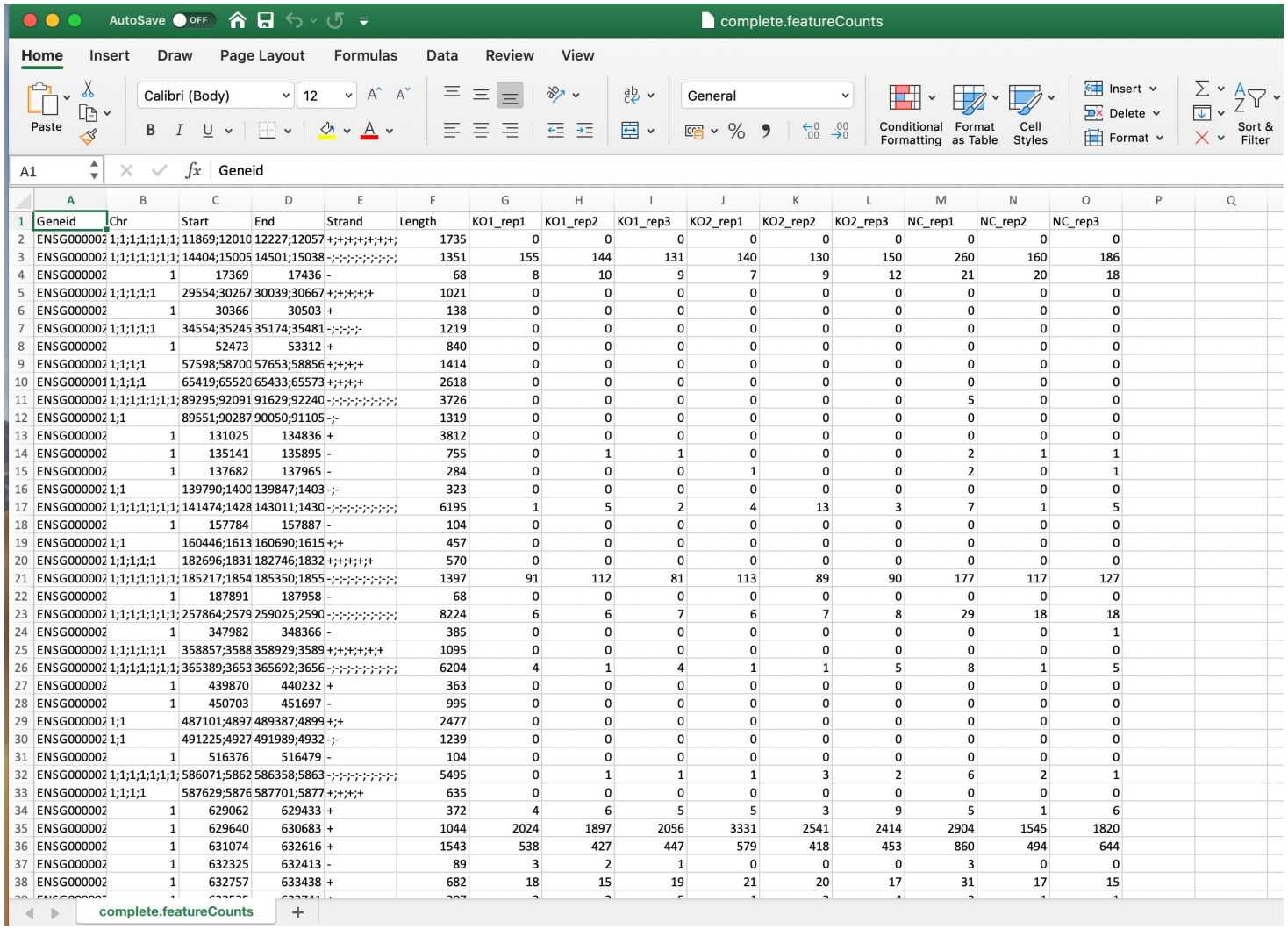
Tools: HTSeq-count, featureCounts, salmon, kallisto, etc.

Be careful about:

- Correct annotation
- What you do with multimappers

	union	intersection_strict	intersection_nonempty
	gene_A	gene_A	gene_A
	gene_A	no_feature	gene_A
	gene_A	no_feature	gene_A
	gene_A	gene_A	gene_A
	gene_A	gene_A	gene_A
	ambiguous (both genes with --nonunique all)	gene_A	gene_A
	ambiguous (both genes with --nonunique all)		
	alignment_not_unique (both genes with --nonunique all)		

Feature count results



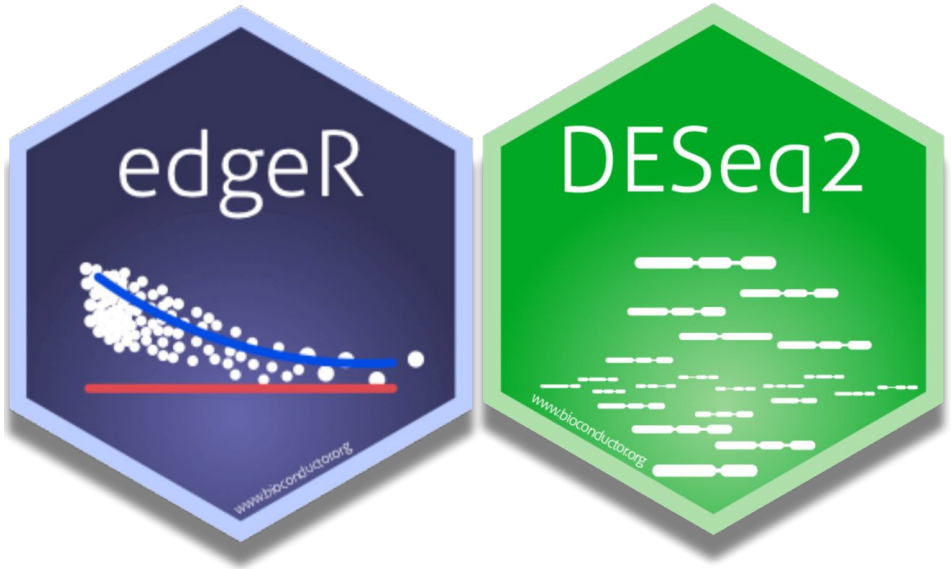
complete.featureCounts

Geneid	Chr	Start	End	Strand	Length	KO1_rep1	KO1_rep2	KO1_rep3	KO2_rep1	KO2_rep2	KO2_rep3	NC_rep1	NC_rep2	NC_rep3
ENSG000002	1;1;1;1;1;1;1;	11869;12010	12227;12057	+;+;+;+;+;+;+;	1735	0	0	0	0	0	0	0	0	0
ENSG000002	1;1;1;1;1;1;1;	14404;15005	14501;15038	-;-;-;-;-;-;-;	1351	155	144	131	140	130	150	260	160	186
ENSG000002	1	17369	17436	-	68	8	10	9	7	9	12	21	20	18
ENSG000002	1;1;1;1;1	29554;30267	30039;30667	+;+;+;+;+	1021	0	0	0	0	0	0	0	0	0
ENSG000002	1	30366	30503	+	138	0	0	0	0	0	0	0	0	0
ENSG000002	1;1;1;1;1	34554;35245	35174;35481	-;-;-;-;-	1219	0	0	0	0	0	0	0	0	0
ENSG000002	1	52473	53312	+	840	0	0	0	0	0	0	0	0	0
ENSG000002	1;1;1;1	57598;58700	57653;58856	+;+;+;+	1414	0	0	0	0	0	0	0	0	0
ENSG000001	1;1;1;1	65419;65520	65433;65573	+;+;+;+	2618	0	0	0	0	0	0	0	0	0
ENSG000002	1;1;1;1;1;1;1;	89295;92091	91629;92240	-;-;-;-;-;-;-;	3726	0	0	0	0	0	0	5	0	0
ENSG000002	1;1	89551;90287	90050;91105	-;-	1319	0	0	0	0	0	0	0	0	0
ENSG000002	1	131025	134836	+	3812	0	0	0	0	0	0	0	0	0
ENSG000002	1	135141	135895	-	755	0	1	1	0	0	0	2	1	1
ENSG000002	1	137682	137965	-	284	0	0	0	1	0	0	2	0	1
ENSG000002	1;1	139790;1400	139847;1403	-;-	323	0	0	0	0	0	0	0	0	0
ENSG000002	1;1;1;1;1;1;1;	141474;1428	143011;1430	-;-;-;-;-;-;-;	6195	1	5	2	4	13	3	7	1	5
ENSG000002	1	157784	157887	-	104	0	0	0	0	0	0	0	0	0
ENSG000002	1;1	160446;1613	160690;1615	+;+	457	0	0	0	0	0	0	0	0	0
ENSG000002	1;1;1;1;1	182696;1831	182746;1832	+;+;+;+;+	570	0	0	0	0	0	0	0	0	0
ENSG000002	1;1;1;1;1;1;1;	185217;1854	185350;1855	-;-;-;-;-;-;-;	1397	91	112	81	113	89	90	177	117	127
ENSG000002	1	187891	187958	-	68	0	0	0	0	0	0	0	0	0
ENSG000002	1;1;1;1;1;1;1;	257864;2579	259025;2590	-;-;-;-;-;-;-;	8224	6	6	7	6	7	8	29	18	18
ENSG000002	1	347982	348366	-	385	0	0	0	0	0	0	0	0	1
ENSG000002	1;1;1;1;1;1	358857;3588	358929;3589	+;+;+;+;+;+	1095	0	0	0	0	0	0	0	0	0
ENSG000002	1;1;1;1;1;1;1;	365389;3653	365692;3656	-;-;-;-;-;-;-;	6204	4	1	4	1	1	5	8	1	5
ENSG000002	1	439870	440232	+	363	0	0	0	0	0	0	0	0	0
ENSG000002	1	450703	451697	-	995	0	0	0	0	0	0	0	0	0
ENSG000002	1;1	487101;4897	489387;4899	+;+	2477	0	0	0	0	0	0	0	0	0
ENSG000002	1;1	491225;4927	491989;4932	-;-	1239	0	0	0	0	0	0	0	0	0
ENSG000002	1	516376	516479	-	104	0	0	0	0	0	0	0	0	0
ENSG000002	1;1;1;1;1;1;1;	586071;5862	586358;5863	-;-;-;-;-;-;-;	5495	0	1	1	1	3	2	6	2	1
ENSG000002	1;1;1;1	587629;5876	587701;5877	+;+;+;+	635	0	0	0	0	0	0	0	0	0
ENSG000002	1	629062	629433	+	372	4	6	5	5	3	9	5	1	6
ENSG000002	1	629640	630683	+	1044	2024	1897	2056	3331	2541	2414	2904	1545	1820
ENSG000002	1	631074	632616	+	1543	538	427	447	579	418	453	860	494	644
ENSG000002	1	632325	632413	-	89	3	2	1	0	0	0	3	0	0
ENSG000002	1	632757	633438	+	682	18	15	19	21	20	17	31	17	15

Post-alignment QC - example

DE analysis tools

- **DESeq2** (https://bioconductor.org/packages/devel/bioc/vignettes/DESeq2/inst/doc/DESeq2.html)
- edgeR (https://bioconductor.org/packages/release/bioc/html/edgeR.html)

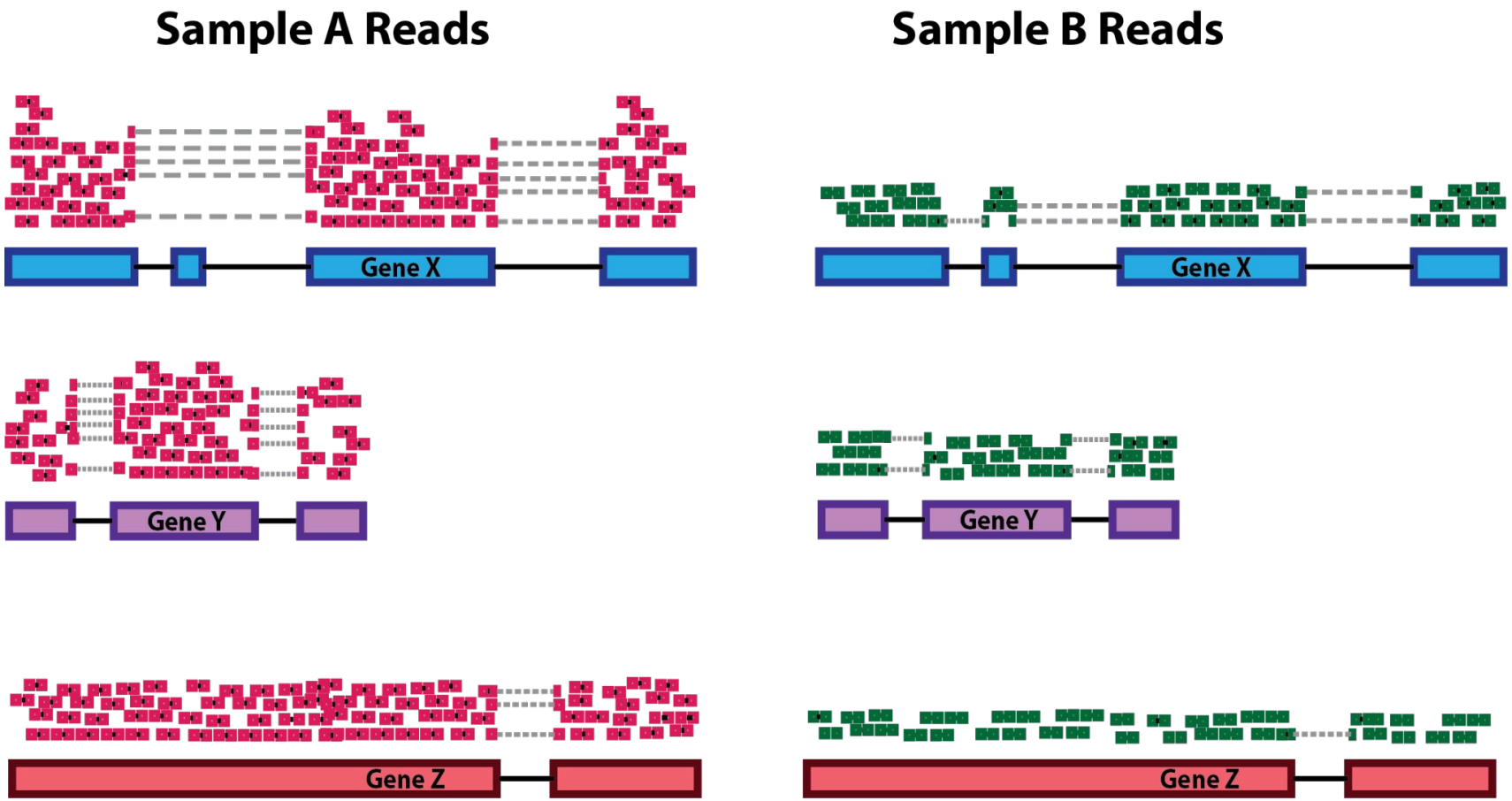


Normalization

Main factors we need to consider:

- Sequencing depth
- Gene length
- Difference in RNA composition

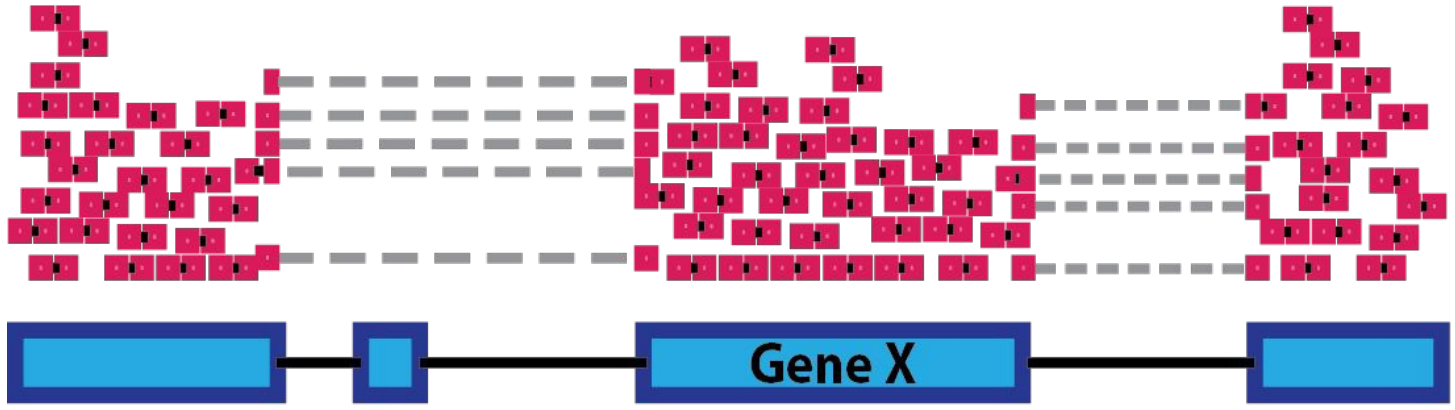
Normalization - sequencing depth

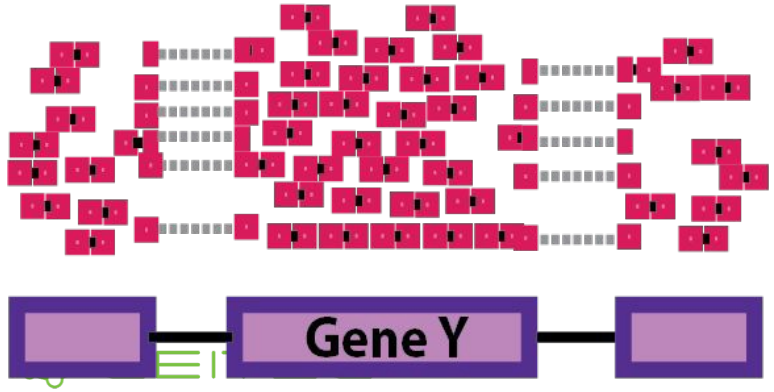


Images on this and following slides: https://github.com/hbctraining/DGE_workshop

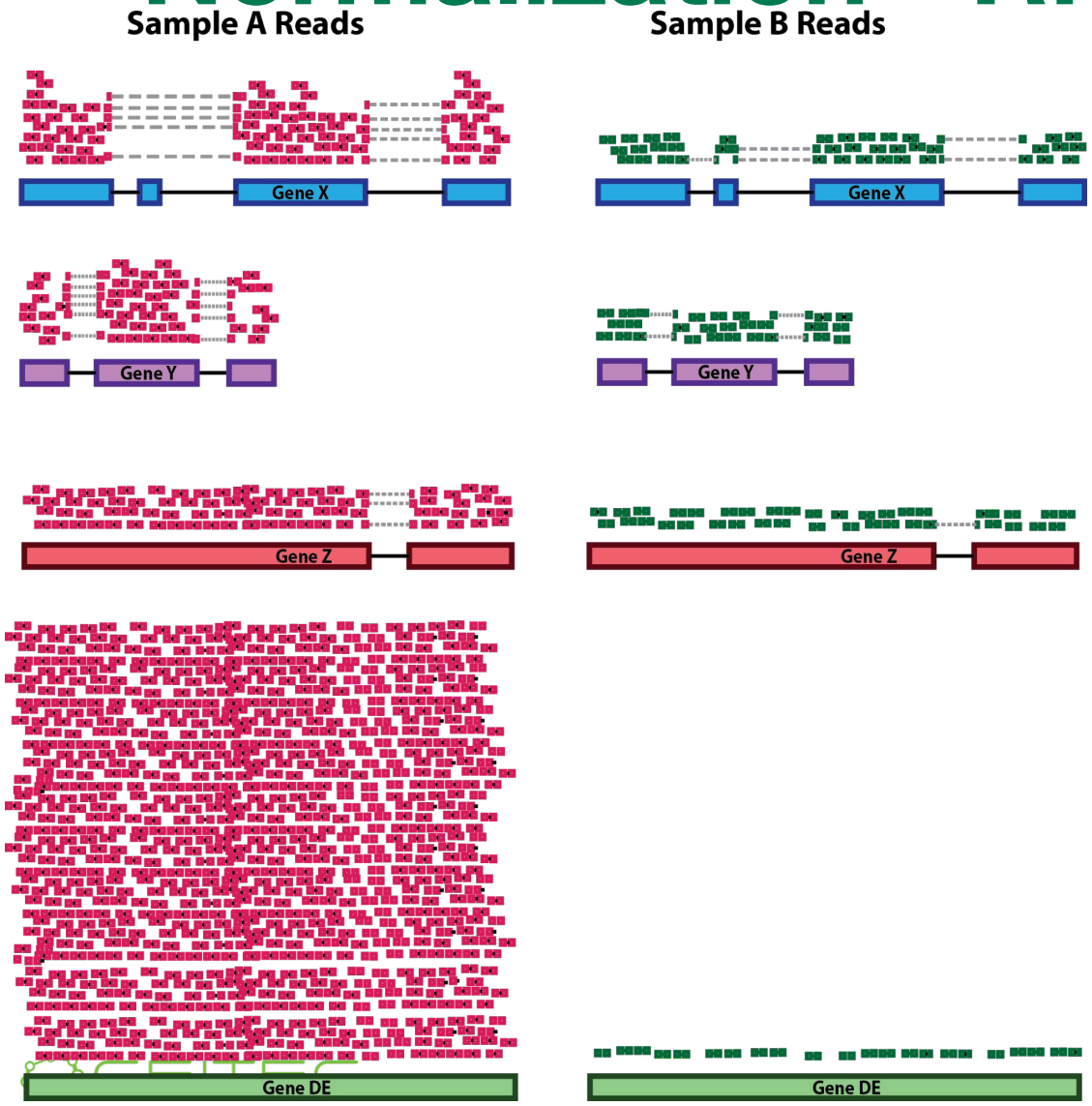
Normalization - gene length

Sample A Reads





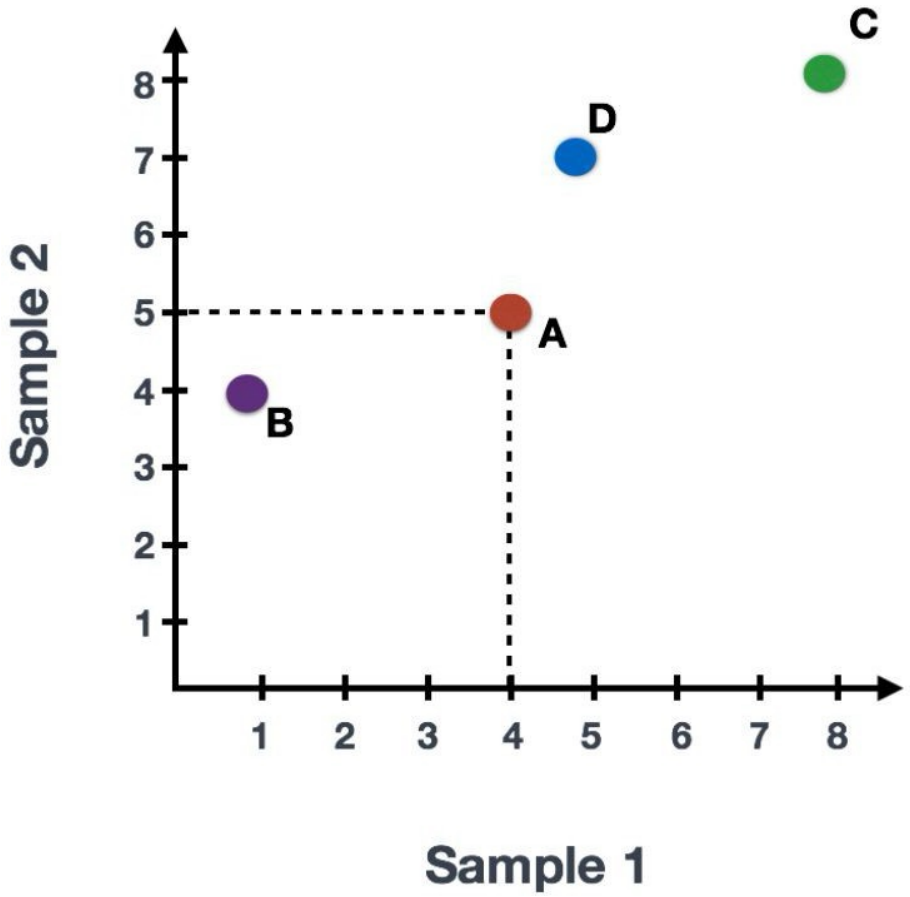
Normalization - RNA composition



Normalization methods

Normalization method	Description	Accounted factors	Recommendations for use
CPM (counts per million)	counts scaled by total number of reads	sequencing depth	gene count comparisons between replicates of the same sample group; **NOT for within sample comparisons or DE analysis**
TPM (transcripts per kilobase million)	counts per length of transcript (kb) per million reads mapped	sequencing depth and gene length	gene count comparisons within a sample or between samples of the same sample group; **NOT for DE analysis**
RPKM/FPKM (reads/fragments per kilobase of exon per million reads/fragments mapped)	similar to TPM	sequencing depth and gene length	gene count comparisons between genes within a sample; **NOT for between sample comparisons or DE analysis**
DESeq2's **median of ratios** [1]	counts divided by sample-specific size factors determined by median ratio of gene counts relative to geometric mean per gene	sequencing depth and RNA composition	gene count comparisons between samples and for **DE analysis**; **NOT for within sample comparisons**
EdgeR's **trimmed mean of M values (TMM)** [2]	uses a weighted trimmed mean of the log expression ratios between samples	sequencing depth, RNA composition	gene count comparisons between samples and for **DE analysis**; **NOT for within sample comparisons**

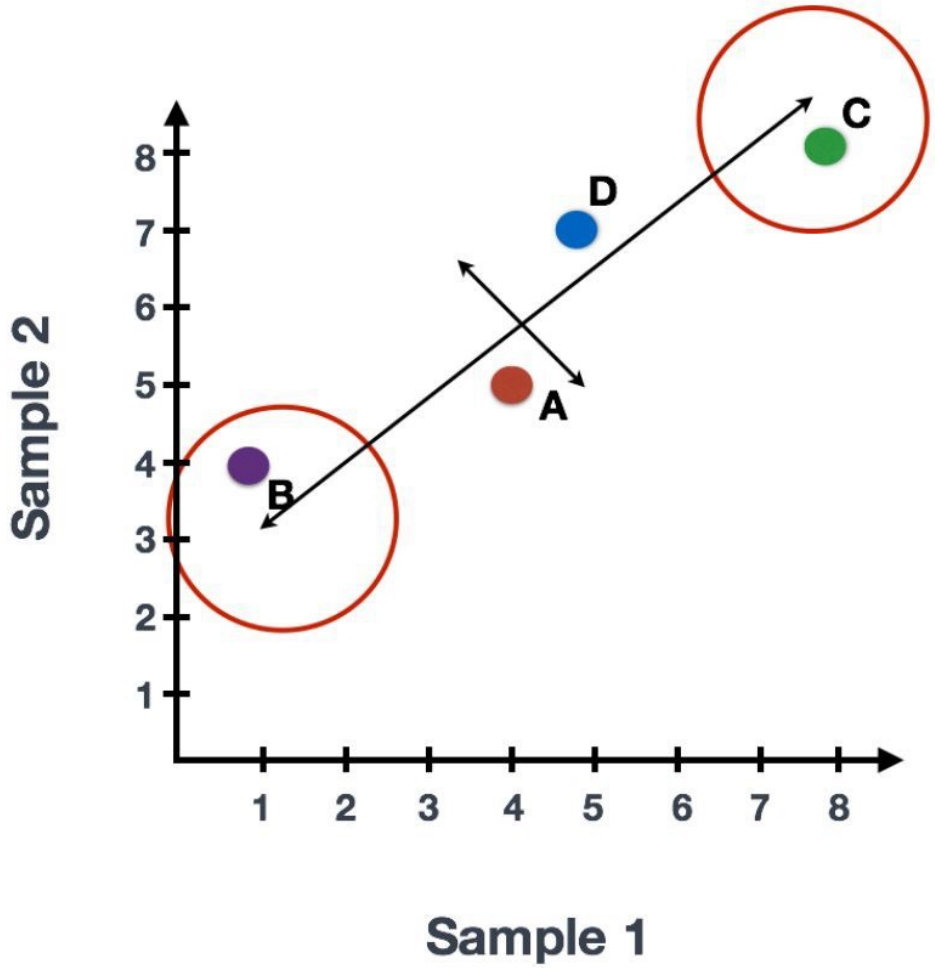
PCA



	Sample 1	Sample 2
Gene A	4	5
Gene B	1	4
Gene C	8	8
Gene D	5	7

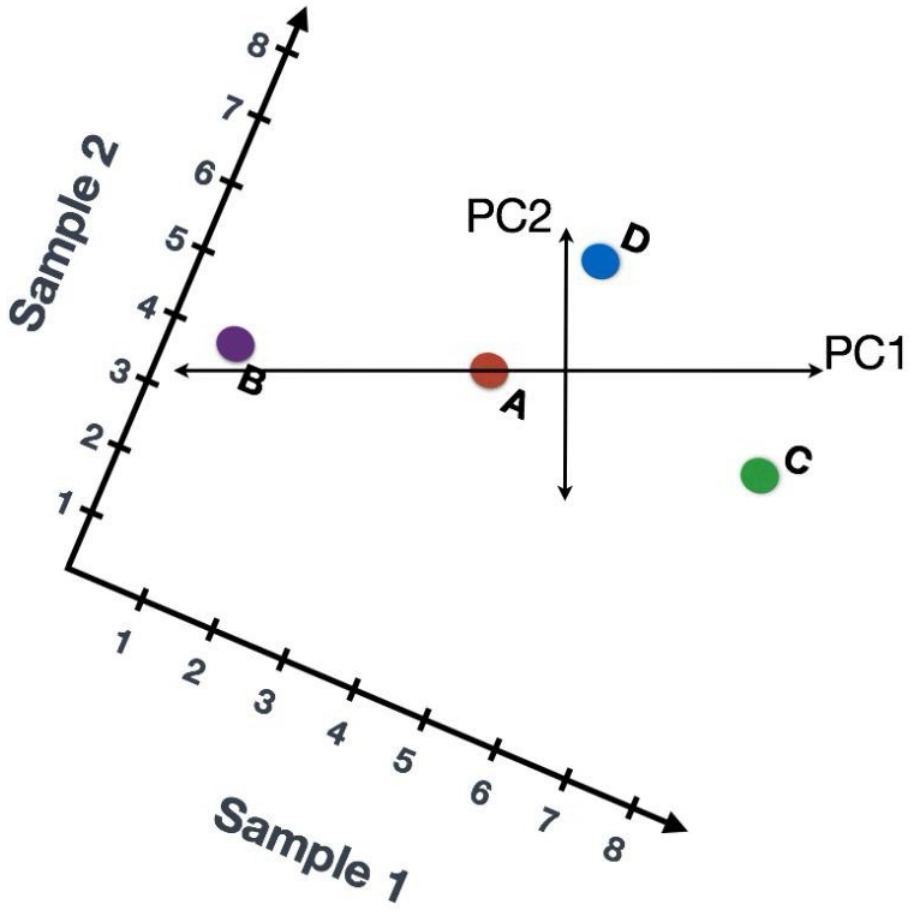
https://hbctraining.github.io/DGE_workshop/lessons/principal_component_analysis.html

PCA



https://hbctraining.github.io/DGE_workshop/lessons/principal_component_analysis.html

PCA

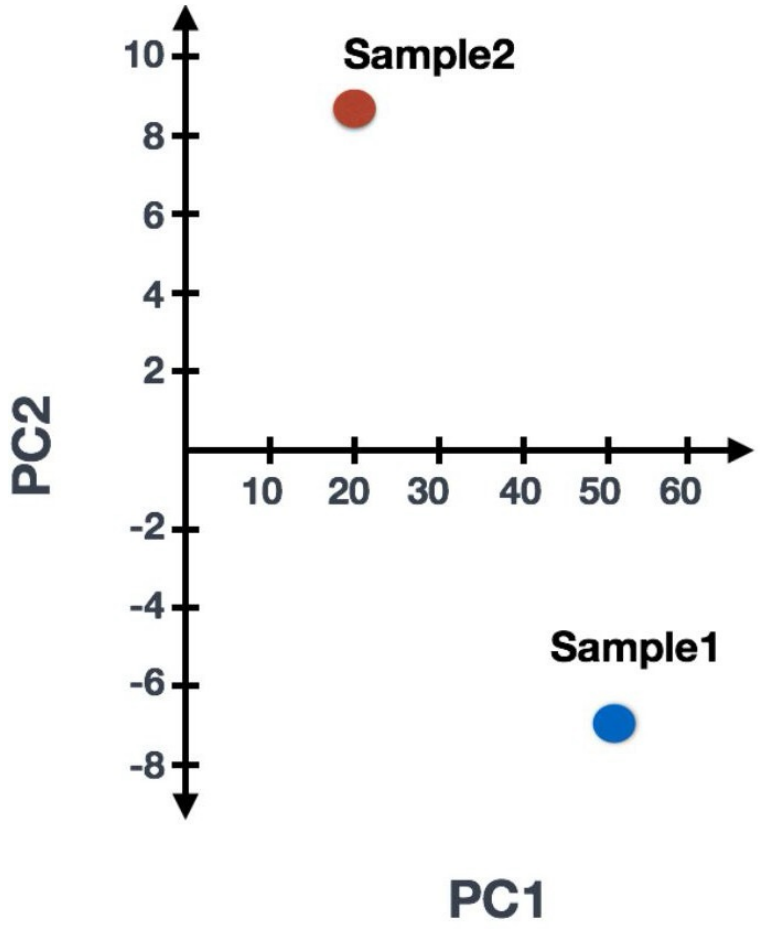


Sample1 PC1 score = (read count * influence) + ... for all genes

	Sample 1	Sample 2	Influence on PC1	Influence on PC2
Gene A	4	5	-2	0.5
Gene B	1	4	-10	1
Gene C	8	8	8	-5
Gene D	5	7	1	6

https://hbctraining.github.io/DGE_workshop/lessons/principal_component_analysis.html

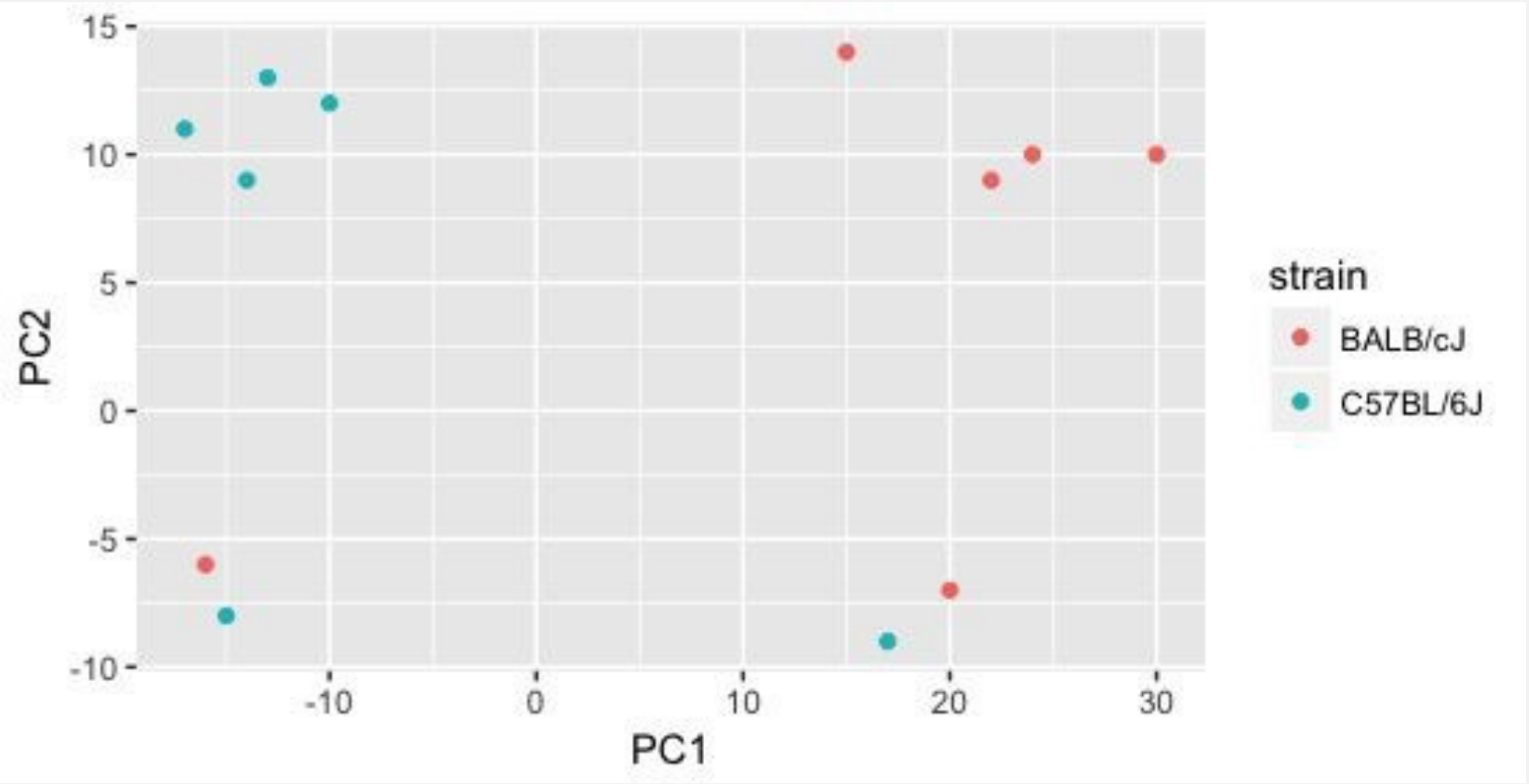
PCA



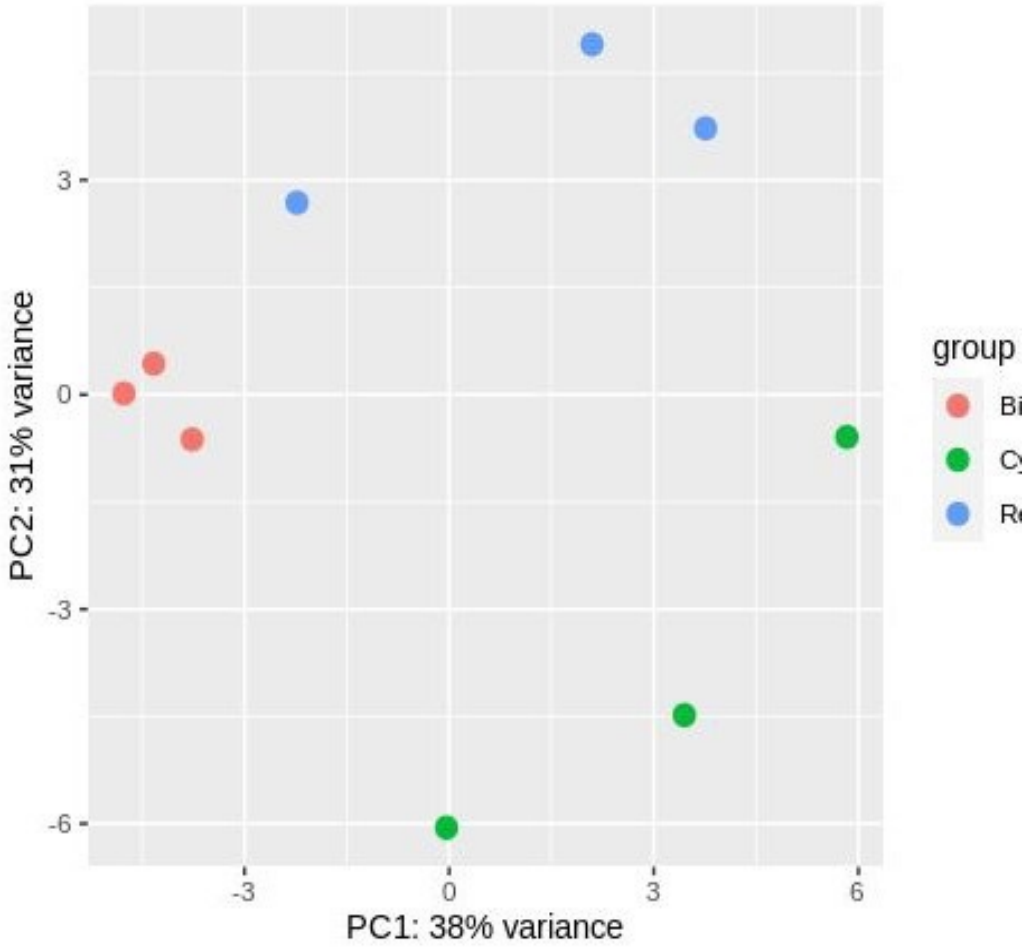
	PC1	PC2
Sample1	51	-7
Sample2	21	8.5

https://hbctraining.github.io/DGE_workshop/lessons/principal_component_analysis.html

PCA

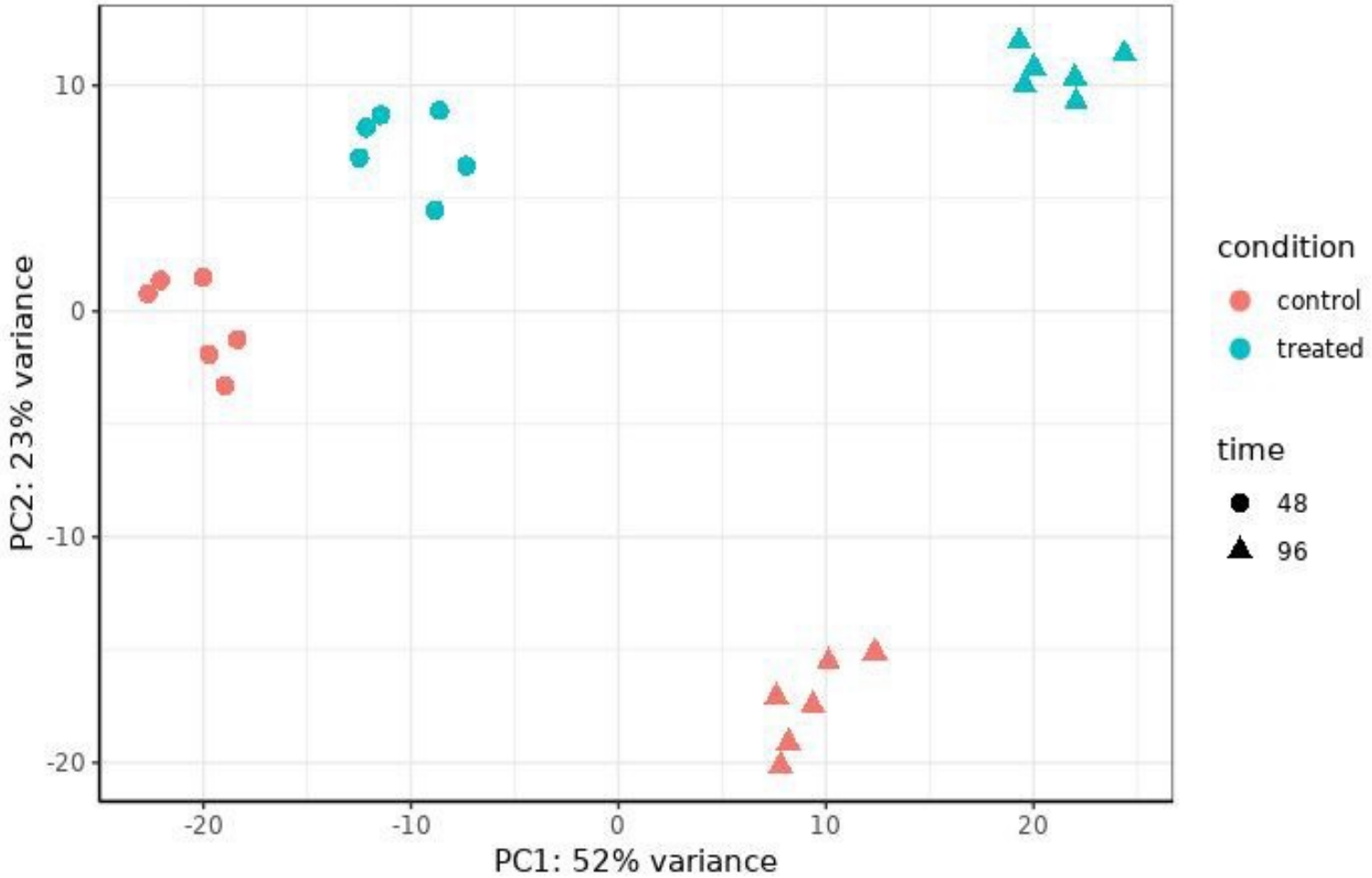


PCA real life examples



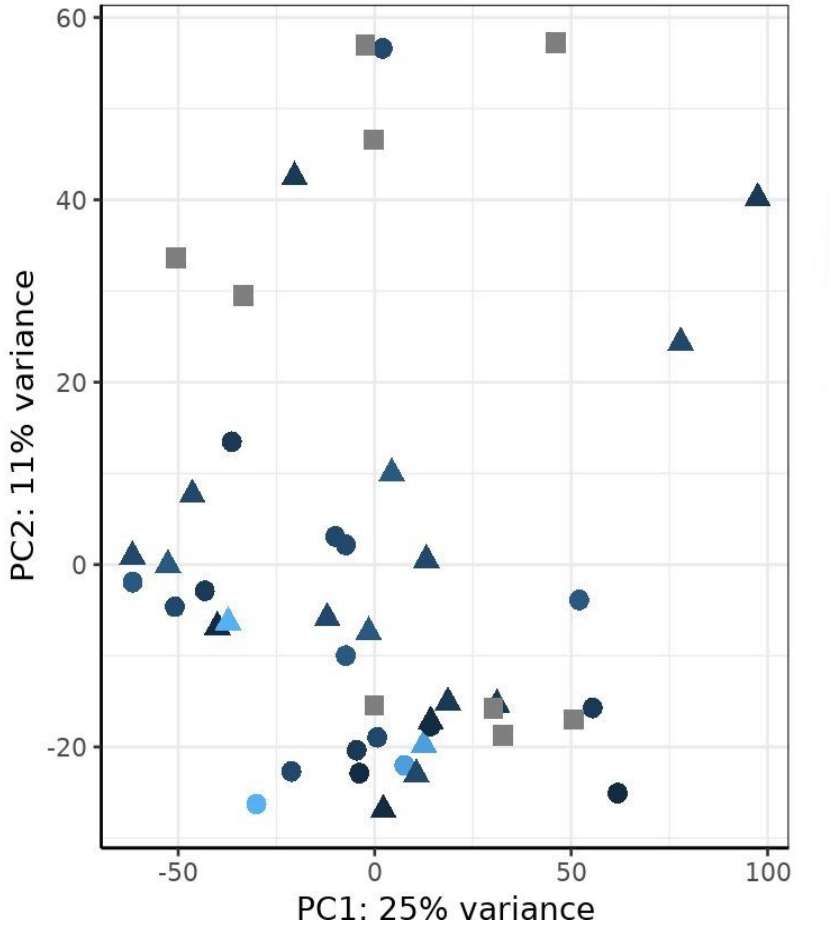
CEITEC

PCA real life examples



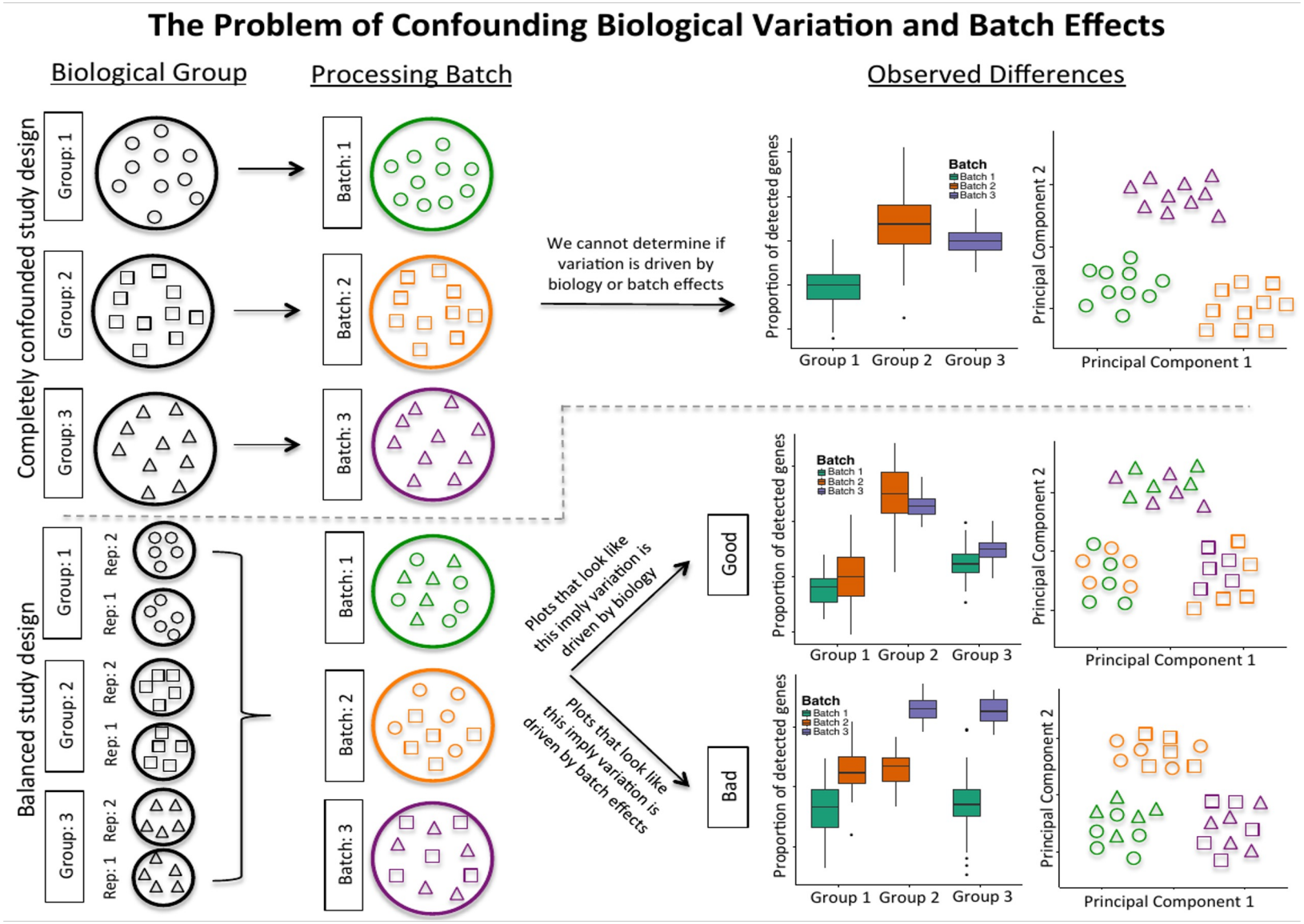
CEITEC

PCA real life examples



Pairing of the samples/batch effect

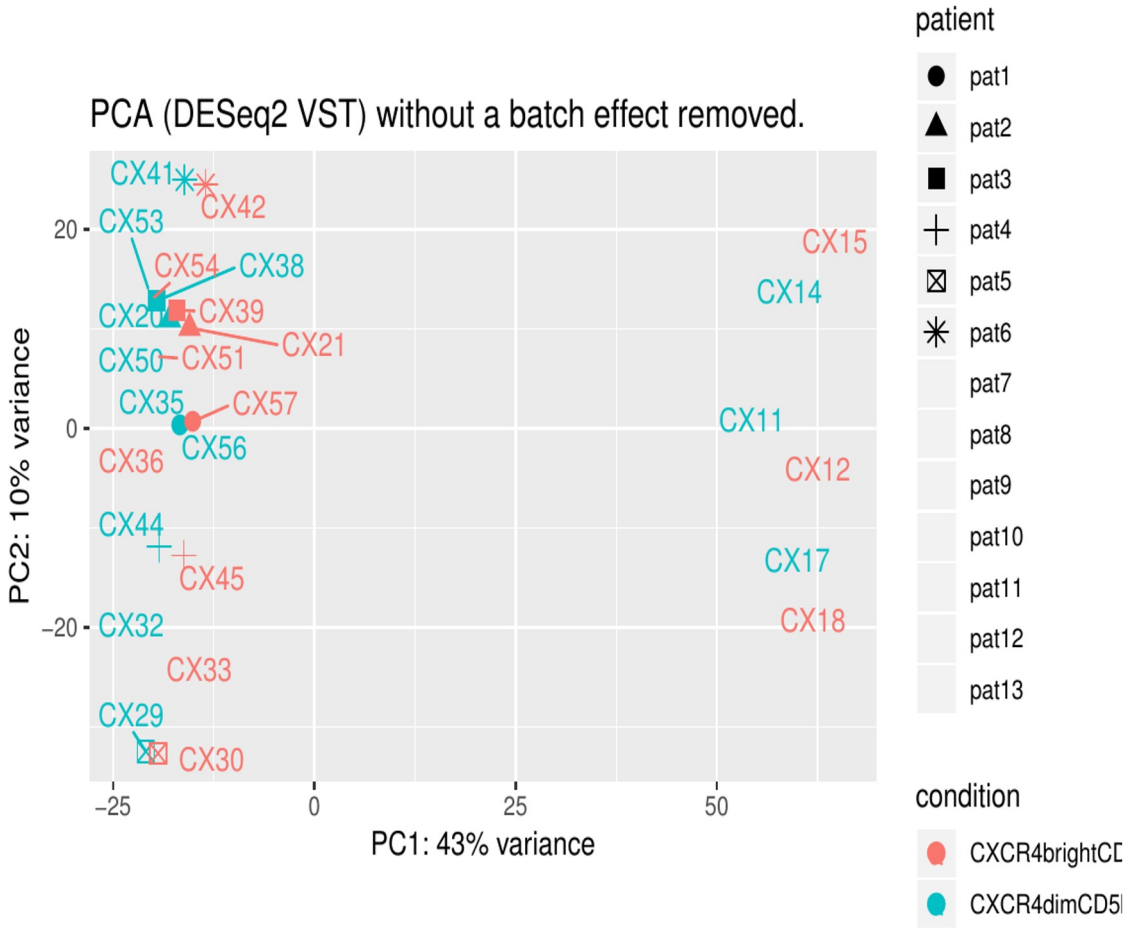
- There is a bad experimental design and a good experimental design
- Very simply - more randomization gives you better results

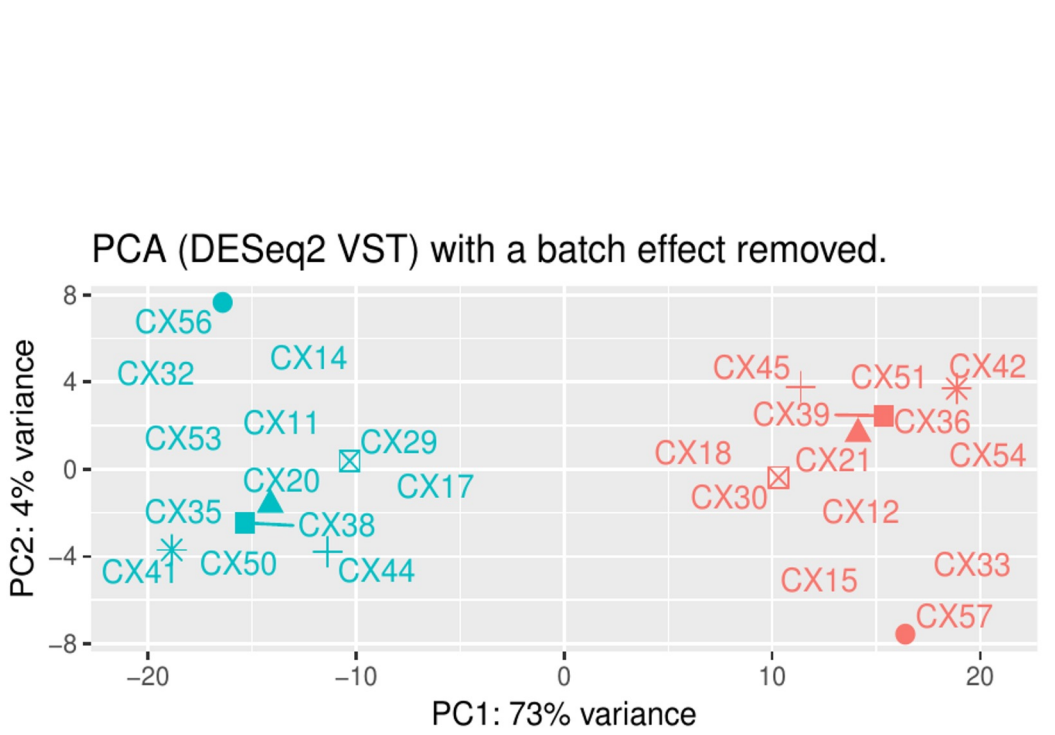


CEITEC

Pairing of the samples/batch effect

- And example pairing of the patients AND different sequencing years - double batch

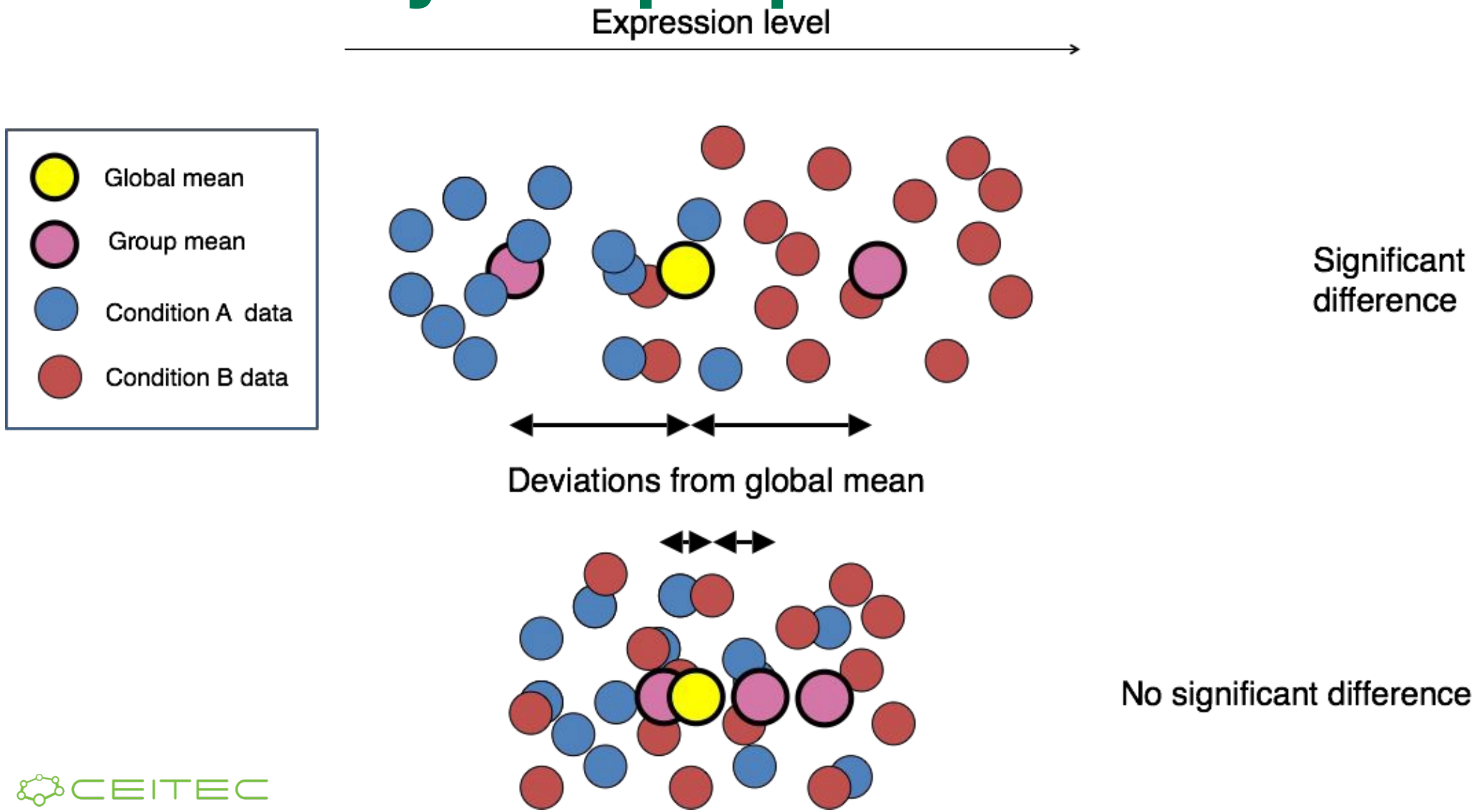




Pairing of the samples/batch effect

- Paired samples are not the same as paired-end sequencing!

DE analysis - proper



DE analysis - proper

We need for DEseq2:

- Table describing all the samples
- Table with raw counts

sample	strain	date	cage	treatment	replicate	sex
B1	BALB/cJ	20180515	1	yes	1	M
B2	C57BL/6J	20180515	2	yes	1	M
B3	BALB/cJ	20180515	3	no	1	M
B4	C57BL/6J	20180515	1	no	1	F
B5	BALB/cJ	20180515	2	yes	2	F
B6	C57BL/6J	20180515	3	yes	2	M
B7	BALB/cJ	20180515	1	no	2	M
B8	C57BL/6J	20180515	2	no	2	M
B9	BALB/cJ	20180515	3	yes	3	F
B10	C57BL/6J	20180307	1	yes	3	F
B11	BALB/cJ	20180307	2	no	3	M
B12	C57BL/6J	20180307	3	no	3	M

Experimental design

- **Biological replicates** represent multiple samples from the same sample group
- **Technical replicates** represent the same sample (i.e. RNA from the same mouse) but with technical steps replicated, or the same sample sequenced multiple times
- Usually **biological variance is much greater than technical** variance, so we do not need to account for technical variance to identify biological differences in expression
- **Don't spend money on technical replicates - biological replicates are much more useful**

Experimental design

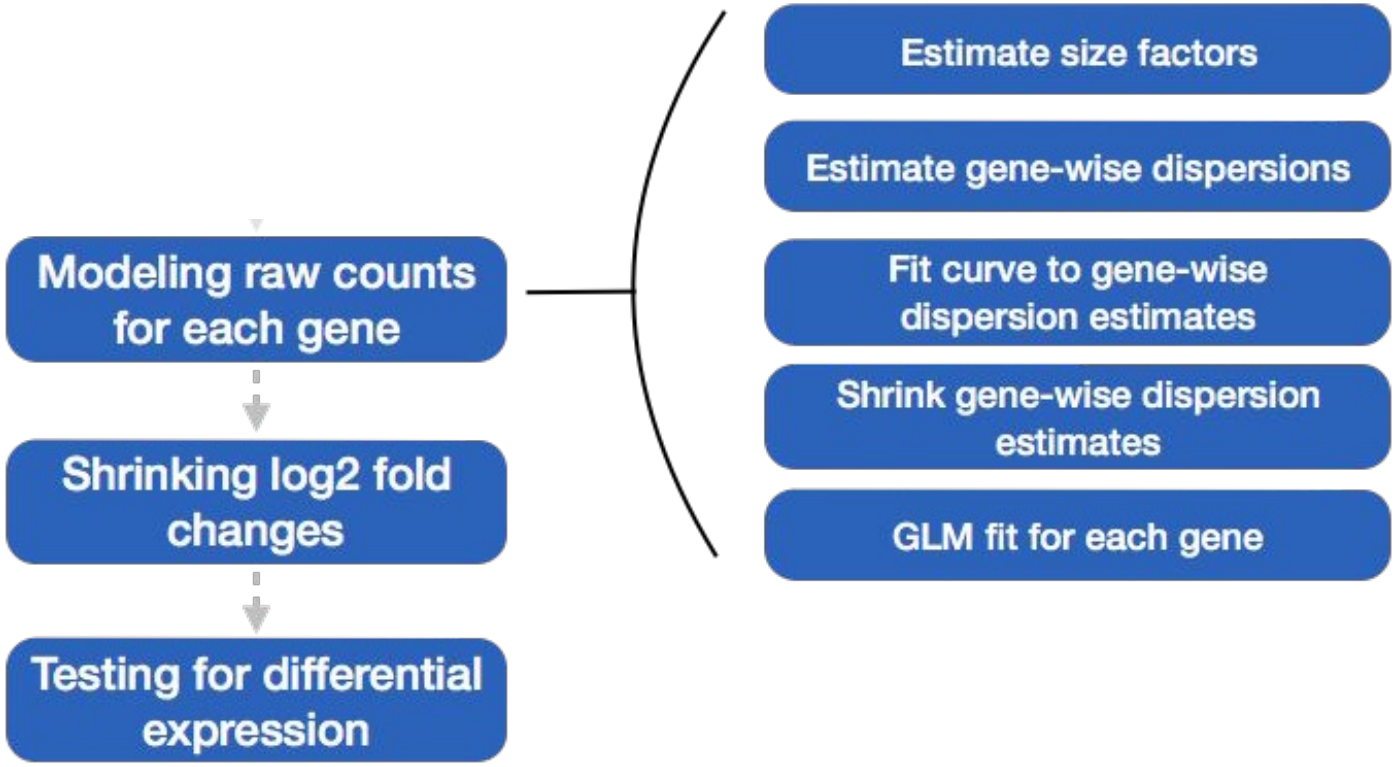
	Replicates per group		
	3	5	10
Effect size (fold change)			
1.25	17 %	25 %	44 %
1.5	43 %	64 %	91 %
2	87 %	98 %	100 %
Sequencing depth (millions of reads)			
3	19 %	29 %	52 %
10	33 %	51 %	80 %
15	38 %	57 %	85 %

Experimental design - confounding variables

A confounded RNA-Seq experiment is one where you cannot distinguish the separate effects of two different sources of variation in the data.

- Do not try to save money by “pooling” variables together!
- Avoid introducing variables
- Keep metadata

DE analysis - proper



DE analysis - result table

```
log2 fold change (MLE): condition treated vs control
Wald test p-value: condition treated vs control
DataFrame with 26596 rows and 6 columns
                    baseMean log2FoldChange     lfcSE      stat    pvalue      padj
                   <numeric>      <numeric> <numeric> <numeric> <numeric> <numeric>
ENSMUSG00000000001 1316.691048       0.345714  0.166648  2.074518 0.0380312  0.205846
ENSMUSG00000000028  811.387199       0.154380  0.131436  1.174561 0.2401705  0.584747
ENSMUSG00000000031    1.327391      -3.197179  2.640511 -1.210818 0.2259651        NA
ENSMUSG00000000037   26.520229      -0.154937  0.802141 -0.193154 0.8468384  0.953352
ENSMUSG00000000049    0.975305      -0.372812  2.563517 -0.145430 0.8843715        NA
...                        ...            ...       ...       ...       ...       ...
ENSMUSG00002076966    0.866043        1.14651   1.76946  0.647947  0.517019        NA
ENSMUSG00002076975    0.407060        2.59607   3.05970  0.848473  0.396175        NA
ENSMUSG00002076981    0.163449        1.53905   3.10112  0.496289  0.619691        NA
ENSMUSG00002076983    0.373390       -1.97875   3.06686 -0.645203  0.518795        NA
ENSMUSG00002076989    0.369877       -1.96782   3.06721 -0.641566  0.521155        NA
```

log2(fold-change)

- **Fold-change** is usually calculated by **average expression of all samples** of **condition 1** vs average expression of all samples of **condition 2**
- **Example:**
    a) geneA expression in **pre is 5**, in **post is 10**; fold-change of post/pre is **2** = gene is **up-regulated 2x**
    b) geneB expression in **pre is 10**, in **post is 5**; fold-change of post/pre is **0.5** = gene is **down-regulated 1/2x** … (O_o)
- **Solution**: Adding **log2** gives us **log2(2) = 1**, **log2(0.5) = -1**
- Nice and even distribution around 0 and clear interpretations

P-value and adjusted p-value

- **P-value tries** to give you “a **number**” saying if the **differences** you are observing are **robust** and the differences are **not** “**random**” between the compared conditions/samples
- **Adjusted p-value** adds a **correction** for the **multiple testing** we are doing - tries to add correction of **getting** a **p-value just by accident**
- But **is** adjusted p-value **0.049** really **better** than **0.051**?
- **Number of replicates highly influences** the **estimates**
    - The **observations might be the same** but the **statistical significance** might be **lower**

How many differentially expressed genes I have?

It depends **how many you want**…:)

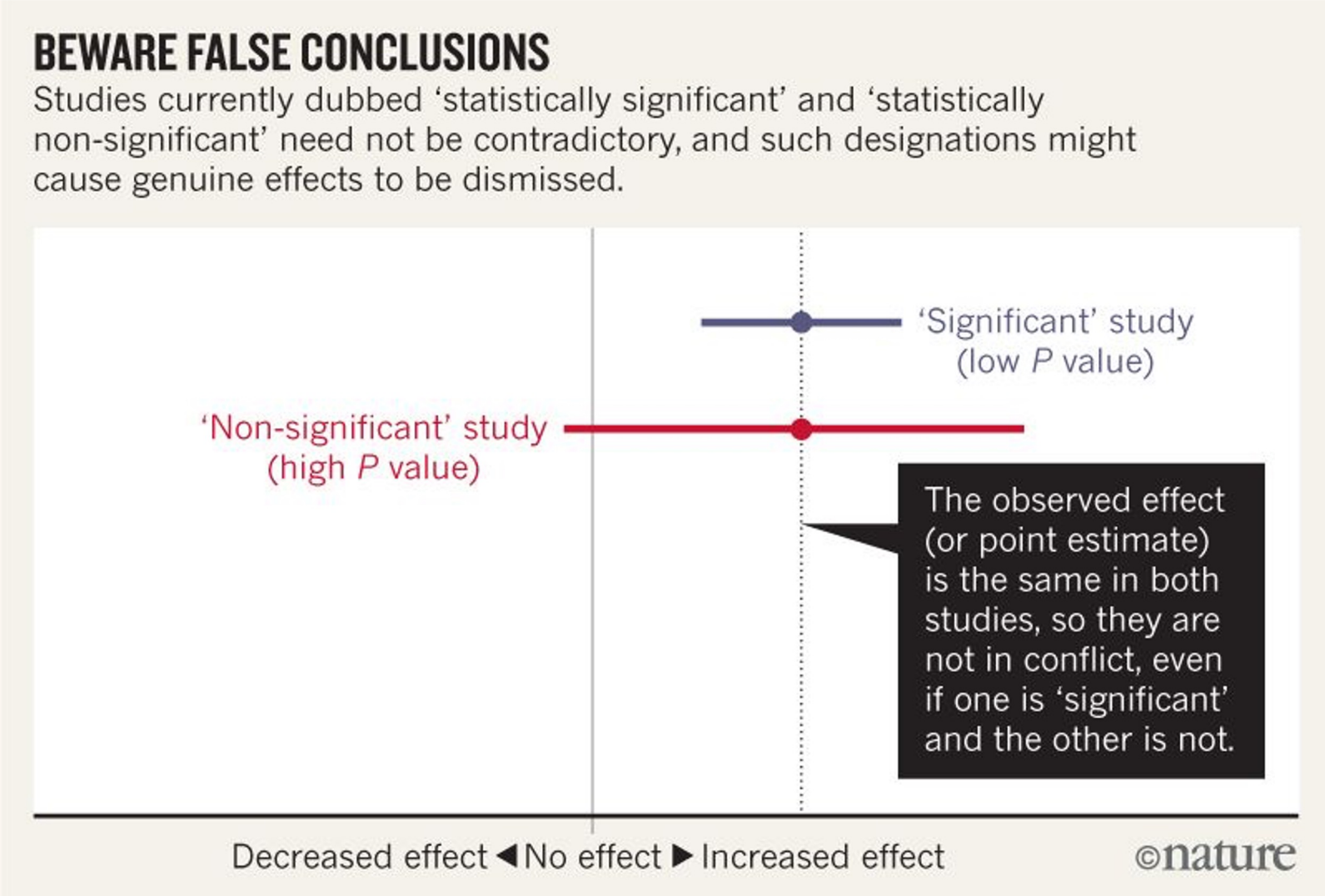
Selection of the **differentially expressed** (DE) gene is **completely up to you**

Some people use **p-value, some adjusted p-value and some people log2fc and their combinations**, some just take top *n* genes

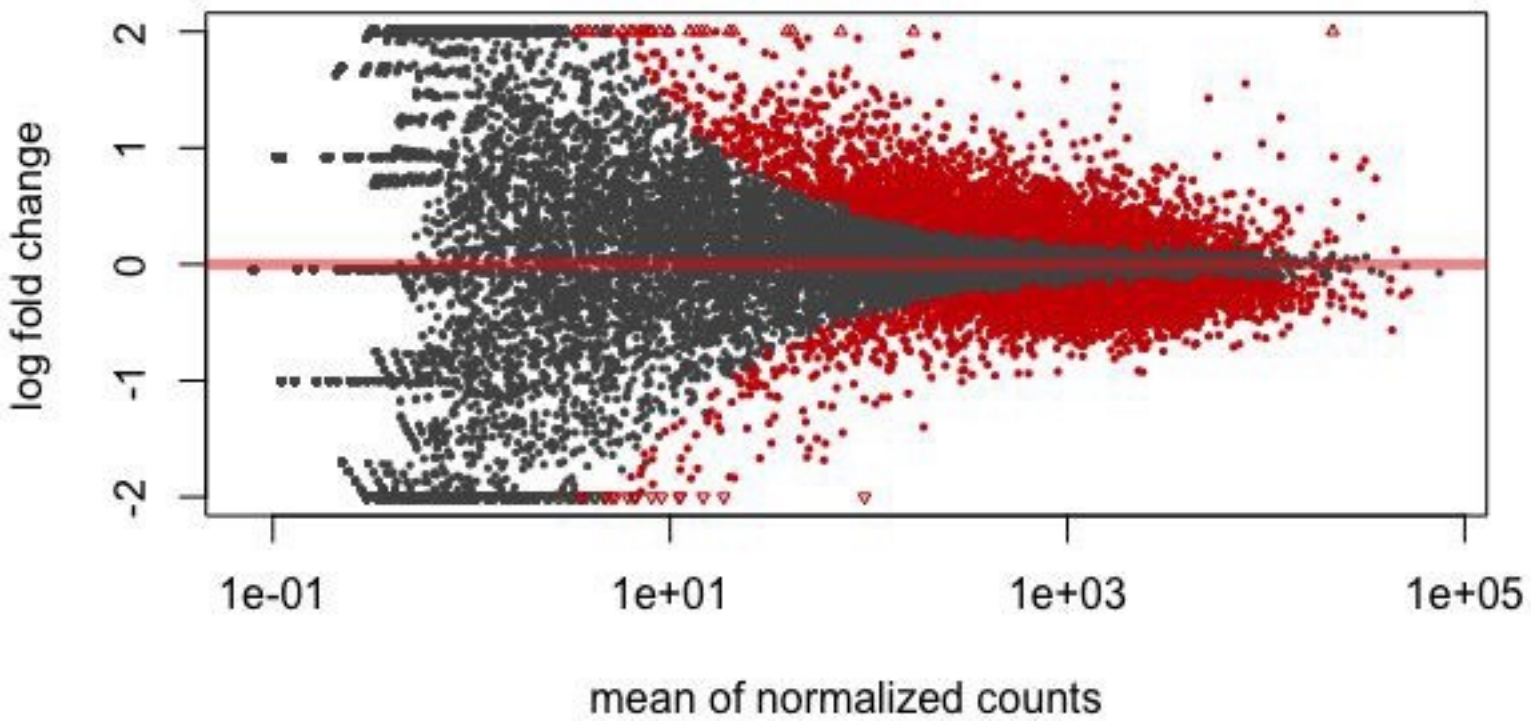
Statistical significance ≠ biological relevance!!!

Scientists rise up against statistical significance, Nature 567, 305-307 (2019), doi: 10.1038/d41586-019-00857-9

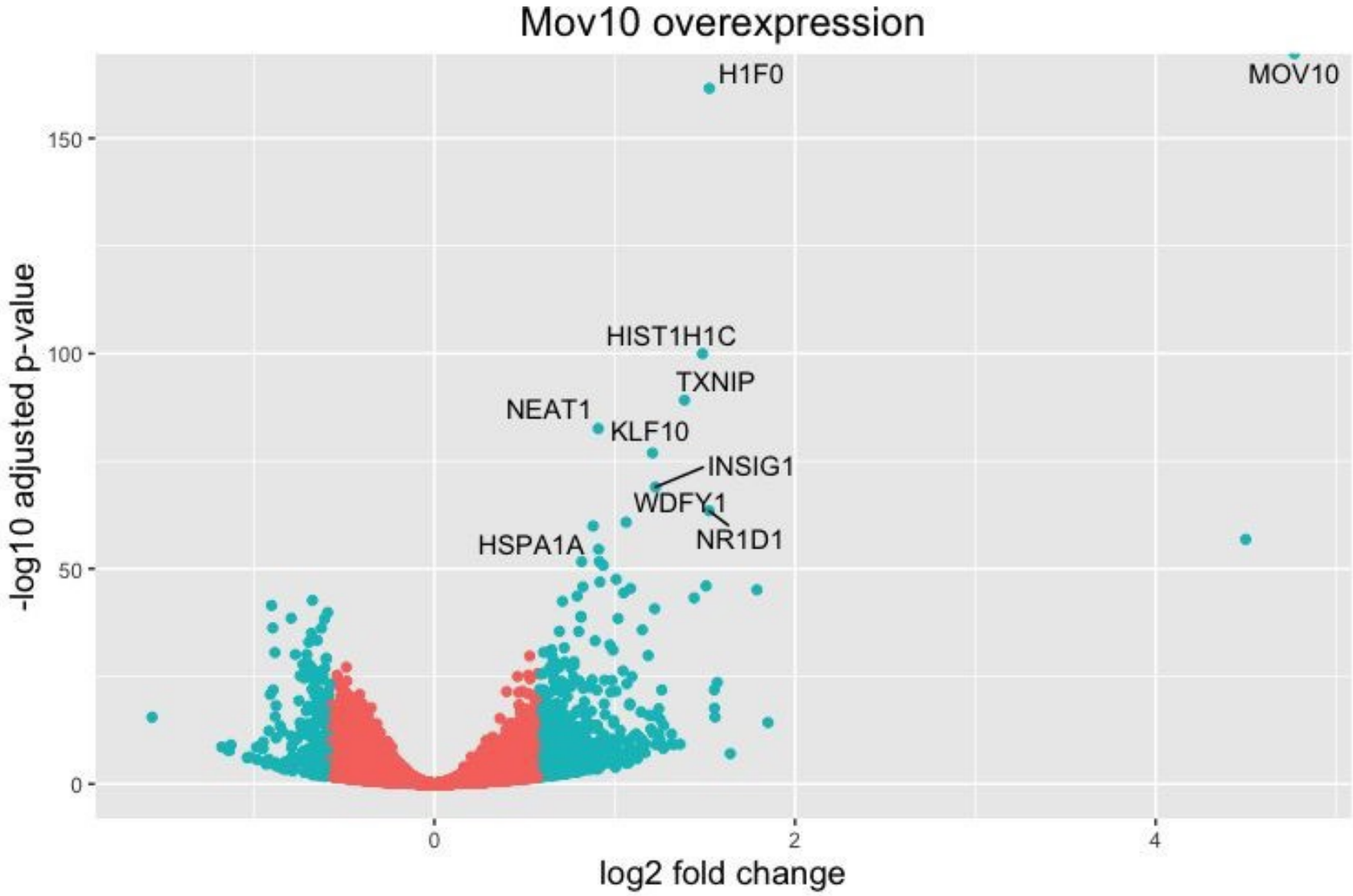
P-value significance



Visualisation - MA plot

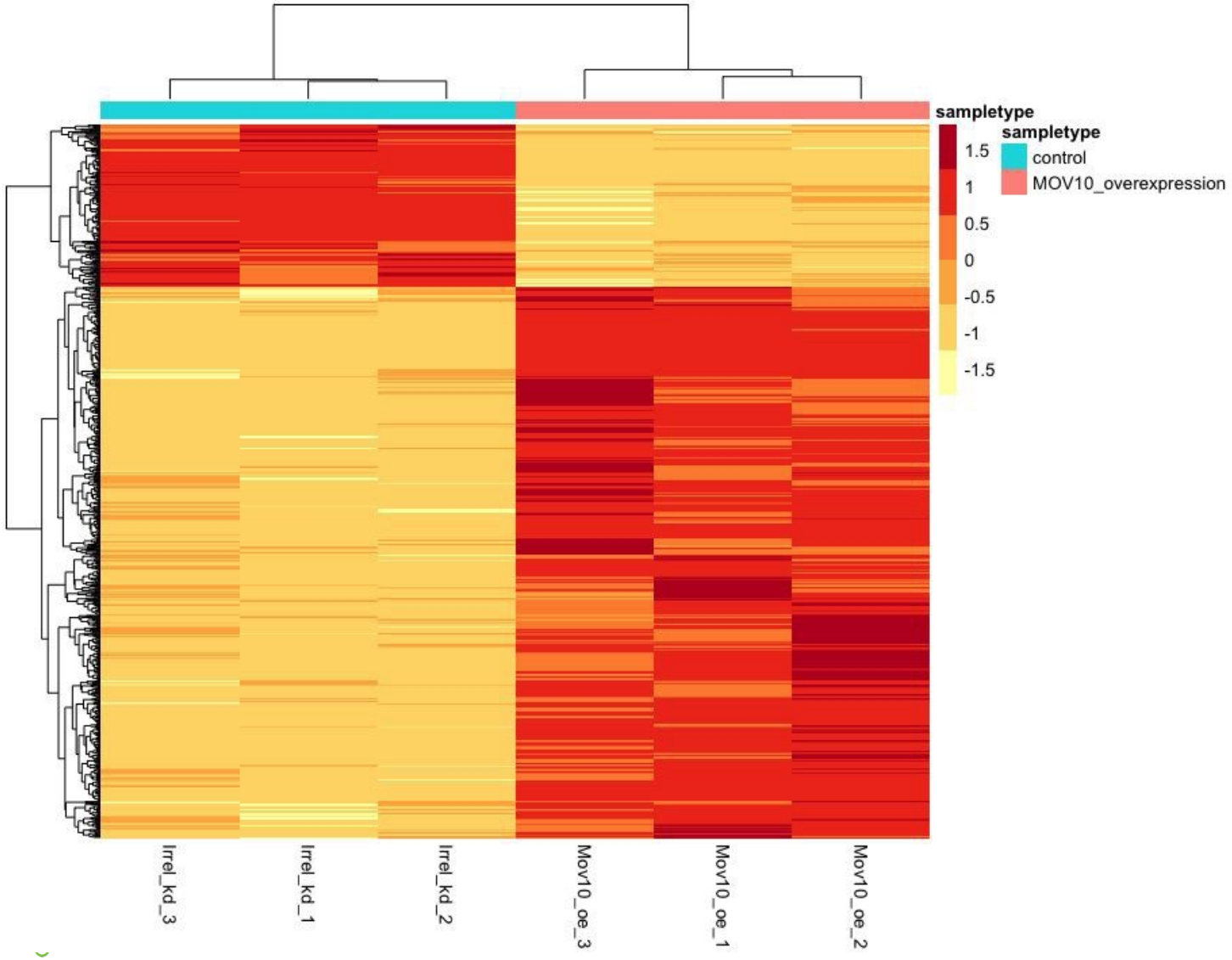


Visualisation - volcano plots



Amazing interactive tool: Glimma - https://bioconductor.org/packages/release/bioc/html/Glimma.html

Visualisation - heatmaps



Main takeaway points

- **Have a question that can be answered with RNA-seq.**
- Avoid confounding variables.
- Have as many biological replicates as you can.
- Don’t bother with technical replicates.
- Rather do more replicates than deeper sequencing. *
- Consult your design with a sequencing facility / bioinformatician ***before*** you start.
- Keep metadata, even if it feels crazy.
- Try to make a pilot study before you commit more resources. (Student’s mental health *is a resource too!*)

*Unless you want some really low expressed genes, then you probably need both replicates *and* depth

CEITEC

CEITEC

@CEITEC_Brno

Thank you for your attention!

Vojta Bystry
vojtech.bystry@ceitec.muni.cz