



CEITEC

Central European Institute of Technology
BRNO | CZECH REPUBLIC

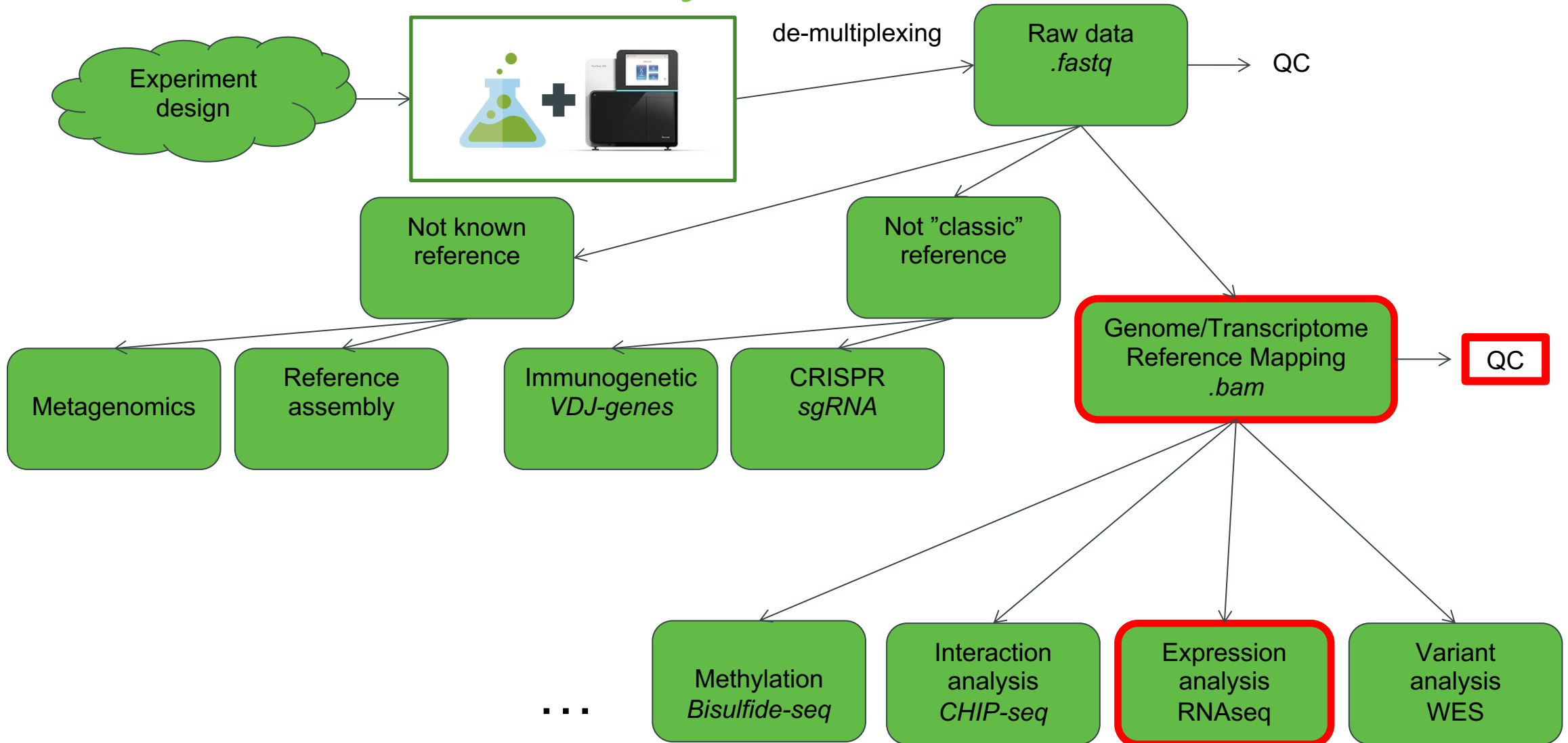


**Modern methods for genome analysis
(PřF:Bi7420)**

Lecture 6 : RNA-seq differential expression

Vojta Bystry
vojtech.bystry@ceitec.muni.cz

NGS data analysis



Feature count results

complete.featureCounts

Home Insert Draw Page Layout Formulas Data Review View

Calibri (Body) 12

Geneid

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

Differential expression

- We have our raw read counts but we need to find the real differences
- We want to figure out the change comparing the before and after treatment
- What are the changed genes? Are there even any? Is there even difference between the samples? And what about the experimental design - paired samples - does it affect the evaluation?
- The tools for the differential expression have to account for different libraries depths, model and “fix” outliers, account for different levels of expressions, and many other things
- Luckily, there are few tools that have all of this and can be used

Differential expression - tools

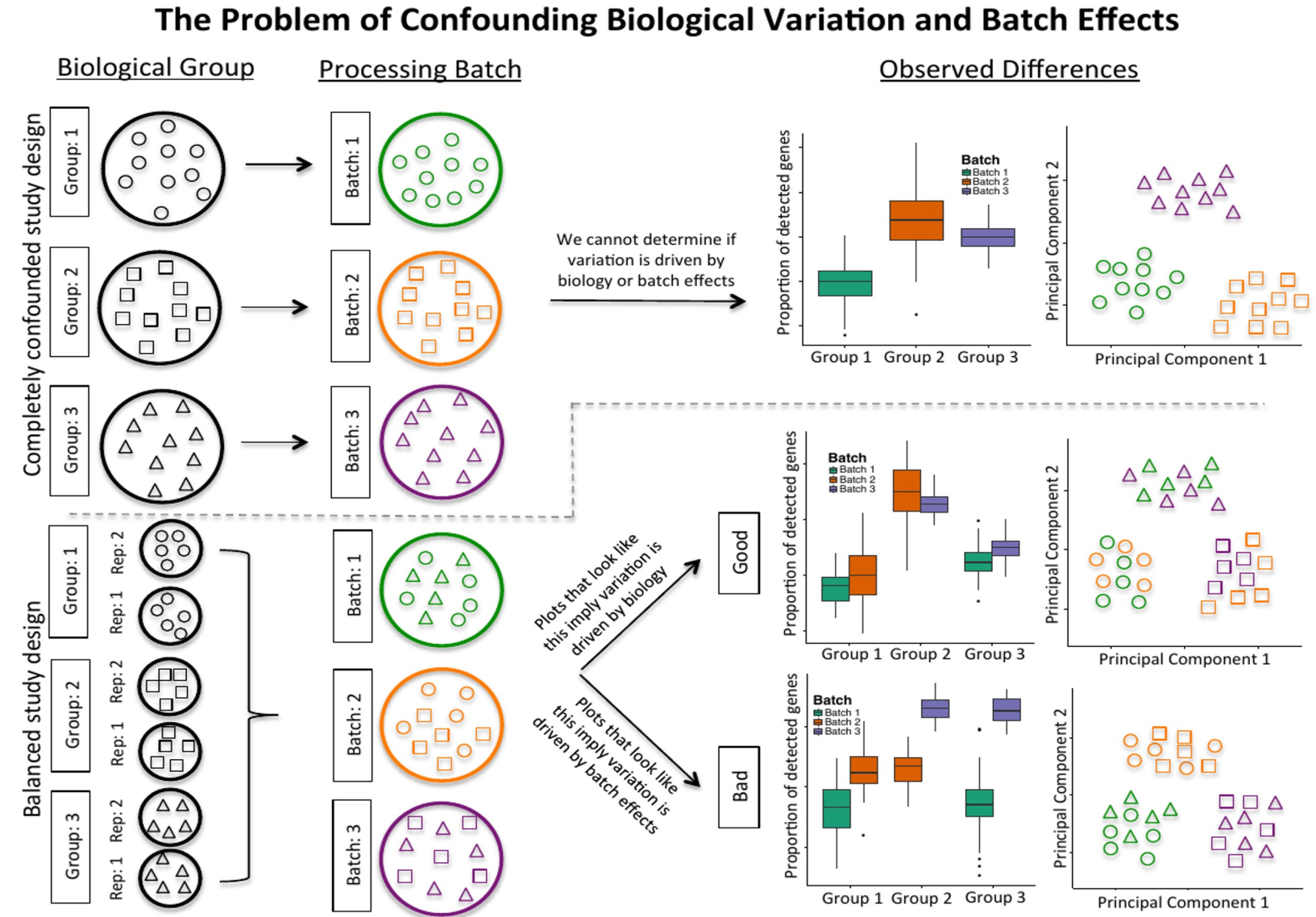
- DESeq2
 - More specific
- edgeR
 - More sensitive
- The important part of the calculation is the **design**
 - Assignment of a group/condition to a sample
 - If the samples are paired (the same patient twice) we have to account for this as well!
 - Technically, the pairing of the samples is a **batch effect** so it is similar to have a technical noise in your data

Pairing of the samples/batch effect

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

Pairing of the samples/batch effect

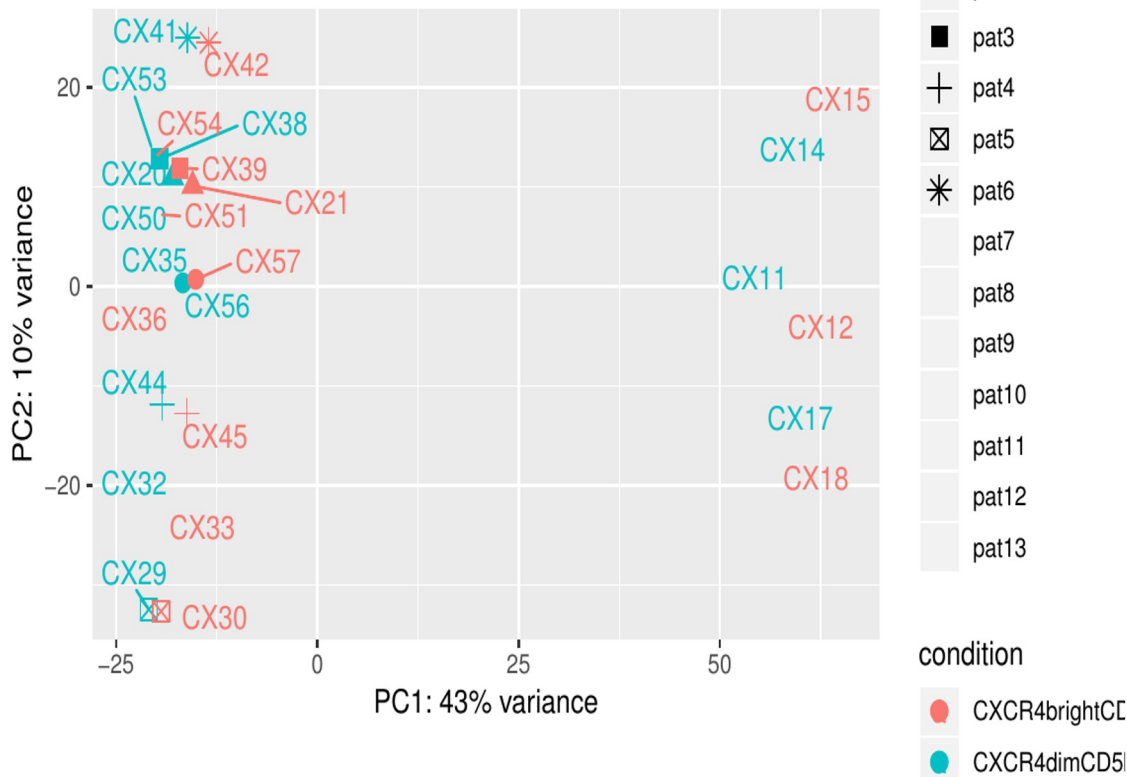
- There is a bad experimental design and a good experimental design
- Very simply - more randomization gives you better results



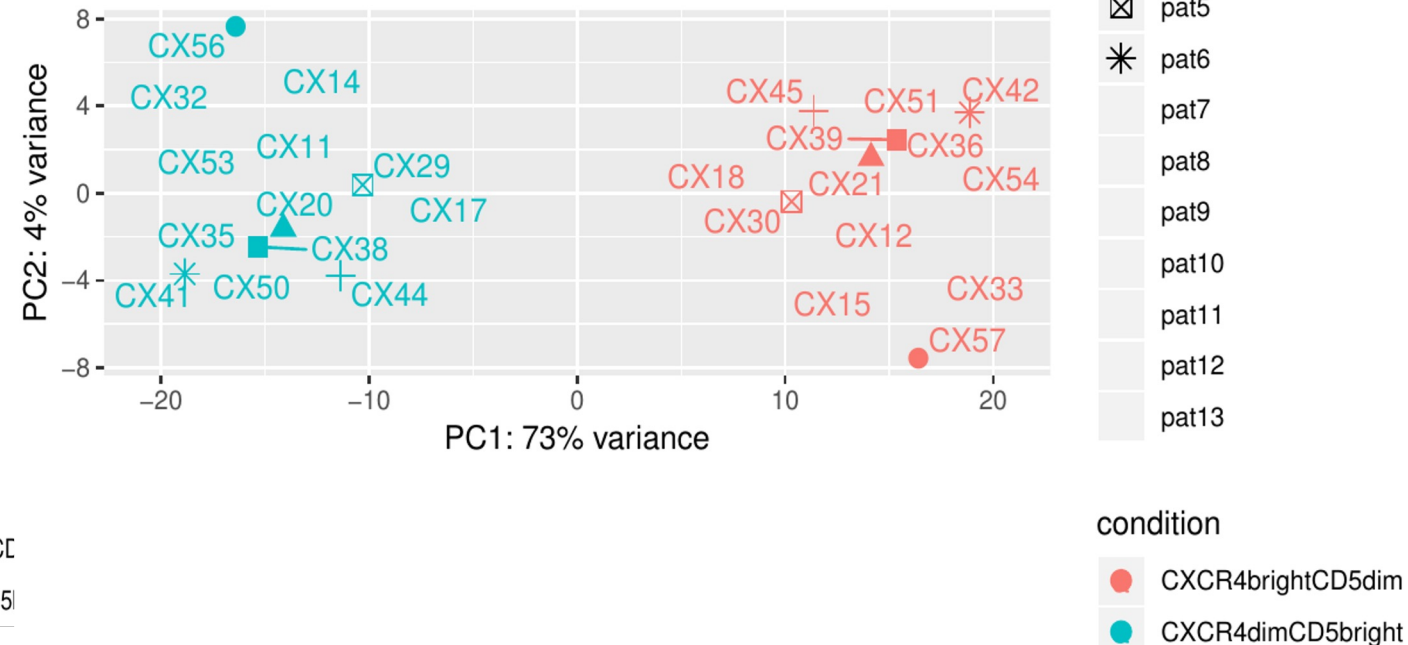
Pairing of the samples/batch effect

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

PCA (DESeq2 VST) without a batch effect removed.



PCA (DESeq2 VST) with a batch effect removed.



Count normalisation

- Normalize to:
 - Gene size
 - Library size
- rpkm - Reads Per Kilobase of transcript per Million mapped reads
- fpkm - Fragments Per Kilobase of transcript per Million mapped reads
- tpm - Transcripts Per Million (TPM)
 - for every 1,000,000 RNA molecules in the RNA-seq sample, x came from this gene/transcript
- Never ever use normalized counts for any comparisons
 - ...except comparing a single gene in a single experiment for the samples
 - If you really, really need to use any kind of normalized counts to compare use TPM

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

log₂(fold-change)

- But it might be **misleading**
- **Large log₂FC on low-expressed genes** are most **likely not biologically relevant**
- **Small log₂FC on highly-expressed genes** might be **biologically relevant**
- Example: “Common” cut-off value of **fold-change of 2x** (log₂FC=+/-1) or **1.5x** (log₂FC=+/-0.58)
 - geneA expression in WT is **10** and in KO is **4**, **log₂FC = -1.32 YES (?)**
 - geneB expression in WT is **1,000,000** and in KO is **500,001**, **log₂FC = -0.99 NO (?)**

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

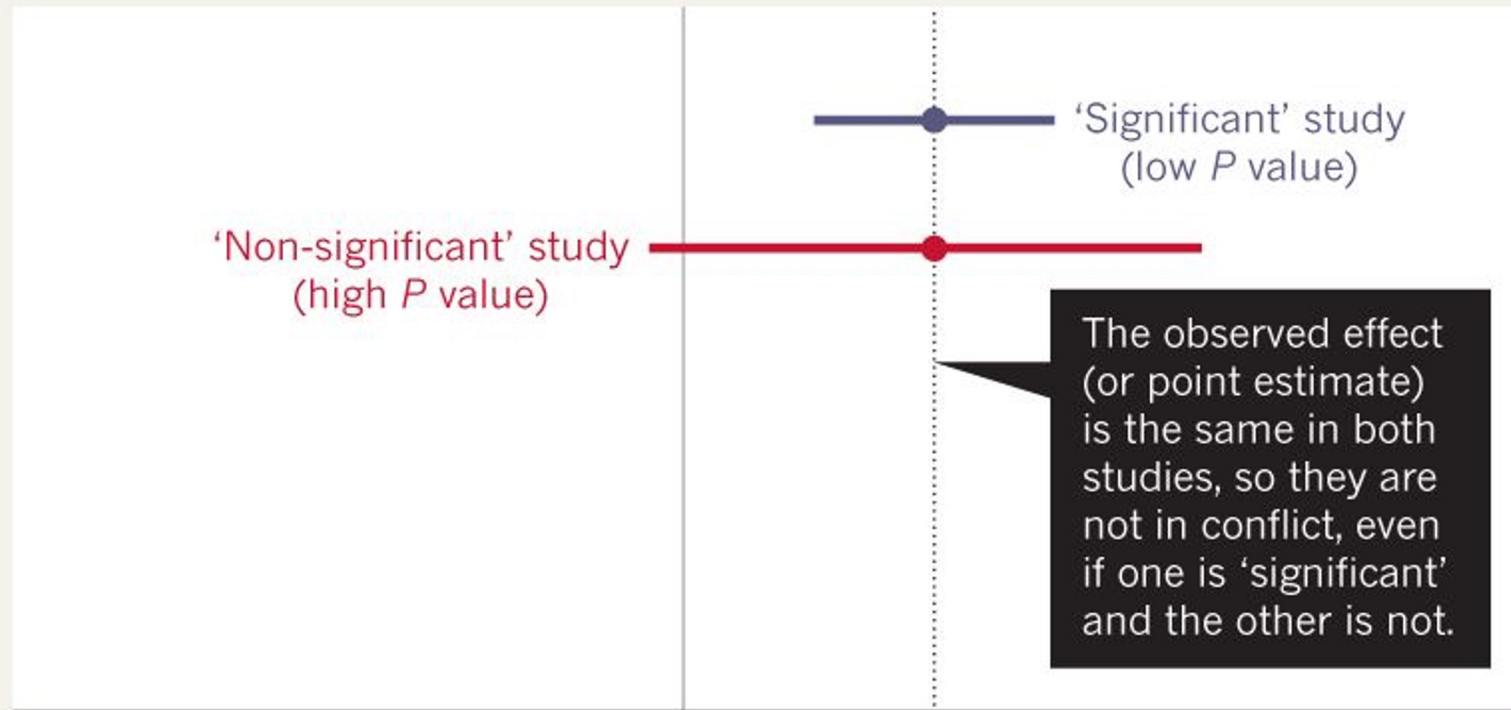
Statistical significance \neq biological relevance!!!

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

P-value significance

BEWARE FALSE CONCLUSIONS

Studies currently dubbed 'statistically significant' and 'statistically non-significant' need not be contradictory, and such designations might cause genuine effects to be dismissed.



Decreased effect ◀ No effect ▶ Increased effect

©nature

Differential expression output

- Example



CEITEC



@CEITEC_Brno

Thank you for your attention!

