CEITEC

Central European Institute of Technology

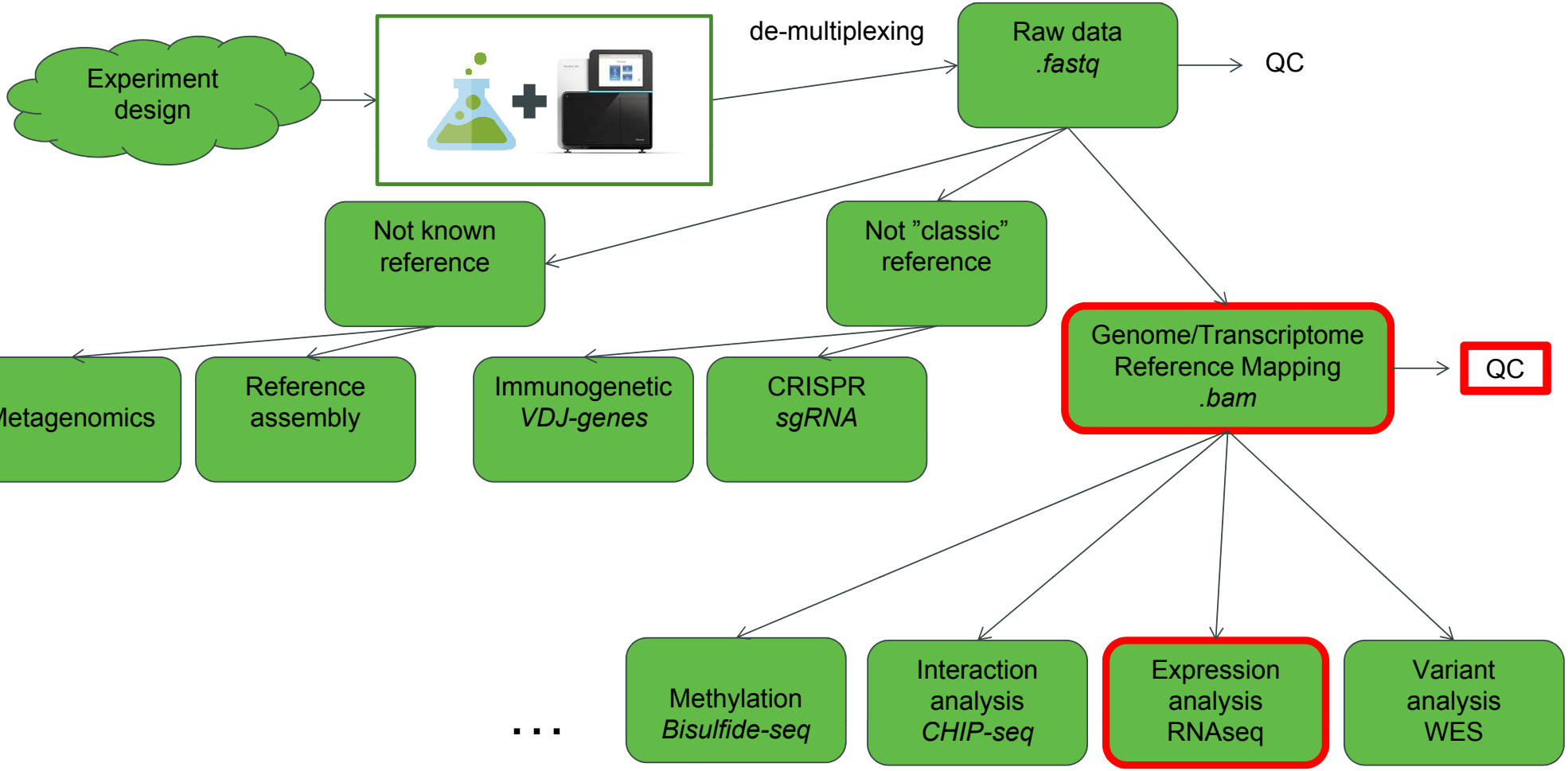
BRNO | CZECH REPUBLIC

Modern Genomic Technologies (LF:DSMGT01)

Lecture 5 : RNA-seq analysis

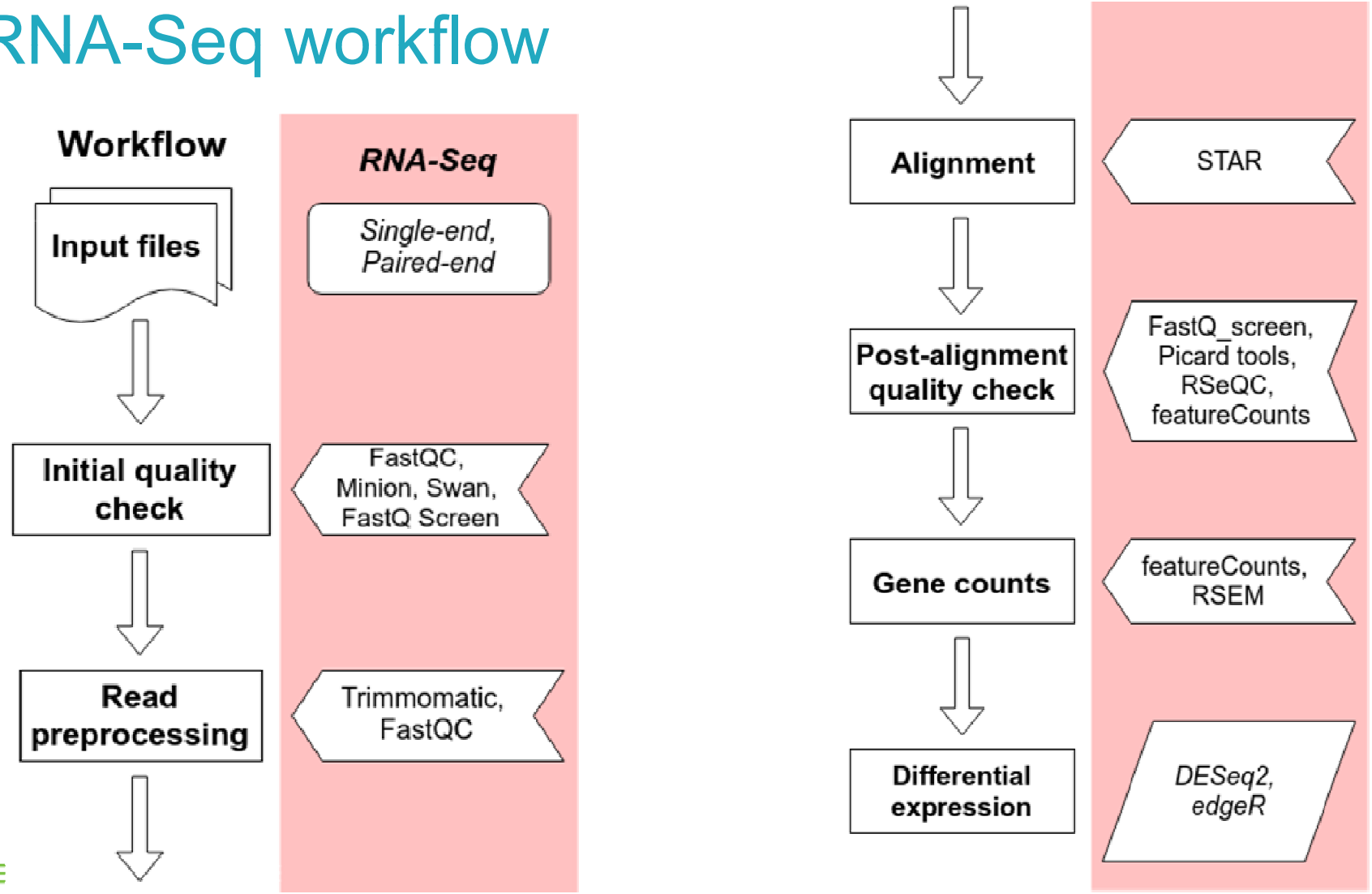
Vojta Bystry
vojtech.bystry@ceitec.muni.cz

NGS data analysis



CEITEC

The RNA-Seq workflow



CEITE

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)

Alignment

- Our choice is the `STAR` aligner
- It performs genome alignment
- Offers a lot of settings to support splicing, soft-clipping, chimeric alignments, ...
- Other techniques (`Salmon` or `Kallisto`) do not use alignment per se and can give you the gene count information right away
  - They use only transcriptome as a reference and are very quick
  - Drawback is you see only what's in the transcriptome and nothing else

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

- Post-alignment QC gives us information about the mapping
  - Number of mapped reads - unique + multi mapped
  - Mapped locations
  - Duplication rates
  - Library strand specificity
  - Captured biotypes
  - rRNA contamination
  - 5’ to 3’ end coverage bias
  - …

Post-alignment QC - Tools

- STAR alignment results - number of mappings and others
- `RSeQC` - mapped locations (`Read Distribution`), library strand specificity
- `featureCounts` biotypes - summary of mappings to gene biotypes
- `FastQ screen` (not exactly Post QC) - residual content of rRNA, tRNA, general mapping percentage to the genome (if selected)
- `Qualimap` - general alignment statistics focused on RNA-Seq (`rnaseq`) including gene body coverage

Post-alignment QC - RSeQC

- RSeQC is a general tool for many QC results
- Few of them are
  - `Read distribution` - calculates assignment of reads to different genomic features
  - `Infer experiment` - test strand specificity of the library
  - `Inner distance` - calculates approximate distance between read pairs

Post-alignment QC - FastQ_screen

- `FastQ screen` is a quick scan of potential mapping locations on different references
- We can use it to do a quick scan of contaminations (various organisms) as well as estimate residual rRNA content
  - In **polyA** selection based libraries we expect to have less then 2% rRNA content
  - In **rRNA** libraries we can have up to 10-15% of rRNA and still consider it a good library
- Biobloom other option more computationally expensive

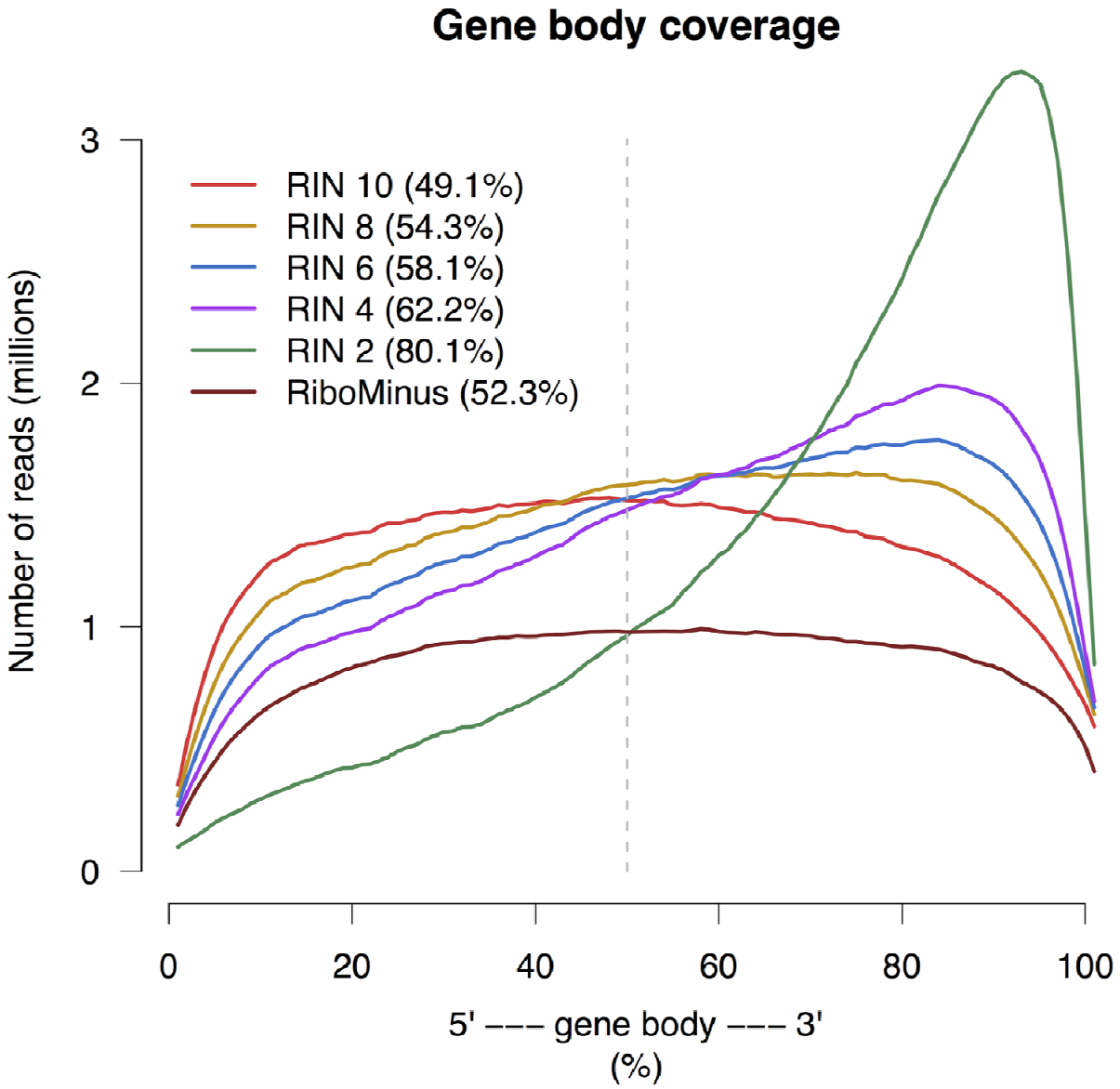
Post-alignment QC - Qualimap

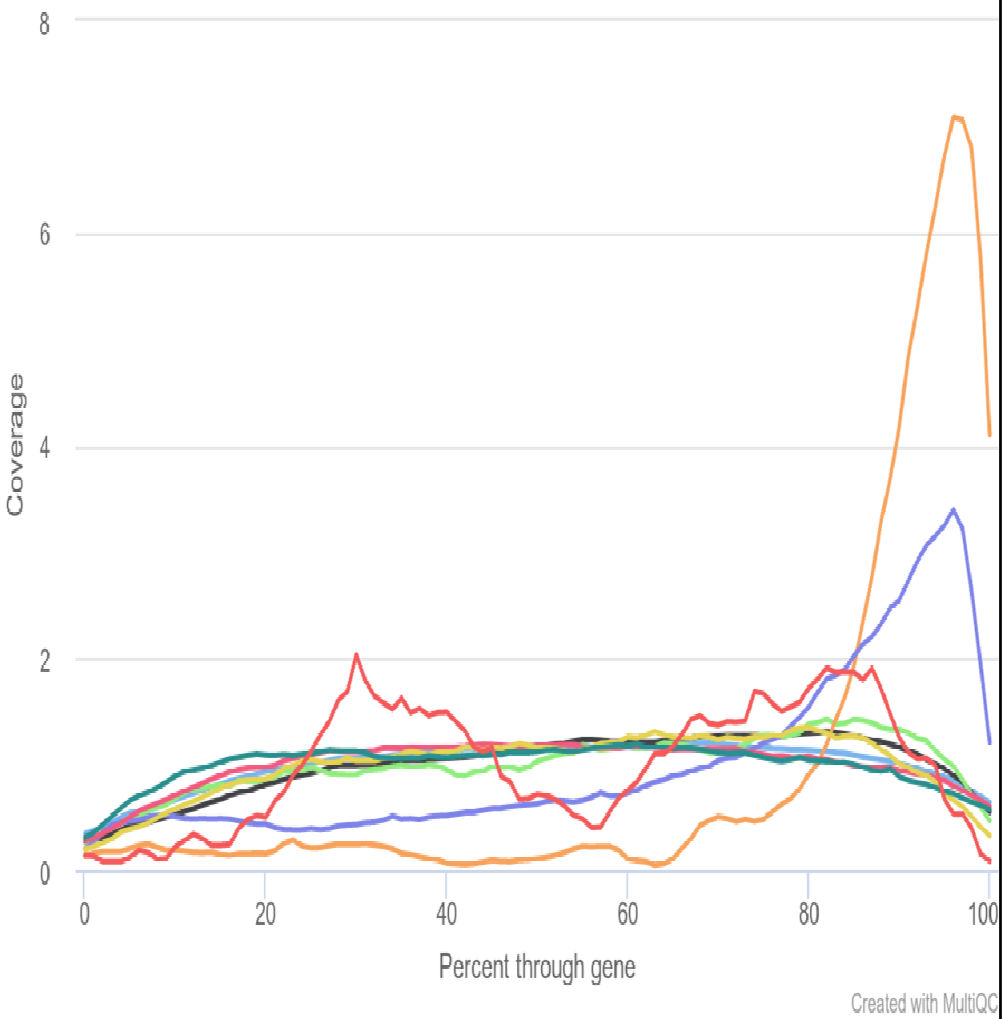
- `Qualimap` performs a numerous checks of the alignment
  - One of the modules is `rnaseq` which is focused directly on RNA-Seq alignments
- One of the main information we can get from this module is the gene body coverage
  - We would like to see a nice and even read mapping coverage along the whole length of the genes
  - The coverage, however, depends on the library fragmentation (low RIN, FFPE samples but also depends on the used library kit (Lexogen QuantSeq)!

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





Mapping QC

Mapping QC

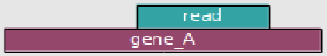
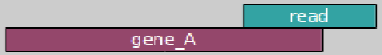
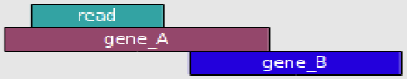
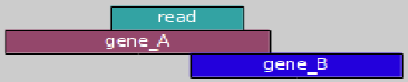
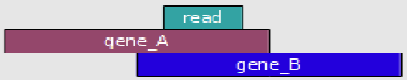
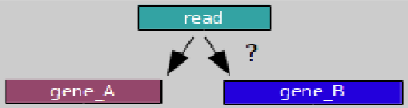
- Examples

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)

	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)		

CEITEC

Feature counting

- The regular settings are - summarize reads mapping to exons (-t exon) and sum them up to gene id (-g gene_id)
- Other possibilities:
  - Count per exons
  - Include introns
  - …

Gene counts - Tools

- `featureCounts` is build around the “classic” read to gene assignment
  - By default, assigns only uniquely mapped reads an only reads uniquely assignable to a single gene (but both can be changed)
  - Gives you **raw read counts** per **gene**
- `RSEM` is efficient in counting also multi mapped reads and can estimate expression of individual gene isoforms
  - Tries to “weight” the probability a mapped position of a multi mapped read and assign it correctly to the real source
  - Gives you **estimated counts** per **gene** as well as per **isoform** and normalized **TPM** = **Transcripts per million transcripts**

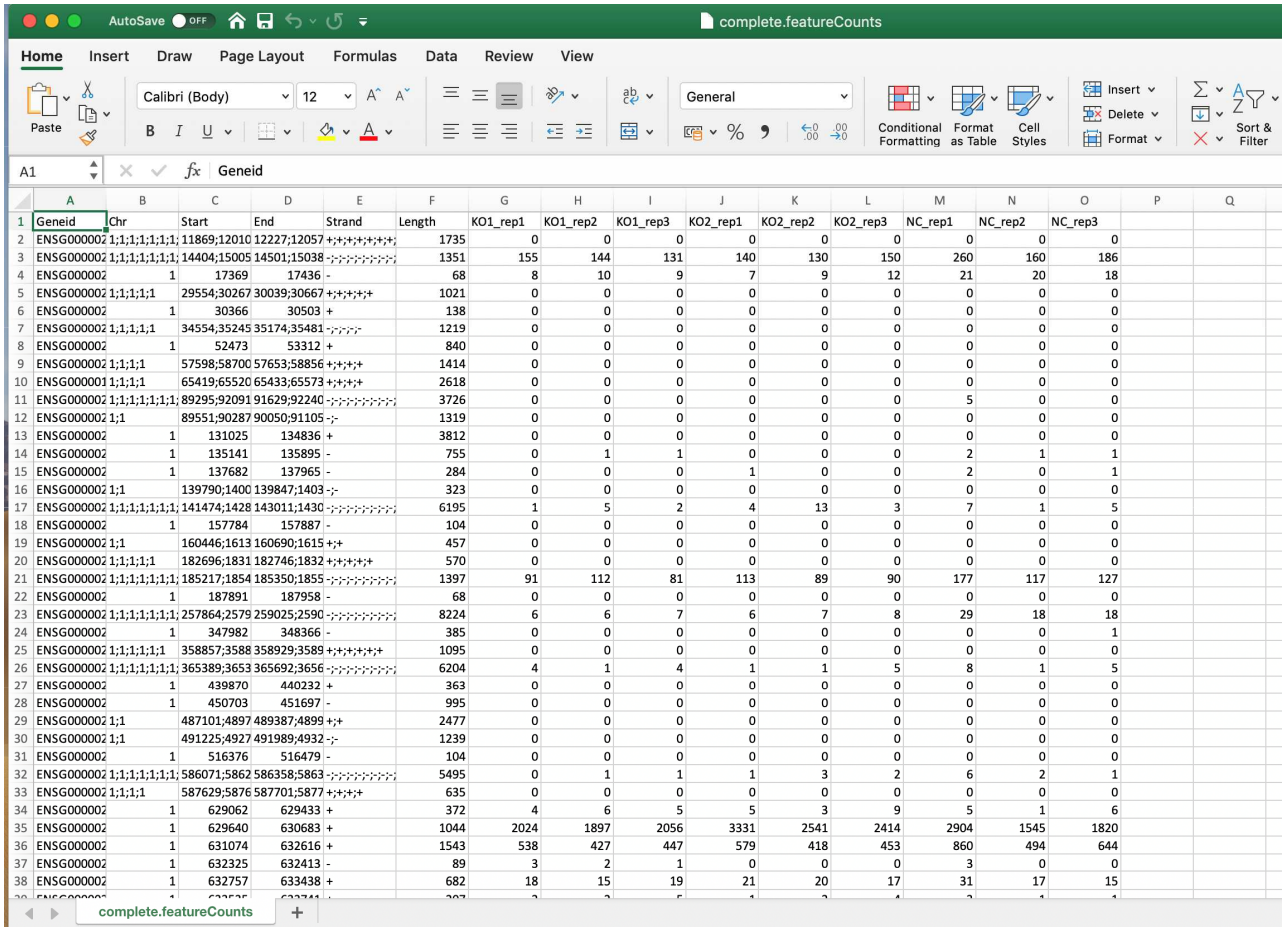
- But, there is a **big differences** in the **minimal required** “good” aligned reads

Source: Roberts et al. Nature Methods 2013

Minimal number of reads and expression I

- RSEM is less precise in low read counts (<40-50M reads) and for low expressed RNAs (difficult to estimate)
- For lower read counts it's safer to go for `featureCounts`
- Our best practices for a minimal read count for each tools:
  - Less than **40-50M aligned reads** (to the good stuff) -> `featureCounts`
  - More than **40-50M aligned reads** (to the good stuff) -> RSEM
- But if you want isoforms!!! -> RSEM

Feature count results



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

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

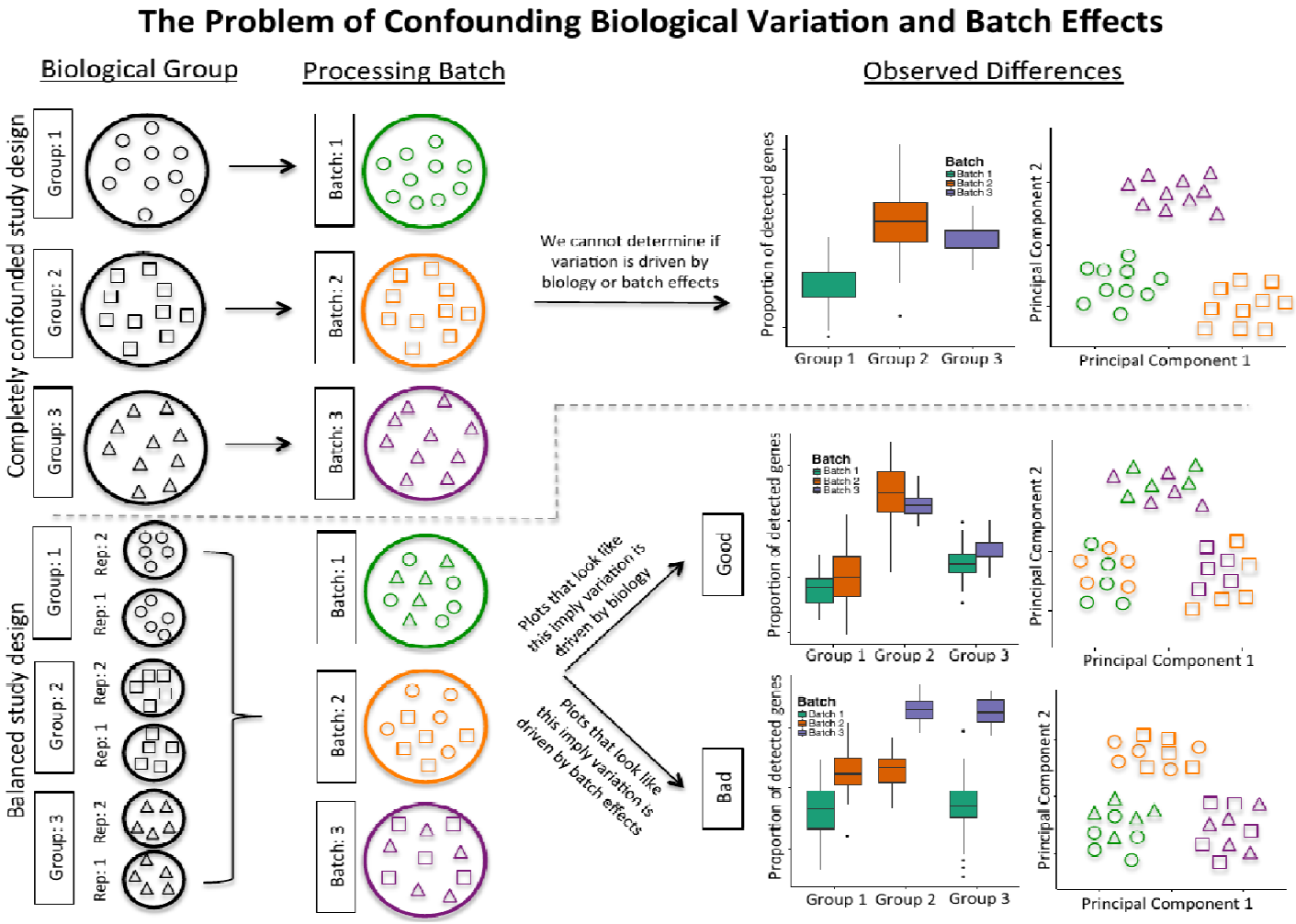
- DESeq2
  - More specific
- edgeR
  - More sensitive
- The important part of the calculation is the **design**
  - Assignment of a group/condition to a sample
  - If the samples are paired (the same patient twice) we have to account for this as well!
  - Technically, the pairing of the samples is a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

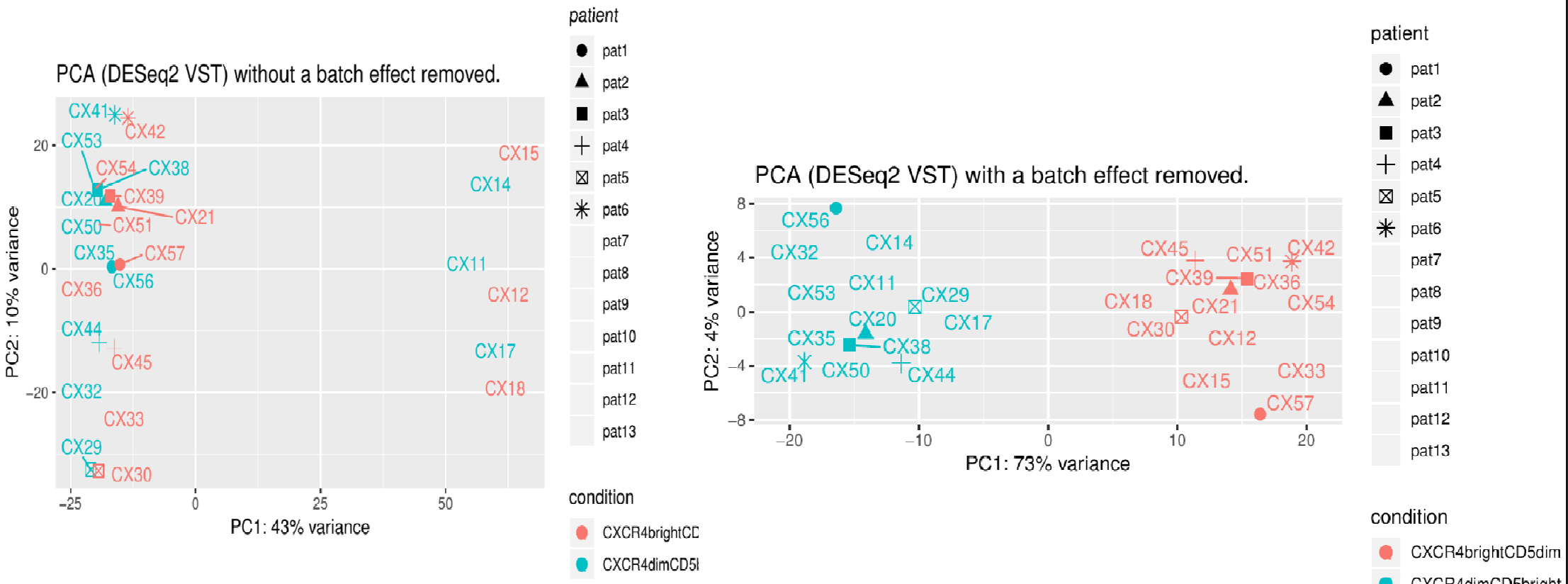
- There is a bad experimental design and a good experimental design
- Very simply - more randomization gives you better results



CEITEC

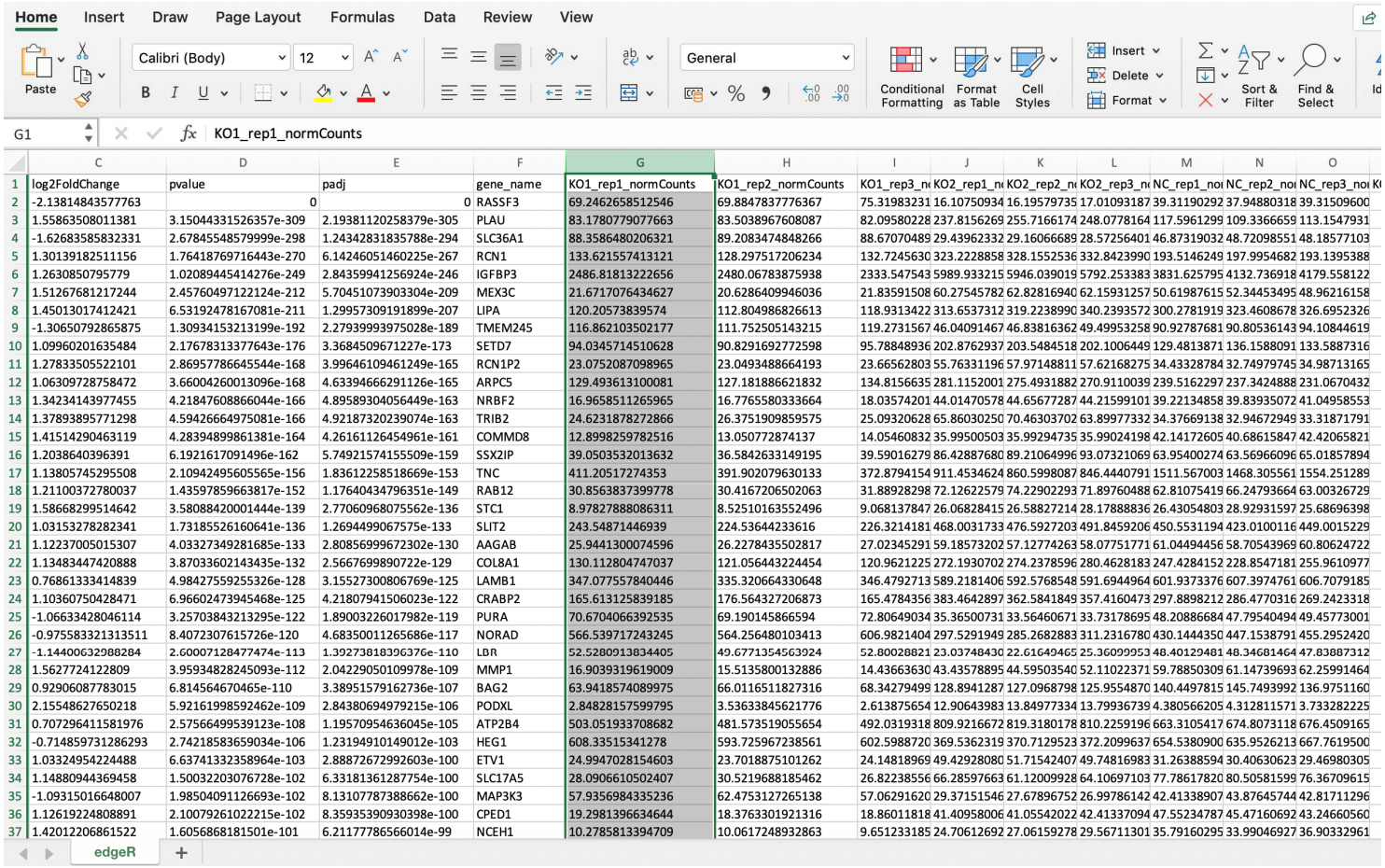
Pairing of the samples/batch effect

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



CEITEC

Differential expression results



Home Insert Draw Page Layout Formulas Data Review View

Paste | Calibri (Body) | 12 | General | Conditional Formatting | Format as Table | Cell Styles | Insert | Delete | Format | Sort & Filter | Find & Select

G1 | KO1_rep1_normCounts

	C	D	E	F	G	H	I	J	K	L	M	N	O
1	log2FoldChange	pvalue	padj	gene_name	KO1_rep1_normCounts	KO1_rep2_normCounts	KO1_rep3_n	KO2_rep1_n	KO2_rep2_n	KO2_rep3_n	NC_rep1_no	NC_rep2_no	NC_rep3_no
2	-2.13814843577763	0	0	RASSF3	69.2462658512546	69.8847837776367	75.31983231	16.10750934	16.19579735	17.01093187	39.31190292	37.94880318	39.31509600
3	1.55863508011381	3.15044331526357e-309	2.19381120258379e-305	PLAU	83.1780779077663	83.5038967608087	82.09580228	237.8156269	255.7166174	248.0778164	117.5961299	109.3366659	113.1547931
4	-1.62683585832331	2.67845548579999e-298	1.24342831835788e-294	SLC36A1	88.3586480206321	89.2083474848266	88.67070489	29.43962332	29.16066689	28.57256401	46.87319032	48.72098551	48.18577103
5	1.30139182511156	1.76418769716443e-270	6.14246051460225e-267	RCN1	133.621557413121	128.297517206234	132.7245630	323.2228858	328.1552536	332.8423990	193.5146249	197.9954682	193.1395388
6	1.2630850795779	1.02089445414276e-249	2.84359941256924e-246	IGFBP3	2486.81813222656	2480.06783875938	2333.547543	5989.933215	5946.039019	5792.253383	3831.625795	4132.736918	4179.558122
7	1.51267681217244	2.45760497122124e-212	5.70451073903304e-209	MEX3C	21.6717076434627	20.6286409946036	21.83591508	60.27545782	62.82816940	62.15931257	50.61987615	52.34453495	48.96216158
8	1.45013017412421	6.53192478167081e-211	1.29957309191899e-207	LIPA	120.20573839574	112.804986826613	118.9313422	313.6537312	319.2238990	340.2393572	300.2781919	323.4608678	326.6952326
9	-1.30650792865875	1.30934153213199e-192	2.27939993975028e-189	TMEM245	116.862103502177	111.752505143215	119.2731567	46.04091467	46.83816362	49.49953258	90.92787681	90.80536143	94.10844619
10	1.09960201635484	2.17678313377643e-176	3.3684509671227e-173	SETD7	94.0345714510628	90.8291692772598	95.78848936	202.8762937	203.5484518	202.1006449	129.4813871	136.1588091	133.5887316
11	1.27833505522101	2.86957786645544e-168	3.99646109461249e-165	RCN1P2	23.0752087098965	23.0493488664193	23.66562803	55.76331196	57.97148811	57.62168275	34.43328784	32.74979745	34.98713165
12	1.06309728758472	3.66004260013096e-168	4.63394666291126e-165	ARPC5	129.493613100081	127.181886621832	134.8156635	281.1152001	275.4931882	270.9110039	239.5162297	237.3424888	231.0670432
13	1.34234143977455	4.21847608866044e-166	4.89589304056449e-163	NRBF2	16.9658511265965	16.7765580333664	18.03574201	44.01470578	44.65677287	44.21599101	39.22134858	39.83935072	41.04958553
14	1.37893895771298	4.59426664975081e-166	4.92187320239074e-163	TRIB2	24.6231878272866	26.3751909859575	25.09320628	65.86030250	70.46303702	63.89977332	34.37669138	32.94672949	33.31871791
15	1.41514290463119	4.28394899861381e-164	4.26161126454961e-161	COMMD8	12.8998259782516	13.050772874137	14.05460832	35.99500503	35.99294735	35.99024198	42.14172605	40.68615847	42.42065821
16	1.2038640396391	6.1921617091496e-162	5.74921574155509e-159	SSX2IP	39.0503532013632	36.5842633149195	39.59016279	86.42887680	89.21064996	93.07321069	63.95400274	63.56966096	65.01857894
17	1.13805745295508	2.10942495605565e-156	1.83612258518669e-153	TNC	411.20517274353	391.902079630133	372.8794154	911.4534624	860.5998087	846.4440791	1511.567003	1468.305561	1554.251289
18	1.21100372780037	1.43597859663817e-152	1.17640434796351e-149	RAB12	30.8563837399778	30.4167206502063	31.88928298	72.12622579	74.22902293	71.89760488	62.81075419	66.24793664	63.00326729
19	1.58668299514642	3.58088420001444e-139	2.77060968075562e-136	STC1	8.97827888086311	8.52510163552496	9.068137847	26.06828415	26.58827214	28.17888836	26.43054803	28.92931597	25.68696398
20	1.03153278282341	1.73185526160641e-136	1.2694499067575e-133	SLIT2	243.54871446939	224.53644233616	226.3214181	468.0031733	476.5927203	491.8459206	450.5531194	423.0100116	449.0015229
21	1.12237005015307	4.03327349281685e-133	2.80856999672302e-130	AAGAB	25.9441300074596	26.2278435502817	27.02345291	59.18573202	57.12774263	58.07751771	61.04494456	58.70543969	60.80624722
22	1.13483447420888	3.87033602143435e-132	2.5667699890722e-129	COL8A1	130.112804747037	121.056443224454	120.9621225	272.1930702	274.2378596	280.4628183	247.4284152	228.8547181	255.9610977
23	0.76861333414839	4.98427559255326e-128	3.15527300806769e-125	LAMB1	347.077557840446	335.320664330648	346.4792713	589.2181406	592.5768548	591.6944964	601.9373376	607.3974761	606.7079185
24	1.10360750428471	6.96602473945468e-125	4.21807941506023e-122	CRABP2	165.613125839185	176.564327206873	165.4784356	383.4642897	362.5841849	357.4160473	297.8898212	286.4770316	269.2423318
25	-1.06633428046114	3.25703843213295e-122	1.89003226017982e-119	PURA	70.6704066392535	69.190145866594	72.80649034	35.36500731	33.56460671	33.73178695	48.20886684	47.79540494	49.45773001
26	-0.975583321313511	8.4072307615726e-120	4.68350011265686e-117	NORAD	566.539717243245	564.256480103413	606.9821404	297.5291949	285.2682883	311.2316780	430.1444350	447.1538791	455.2952420
27	-1.14400632988284	2.60007128477474e-113	1.39273818396376e-110	LBR	52.5280913834405	49.6771354563924	52.80028821	23.03748430	22.61649465	25.36099953	48.40129481	48.34681464	47.83887312
28	1.5627724122809	3.95934828245093e-112	2.04229050109978e-109	MMP1	16.9039319619009	15.5135800132886	14.43663630	43.43578895	44.59503540	52.11022371	59.78850309	61.14739693	62.25991464
29	0.92906087783015	6.814564670465e-110	3.38951579162736e-107	BAG2	63.9418574089975	66.0116511827316	68.34279499	128.8941287	127.0968798	125.9554870	140.4497815	145.7493992	136.9751160
30	2.15548627650218	5.92161998592462e-109	2.84380694979215e-106	PODXL	2.84828157599795	3.53633845621776	2.613875654	12.90643983	13.84977334	13.79936739	4.380566205	4.312811571	3.733282225
31	0.707296411581976	2.57566499539123e-108	1.19570954636045e-105	ATP2B4	503.051933708682	481.573519055654	492.0319318	809.9216672	819.3180178	810.2259196	663.3105417	674.8073118	676.4509165
32	-0.714859731286293	2.74218583659034e-106	1.23194910149012e-103	HEG1	608.33515341278	593.725967238561	602.5988720	369.5362319	370.7129523	372.2099637	654.5380900	635.9526213	667.7619500
33	1.03324954224488	6.63741332358964e-103	2.88872672992603e-100	ETV1	24.9947028154603	23.7018875101262	24.14818969	49.42928080	51.71542407	49.74816983	31.26388594	30.40630623	29.46980305
34	1.14880944369458	1.50032203076728e-102	6.33181361287754e-100	SLC17A5	28.0906610502407	30.5219688185462	26.82238556	66.28597663	61.12009928	64.10697103	77.78617820	80.50581599	76.36709615
35	-1.09315016648007	1.98504091126693e-102	8.13107787388662e-100	MAP3K3	57.9356984335236	62.4753127265138	57.06291620	29.37151546	27.67896752	26.99786142	42.41338907	43.87645744	42.81711296
36	1.12619224808891	2.10079261022215e-102	8.35935390930398e-100	CPED1	19.2981396634644	18.3763301921316	18.86011818	41.40958006	41.05542022	42.41337094	47.55234787	45.47160692	43.24660560
37	1.42012206861522	1.6056868181501e-101	6.21177786566014e-99	NCEH1	10.2785813394709	10.0617248932863	9.651233185	24.70612692	27.06159278	29.56711301	35.79160295	33.99046927	36.90332961

edgeR

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**…:)

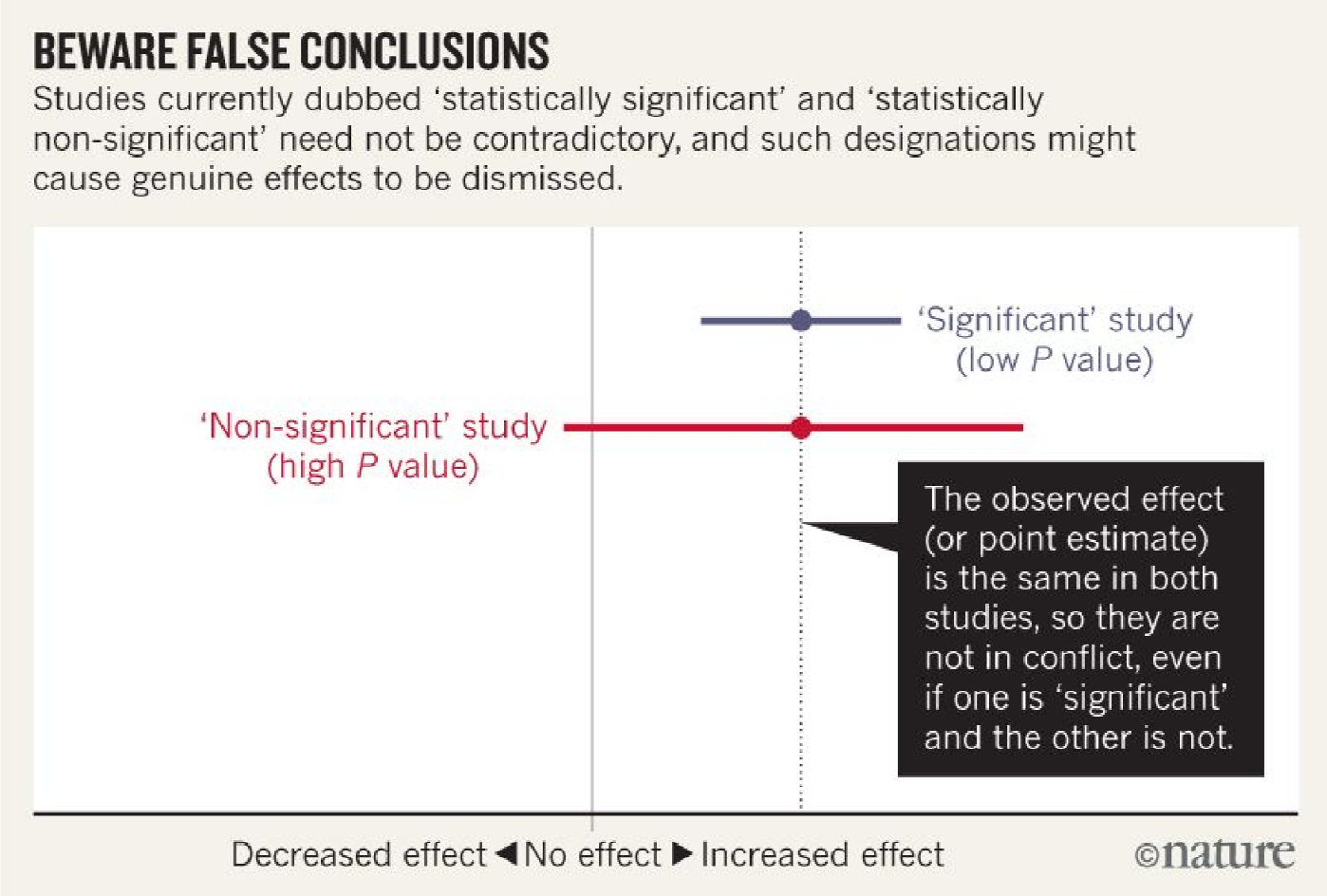
Selection of the **differentially expressed** (DE) gene is **completely up to you**

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

Statistical significance ≠ biological relevance!!!

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

P-value significance



Differential expression output

- Example

CEITEC

@CEITEC_Brno

Thank you for your attention!

Vojta Bystry
vojtech.bystry@ceitec.muni.cz

CEITEC

www.ceitec.eu