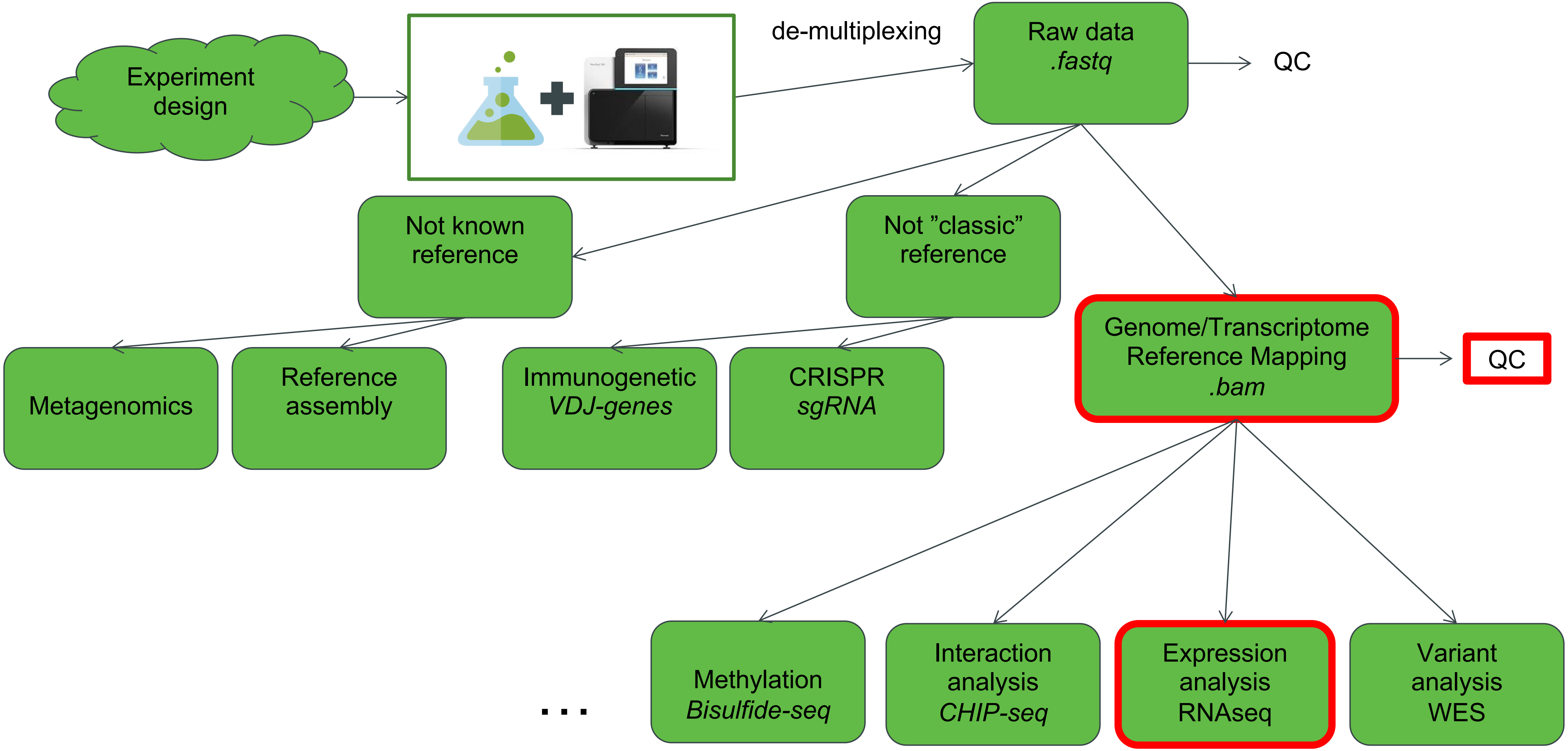
CEITEC
Central European Institute of Technology
BRNO | CZECH REPUBLIC

Modern methods for genome analysis (PřF:Bi7420)

Lecture 5 : RNA-seq primary analysis

Vojta Bystry
vojtech.bystry@ceitec.muni.cz

NGS data analysis



Experiment design

de-multiplexing

Raw data
.fastq

QC

Not known reference

Not ”classic” reference

Genome/Transcriptome Reference Mapping
.bam

QC

Metagenomics

Reference assembly

Immunogenetic
VDJ-genes

CRISPR
sgRNA

. . .

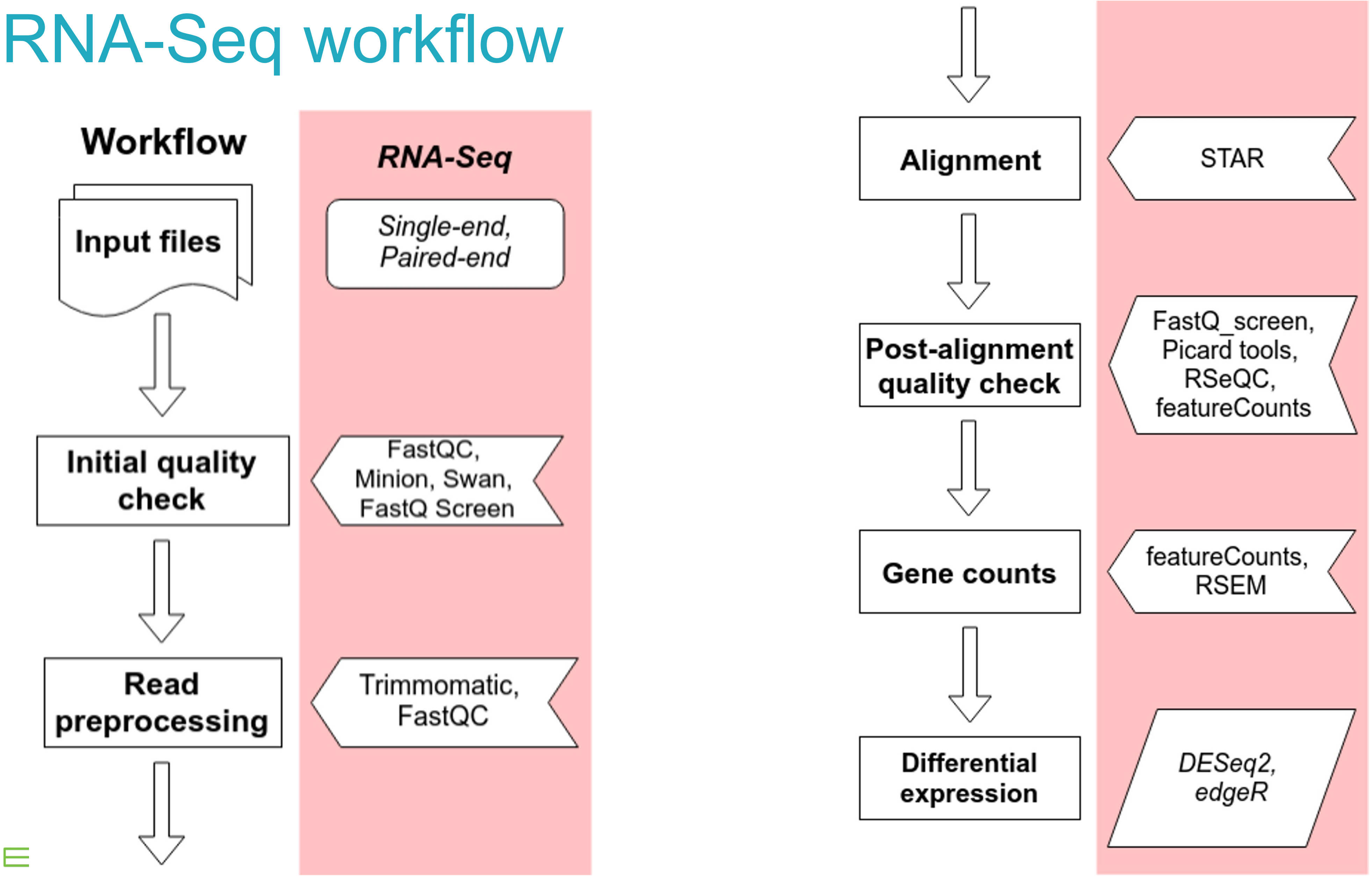
Methylation
Bisulfide-seq

Interaction analysis
CHIP-seq

Expression analysis
RNAseq

Variant analysis
WES

The RNA-Seq workflow



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

- 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

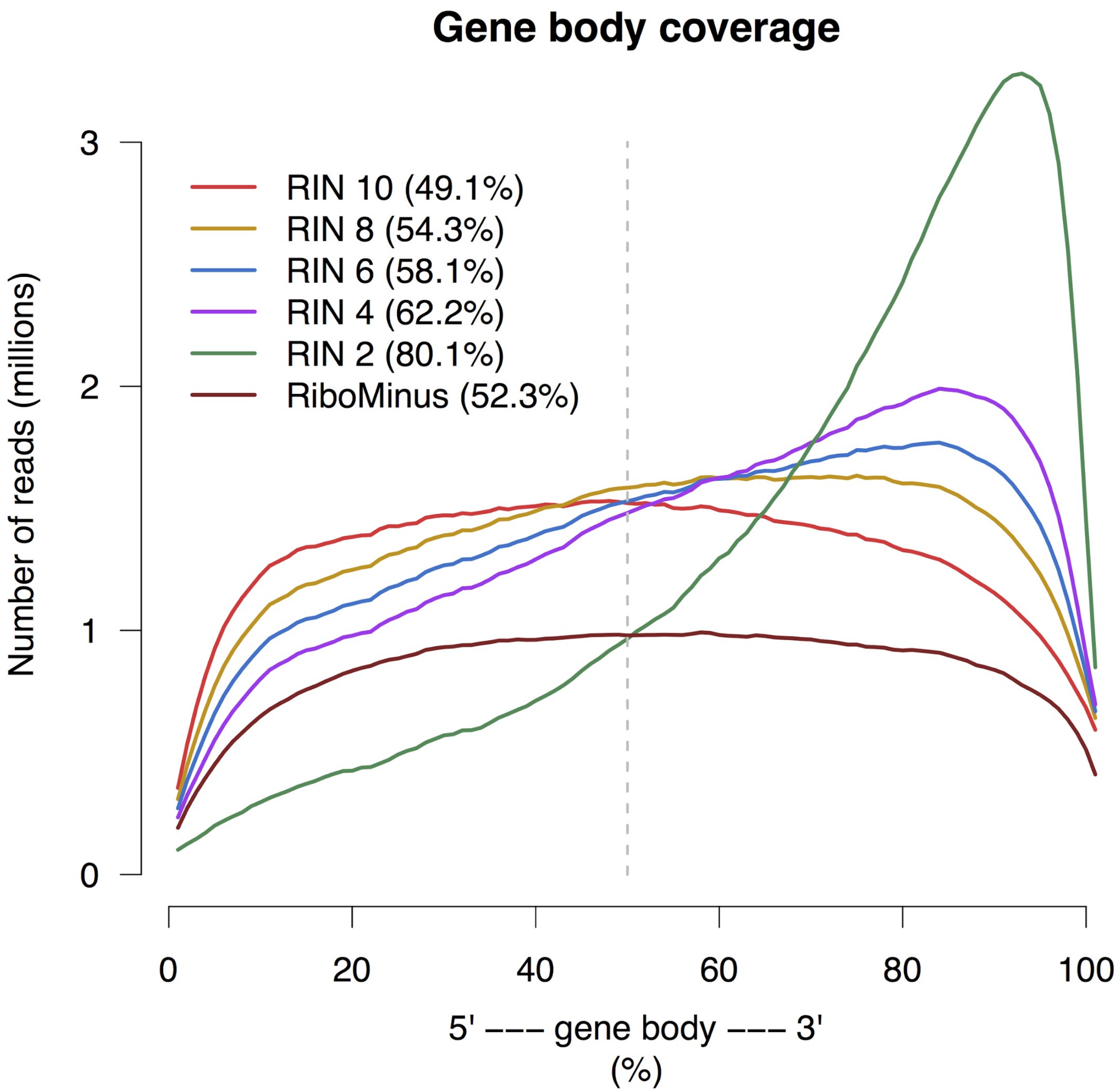
Post-alignment QC - Tools

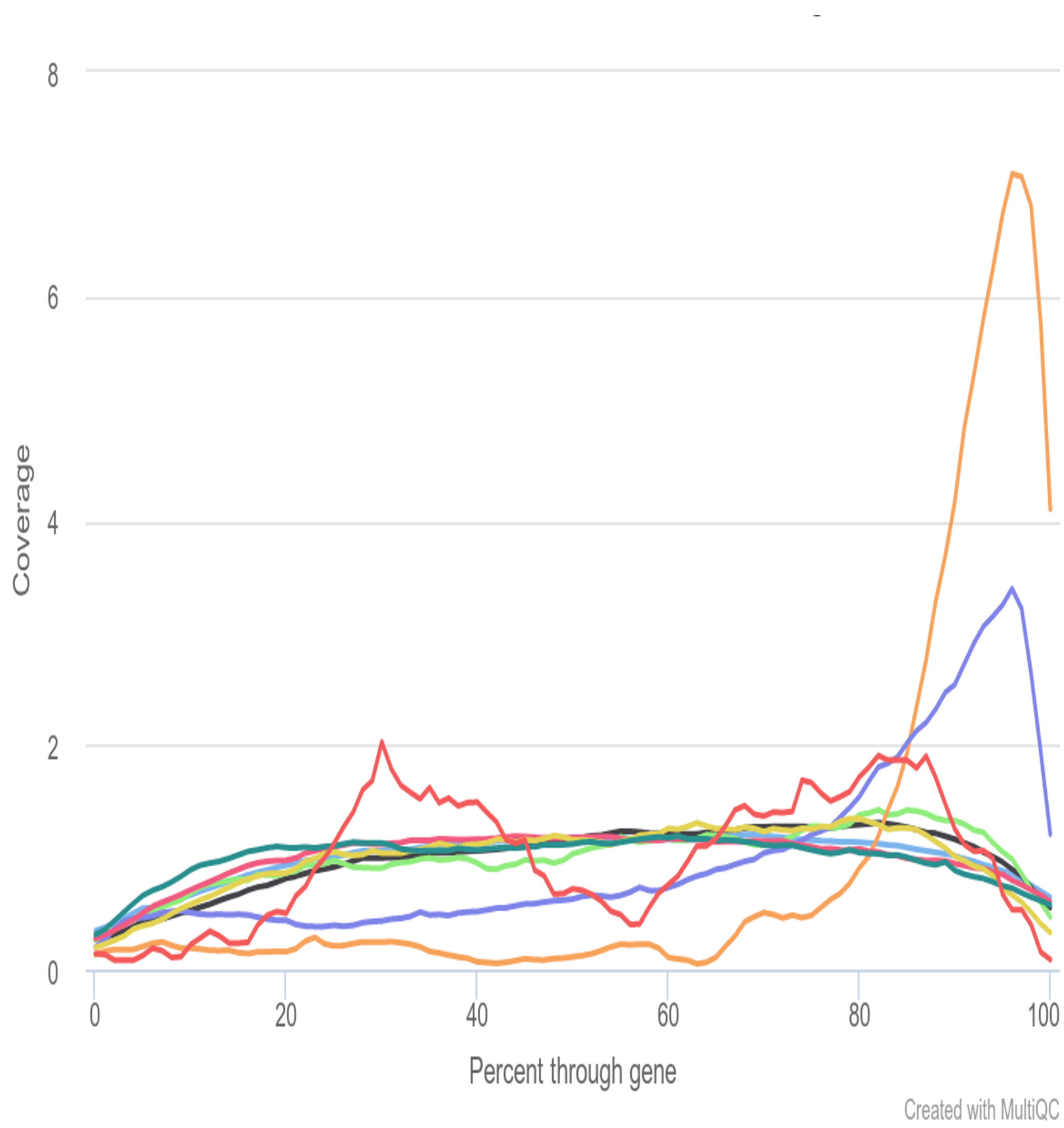
- Aligner report
  - STAR – most direct assesment
- General QC tools
  - RSeQC
  - Picard
  - Qualimap
- Feature counting tools
  - featureCounts
  - RSEM
- Non-aligment tools
  - FastQ screen
  - Biobloom

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)

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**

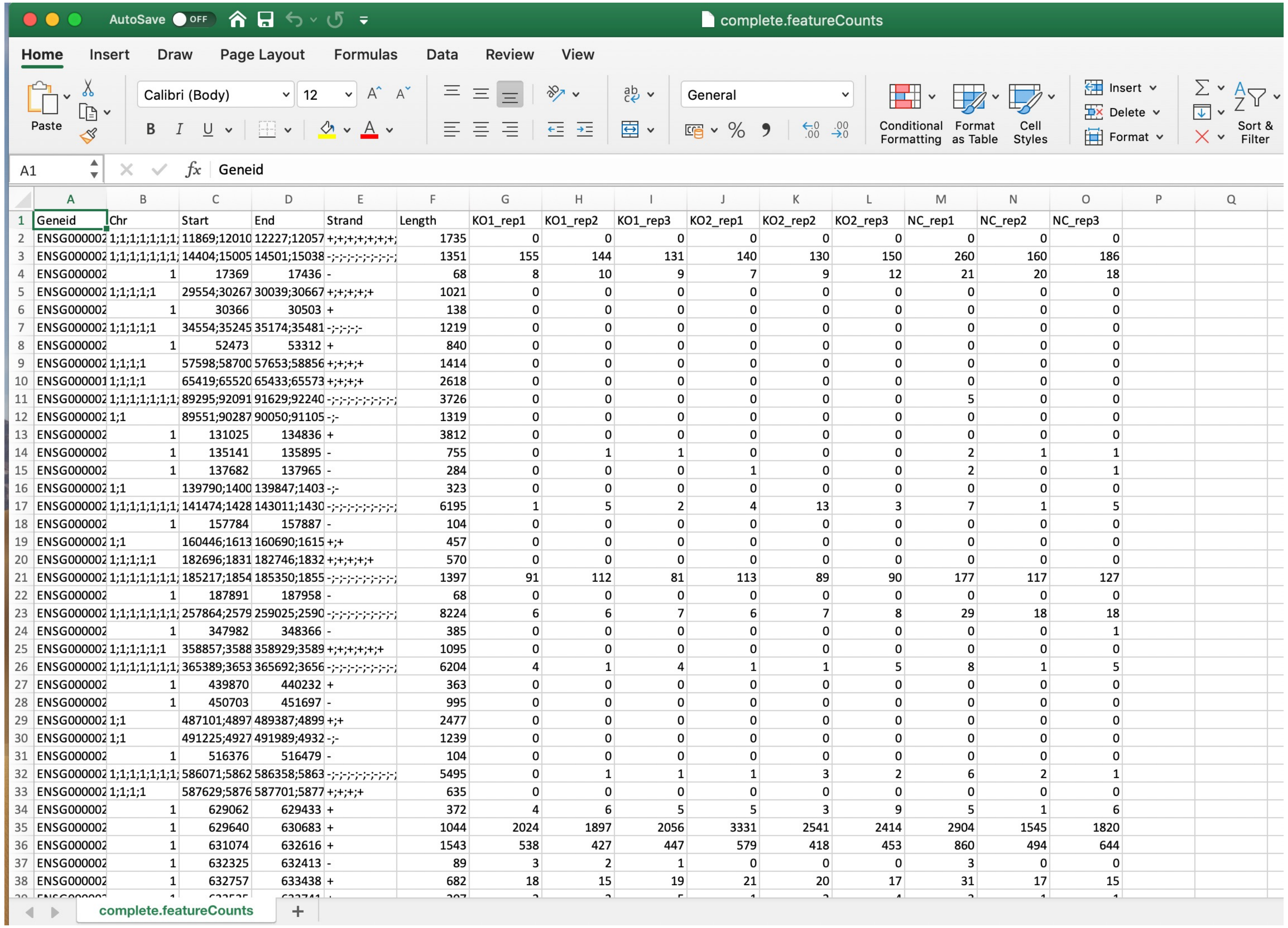
- But, there is a **big differences** in the **minimal required** “good” aligned reads

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

Post-alignment QC - example

CEITEC

@CEITEC_Brno

Thank you for your attention!

Vojta Bystry
vojtech.bystry@ceitec.muni.cz

www.ceitec.eu