RNA sequencing

- part 1 -

Deutsches Krebsforschungszentrum
Angewandte Bioinformatik (Prof. Dr. Benedikt Brors)
Dr. Siao-Han Wong (s.wong@dkfz.de)

Disclaimer

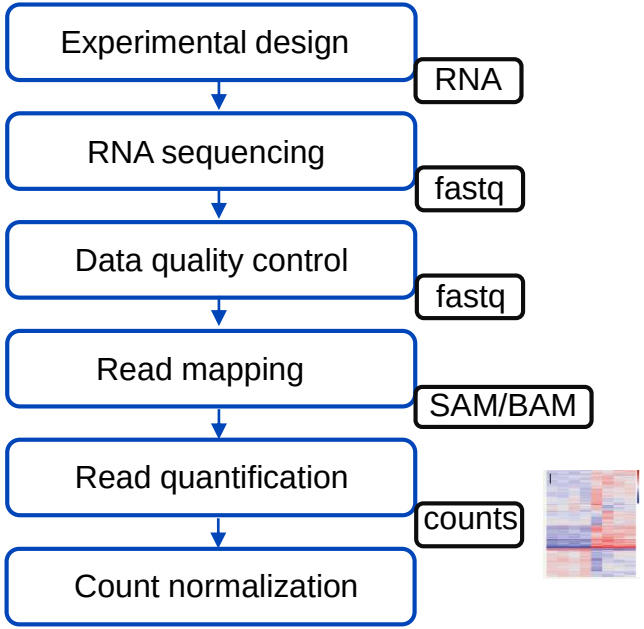
- Based on material from:
  - Benedikt Brors
  - Matthias Schlesner
  - Lena Voithenberg
  - Óscar González-Velasco
  - Roman Kurilov
  - et al.

Outline

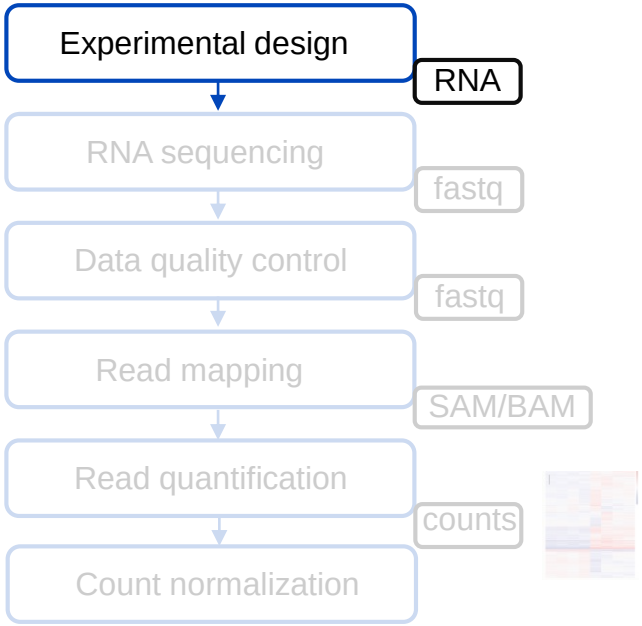
Dr. Siao-Han Wong

Dr. Óscar González-Velasco

Dr. Siao-Han Wong



Outline


Experimental design
RNA
RNA sequencing
fastq
Data quality control
fastq
Read mapping
SAM/BAM
Read quantification
counts
Count normalization

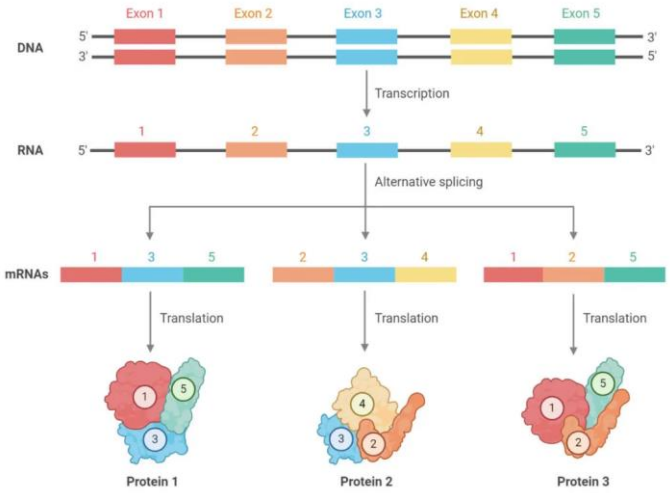
Considerations

What are the goals of the experiment?

- differential expression analysis
- identification of rare transcripts
- detection of splice junctions
- transcriptome assembly

What are the characteristics of the system?

- RNA composition, e.g. rRNA, mRNA, miRNA
- introns
- high degree of alternative splicing
- reference genome/transcriptome available?



https://microbenotes.com/rna-splicing/

Experimental design

Comparison/Control

- appropriate control condition

Account for confounding factors

- groups based on level of confounding factor

Replicates

- sample size
- biological (and technical) replicates

Randomization

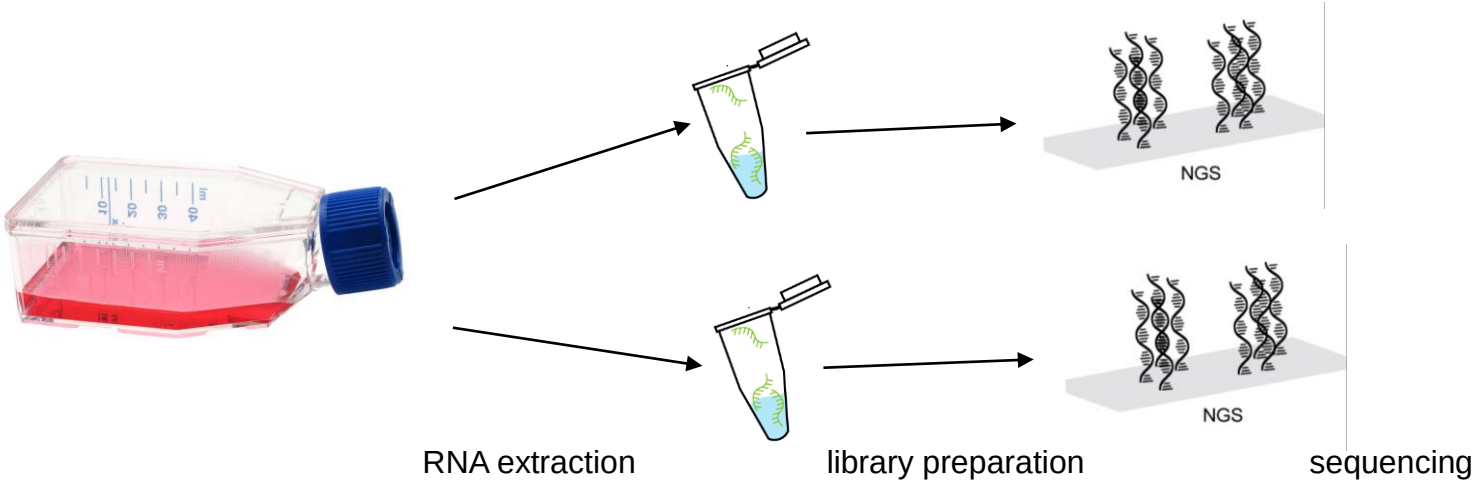
Replicates

- Replicates reduce variance in point estimates
- Give a more realistic answer in heterogeneous settings
- Different levels of replication
  - Technical replicates (e.g. sequence same sample twice)
  - Biological replicates (sequence different samples from each condition)
  - Independent samples
- Biological variation typically much higher then technical „noise“

Slide by Matthias Schlesner

Replicates

- Technical replicates: Prepare several libraries from the same sample
  - control for measurement accuracy



https://www.dreamstime.com/photos-images/tissue-culture-flask.html

Replicates

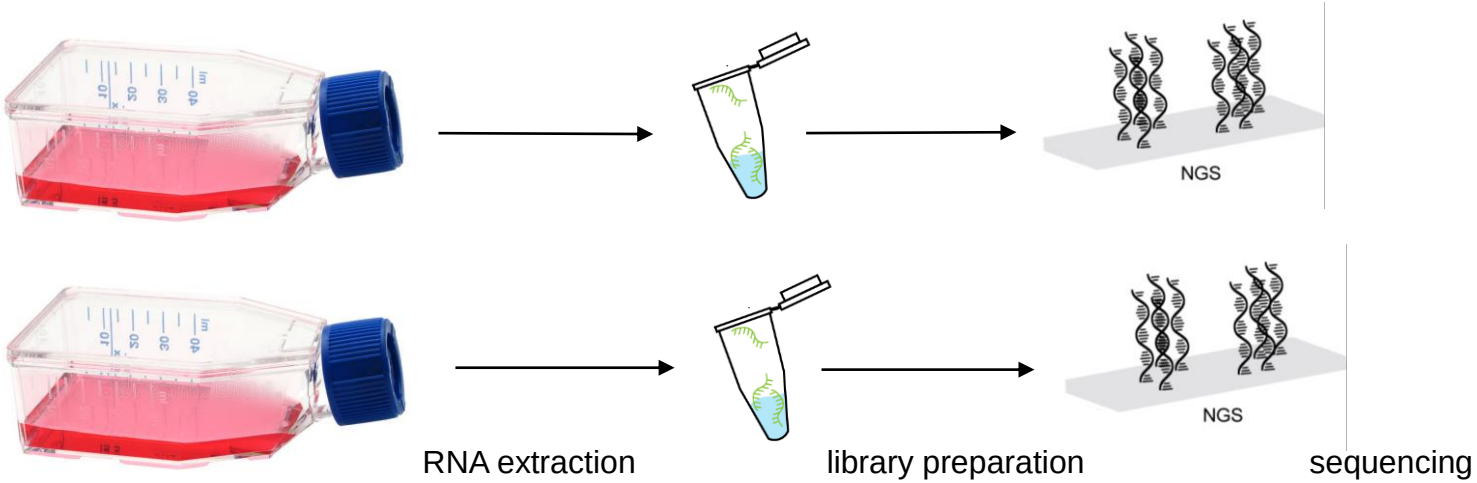
- Technical replicates: Prepare several libraries from the same sample
  - control for measurement accuracy

➡ **allows for conclusions about just this sample/ preparation and measurement accuracy**
➡ **usually not needed for RNAseq**

Adapted from slide by Matthias Schlesner

Replicates

- Biological replicates: several samples from the same cell line
  - control for measurement accuracy and variations in environment and the cells' response to them



https://www.dreamstime.com/photos-images/tissue-culture-flask.html

Replicates

- Biological replicates: several samples from the same cell line
  - control for measurement accuracy and variations in environment and the cells' response to them

➡ **allows for conclusions about the specific cell line**

Adapted from slide by Matthias Schlesner

Replicates

- Biological replicates: samples from multiple individuals
  controls for
  - measurement accuracy,
  - variations in environment, and
  - variations in gene expression levels (between cells), or
  - variations in genotype (between individuals)

➡ **allows for conclusions about the species**

Adapted from slide by Matthias Schlesner

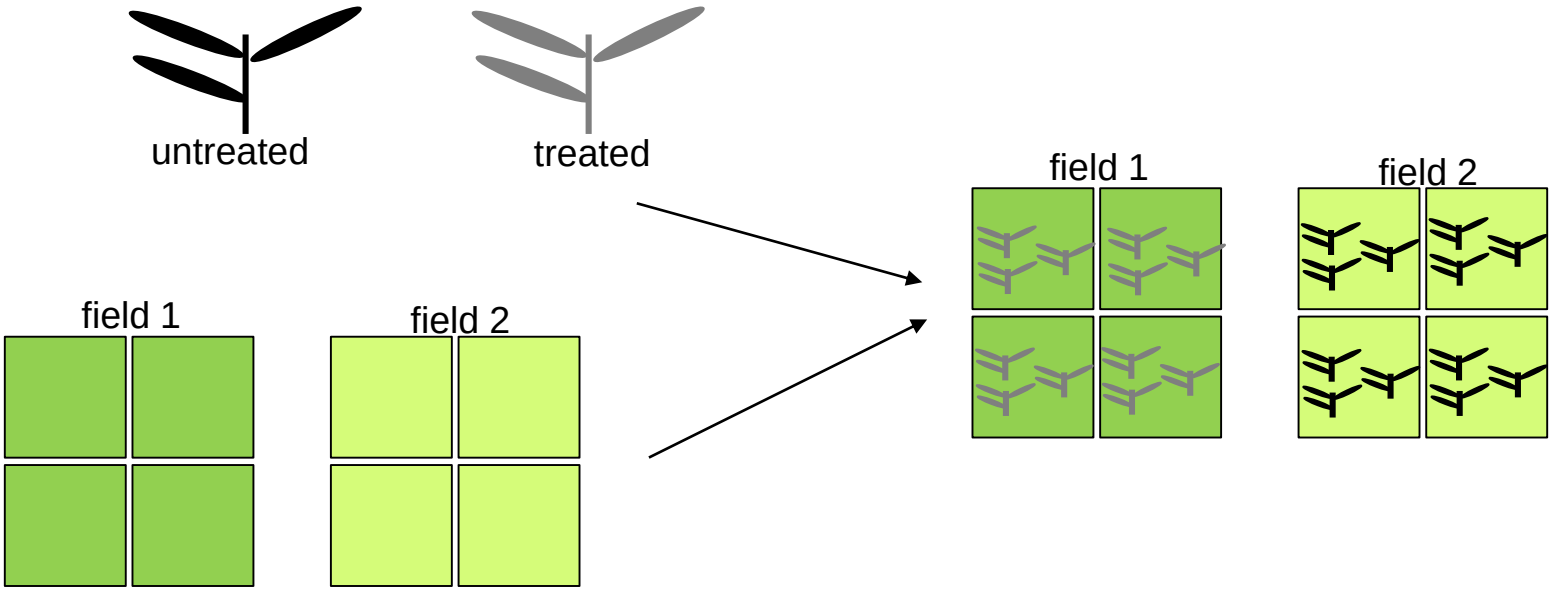
Number of replicates

- Two replicates permit to
  - globally estimate variation
- Sufficiently many replicates permit to
  - estimate variation for each gene
  - randomize out unknown covariates
  - spot outliers
  - improve precision of expression and fold-change estimates
- Statistical rule of thumb: at least 6 per condition
  - But depends: in isogenic cell lines, less may be needed
  - In heterogeneous patient cohorts, many more are advisable
- Systematic analysis: Schurch et al., 2016 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4878611/)

Slide by Matthias Schlesner

Replicates: Example I

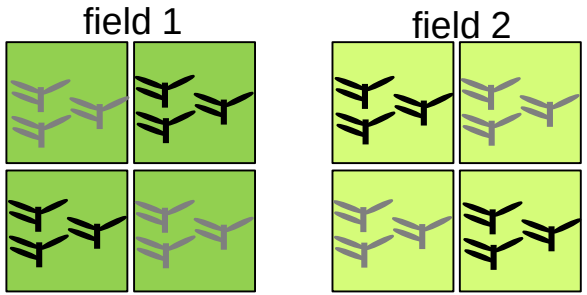
Research question: are there any differences between untreated and treated plants?

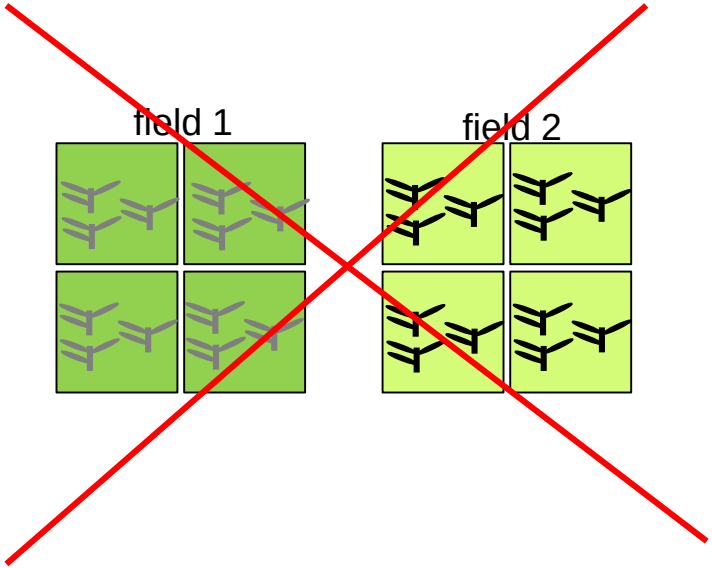


Inspired by Functional Genomics Center Zurich

Replicates: Example I

Research question: are there any differences between untreated and treated plants?





Inspired by Functional Genomics Center Zurich

Avoid batch effects

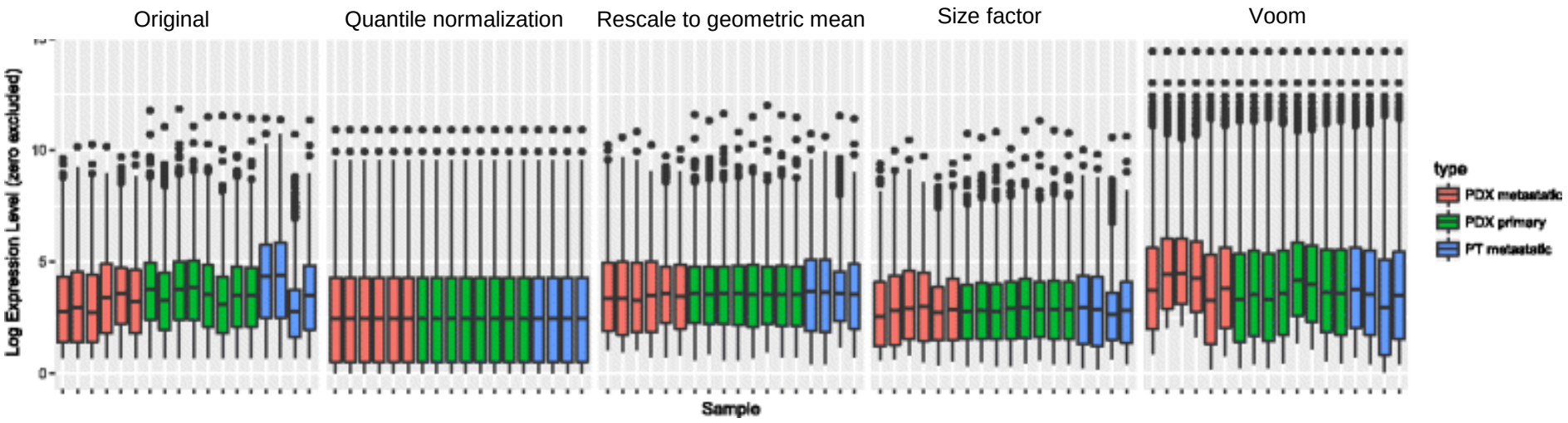
Batch effects

- technical biases
- experimental conditions are confounded (e.g.
    - all untreated plants on one field, or
    - RNA from all treated cells extracted on day 1 and RNA from all controls extracted on day 2
    - controls and treated cells in different library preparation batches/sequencing runs
    - cases from different sample sites/laboratories)

=> reduce confounding factors:

- design your experiment (discuss with biostatisticians or bioinformaticians)
- randomize sample handling as much as possible
- during data quality control check for batch effects (principle component analysis, cluster analysis…)

Avoid batch effects - normalization



Zhu *et al.* (2017) *Genome Medicine* 9:108

Sequencing conditions

Sequencing depth

- dependent on scientific question (e.g. special interest in genes with low abundance)
- for transcript quantification: ~10-30 Mio reads
- for transcript reconstruction: ~200 Mio reads

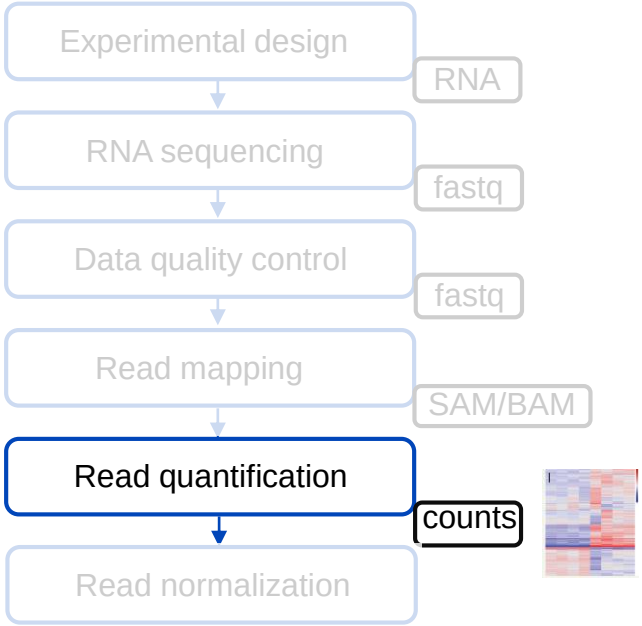
http://genome.ucsc.edu/ENCODE/protocols/dataStandards/RNA_standards_v1_2011_May.pdf

Single-end or paired-end?

- single-end sequencing sufficient for expression level quantification
- paired-end: better mapping in low complexity regions

Adapted from slide by Matthias Schlesner

Outline

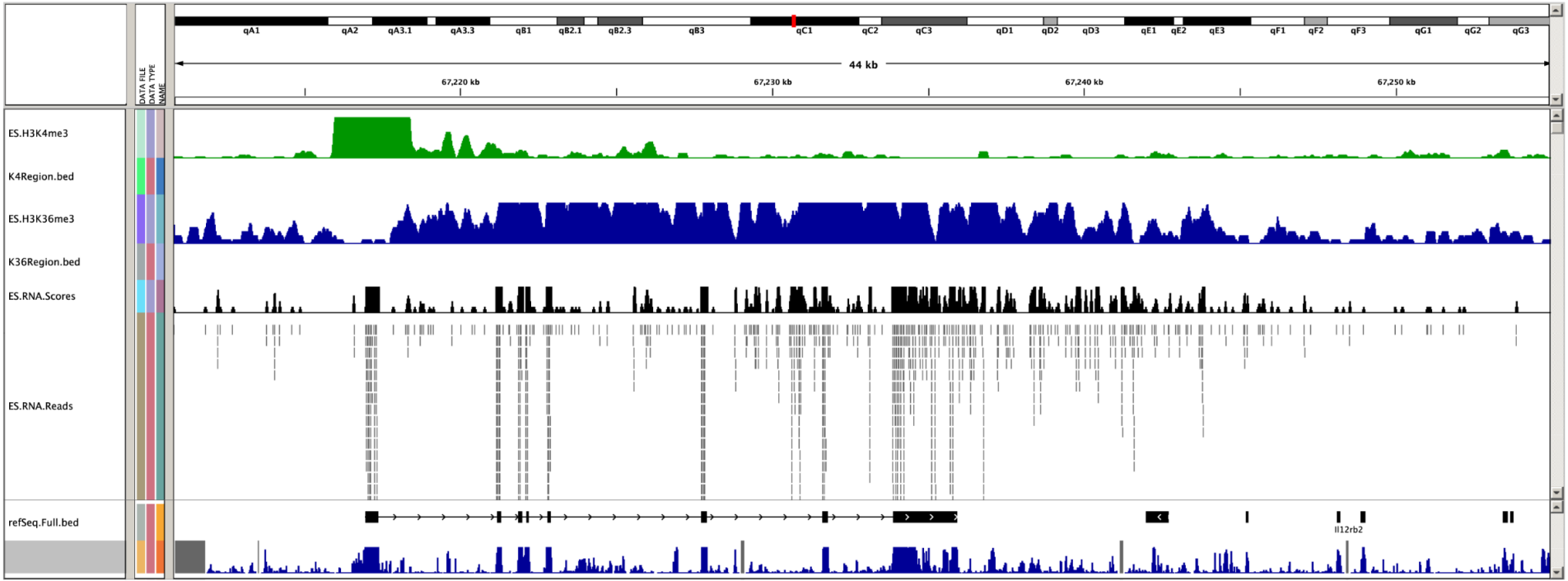

Experimental design
RNA
RNA sequencing
fastq
Data quality control
fastq
Read mapping
SAM/BAM
Read quantification
counts
Read normalization

Visualize mapped reads

Haas and Zody (2010) Nat. Biotechnology 28:421

Genome browser

- UCSC Genome Browser: https://genome.ucsc.edu/
- Integrative Genomics Viewer: http://software.broadinstitute.org/software/igv/



Robinson et al. (2011) Nat. Biotechnology 29:24

Introduced yesterday

Quantification of gene expression

=> detect amount of sequenced reads mapped to a specific gene or transcript

Levels of quantification

Gene level
- simple
- does not consider differentially spliced transcripts

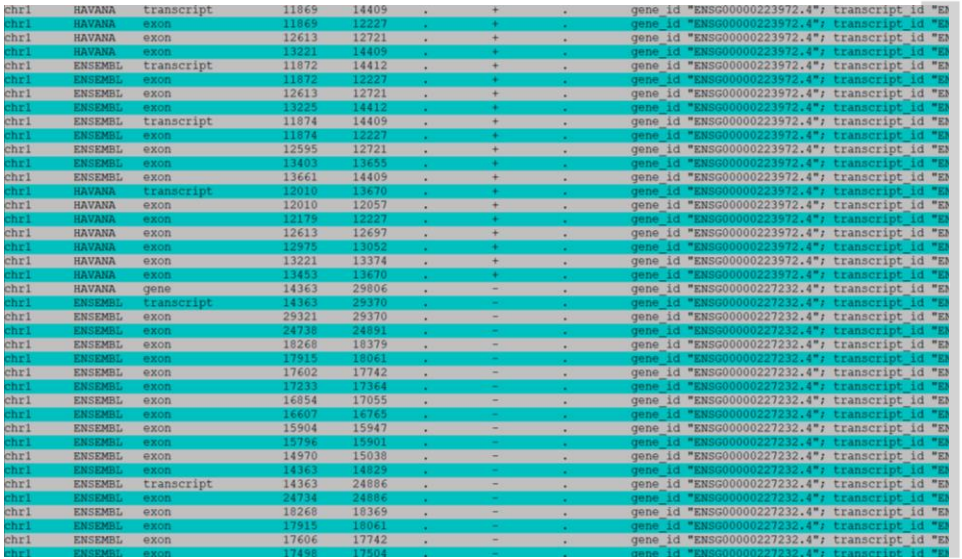
Exon level
- allows one to detect differential exon usage

Transcript/Isoform level
- detection of isoform-specific reads

Quantification of gene expression

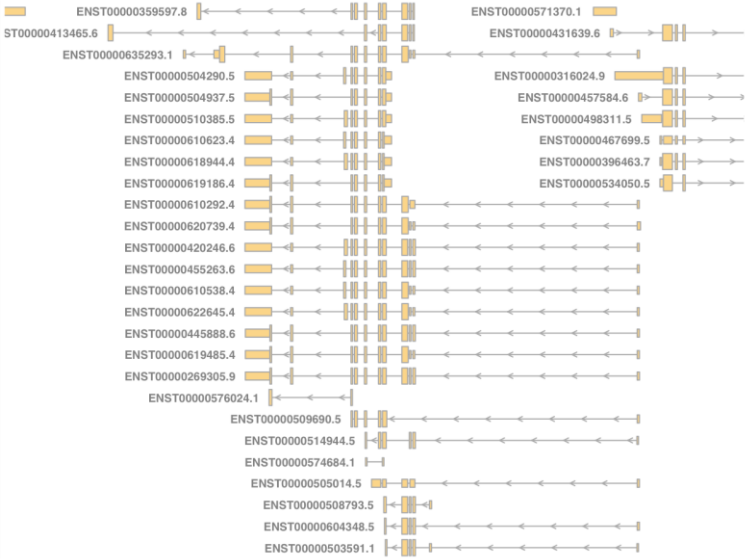
Gene models

- region of a gene thought to be transcribed into RNA
- source: Ensembl, UCSC, GENCODE, RefSeq, etc.
- formats: GTF, GFF, etc.



chr1	HAVANA	transcript	11869	14409	.	+	.	gene_id "ENSG00000223972.4"; transcript_id "EN
chr1	HAVANA	exon	11869	12227	.	+	.	gene_id "ENSG00000223972.4"; transcript_id "EN
chr1	HAVANA	exon	12613	12721	.	+	.	gene_id "ENSG00000223972.4"; transcript_id "EN
chr1	HAVANA	exon	13221	14409	.	+	.	gene_id "ENSG00000223972.4"; transcript_id "EN
chr1	ENSEMBL	transcript	11872	14412	.	+	.	gene_id "ENSG00000223972.4"; transcript_id "EN
chr1	ENSEMBL	exon	11872	12227	.	+	.	gene_id "ENSG00000223972.4"; transcript_id "EN
chr1	ENSEMBL	exon	12613	12721	.	+	.	gene_id "ENSG00000223972.4"; transcript_id "EN
chr1	ENSEMBL	exon	13225	14412	.	+	.	gene_id "ENSG00000223972.4"; transcript_id "EN
chr1	ENSEMBL	transcript	11874	14409	.	+	.	gene_id "ENSG00000223972.4"; transcript_id "EN
chr1	ENSEMBL	exon	11874	12227	.	+	.	gene_id "ENSG00000223972.4"; transcript_id "EN
chr1	ENSEMBL	exon	12595	12721	.	+	.	gene_id "ENSG00000223972.4"; transcript_id "EN
chr1	ENSEMBL	exon	13403	13655	.	+	.	gene_id "ENSG00000223972.4"; transcript_id "EN
chr1	ENSEMBL	exon	13661	14409	.	+	.	gene_id "ENSG00000223972.4"; transcript_id "EN
chr1	HAVANA	transcript	12010	13670	.	+	.	gene_id "ENSG00000223972.4"; transcript_id "EN
chr1	HAVANA	exon	12010	12057	.	+	.	gene_id "ENSG00000223972.4"; transcript_id "EN
chr1	HAVANA	exon	12179	12227	.	+	.	gene_id "ENSG00000223972.4"; transcript_id "EN
chr1	HAVANA	exon	12613	12697	.	+	.	gene_id "ENSG00000223972.4"; transcript_id "EN
chr1	HAVANA	exon	12975	13052	.	+	.	gene_id "ENSG00000223972.4"; transcript_id "EN
chr1	HAVANA	exon	13221	13374	.	+	.	gene_id "ENSG00000223972.4"; transcript_id "EN
chr1	HAVANA	exon	13453	13670	.	+	.	gene_id "ENSG00000223972.4"; transcript_id "EN
chr1	HAVANA	gene	14363	29806	.	-	.	gene_id "ENSG00000227232.4"; transcript_id "EN
chr1	ENSEMBL	transcript	14363	29370	.	-	.	gene_id "ENSG00000227232.4"; transcript_id "EN
chr1	ENSEMBL	exon	29321	29370	.	-	.	gene_id "ENSG00000227232.4"; transcript_id "EN
chr1	ENSEMBL	exon	24738	24891	.	-	.	gene_id "ENSG00000227232.4"; transcript_id "EN
chr1	ENSEMBL	exon	18268	18379	.	-	.	gene_id "ENSG00000227232.4"; transcript_id "EN
chr1	ENSEMBL	exon	17915	18061	.	-	.	gene_id "ENSG00000227232.4"; transcript_id "EN
chr1	ENSEMBL	exon	17602	17742	.	-	.	gene_id "ENSG00000227232.4"; transcript_id "EN
chr1	ENSEMBL	exon	17233	17364	.	-	.	gene_id "ENSG00000227232.4"; transcript_id "EN
chr1	ENSEMBL	exon	16854	17055	.	-	.	gene_id "ENSG00000227232.4"; transcript_id "EN
chr1	ENSEMBL	exon	16607	16765	.	-	.	gene_id "ENSG00000227232.4"; transcript_id "EN
chr1	ENSEMBL	exon	15904	15947	.	-	.	gene_id "ENSG00000227232.4"; transcript_id "EN
chr1	ENSEMBL	exon	15796	15901	.	-	.	gene_id "ENSG00000227232.4"; transcript_id "EN
chr1	ENSEMBL	exon	14970	15038	.	-	.	gene_id "ENSG00000227232.4"; transcript_id "EN
chr1	ENSEMBL	exon	14363	14829	.	-	.	gene_id "ENSG00000227232.4"; transcript_id "EN
chr1	ENSEMBL	transcript	14363	24886	.	-	.	gene_id "ENSG00000227232.4"; transcript_id "EN
chr1	ENSEMBL	exon	24734	24886	.	-	.	gene_id "ENSG00000227232.4"; transcript_id "EN
chr1	ENSEMBL	exon	18268	18369	.	-	.	gene_id "ENSG00000227232.4"; transcript_id "EN
chr1	ENSEMBL	exon	17915	18061	.	-	.	gene_id "ENSG00000227232.4"; transcript_id "EN
chr1	ENSEMBL	exon	17606	17742	.	-	.	gene_id "ENSG00000227232.4"; transcript_id "EN
chr1	ENSEMBL	exon	17498	17504	.	-	.	gene_id "ENSG00000227232.4"; transcript_id "EN

GTF file content



https://davetang.github.io/muse/read_gtf.html

Introduced yesterday

Software for counting reads per genes or transcripts

- HTSeq
- featureCounts

- Cufflinks
- StringTie

- Kallisto
- Salmon

Many more…

HTSeq counting mode

	union	intersection _strict	intersection _nonempty
read, gene_A	gene_A	gene_A	gene_A
read, gene_A	gene_A	no_feature	gene_A
read, gene_A, gene_A	gene_A	no_feature	gene_A
read, read, gene_A, gene_A	gene_A	gene_A	gene_A
read, gene_A, gene_B	gene_A	gene_A	gene_A
read, gene_A, gene_B	ambiguous (both genes with --nonunique all)	gene_A	gene_A
read, gene_A, gene_B	ambiguous (both genes with --nonunique all)		
read, ?, gene_A, gene_B	alignment_not_unique (both genes with --nonunique all)		

Result of counts quantification

Counts per gene cannot be interpreted as the gene‘s expression level

Observation: Gene 5 has twice more counts than Gene 1

Read counts	Sample A	Sample B	Sample C
Gene 1	64	60	74
Gene 2	321	753	365
Gene 3	42	60	54
Gene 4	23	53	27
Gene 5	128	131	129

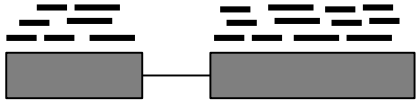
=> Is Gene 5 expression twice as high as the one from Gene 1?

Gene count matrix

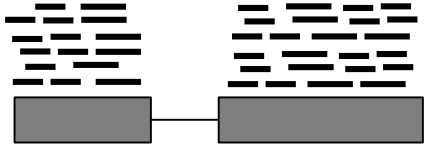
Contributing factors

1. Expression level

gene 1



gene 5

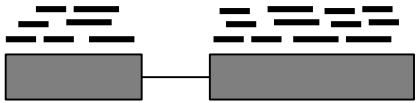


Gene count matrix

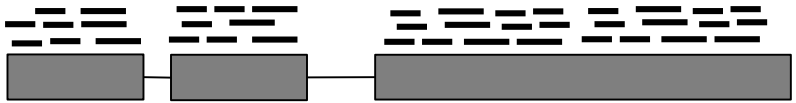
Contributing factors

1. **Expression level**
2. **Length of gene**

gene 1



gene 5



Output read counting

Counts per gene cannot be interpreted as the gene‘s expression level

Observation: Gene 2 has twice more counts in Sample B than in Sample C

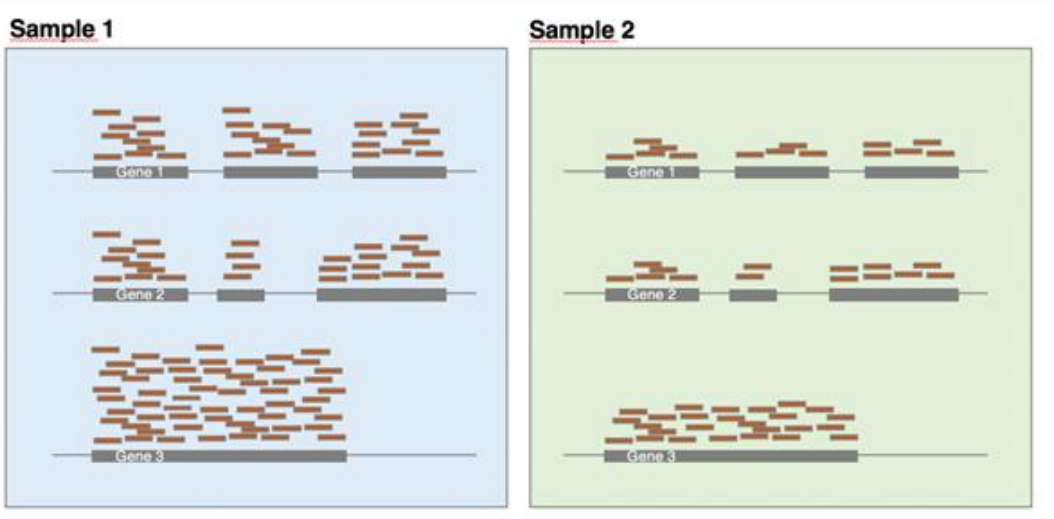
Read counts	Sample A	Sample B	Sample C
Gene 1	64	60	74
Gene 2	321	753	365
Gene 3	42	60	54
Gene 4	23	53	27
Gene 5	128	131	129

=> Is expression of Gene 2 twice as high in Sample B than in Sample C?

Output read counting

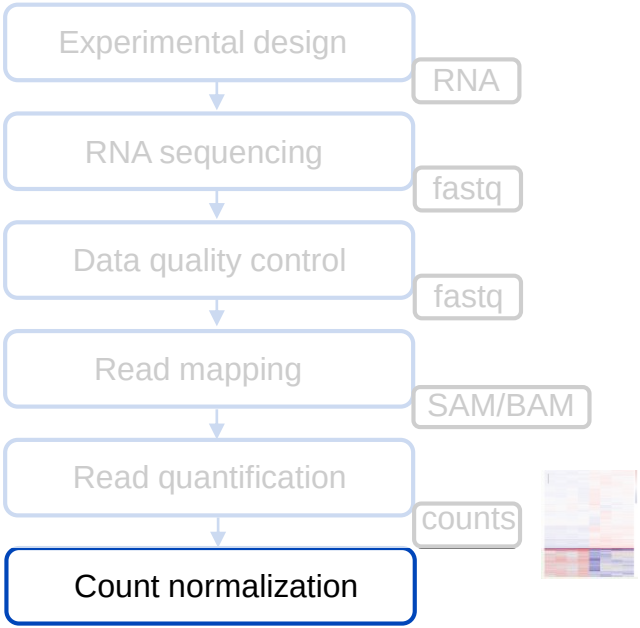
Contributing factors

- **Expression level**
- **Differences in sequencing depth between the different samples**



https://uclouvain-cbio.github.io/WSBIM2122/sec-rnaseq.html

Outline


Experimental design
RNA
RNA sequencing
fastq
Data quality control
fastq
Read mapping
SAM/BAM
Read quantification
counts
Count normalization

From counts to expression measures

We want to normalize read counts for:

1. The sequencing depth
   - Sequencing runs with more depth will have more reads mapping to each gene
2. The length of the gene
   - Longer genes will have more reads mapping to them

Expression measures:

- RPKM (Reads Per Kilobase Million) or FPKM (Fragments Per Kilobase Million)
- TPM (Transcripts Per Million)

https://uclouvain-cbio.github.io/WSBIM2122/sec-rnaseq.html

Introduced yesterday

Expression measures: RPKM and FPKM

- RPKM: Reads per kilobase per million mapped reads
- FPKM: Fragments per kilobase per million mapped reads (for paired-end data)

$$\mathrm{RPKM}_g = \frac{r_g \times 10^9}{\mathrm{fl}_g \times R}$$

r_g: reads mapped for each gene
R: total number of mapped reads for the sample $\sum r_g$
fl_g: feature length of each gene

Explanation:

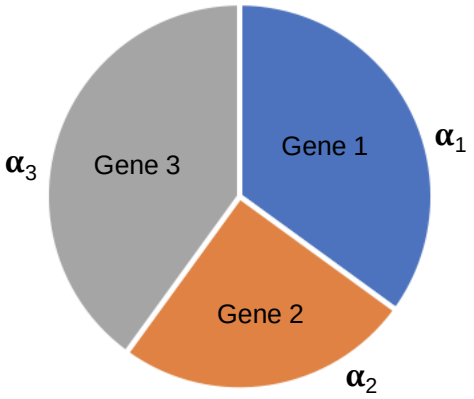
Normalize for gene length (”reads per kilobase”): $\frac{\mathrm{r_g}}{\mathrm{fl_g}} 10^3$

Normalize for total number of reads (“per million Mapped reads”): $\frac{\mathrm{R}}{10^6}$

$$\mathrm{RPKM}_g = \frac{\frac{r_g 10^3}{\mathrm{fl}_g}}{\frac{R}{10^6}} = \frac{r_g \times 10^9}{\mathrm{fl}_g \times R}$$

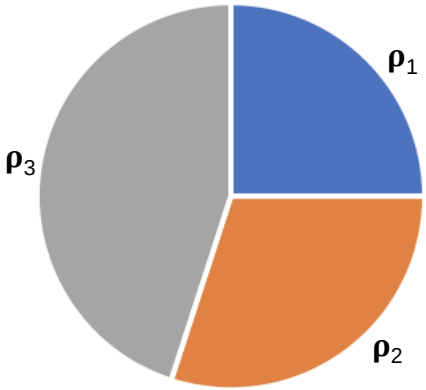
Normalization: RPKM/FPKM

- Problem: the probability of picking a read from a target does not directly relate to the relative target abundance in the sample



probabilities of picking reads from each target

$$\rho_i = \frac{\frac{\alpha_i}{l_i}}{\sum_j \frac{\alpha_j}{l_j}}$$



relative target abundance in sample

Introduced yesterday

TPM: transcripts per million

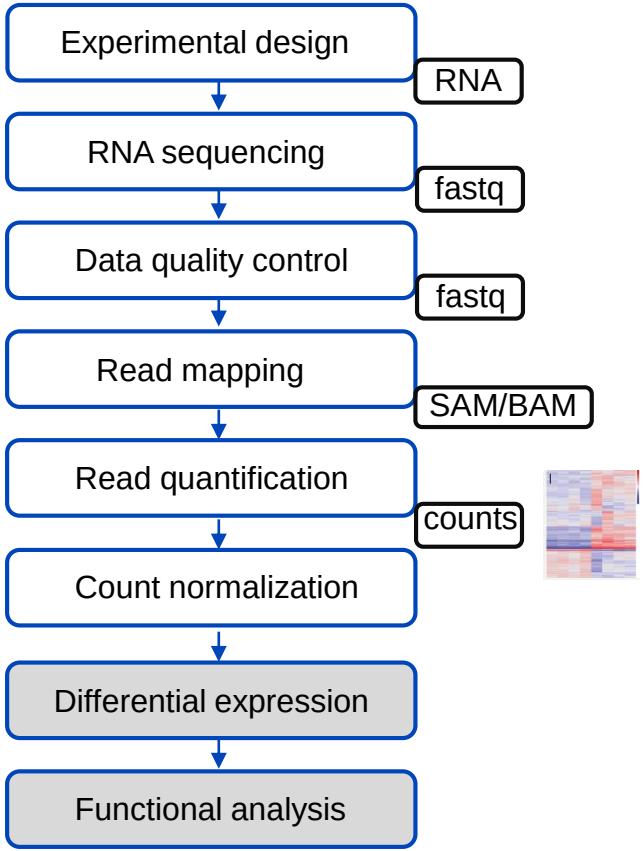
$$\mathrm{TPM} = \frac{r_g \times \mathrm{rl} \times 10^6}{\mathrm{fl}_g \times T} \qquad T = \sum_{g \in G} \frac{r_g \times \mathrm{rl}}{\mathrm{fl}_g}$$

- r_g: number of reads for gene g
- l_r: read length
- fl_g : length of gene/transcript/exon
- T: total number of transcripts sampled in a sequencing run

- Proportional to RPKM, but with a sample-specific scaling factor; T estimate for #transcripts derived from #mapped reads per gene normalized by length of gene

$$\mathrm{TPM}_i = \left(\frac{FPKM_i}{\sum_j FPKM_j}\right) \cdot 10^6$$

Summary


Experimental design
RNA
RNA sequencing
fastq
Data quality control
fastq
Read mapping
SAM/BAM
Read quantification
counts
Count normalization
Differential expression
Functional analysis