RNA-Seq analysis recap

Nicolas Dehomme, Bastian Schifthaler

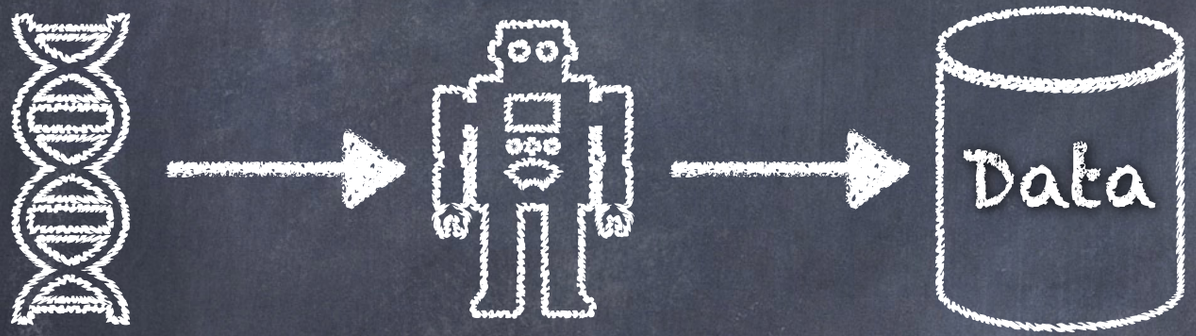
What we learned before the summer break

- High-throughput sequencing (HTS) technologies (arguably Illumina-centric)
- HTS data pre-processing, rRNA sorting, trimming, quantifying, quality control (technical)
- RNA-Seq data exploratory analysis (biological QC)
- RNA-Seq data analysis, the most common, namely differential expression (DE) analysis

What's left on the agenda for the fall

- Downstream analyses
  - Gene Set Enrichment Analyses
  - Pathway analyses
  - Gene network inference
- Single-cell data and analysis specifics

Recap, ask questions!



Recap, ask questions!

QC

- Look for technical issues, commonly used tool: FastQC - keep in mind it was developed for DNA-Seq
  - assess read quality per cycle and per read
  - assess GC content
  - assert adapters presence/absence
  - assert over-represented sequences preponderance

Recap, ask questions!

Cleansing (if needed)

- If the QC reveals anything suspect
  - trim the reads
  - filter the reads
  - do both
- or neither!

Recap, ask questions!

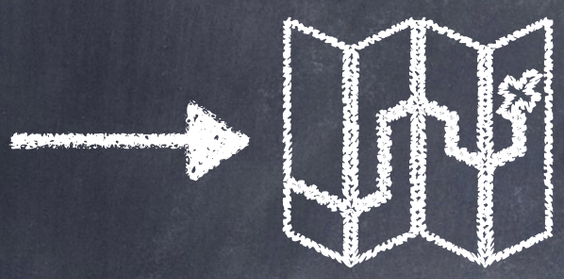
Quantification

- genome or transcriptome based
  - genome:
    - less accurate quantification ⊖
    - novel genes detection ⊕
  - transcriptome:
    - more accurate expression profiling when using latest tools (e.g. Salmon) ⊕
    - limited by the quality of the transcriptome ⊖

Quantification (2)

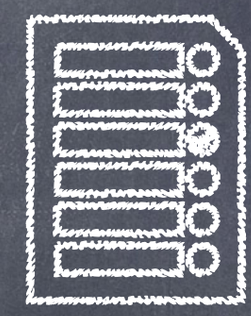
- transcriptome based quantification is very actively developed
  - even more robust: e.g. Fishpond: differential transcript and gene expression with inferential replicates
  - even more comprehensive: terminus (https://github.com/COMBINE-lab/terminus) data driven approach that summarises expression at the gene level whenever the transcript level is unreliable

Recap, ask questions!



(pseudo-)
mapping

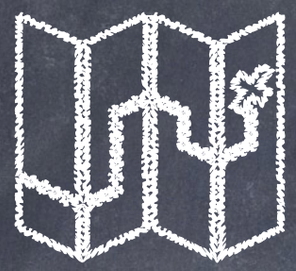
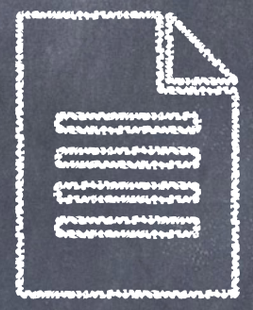
QC summary



Yes, again!

MultiQC is a
convenient tool

Recap, ask questions!

→  (pseudo-) mapping →  expression quantities

→ Differential Gene Expression

→ Differential Transcript Expression

→ Differential Transcript Usage

Recap, ask questions!

→ Differential Gene Expression +
→ Differential Transcript Expression +
→ Differential Transcript Usage =

Gene Differential Expression

https://liorpachter.wordpress.com/2018/02/15/gde²-dge²-dtu²-dte$_1$²-dte$_2$²/

QC again!

- Perform the biological QA
  - check samples' read counts, distributions

- does the data match the expectations (studies design): PCA, expression of gene of interest

Differential Expression

- From the literature:
  - best tools: DESeq2, edgeR
  - do NOT use R/FPKM

- From my opinion
  - use TPM (if at all) only for visualisation
  - best, use data normalised by either afore mentioned package, and keep in mind heteroscedasticity if not stabilising the variance

Differential Expression (2)

- Modelling the data
  - start with (a) simple model(s)
  - progress stepwise to the most complex model that is still easy to interpret
  - double-check the results (heat maps, expression profiles, etc.)
  - Charlotte's slides as a reference!

Check-out

- our public repository on GitHub: https://github.com/UPSCb/UPSCb-common
- It contains code to run frequently used pre-processing tools, to

Minimal tools we'd recommend

- FastQC
- Trimmomatic
- SortMeRNA (or kraken2 as a faster alternative)
- Salmon
- MultiQC
- R packages DESeq2