Docker Course

Bioinformatics software tests in docker containers

- Alignment and quantification tools for RNA sequencing data -

DTU Biosustain

Data Science Platform

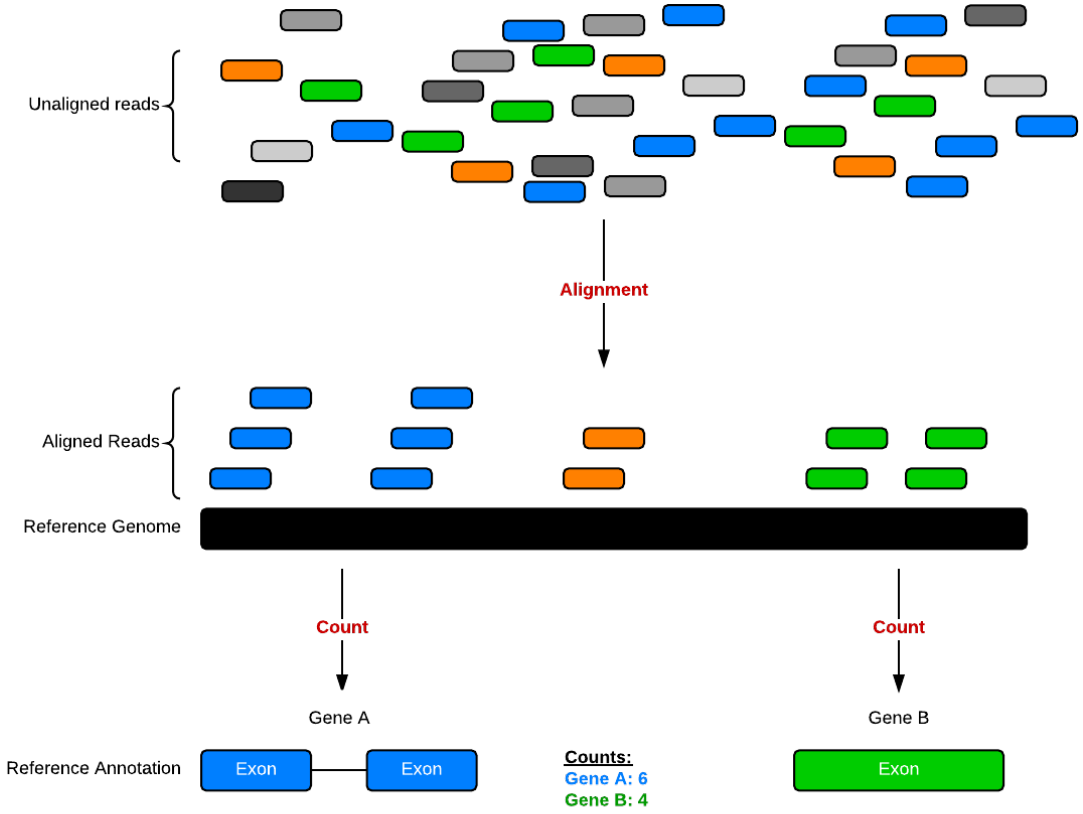
Sebastian Schulz

Senior Data Scientist

sebschu@dtu.dk

26 November 2025

Major bioinformatic steps in RNA sequencing (RNAseq)



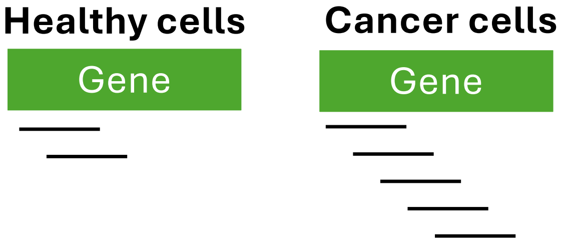
https://bioconnector.github.io/bims8382/r-rnaseq-airway.html

Major steps in RNAseq processing

- Alignment: map reads to a reference sequence (e.g. genome)
- Read quantification: count how many reads are aligned to a specific genomic region (feature)

One major downstream analysis goal

- Differential expression (DE) analysis comparing read counts of a specific genomic feature between biological experiments



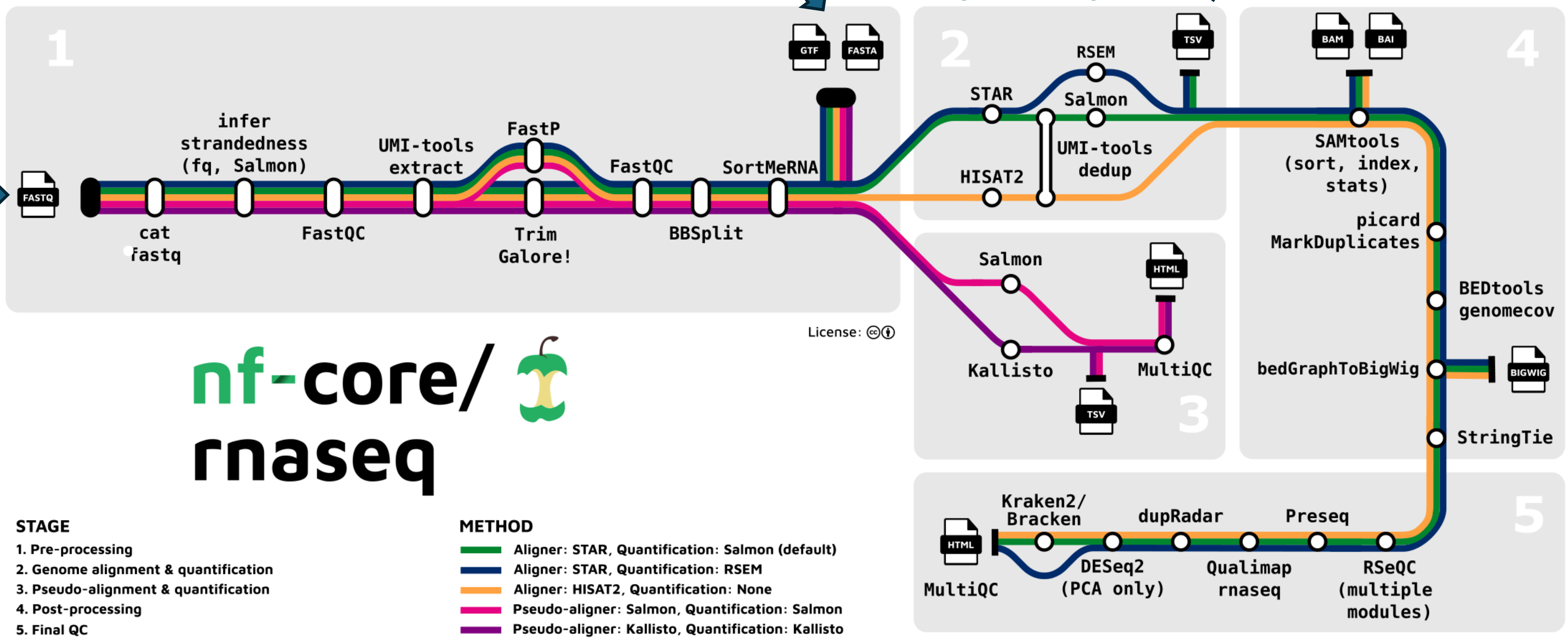
Data processing – nf-core/rnaseq pipeline

Genome .fasta
Annotation .gtf

Alignment (STAR) and quantification (Salmon)

Count table (.tsv)

.fastq read files



STAGE

1. Pre-processing
2. Genome alignment & quantification
3. Pseudo-alignment & quantification
4. Post-processing
5. Final QC

METHOD

- 🟩 Aligner: STAR, Quantification: Salmon (default)
- 🟦 Aligner: STAR, Quantification: RSEM
- 🟧 Aligner: HISAT2, Quantification: None
- 🟪 Pseudo-aligner: Salmon, Quantification: Salmon
- 🟪 Pseudo-aligner: Kallisto, Quantification: Kallisto

https://nf-co.re/rnaseq/3.16.1

Complex pipeline with many very useful software tools!

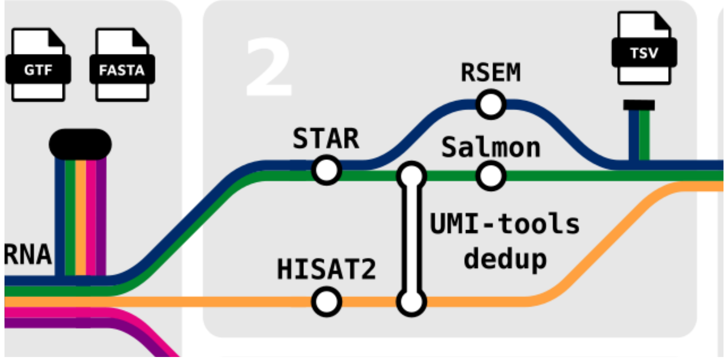
Pro- and eukaryotic data processing with the nf-core/ rnaseq pipeline

- The pipeline **succeeds with eukaryotic data sets** (e.g. mouse)

- The pipeline **<u>fails</u> with prokaryotic data sets** (e.g. *E. coli*) due to
  - Suboptimal quality of bacterial gtf annotation files and therefore an incompatibility issue of STAR and Salmon
    - Successfully addressed by modifying the gtf file and pipeline parametrization (github issue #1512)

- The aligner-quantifier tool set, STAR + Salmon, is **not necessarily the first choice for the processing of prokaryotic data set**
  - STAR was developed for read alignment using eukaryotic data sets
- Is the combination of STAR + Salmon still applicable to prokaryotic data sets?
  - How do the results obtained from STAR + Salmon compare to those from tool sets that are (commonly) used in prokaryotic transcriptomics, like Bowtie2 + featureCounts?

Alignment (STAR) and quantification (Salmon)

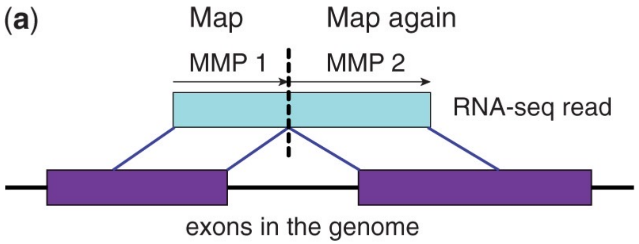


https://nf-co.re/rnaseq/3.16.1

Alignment and quantification software

Read alignment

	Splice-junction detection	Application
STAR Dobin *et al.*, 2012	yes	*De novo* splice juntion detection (eukaryotes)
Bowtie/ Bowtie2 Langmead *et al.*, 2009 Langmead and Salzberg, 2012	no	Established aligner for prokaryotes (also for eukaryotes)



MMP: Maximal Mappable Prefix
From Dobin *et al.*, 2012

Read quantification

	Quantification model	Application
Salmon Patro *et al.*, 2017	Complex statistical model considering 5' and 3' bias, GC-content bias, etc	More sophisticated quantification model for more precise results considering different biases
featureCounts Liao *et al.*, 2014	Conceptually simpler counting mechanism	Performs well on one-isoform genes (Parelo *et al.*, 2014; tested on eukaryotic data sets), which would be desirable for prokaryotic data

Dobin *et al.*, 2012, https://doi.org/10.1093/bioinformatics/bts635 (STAR)
Langmead *et al.*, 2009, https://doi.org/10.1186/gb-2009-10-3-r25 (Bowtie)
Langmead and Salzberg, 2012, https://doi.org/10.1038/nmeth.1923 (Bowtie2)
Liao *et al.*, 2014, https://doi.org/10.1093/bioinformatics/btt656 (featureCounts)
Patro *et al.*, 2017, https://doi.org/10.1038/nmeth.4197 (Salmon)
Parelo *et al.*, 2024, https://doi.org/10.1093/nargab/lqae020

Comparison of alignment and quantification software

Goal

- To test how different aligner-quantifier combinations perform on prokaryotic data sets **outside of the nf-core/rnaseq pipeline**
    - Example: Does Bowtie2 + featureCounts perform similar or different compared to STAR + Salmon?

Approach

- Substituting software tools directly in the nf-core/rnaseq pipeline just for testing purposes would be too labour-intensive and time-consuming
- Test different aligner-quantifier combinations **in Docker containers** to ensure **maximal reproducibility** for us and most importantly when **shared with the nf-core community**
    - As these analysis should form the basis for the potential further development of the nf-core/rnaseq pipeline, we aimed at ensuring a high-level of standardization in our analyses

Quantifier

Aligner

	Salmon	**featureCounts**
STAR	Pipeline default	
Bowtie		
Bowtie2		This Docker course

Reduced version of entire test plan

DTU

The Novo Nordisk Foundation Center for Biosustainability

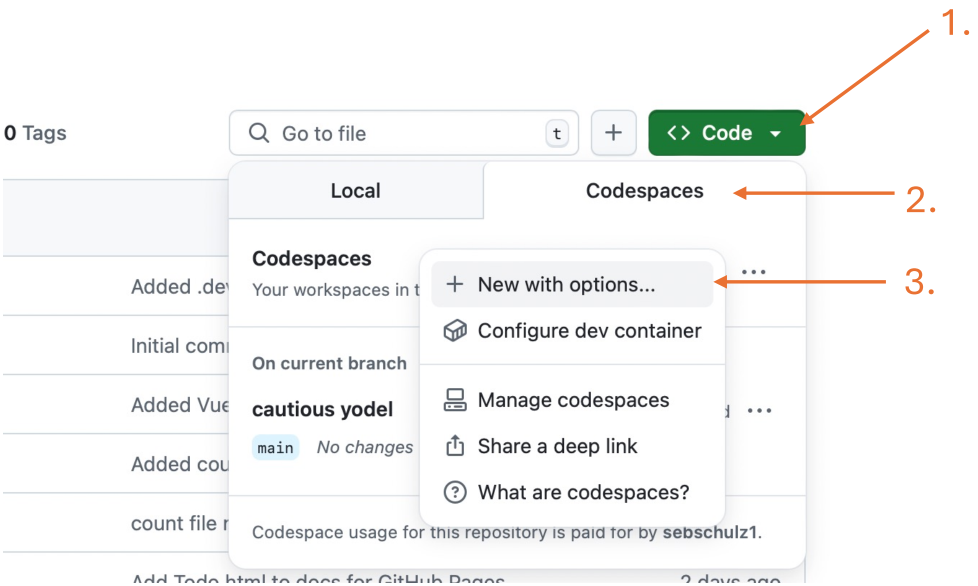
ACKNOWLEDGEMENTS

novo
nordisk
fonden

Supplementary slides

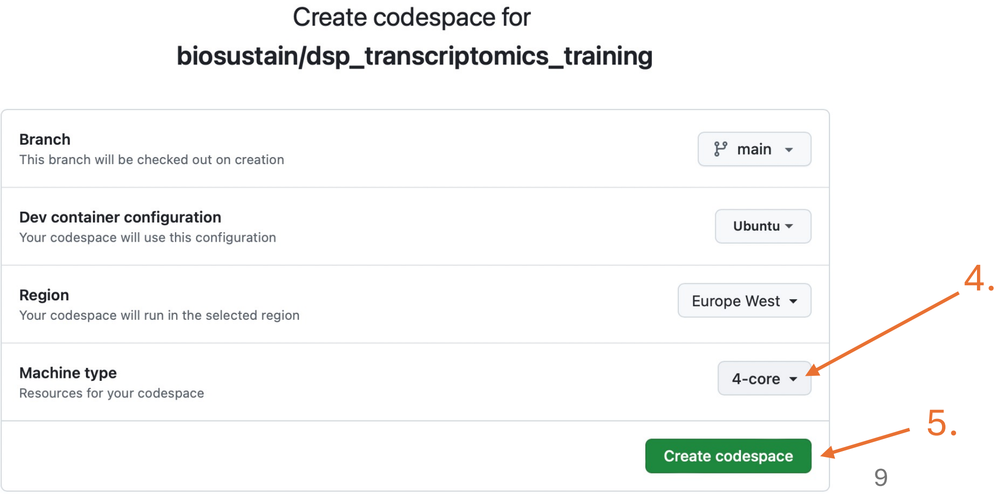
Create the codespace

- Log into your github account
- Open the training repo: https://github.com/biosustain/dsp_transcriptomics_training
- Create codespace **with 4 cores** which we will use to run code



Codespace usage for this repository is paid for by **sebschulz1**.

Select 4-cores for machine type before creating the codesapce



Create codespace for
biosustain/dsp_transcriptomics_training

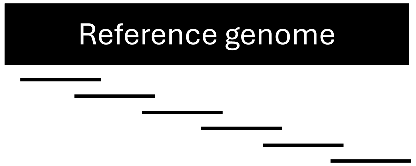
Branch
This branch will be checked out on creation: main

Dev container configuration
Your codespace will use this configuration: Ubuntu

Region
Your codespace will run in the selected region: Europe West

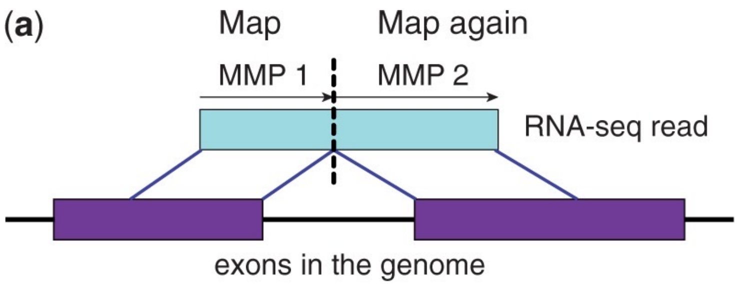
Machine type
Resources for your codespace: 4-core

1. Data processing - Read alignment with STAR



- Spliced Transcripts Alignment to a Reference (STAR)
- Very fast aligner
- Alignment against genome (or transcriptome)
- *De novo* detection of splice junctions (no prior annotation of splicing event required)

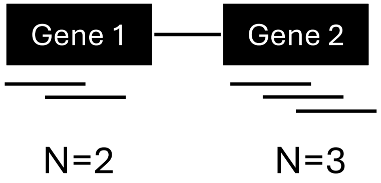
Detection of splice junctions with STAR



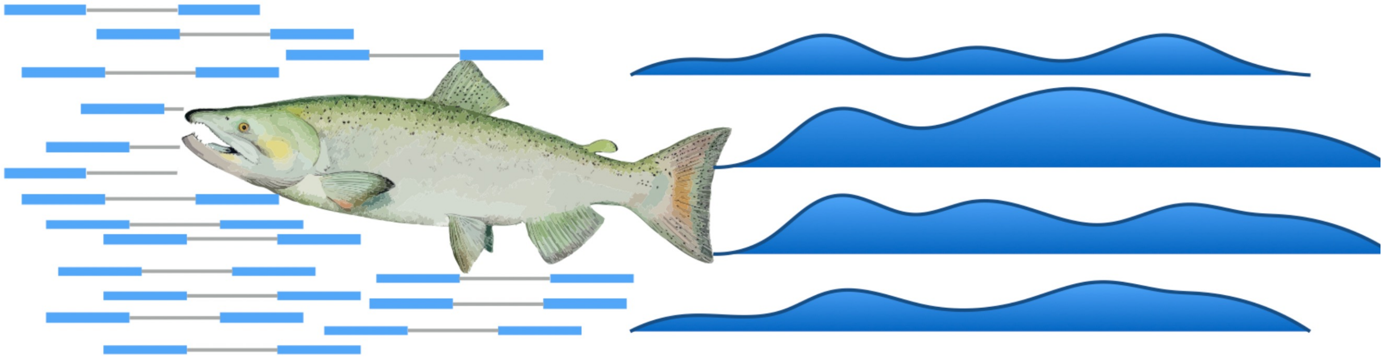
MMP: Maximal Mappable Prefix

https://doi.org/10.1093/bioinformatics/bts635

1. Data processing - Read quantification with Salmon



- More complex quantification procedure compared to other read counters/quantifiers
- Handles counting of multi-mapping reads
- More accurate quantification of reads achieved by considering sample-specific parameters and biases of RNAseq data:
  - positional biases in coverage
  - sequence-specific biases at the 5' and 3' ends of sequenced fragments
  - fragment-level GC bias
  - strand-specific protocols
  - fragment length distribution



https://combine-lab.github.io/salmon/about/ https://doi.org/10.1038/nmeth.4197