RNA-seq Best Practices & The Pipeline

Erin Osborne Nishimura

DSCI 512: RNA sequencing data analysis

November 7, 2024

These slides

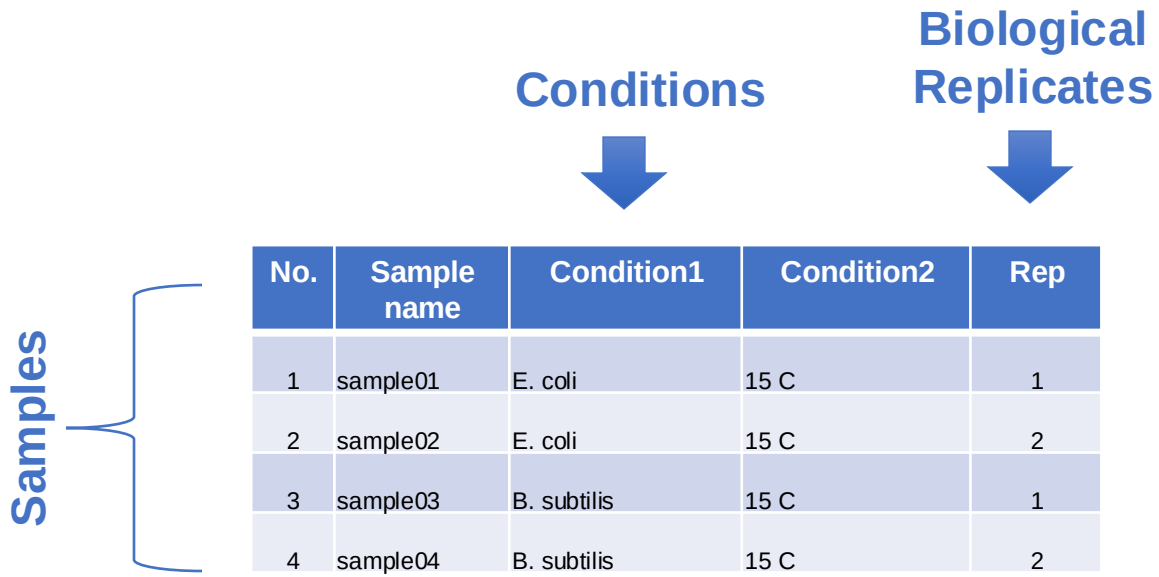
1. Best practices in experimental design
2. An intro to the course project
3. The RNA-seq data analysis pipeline

1. Best practices In experimental design

Experimental design in RNA-seq projects is very import and should be driven by the biological question

- Biological question?
  - Conditions?
  - What characteristics of the transcriptome do you expect will be different?
  - What will you do next with this information?
- Resources?
  - How many ***biological*** replicates?
  - How much material?
  - How much funding?
- How can batch effects be reduced?
- Who will analyze the data?

On vocabulary – samples, conditions, replicates



No.	Sample name	Condition1	Condition2	Rep
1	sample01	E. coli	15 C	1
2	sample02	E. coli	15 C	2
3	sample03	B. subtilis	15 C	1
4	sample04	B. subtilis	15 C	2

- **Samples** – Each RNA-seq library that was prepped and sequenced is a sample
- **Conditions** – Represent the differences we are trying to assess
  - Can be pairwise such as mutant v. wild-type, disease v. healthy
  - Can be ordered – developmental time, temperature, salt conc.
- **Replicates** – these are duplicates
  - Biological replicates – different organisms were harvested. Very important.
  - Technical replicates – the same organisms were harvested. Not advised.

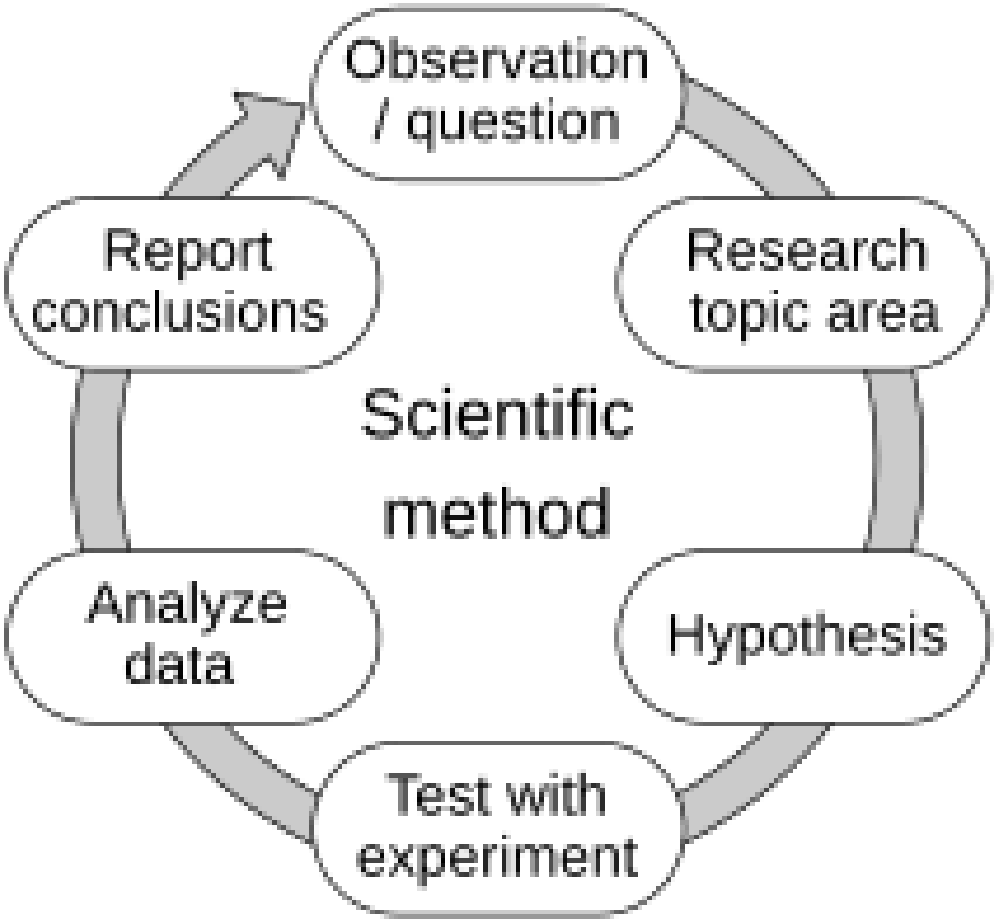
Ideal practices

- A minimum of ***FOUR*** biological replicates per condition for differential abundance experiments
- ***Defend against batch effects*** when performing biological replicates
- Consult with your sequencing facility and analytical team ***BEFORE*** doing the experiment
- Think deeply about positive and negative ***controls***

Key limitations and assumptions of RNA-seq

- For differential expression analysis, protocols ***assume*** at least ***90 % of RNAs are NOT changing*** in abundance between conditions
- mRNA abundance is ***not always a good marker for protein*** production or transcriptional activity
- Most samples are mixture of ***heterogeneous*** cells or tissue types
- Differences between conditions can be ***direct or indirect*** effects
- Individual cells are inherently ***variable***
- Some important genes are ***lowly expressed*** or have ***modest changes*** in RNA-seq

Where does transcriptome profiling fit in the scientific method?

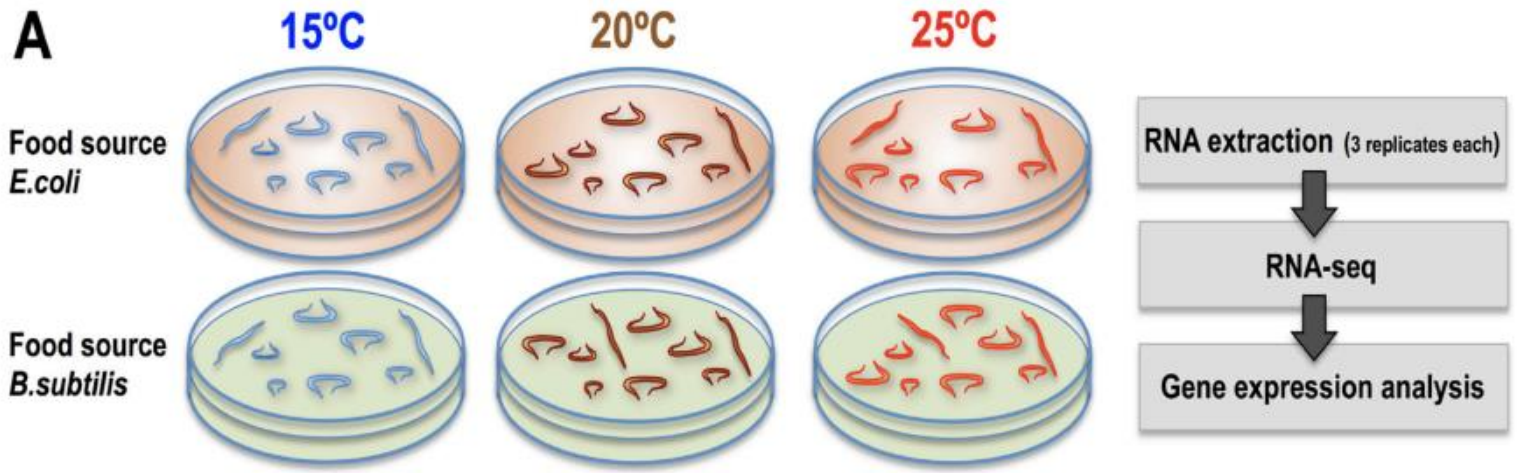


Have a plan for the next steps

- **Confirm** results with an alternative method
  - Microscopy
  - GFP reporter
  - Western blot
- **Zoom in** on key genes to study
  - Build a story
- **Follow up** on the hypotheses generated

2. An Intro to the Course Project

How do diet and temperature impact gene expression patterns in the *Caenorhabditis nematode worm?*



- Gómez-Orte, et al., (2017) Effect of the diet type and temperature on the C. elegans transcriptome. Oncotarget. 2018 Feb 9; 9(11): 9556–9571.

The data structure of the Gómez-Orte project

No.	Sample name	Condition1	Condition2	Rep
1	sample01	E.coli	15 C	1
2	sample02	E.coli	15 C	2
3	sample03	E.coli	15 C	3
4	sample04	E.coli	20 C	1
5	sample05	E.coli	20 C	2
6	sample06	E.coli	20 C	3
7	sample07	E.coli	25 C	1
8	sample08	E.coli	25 C	2
9	sample09	E.coli	25 C	3
10	sample10	B. subtilis	15 C	1
11	sample11	B. subtilis	15 C	2
12	sample12	B. subtilis	15 C	3
13	sample13	B. subtilis	20 C	1
14	sample14	B. subtilis	20 C	2
15	sample15	B. subtilis	20 C	3
16	sample16	B. subtilis	25 C	1
17	sample17	B. subtilis	25 C	2
18	sample18	B. subtilis	25 C	3

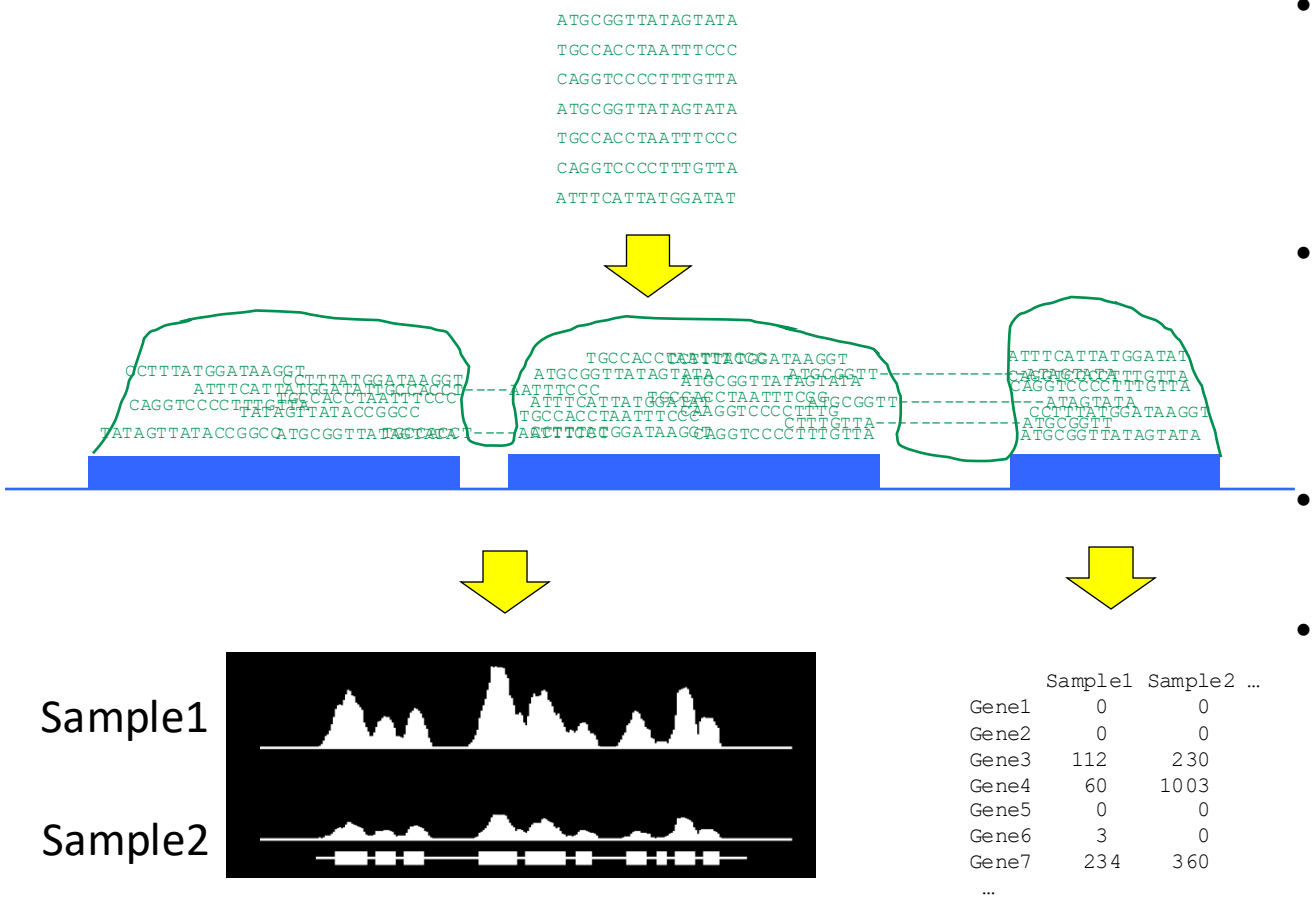
3. The RNA-seq data analysis pipeline

What is a pipeline?

- A series of data processing operations in which the output of one process is the input for the next process

Data analysis

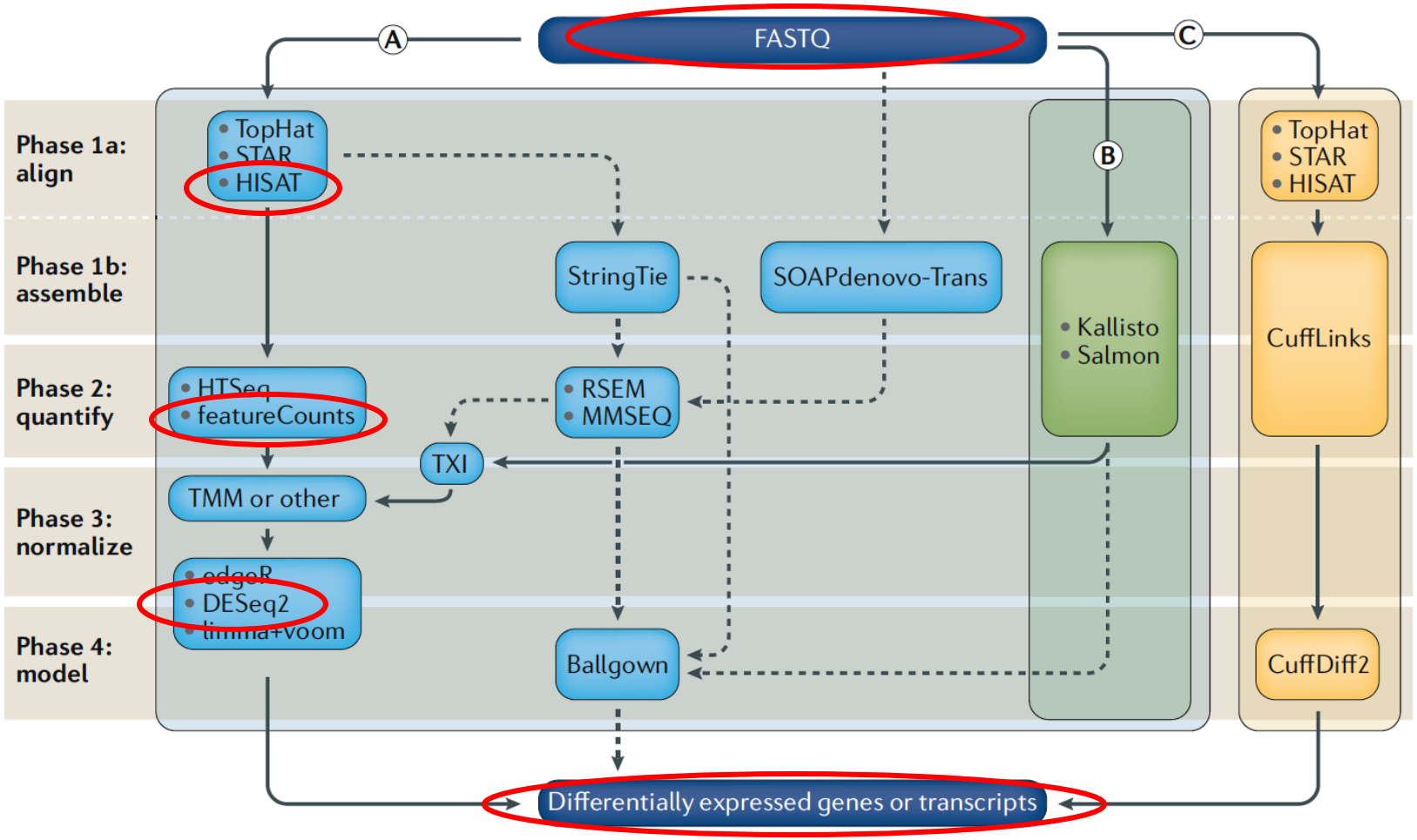
Use computer algorithms to align sequences to the genome



	Sample1	Sample2 …
Gene1	0	0
Gene2	0	0
Gene3	112	230
Gene4	60	1003
Gene5	0	0
Gene6	3	0
Gene7	234	360
…		

- **Step 1** – Align sequenced fragments (called reads) to the genome
- **Step 2** – Quantify the number of reads associated with each gene
- **Step 3** – Normalize the samples
- **Step 4** – use modeling and statistics to identify differentially expressed genes

There are many tools available for RNA-seq analysis



Stark et al., 2019. RNA sequencing: the teenage years

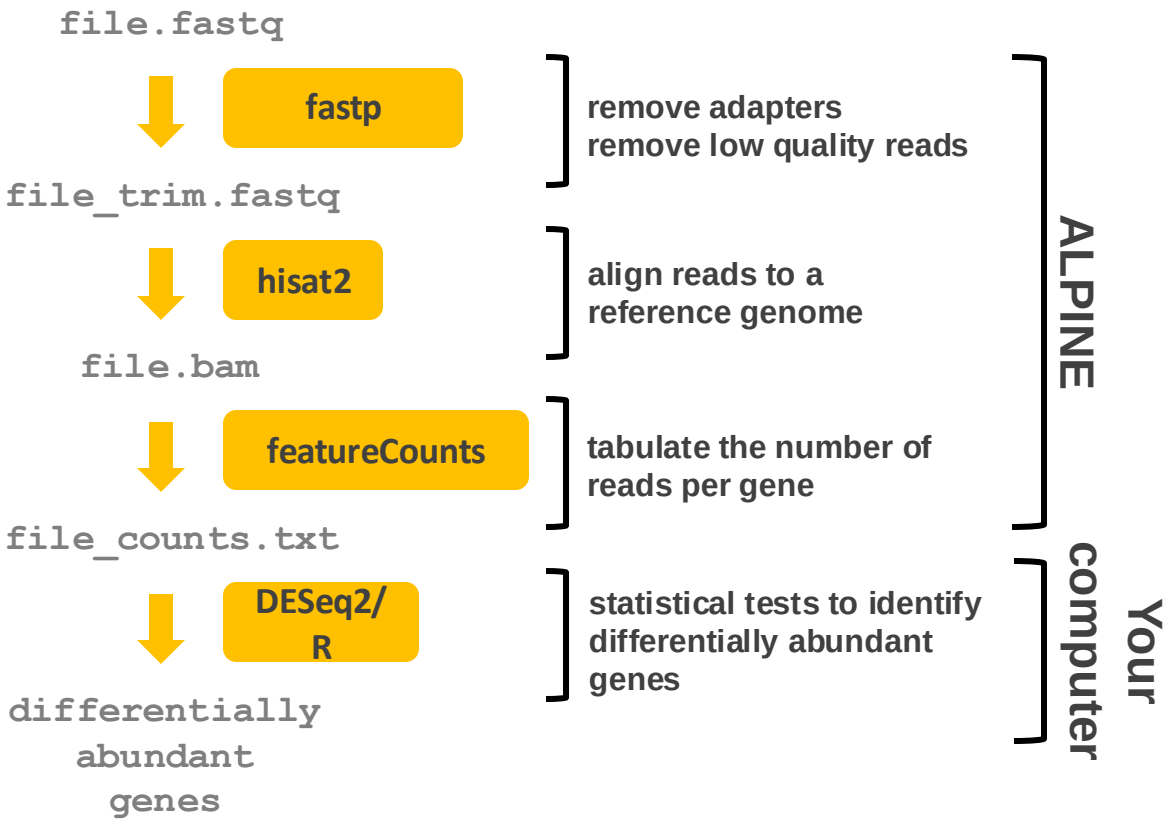
Which tools are best?

- The tools you use will depend on your research question in some cases

- The good news: Alignment is robust
  - "… [our analysis] indicates that the quality of (spliced) aligners may have reached a point where it does not appear to make a big difference which one is used in the context of gene profiling analysis." Fonseca, 2014

- The recent good news:
  - "We did not identify among the evaluated methods a tool that obtained optimum results in all performance measures, for the evaluated experimental conditions. The NOIseq, DESeq2 and limma+vomm methods present the best individual results with 95%, 95% and 93% of Specificity and 80%, 84% and 81% of True Positive Rate, respectively." Silva Costa, 2017

The pipeline for this course



The full pipeline for this course