

Intro to Computational Biology

Omar Salem, PhD

Senior Computational Biologist, Roche Canada.

July 9th 2024

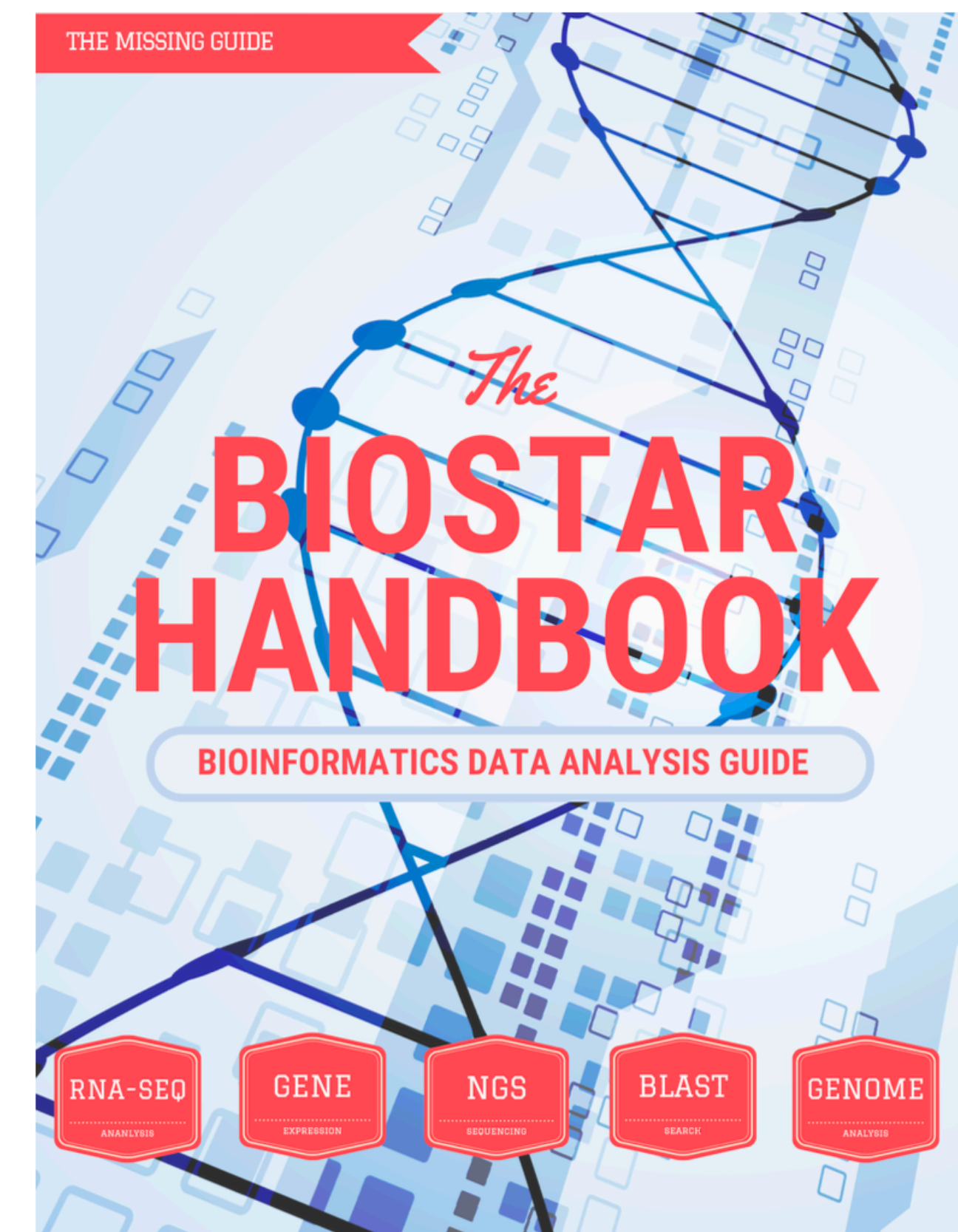
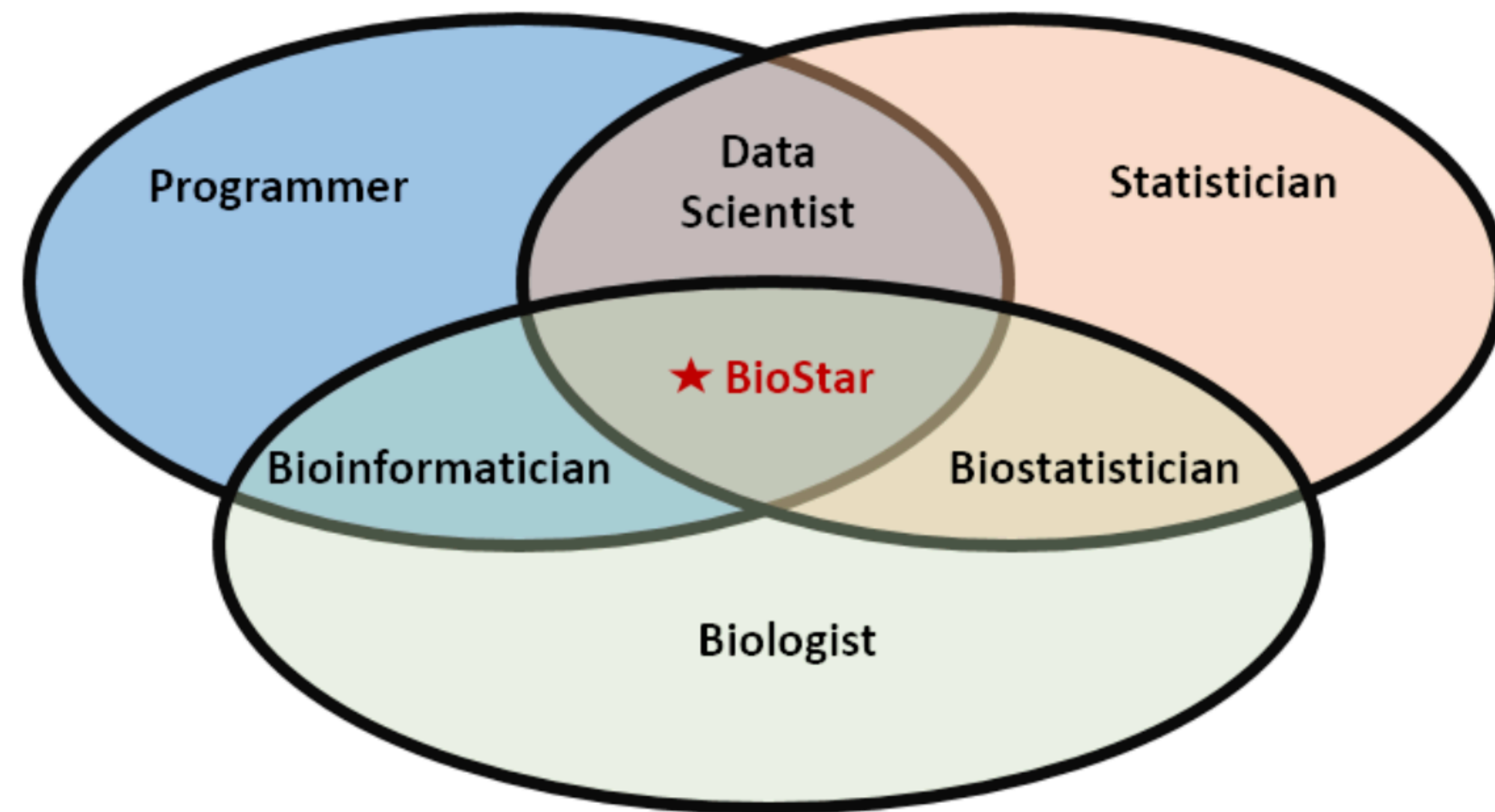
Slides, code and everything we go over will be available to consume later, so use this time to listen, understand, absorb, interact and learn !


My personal journey



The What !

Who is a Computational biologist / scientist ?



My  : Focus on becoming a good biologist / scientist first and work on your computational toolkit as you go

The Why !

There are some biological questions scientists
cannot answer at the bench !

Can you think of any ?

Metabolomics

Proteomics

Epigenomics

Population Genetics
and Evolutionary
Biology

Genomics

Structural Biology

Transcriptomics

Spatial -omics

Systems Biology

The How!

Resources for learning computational biology

- <https://www.biostars.org/>
- Introduction to Bioinformatics and Computational Biology: <https://liulab-dfci.github.io/bioinfo-combio/>
- StatQuest with Josh Starmer: <https://www.youtube.com/@statquest>
- R for Data Science (2e): <https://r4ds.hadley.nz/>
- <https://bioinformatics.ca/>

Academia vs. Industry

RNA sequencing

What!

Sequence total mRNA within a tissue / cell

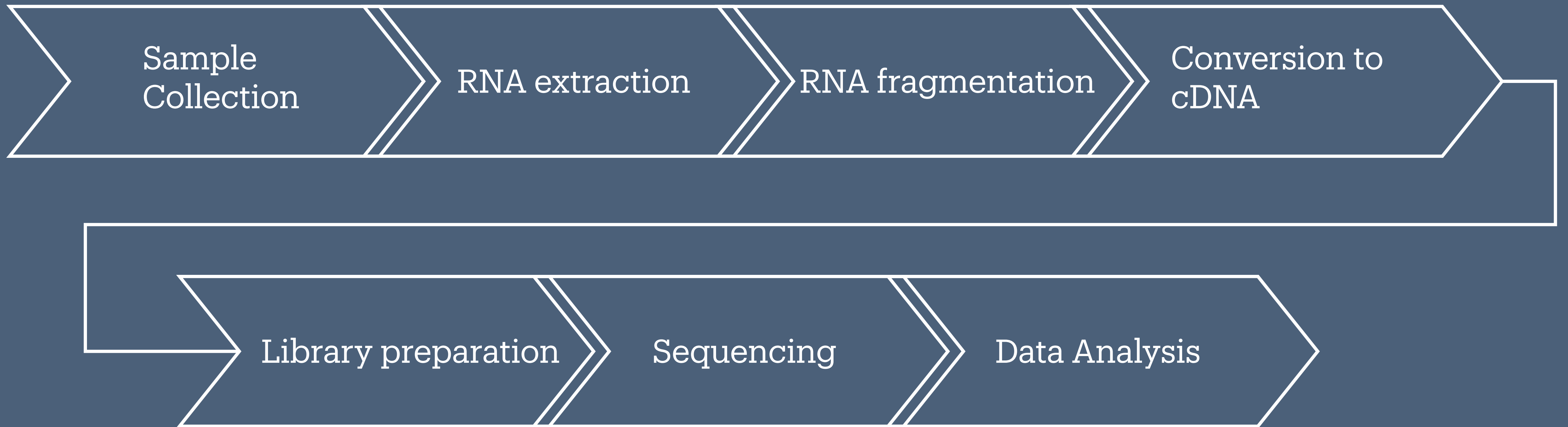
Why!

DNA -> mRNA -> protein

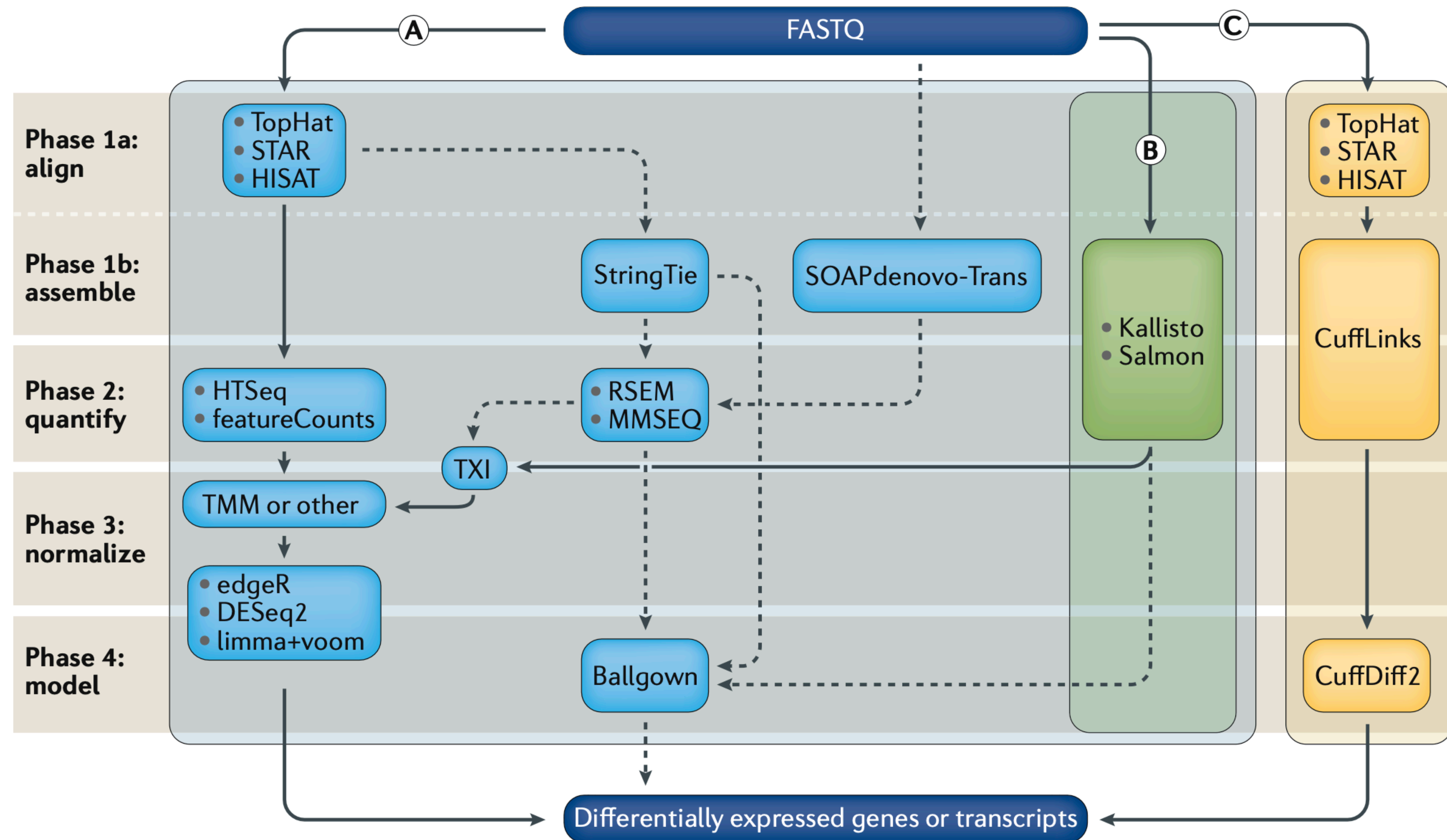
mRNA is a proxy for studying cellular processes
and functions

Primary application: Identify DEGs between
samples / conditions

How!



RNA seq data analysis



Mapping

- Splice-aware aligners e.g. STAR / HISAT / TopHat
- Splice-unaware aligners e.g. Bowtie2, BWA

A quantifier is necessary when using any of these tools!

- Quasi-mappers / alignment-free mappers e.g. Kallisto, Salmon

Normalization

- RPKM/FPKM
- TPM
- CPM
- DESeq2 / EdgeR

Differential gene Expression

- DESeq2
- EdgeR
- Voom-limma

Let's run some code!

scRNA sequencing

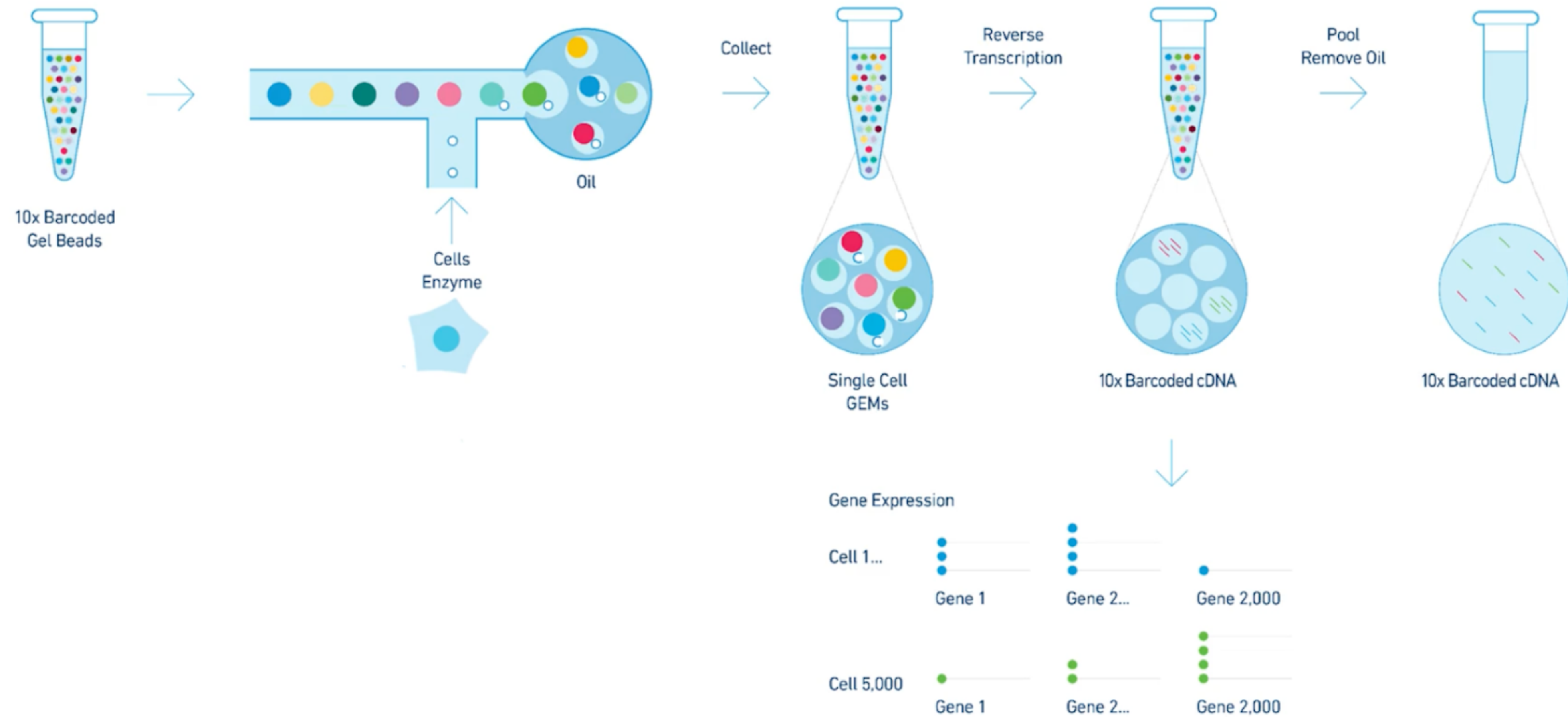
What!

Sequence total mRNA within a single cell

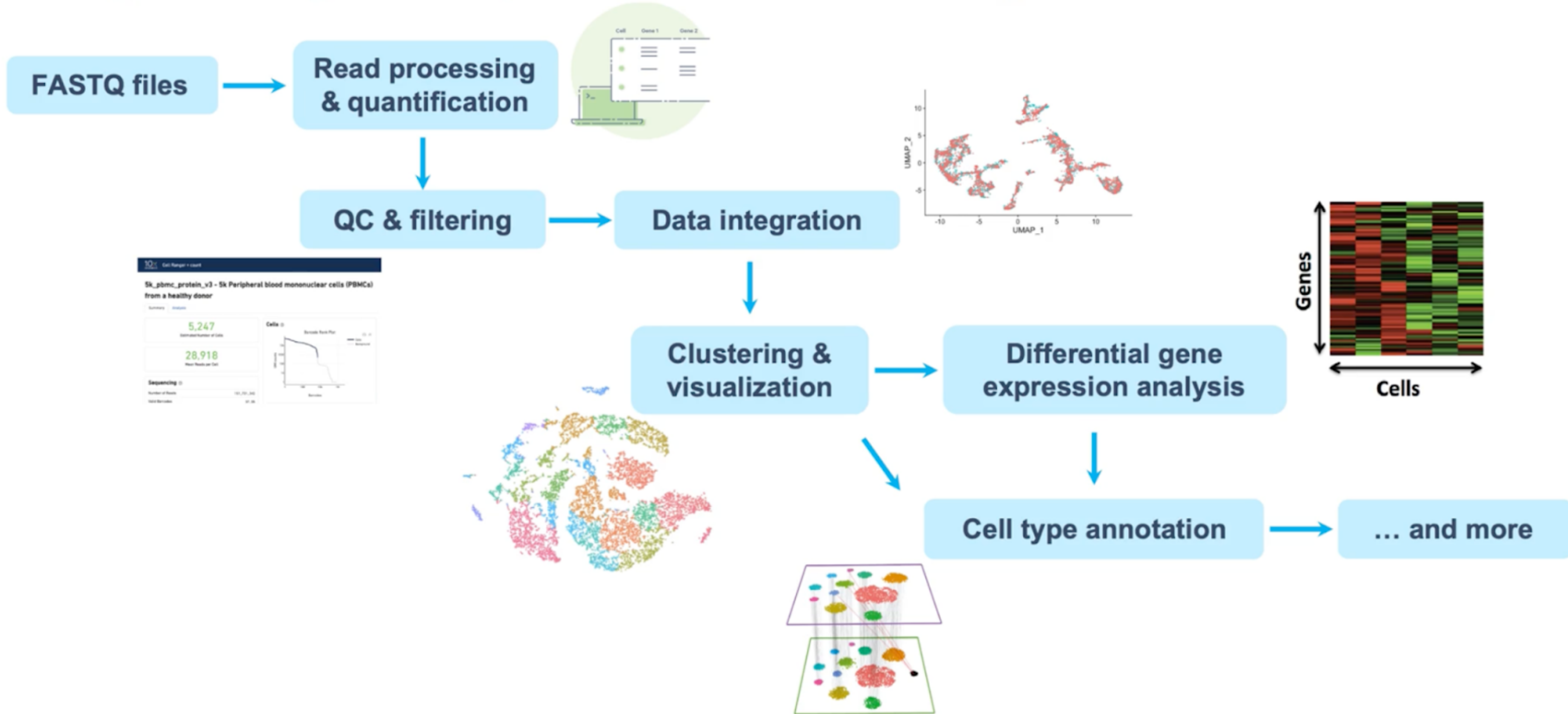
Why!

- Study cellular heterogeneity
- Study cell differentiation (trajectory analysis)
- Study cell-cell interactions (receptor - ligand analysis)

How!



Single cell gene expression - data analysis flow



QC

- Low-quality cells / empty droplets
- Cell doublets or multiplets
- Mitochondrial genes

Normalization

- **Log-normalization**
- **SCtransform**: <https://genomebiology.biomedcentral.com/articles/10.1186/s13059-021-02584-9>
- Fun read: <https://www.nature.com/articles/s41592-023-01814-1>

Data integration

- Remove batch / tissue specific effects
- Can be computationally expensive
- Different tools have different results
- Examples beyond Seurat: Harmony, Scanorama, scVI, MNN

Downstream analysis

- Clustering - Resolution is tricky to get right
- Annotations - possibly the most challenging part of the analysis
- Answer your biological questions!

Let's run some code!