INF-BIOx121 2017

RNA-seq

differential expression analysis

Arvind Sundaram

Sep 18-20, 2017

RNA-seq analysis

Introduction

Arvind Sundaram
Sep 18, 2017

Outcome

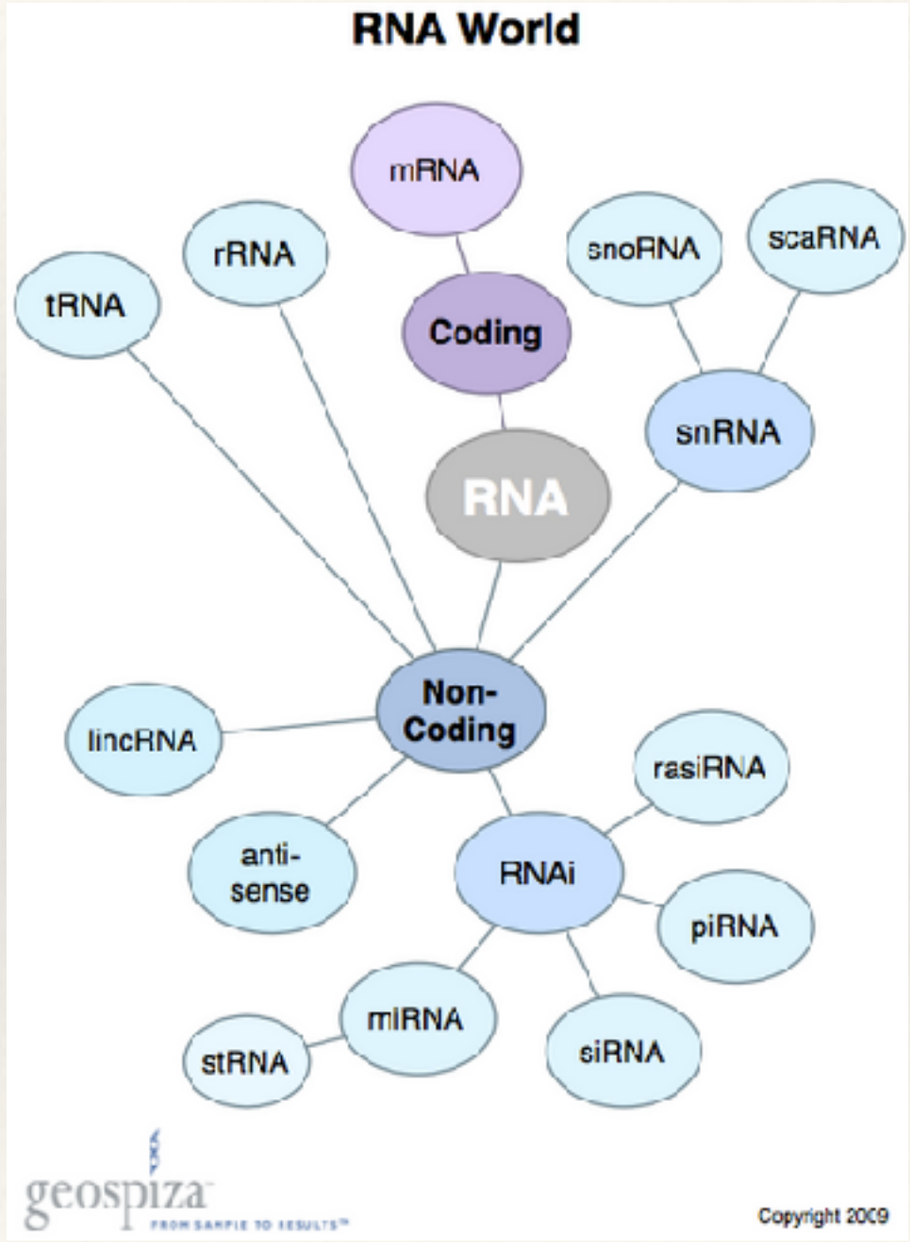
- Explain what is RNA-seq
- Get an insight on how the data is analysed
- List different types of RNA-seq methods
- Tailor data analysis pipeline based on the organism and the hypothesis being investigated

Outcome

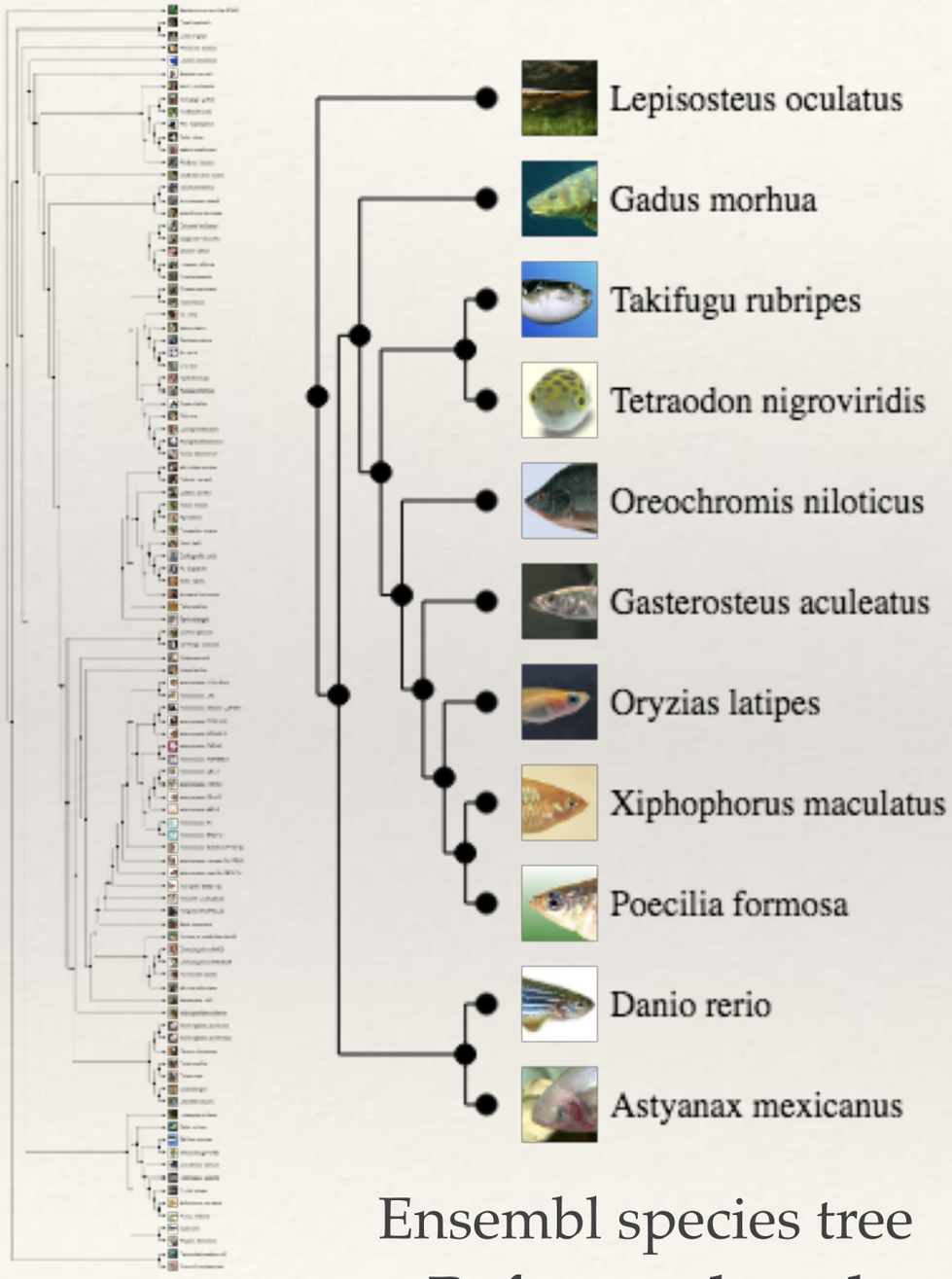
- After this module you should be able to perform:
  - Differential gene expression analysis within pair-wise comparison using a reference genome
  - Understand overall statistics such as mapping percentage, find potential outliers, list DE genes
  - Extract and present biological meaning of the DE genes

Transcriptome

A transcriptome is a snapshot in time of all RNAs present in a sample isolated from a given cell, tissue or organism

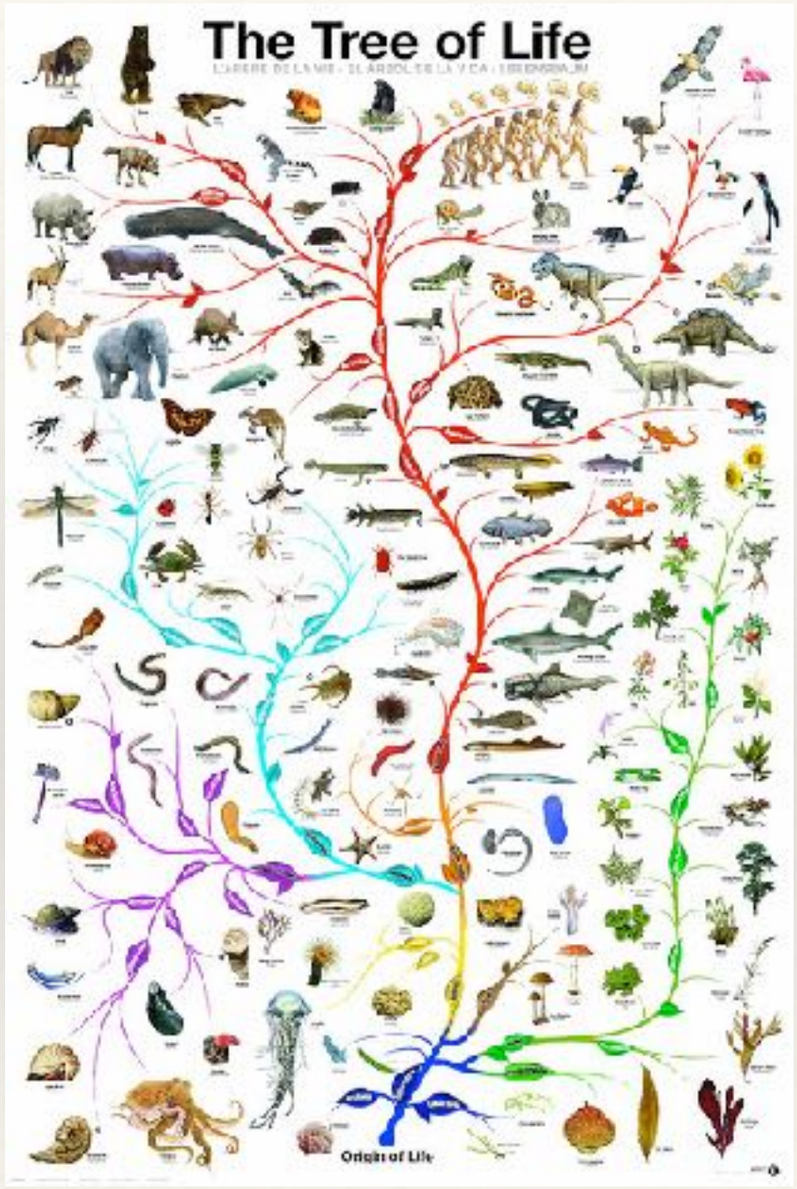


Limit…



Ensembl species tree

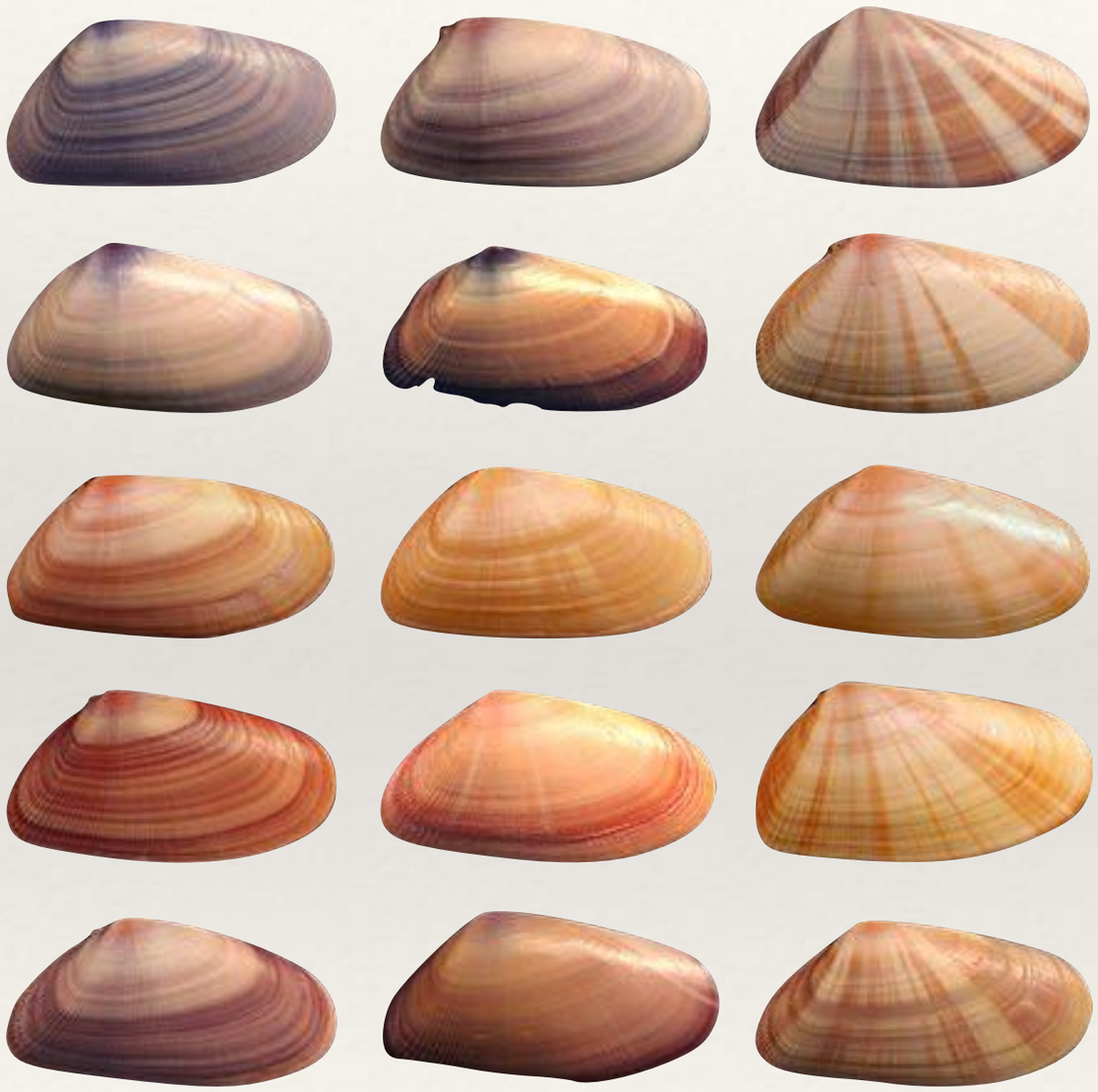
Reference based



***de novo* approach**

http://www.ensembl.org/info/about/speciestree.html
https://www.thinglink.com/scene/645083259847311362

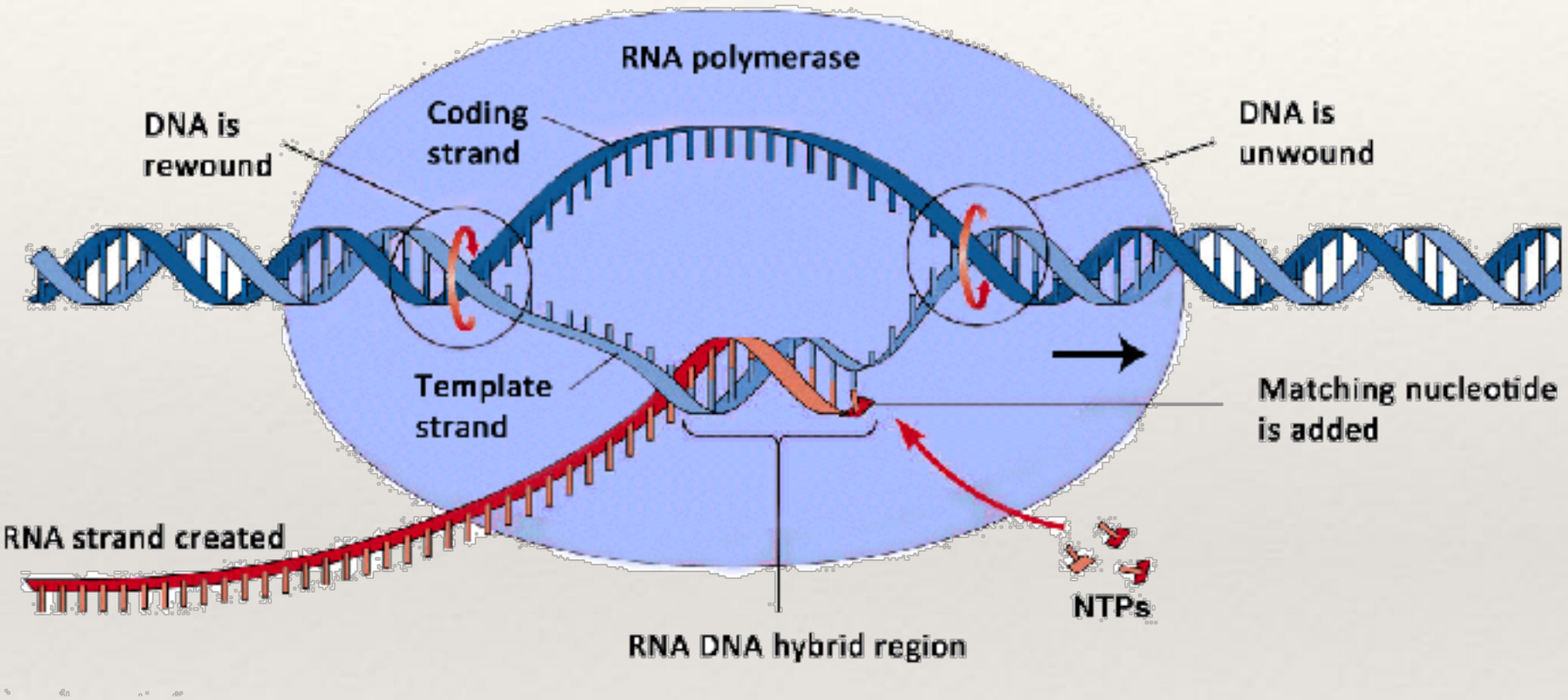
Transcriptome



Study individual variation

By Debivort

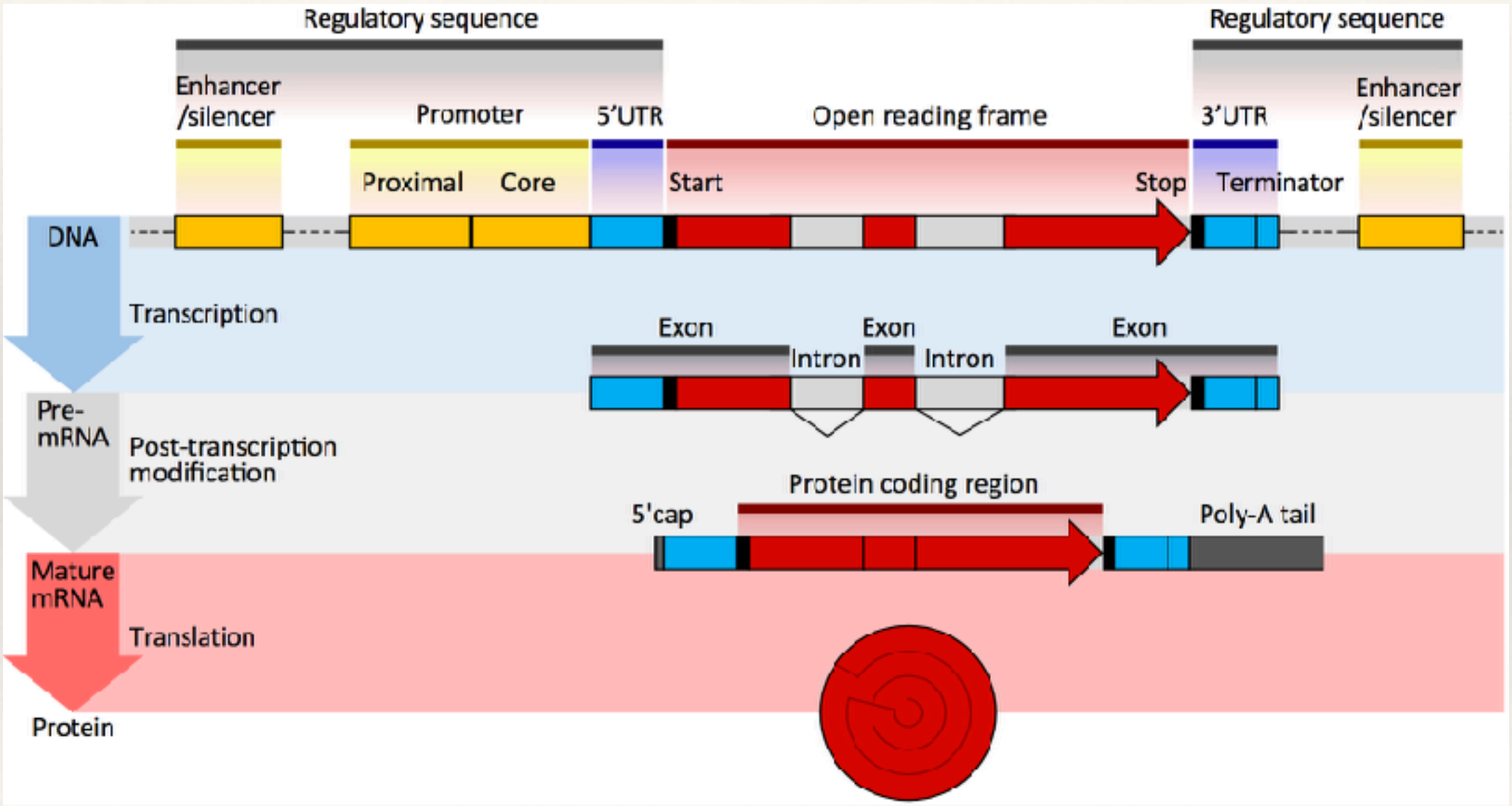
Transcription



Copying information from DNA to a RNA molecule for regulation or translation to protein

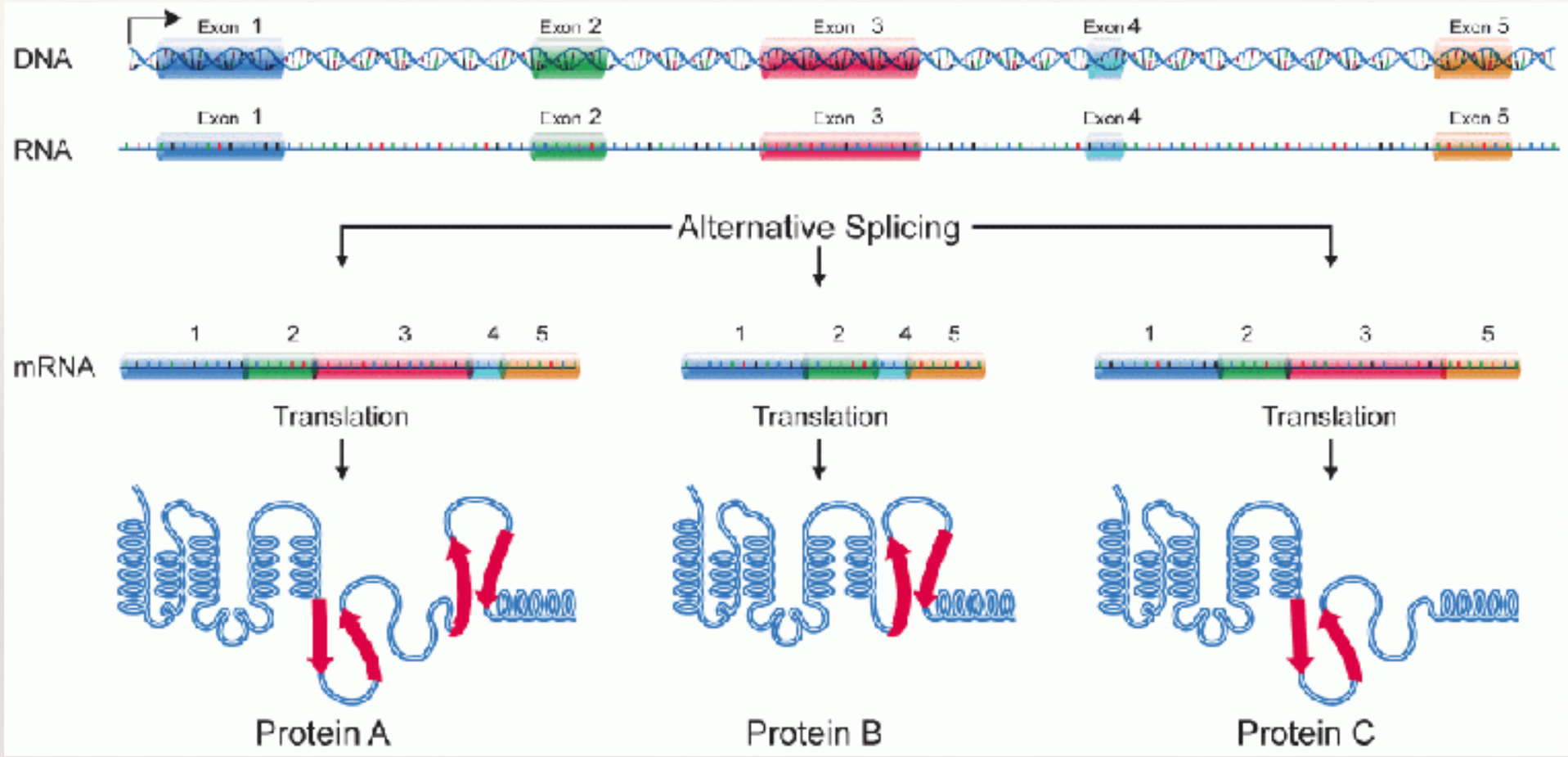
https://www.thinglink.com/scene/618815759912534018

Eukaryotic mRNA processing



https://en.wikipedia.org/wiki/Post-transcriptional_modification

Eukaryotic mRNA processing

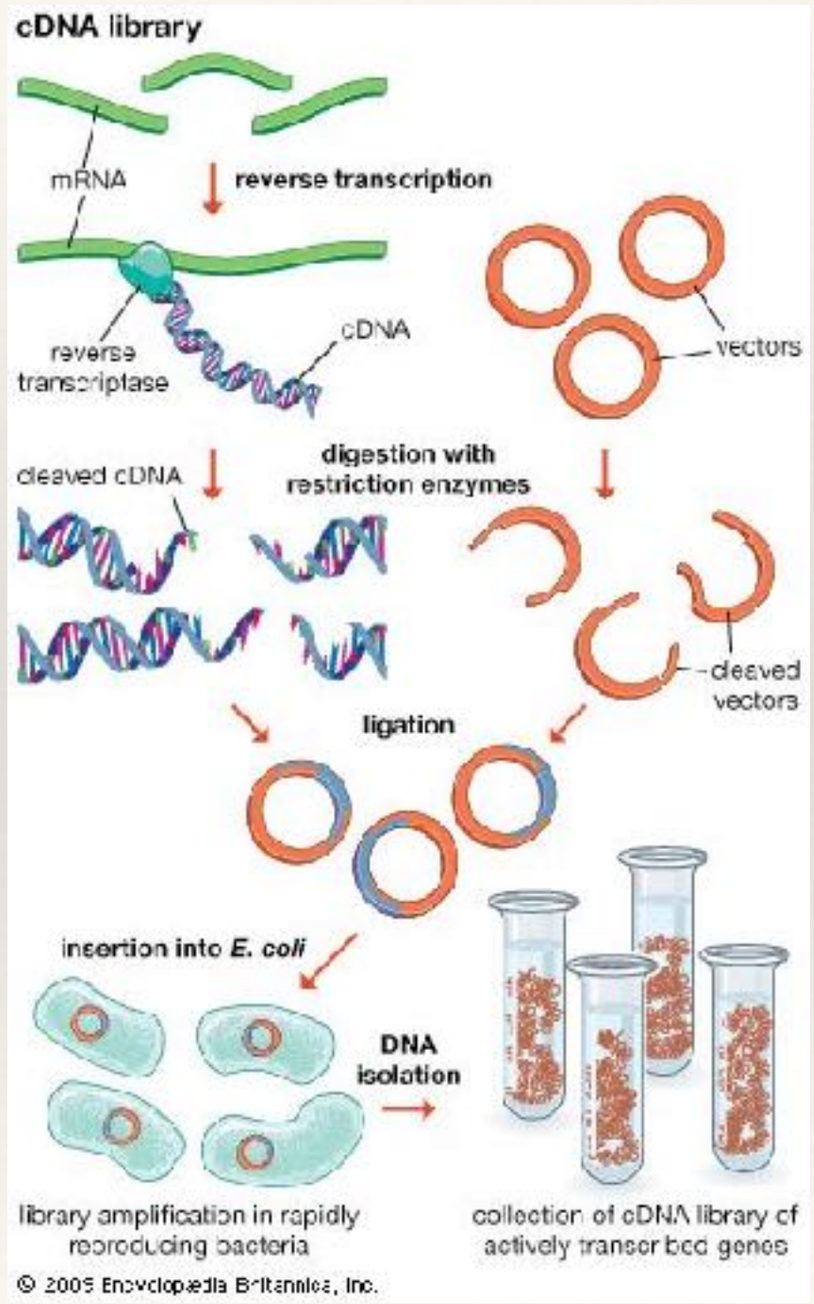


Splicing

https://en.wikipedia.org/wiki/Alternative_splicing

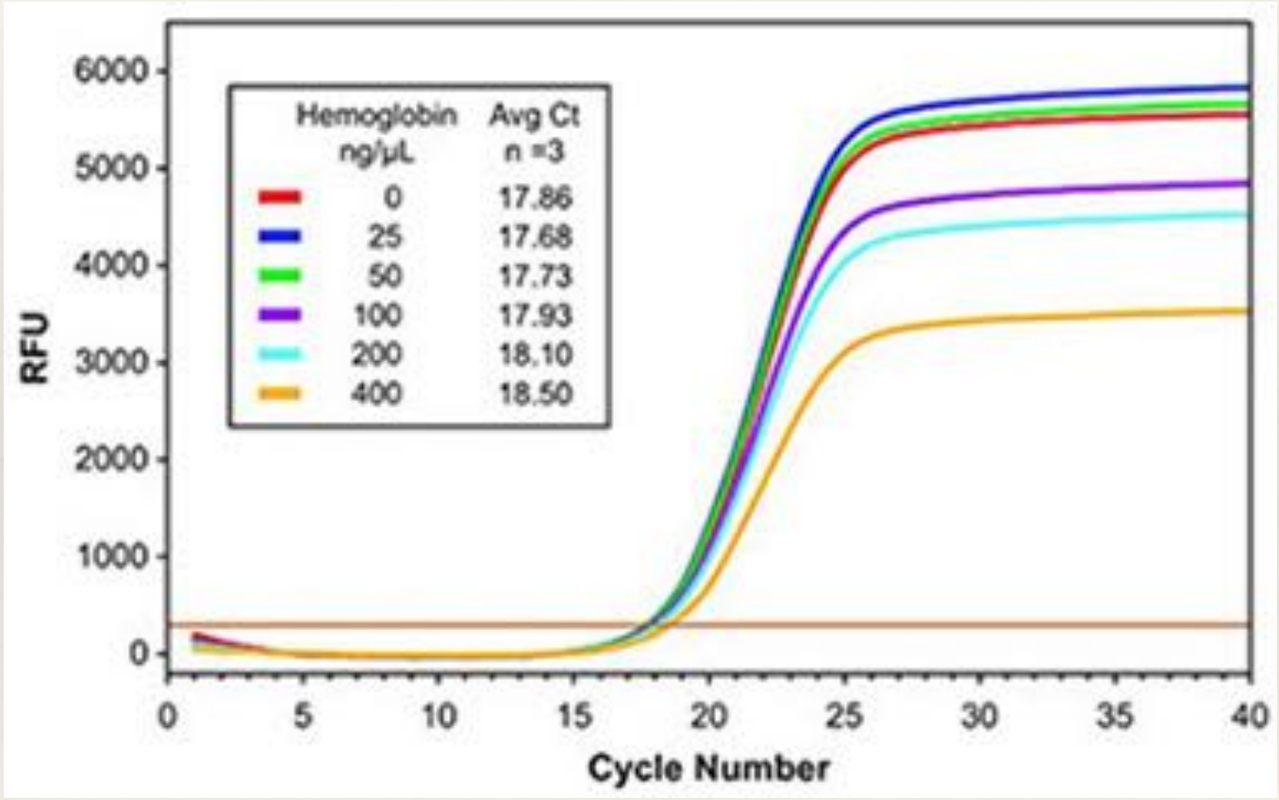
Obtaining transcriptome

- Sanger sequencing
  - mRNA converted to the more stable cDNA
  - cDNA cleaved and ligated into vectors
  - Vectors amplified (cloned) in *E. coli*
  - DNA isolated = cDNA library
  - Sequenced on Sanger
  - Low throughput
  - High accuracy



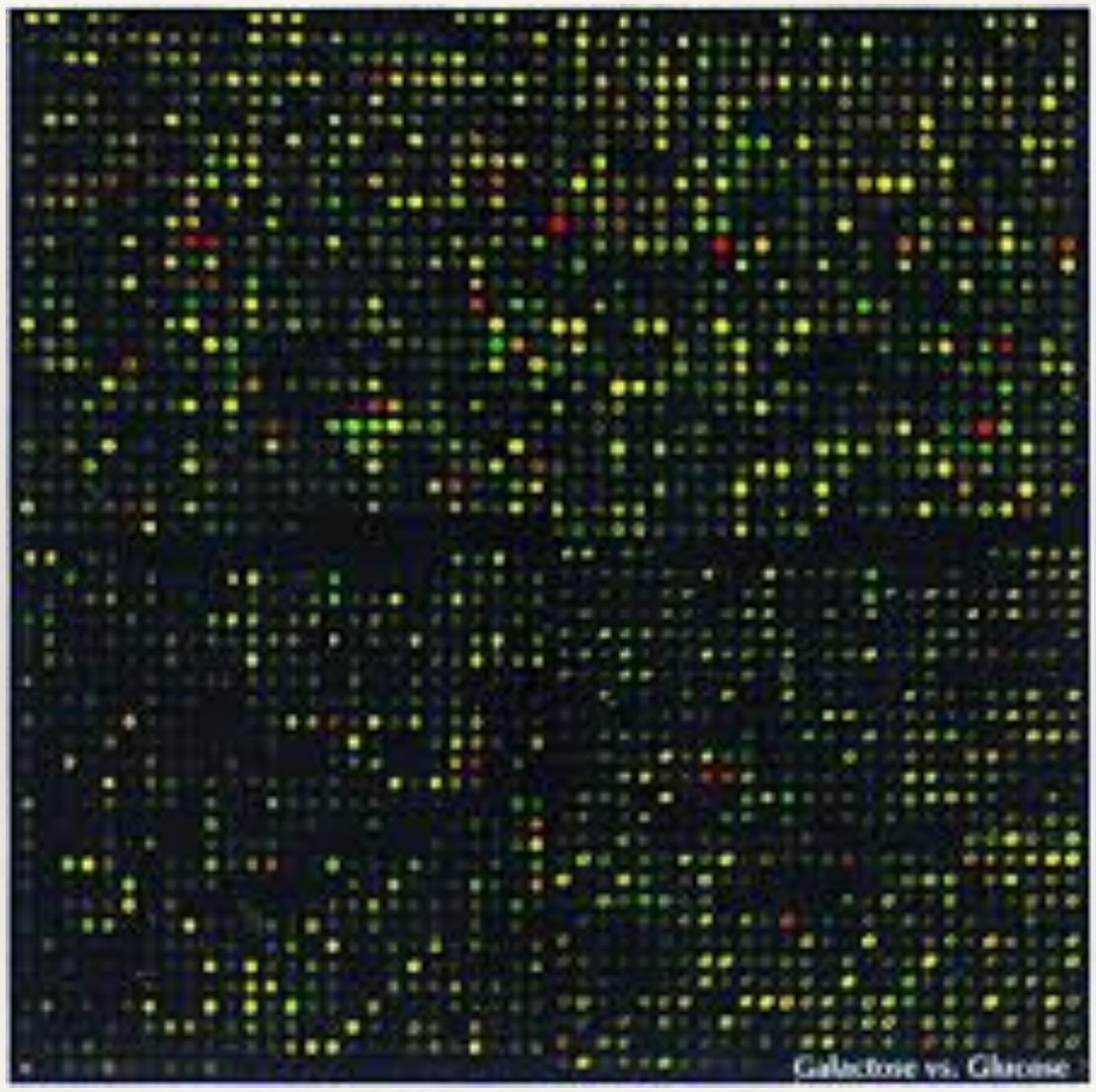
Quantifying expression

- Quantitative RT-PCR
  - qRT-PCR requires knowledge of gene sequence
  - Hard manual work
  - Low throughput
  - Expression level relative to control (house-keeping gene)



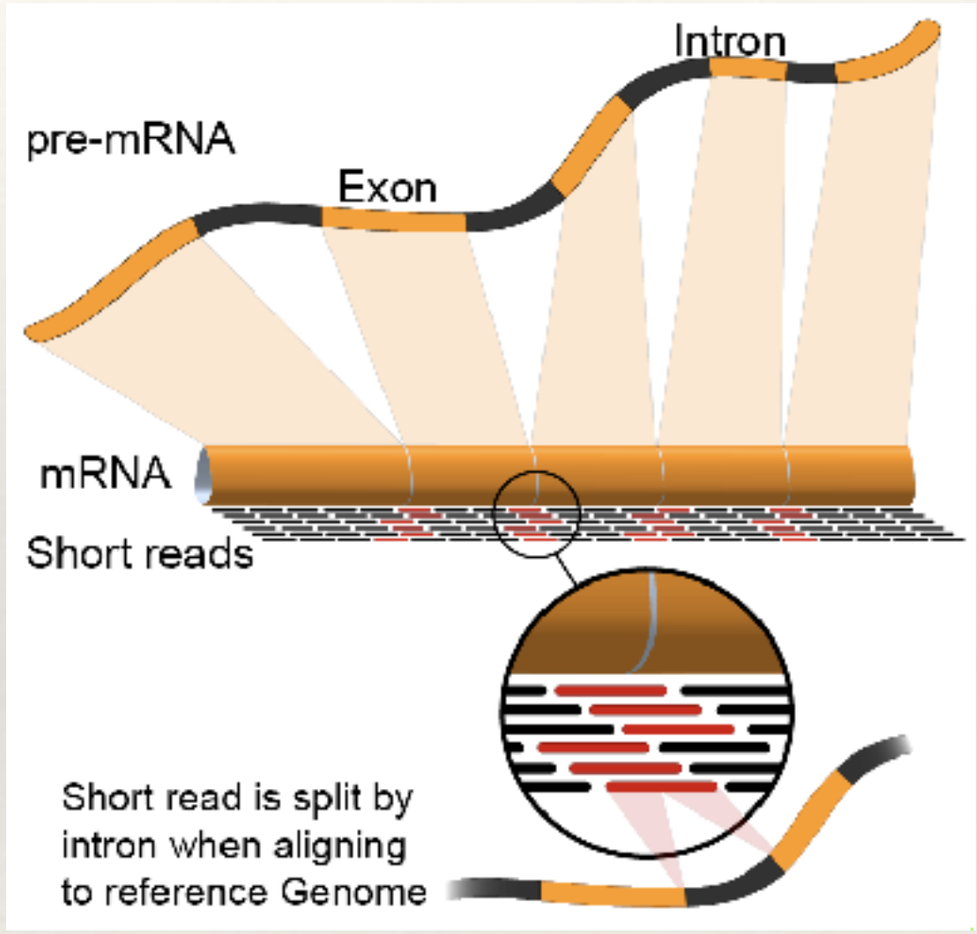
Quantifying expression

- Microarray
  - Requires gene sequences for probe design
  - High throughput compared to qRT-PCR
  - Possibility of outsourcing
  - Expression results relative to all probes



Quantifying expression

- RNA-seq
  - Transcriptome and expression in one go
  - No need for gene sequence information
  - High throughput
  - Can be outsourced
  - Costly, but effective
  - Expression results relative to all transcripts



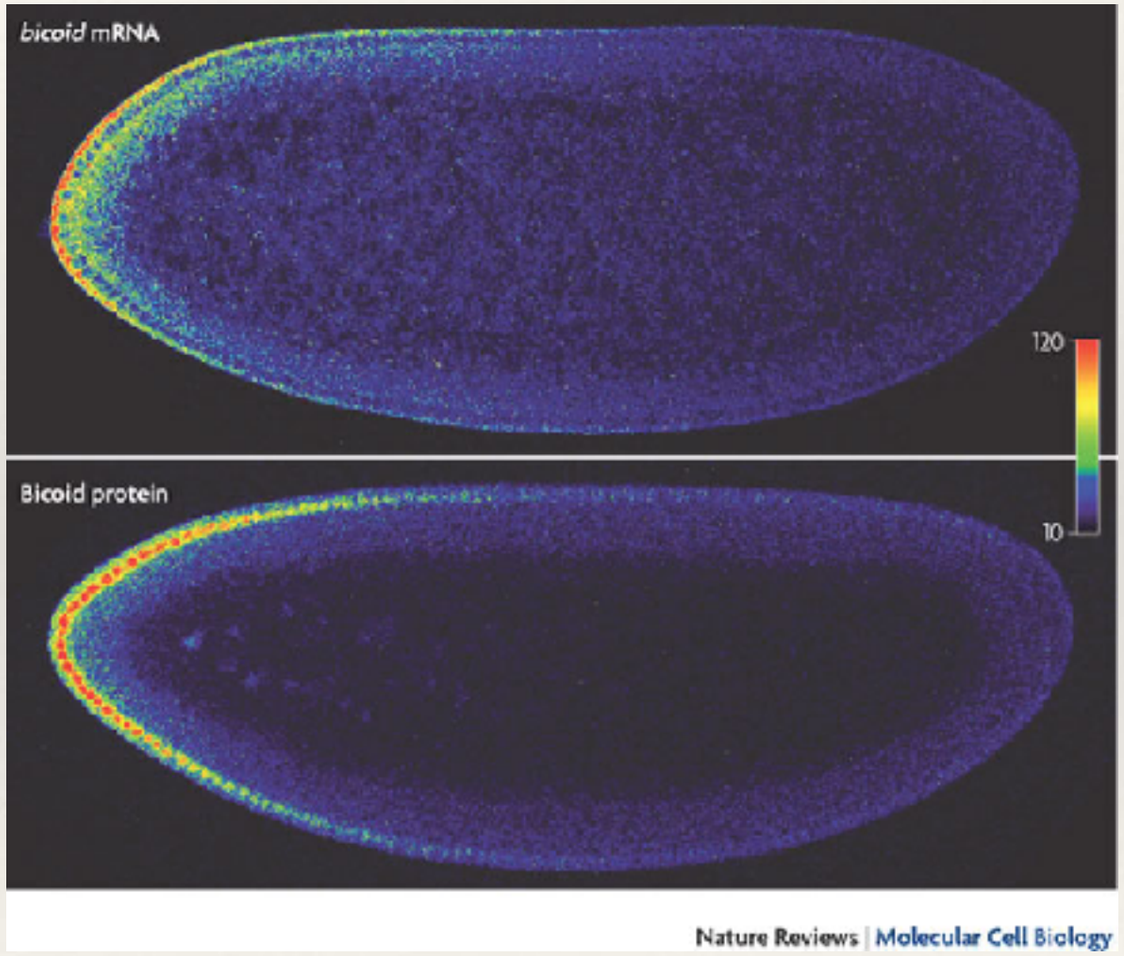
- Needs a different mindset

https://en.wikipedia.org/wiki/RNA-Seq

Quantifying expression

- Transcriptome = mRNA
- mRNA = Protein
- Protein = Biological relevance

- Things are seldom as simple as clear cut...



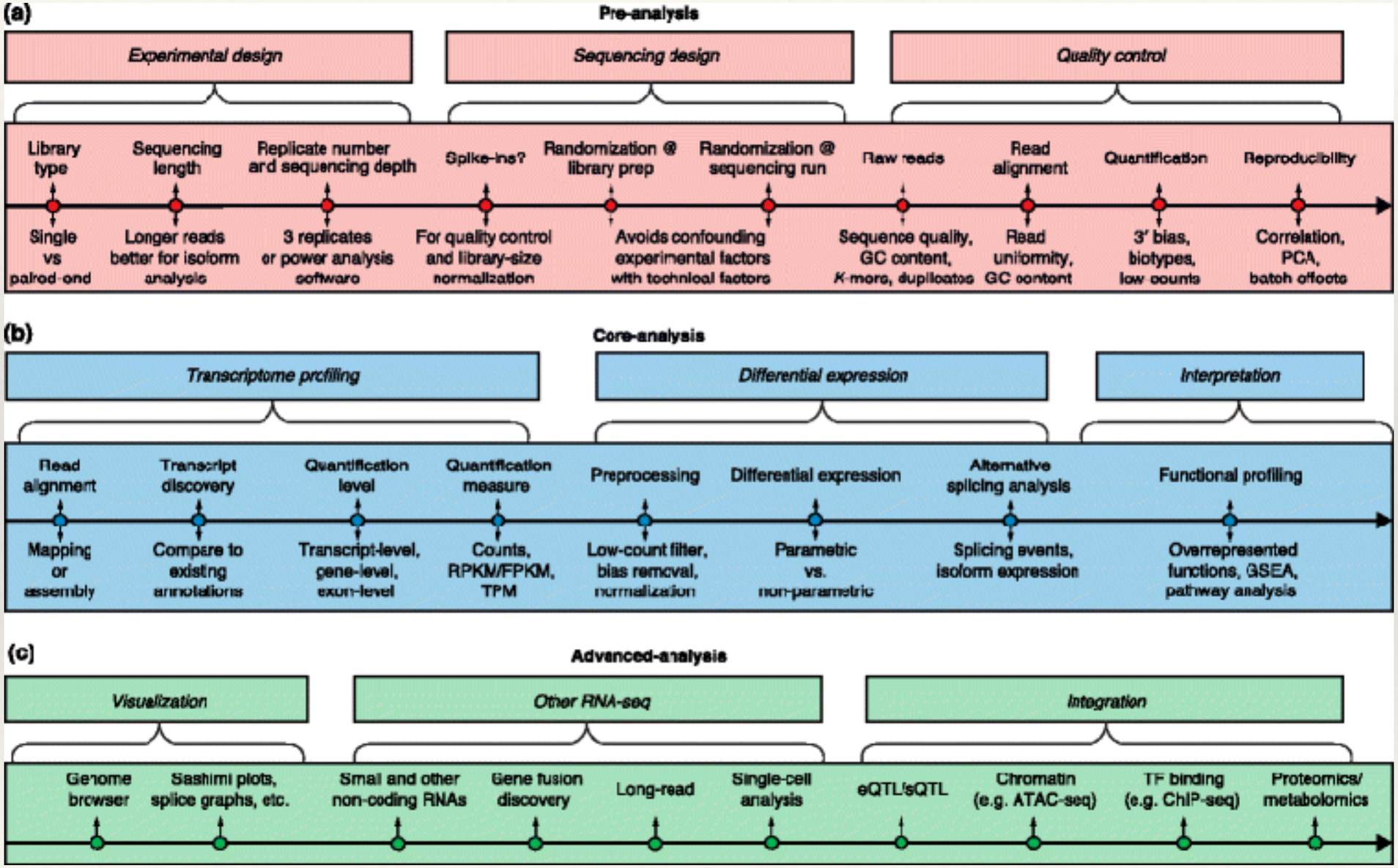
http://doi.org/10.1242/dev.031195

Things to remember

- RNA decay
- RNA editing
- RNA splicing
- Translation regulation
- RNA interference

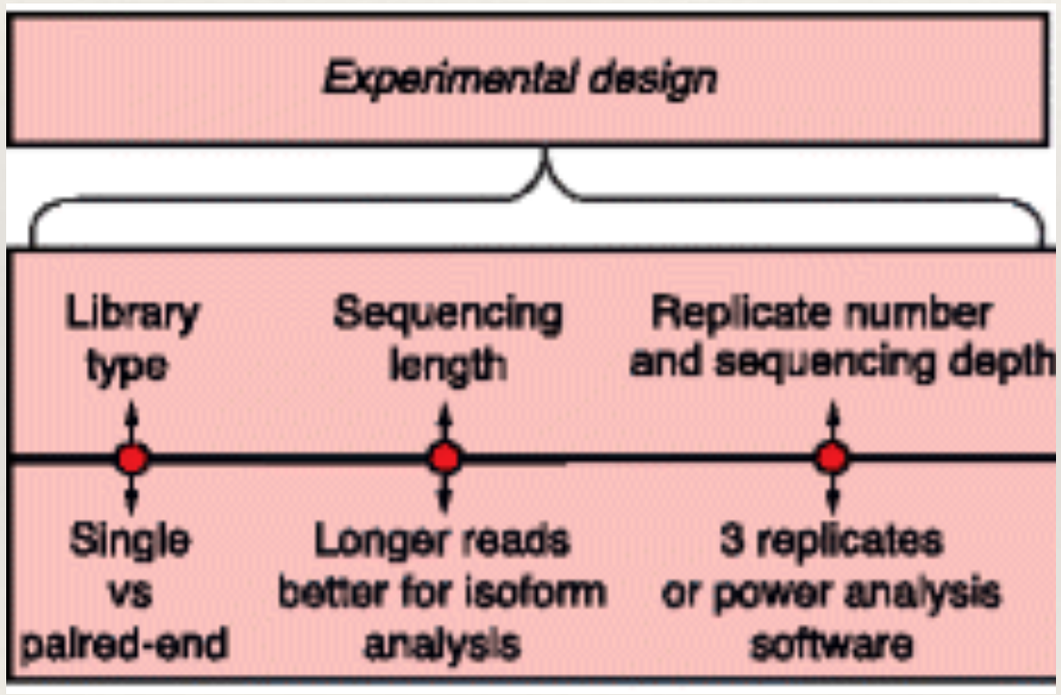
- Heavily dependent on proper experimental design
- Enormous amounts of data
- No straight forward analysis
- Usually no clear-cut story from individual gene expressions

Pipeline(s) - too many



A survey of best practices for RNA-seq data analysis
https://doi.org/10.1186/s13059-016-0881-8

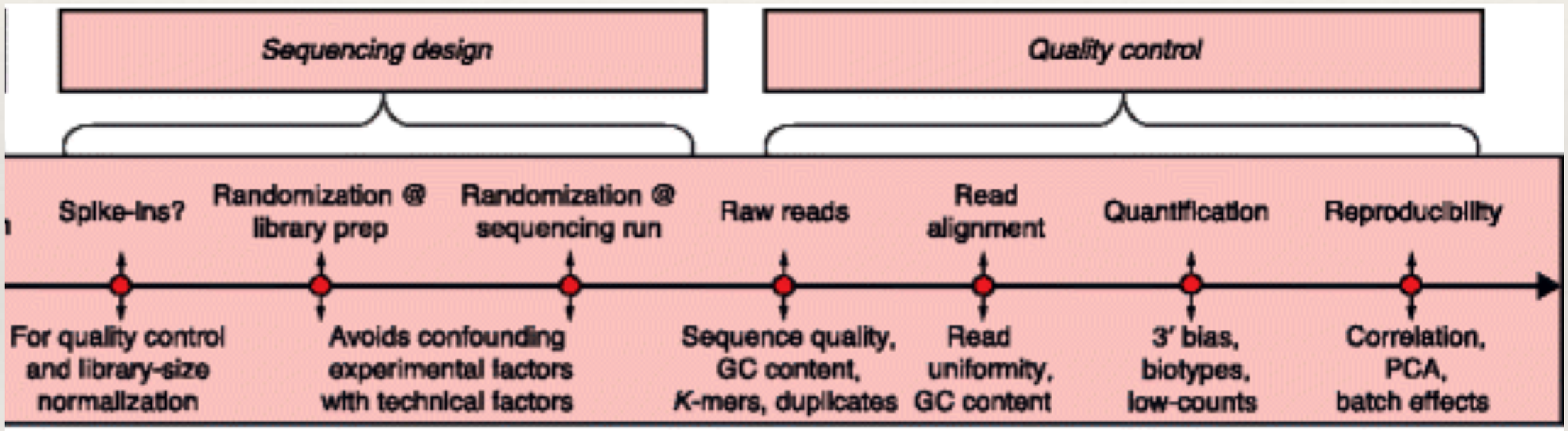
Pipeline(s) - too many



Experimental design is very important.

A survey of best practices for RNA-seq data analysis
https://doi.org/10.1186/s13059-016-0881-8

Pipeline(s) - too many

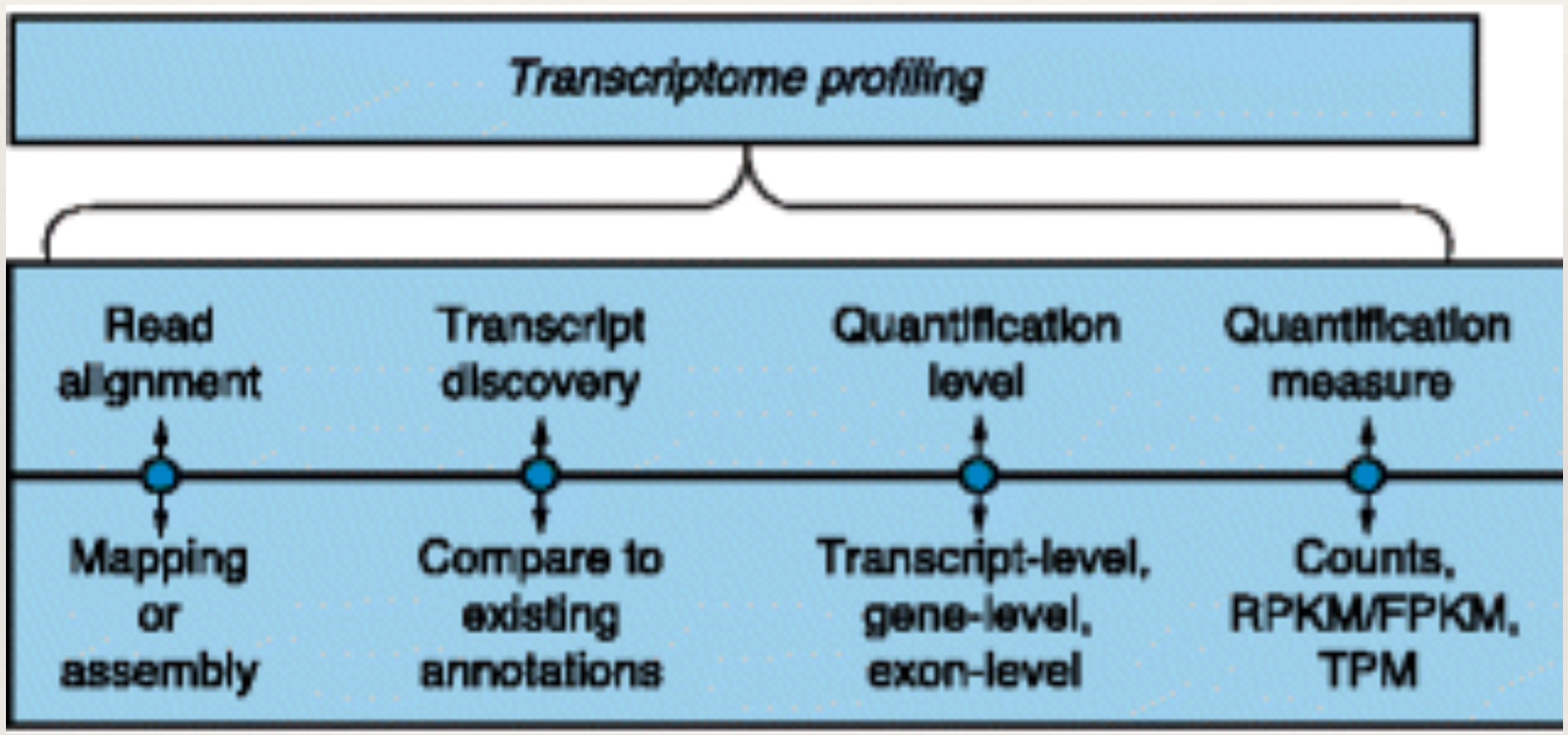


ERCC https://tools.thermofisher.com/content/sfs/manuals/cms_086340.pdf

PhiX

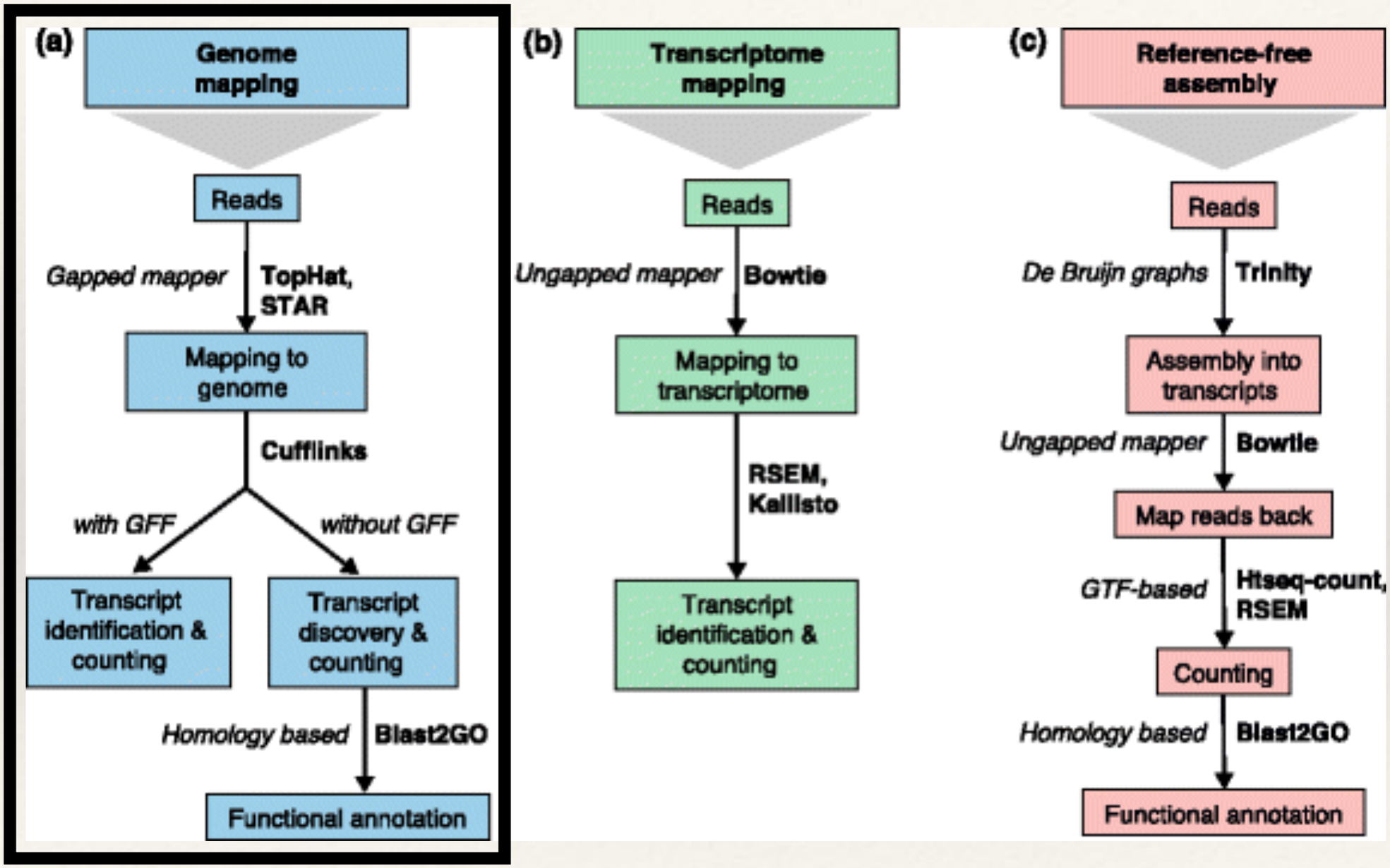
A survey of best practices for RNA-seq data analysis
https://doi.org/10.1186/s13059-016-0881-8

Pipeline(s) - too many



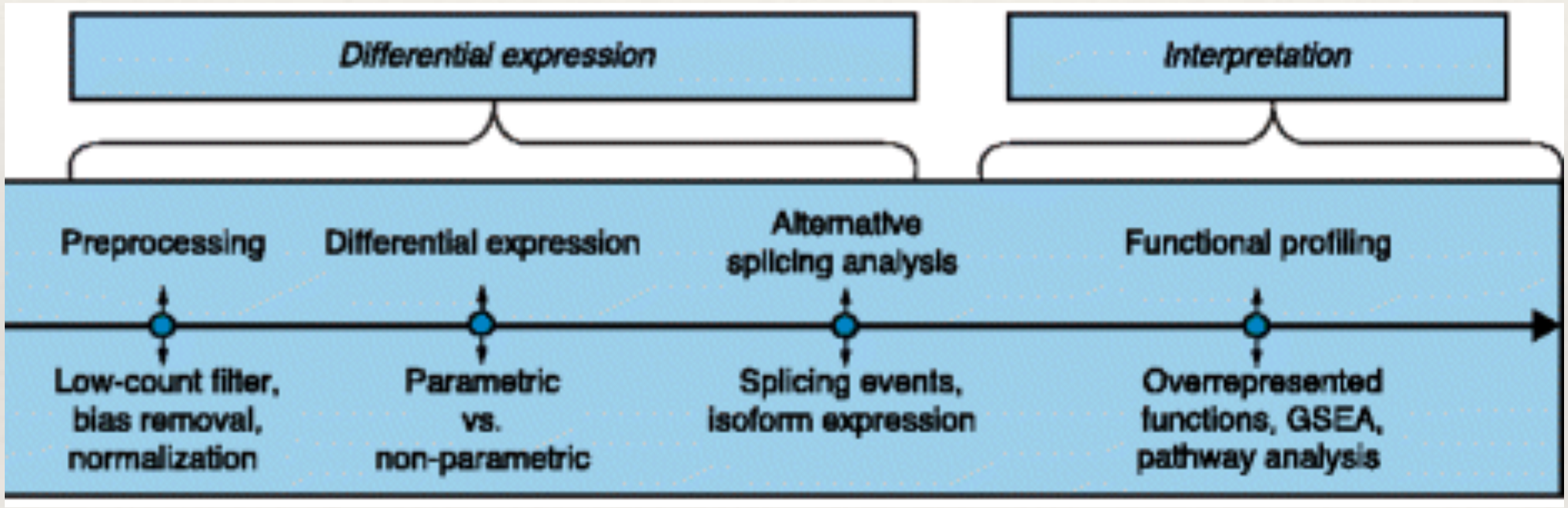
A survey of best practices for RNA-seq data analysis
https://doi.org/10.1186/s13059-016-0881-8

Mapping sequence data



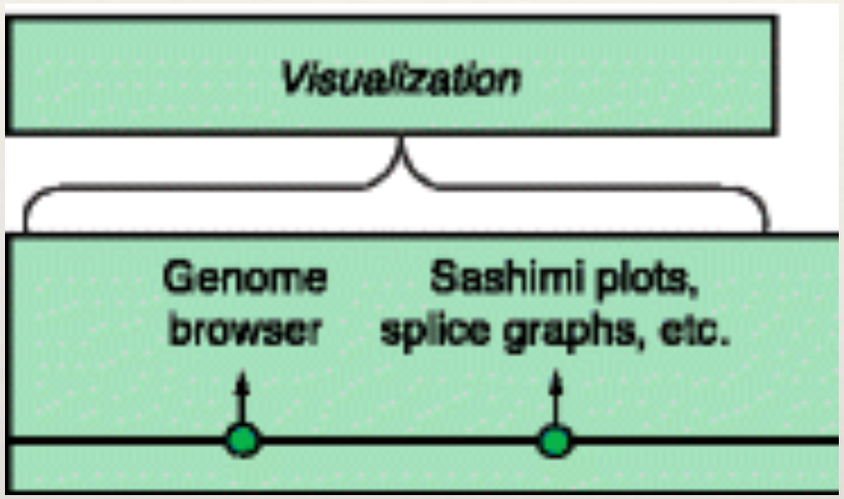
A survey of best practices for RNA-seq data analysis
https://doi.org/10.1186/s13059-016-0881-8

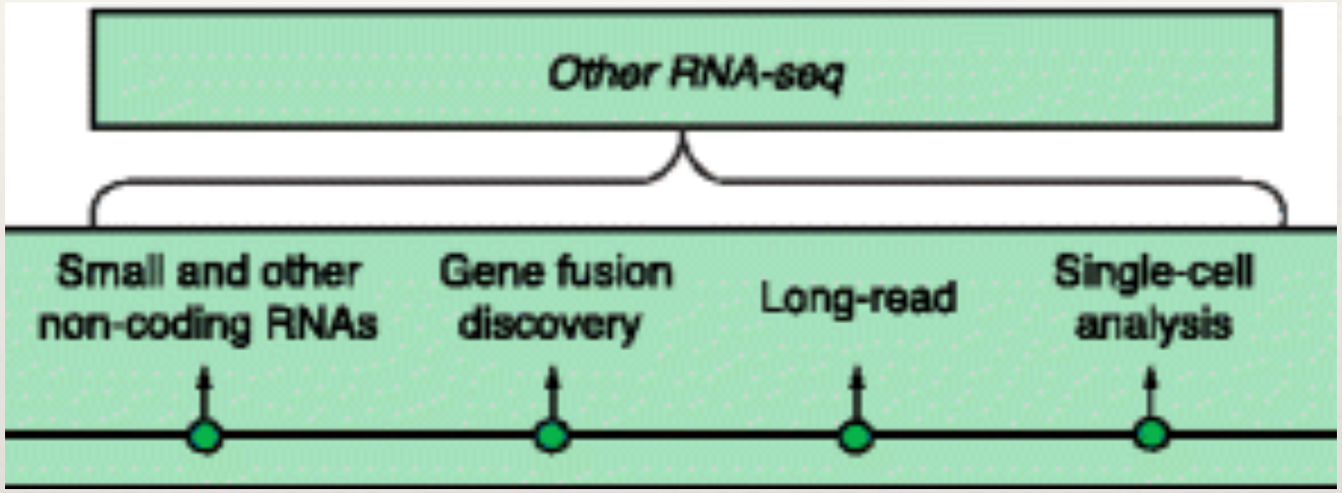
Pipeline(s) - too many

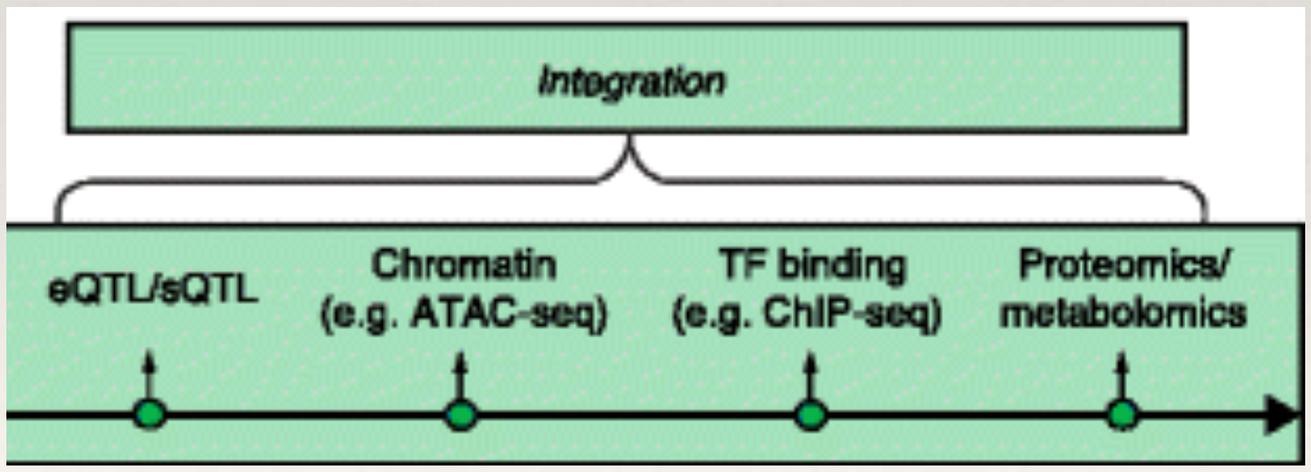


A survey of best practices for RNA-seq data analysis
https://doi.org/10.1186/s13059-016-0881-8

Pipeline(s) - too many







A survey of best practices for RNA-seq data analysis
https://doi.org/10.1186/s13059-016-0881-8