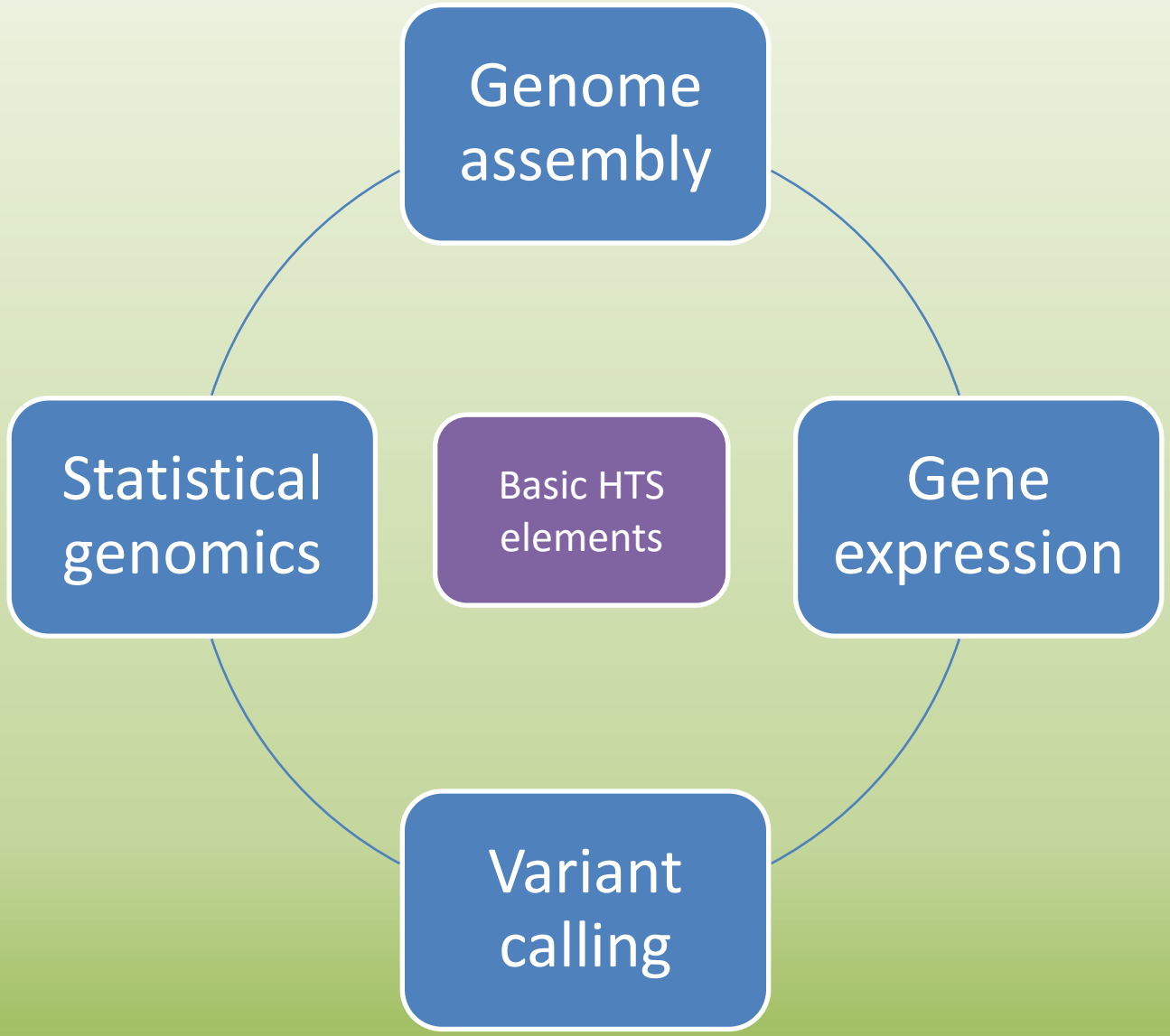
RNA seq: differential expression analysis

For INF-BIO 4121/9121
Fall semester 2016

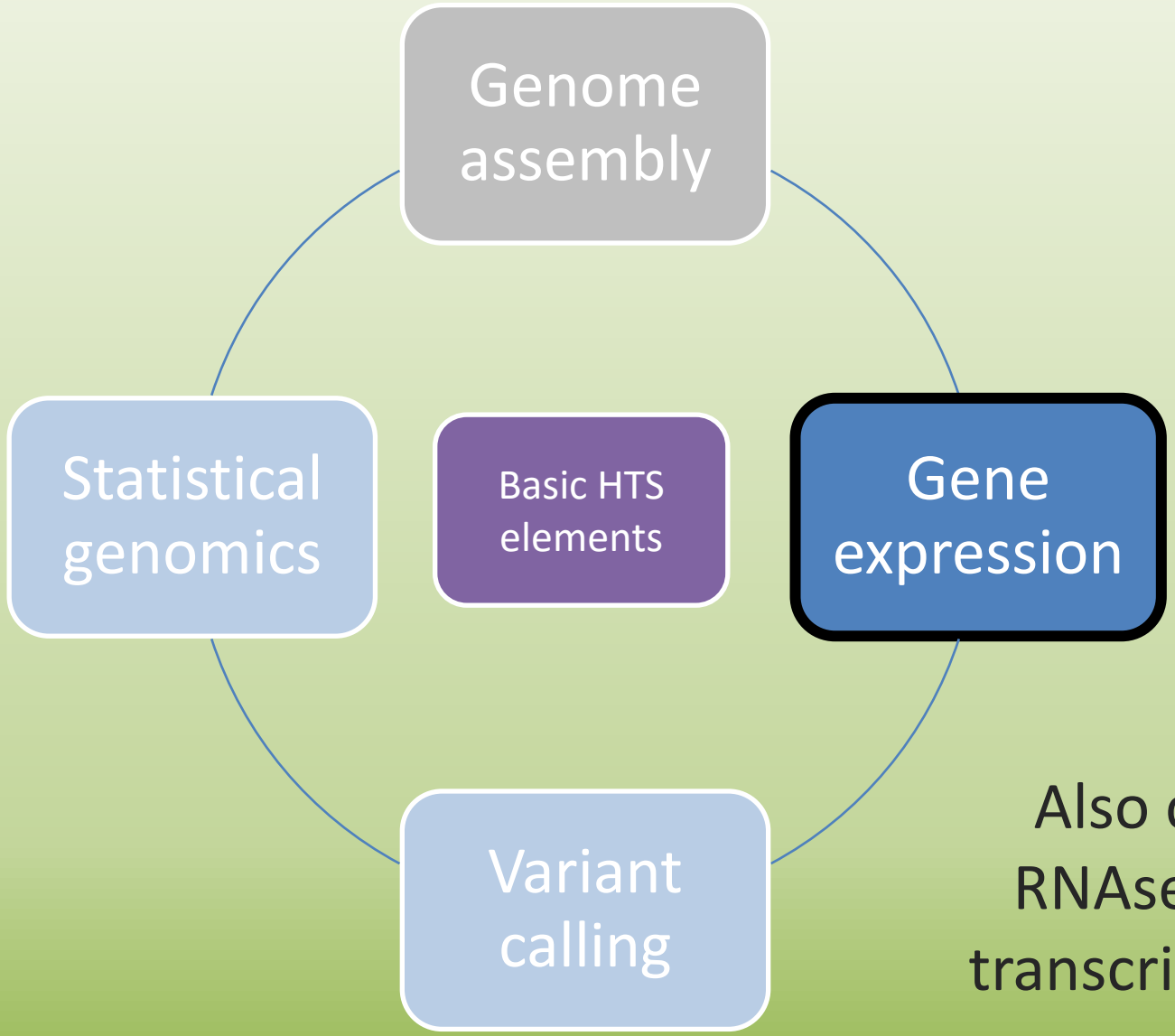
Rebekah Oomen / Monica Hongrø Solbakken
r.a.oomen@ibv.uio.no / m.h.solbakken@ibv.uio.no

UiO : Centre for Ecological and Evolutionary Synthesis
University of Oslo

INFBIO x121


Genome assembly
Gene expression
Variant calling
Statistical genomics
Basic HTS elements

INFBIO x121



Also called RNAseq and transcriptomics

Aims I

- You should be able to tell us:
  - Overall:
    - What is RNAseq?
    - Are there different kinds of RNAseq?
    - What is RNAseq used for?
  - In more depth:
    - How to design a RNAseq experiment for differential expression analysis
      - How to chose analysis strategy
      - Pitfalls in differential expression analysis (sequencing depth, batch effects, statistical approach etc.)

Aims I

- You should be able to tell us:
  - Overall:
    - What is RNAseq?
    - Are there different kinds of RNAseq?
    - What is RNAseq used for?
  - In more depth:
    - How to design a RNAseq experiment for differential expression analysis
      - How to chose analysis strategy
      - Pitfalls in differential expression analysis (sequencing depth, batch effects, statistical approach etc.)

Aims II

- You should be able to perform:
  - A reference based differential gene expression analysis with a pair-wise comparison involving several biological replicates
  - Present the overall statistics from that analysis (mapping percentage, variance, potential outliers, number of differentially expressed genes etc.)
  - Extract and present the main biological result(s) based on the annotation of the differentially expressed genes

Aims II

- You should be able to perform:
  - A reference based differential gene expression analysis with a pair-wise comparison involving several biological replicates
  - Present the overall statistics from that analysis (mapping percentage, variance, potential outliers, number of differentially expressed genes etc.)
  - Extract and present the main biological result(s) based on the annotation of the differentially expressed genes

Aims II

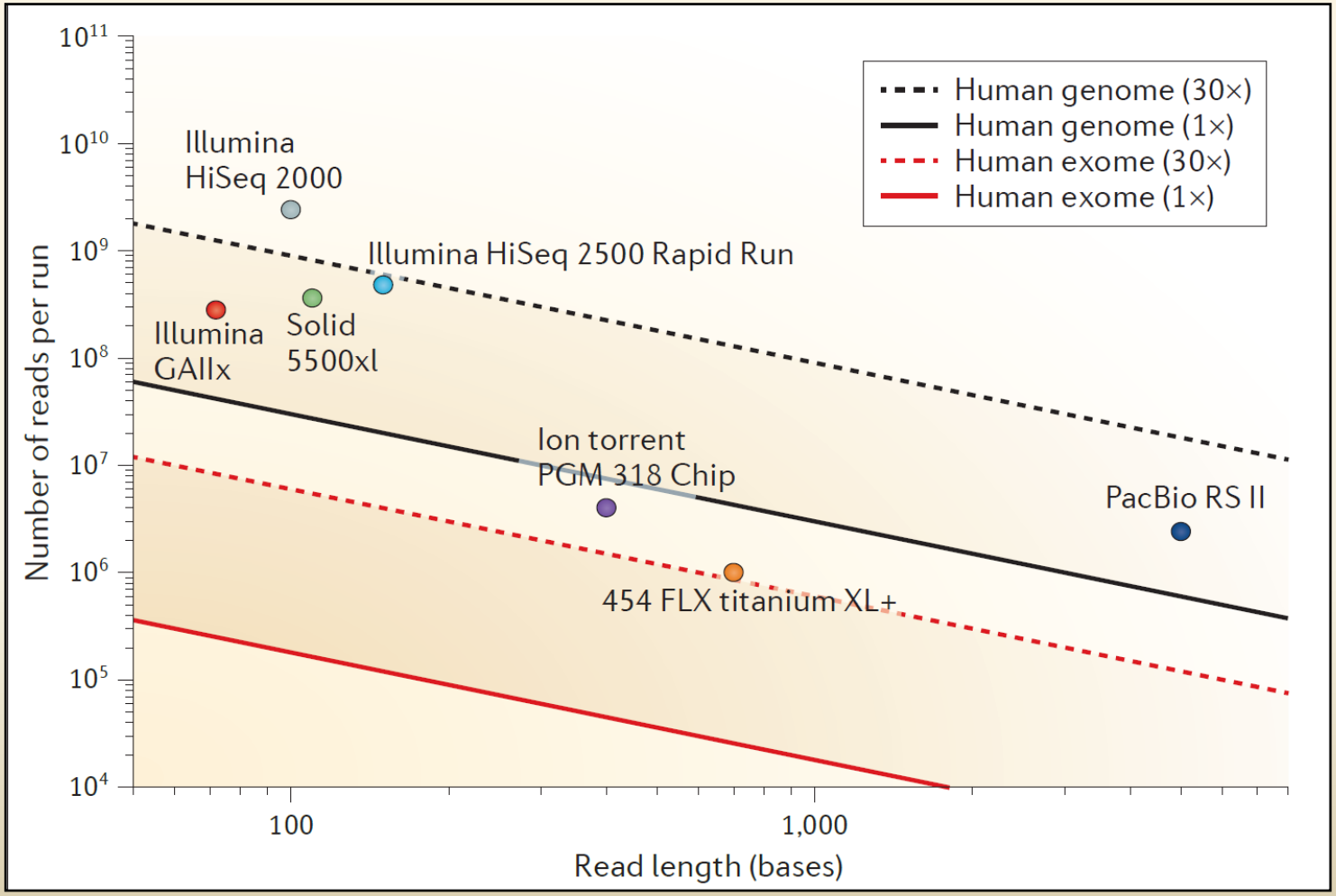
- You should be able to perform:
  - A reference based differential gene expression analysis with a pair-wise comparison involving several biological replicates
  - Present the overall statistics from that analysis (mapping percentage, variance, potential outliers, number of differentially expressed genes etc.)
  - Extract and present the main biological result(s) based on the annotation of the differentially expressed genes

Outline I

- RNAseq module day 1
  - Introducing RNAseq
  - Experimental design and considerations

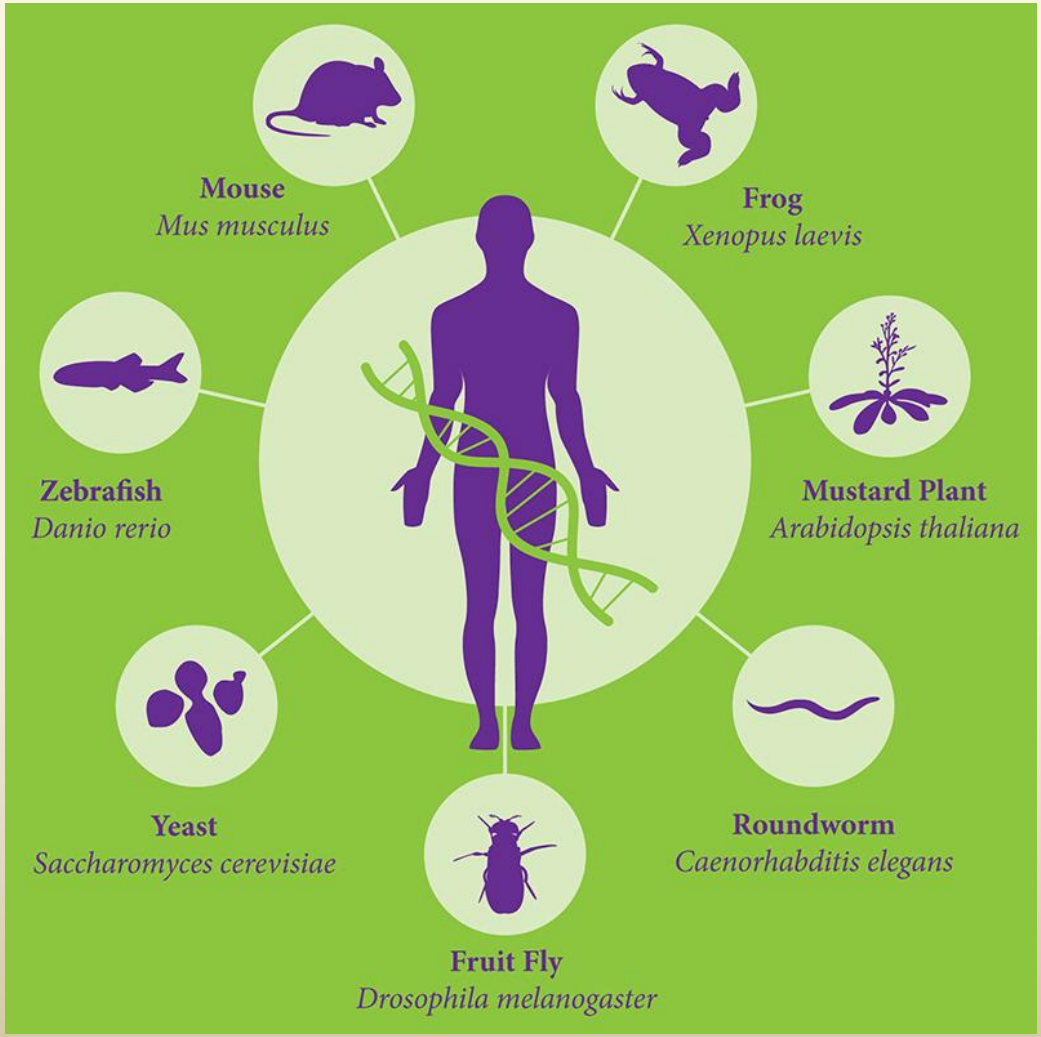
    - Experimental design exercise
    - The first steps of RNAseq exercise

Next generation sequencing and new possibilities

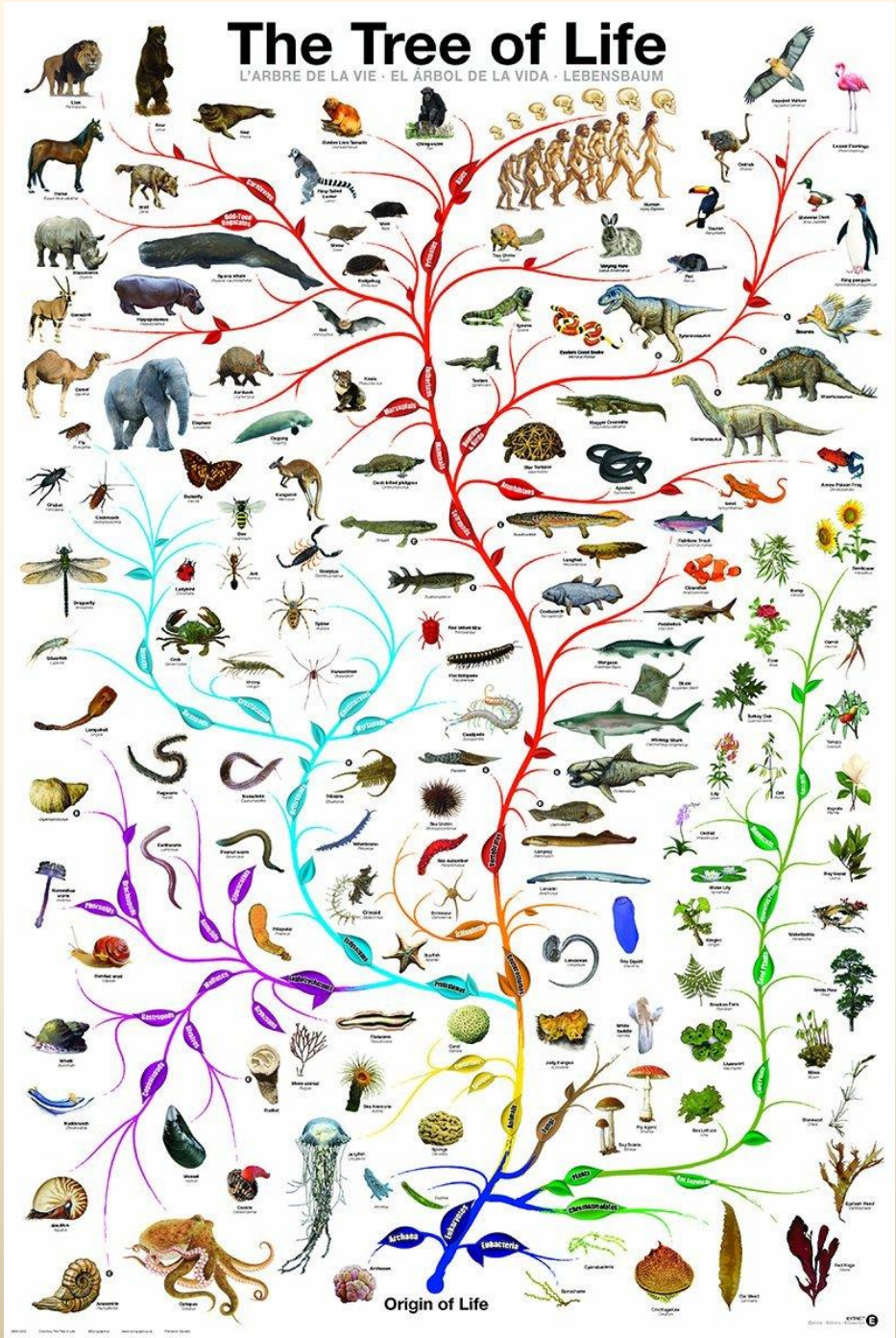


http://www.nature.com/nrg/journal/v15/n2/full/nrg3642.html

Moving away from model systems



https://biology.uiowa.edu/model-organisms

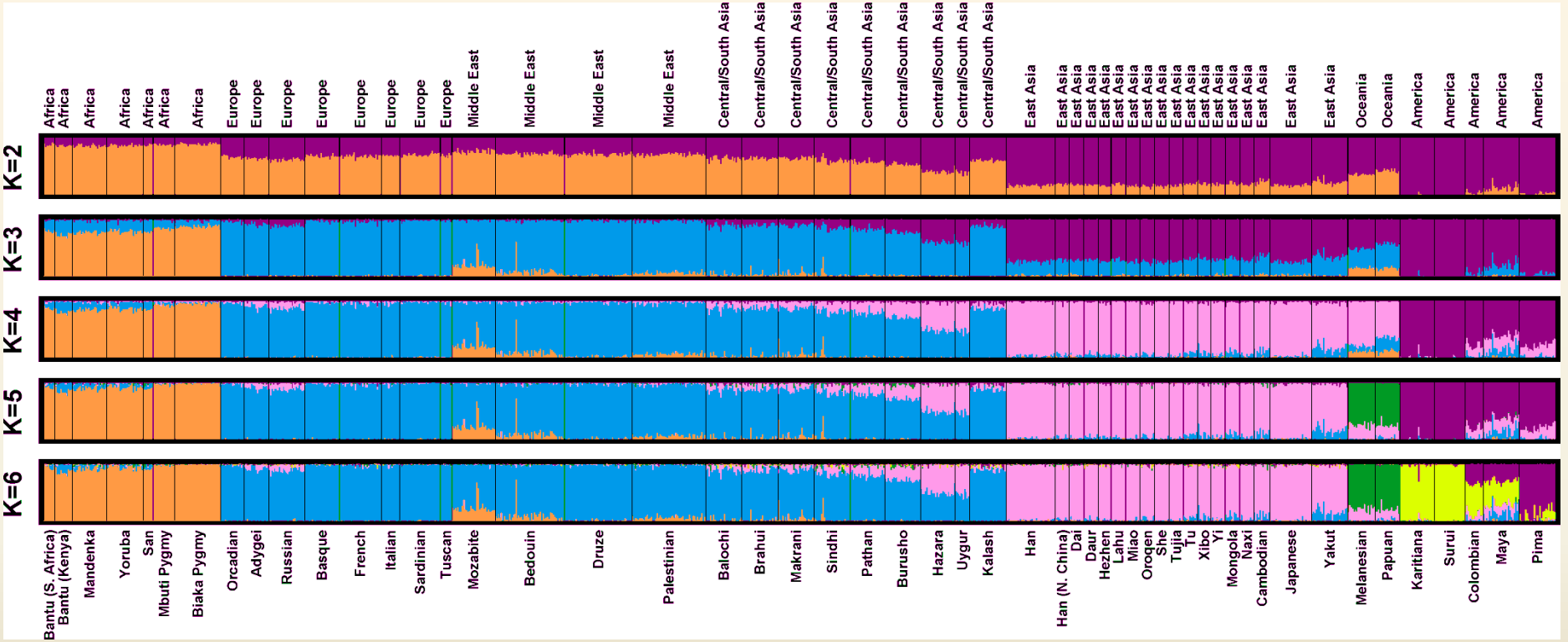


Non-model species

The tree is the limit…

https://www.thinglink.com/scene/645083259847311362

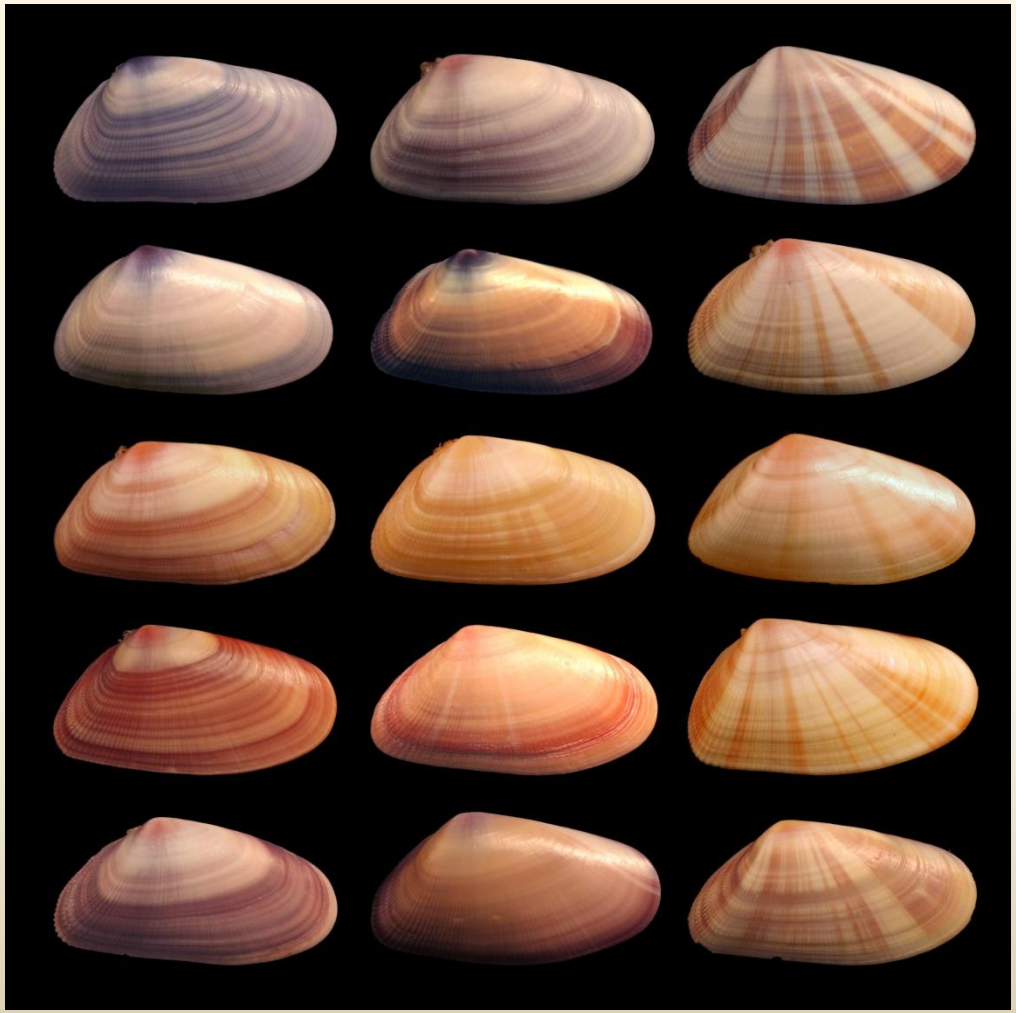
Population variation



Inferred Population Structure Based on 1,048 Individuals and 993 Markers, Assuming Correlations among Allele Frequencies across Clusters.
Each individual is represented by a thin line partitioned into K colored segments that represent the individual's estimated membership fractions in K clusters.

http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.0010070

Individual variation

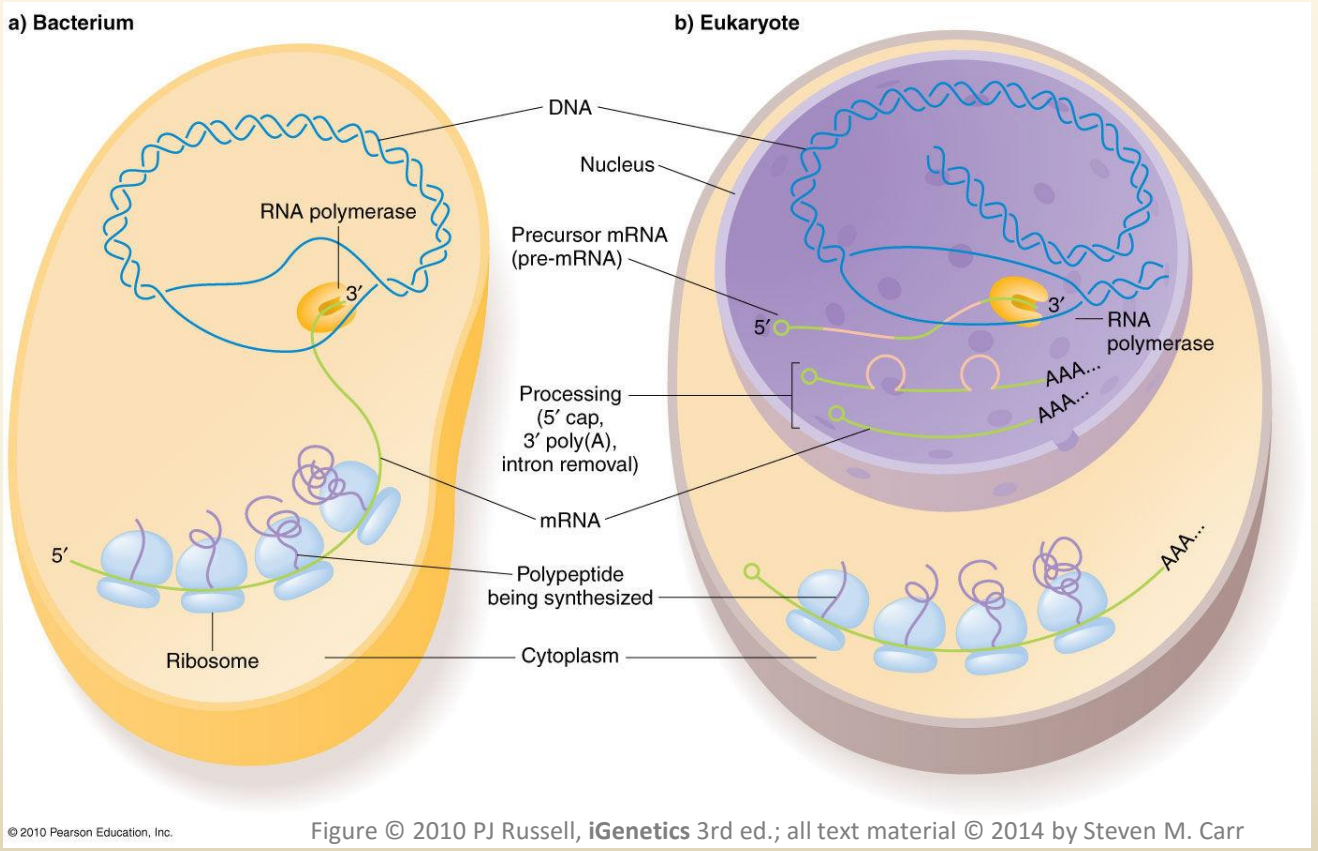


A range of variability in the mussel *Donax variabilis*

Quickly about transcription and old-school transcriptomics

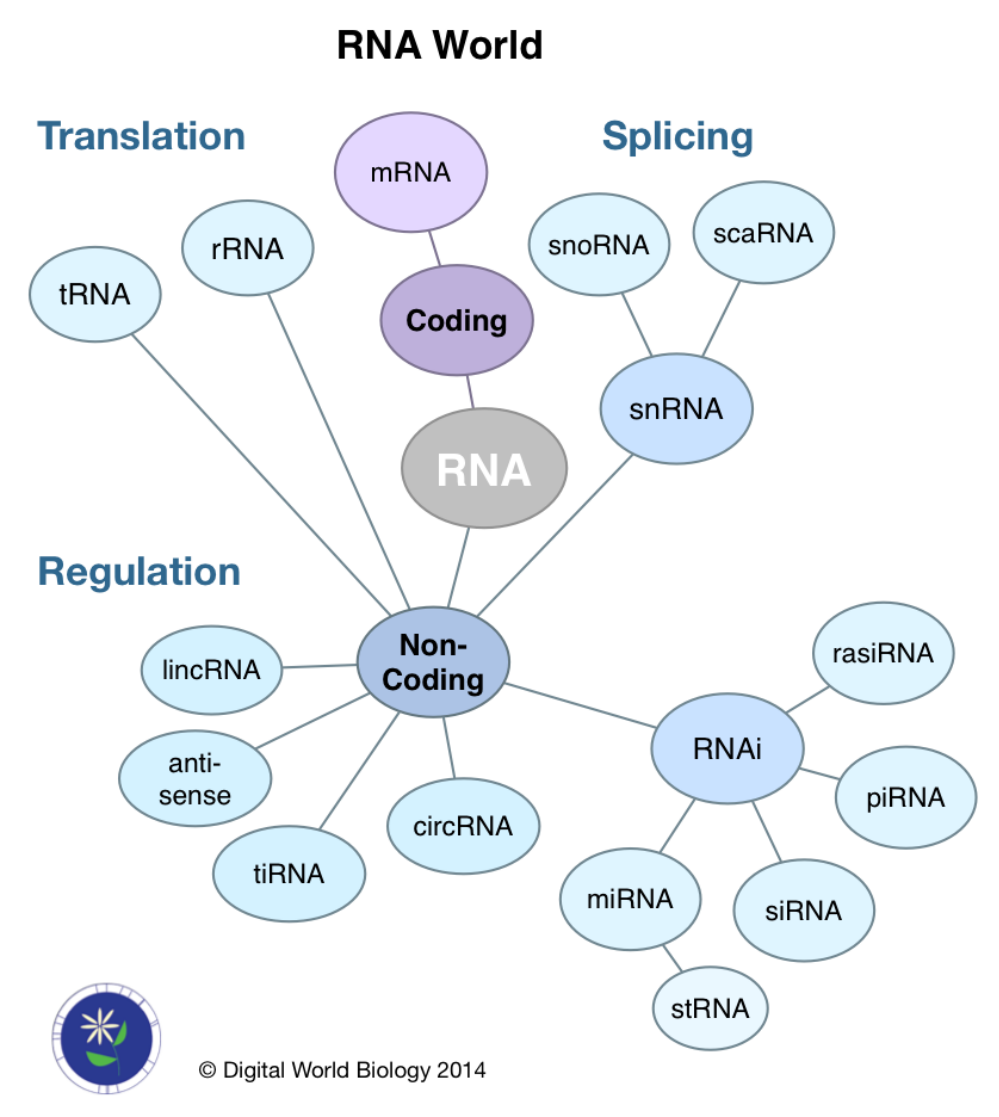
Transcription

Copying information from DNA to a mobile RNA for regulatory or protein coding purposes

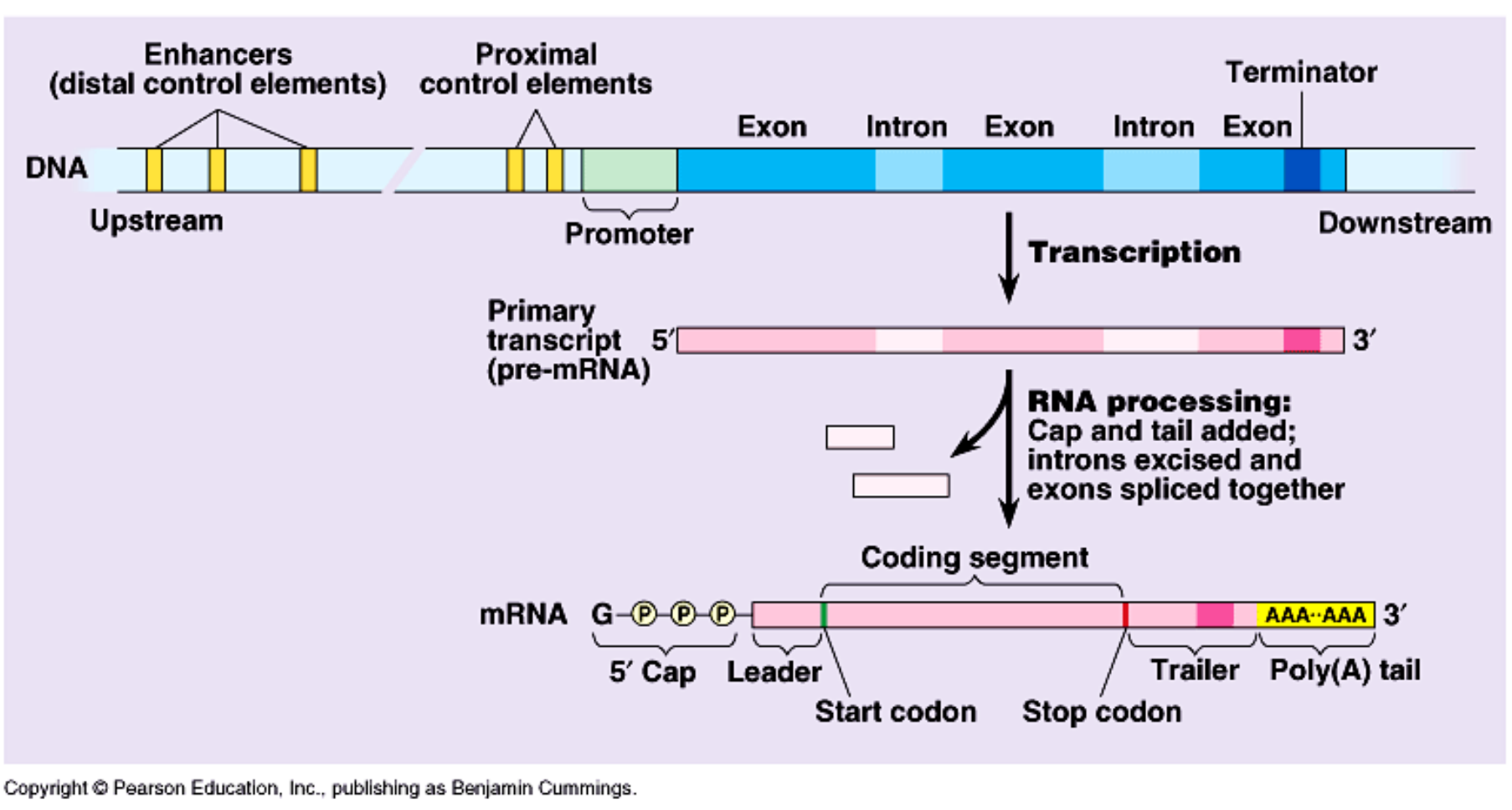


Transcription – all the RNAs

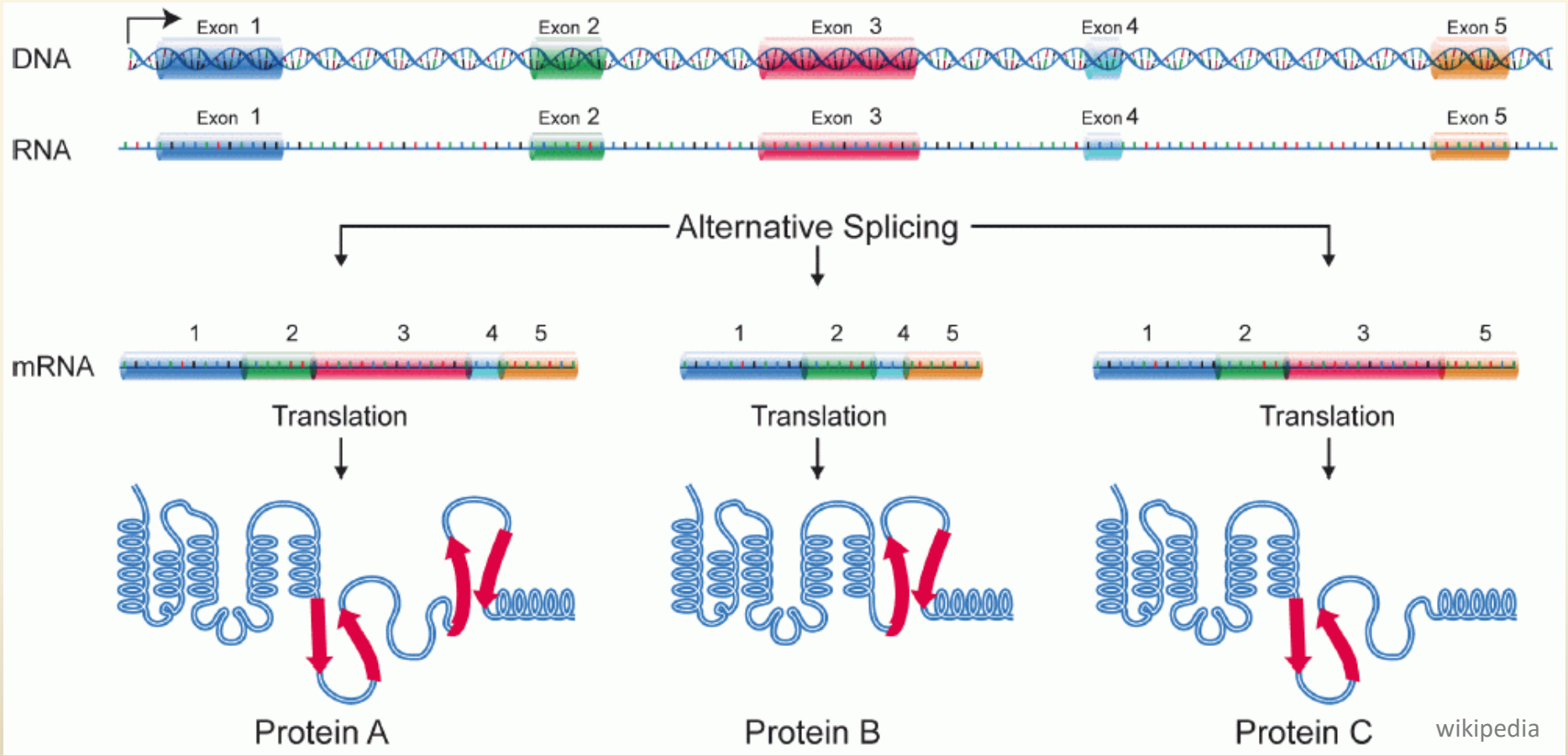
A transcriptome is a snapshot in time of **all RNAs** present in a sample isolated from a given cell, tissue or organism



Transcription eukaryote mRNA



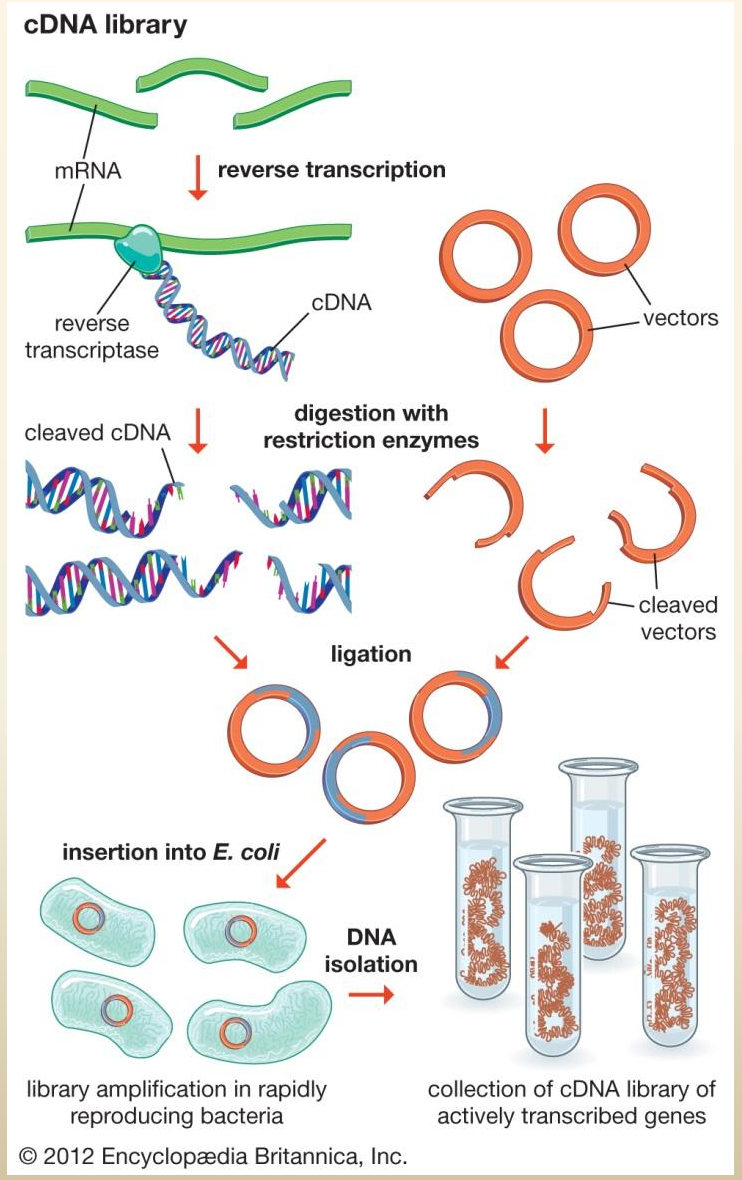
Splice variation eukaryotic mRNA



Obtaining transcriptomes I

Sanger cDNA library sequencing

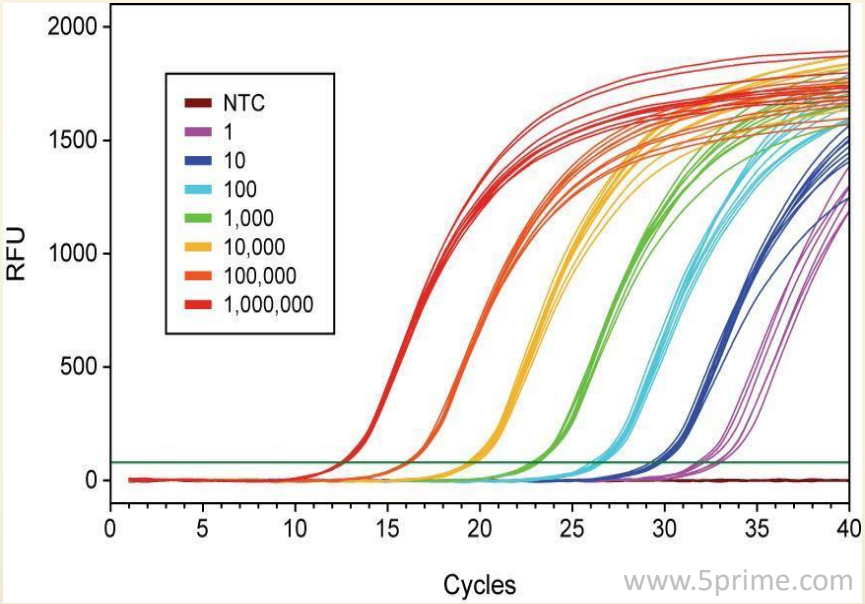
- mRNA converted to the more stable cDNA
- cDNA cleaved and ligated into vectors
- Vectors amplified (cloned) in *E. coli*
- DNA isolated = cDNA library
- Sequenced on Sanger
- Low throughput
- High accuracy



Obtaining expression I

Quantitative RT-PCR

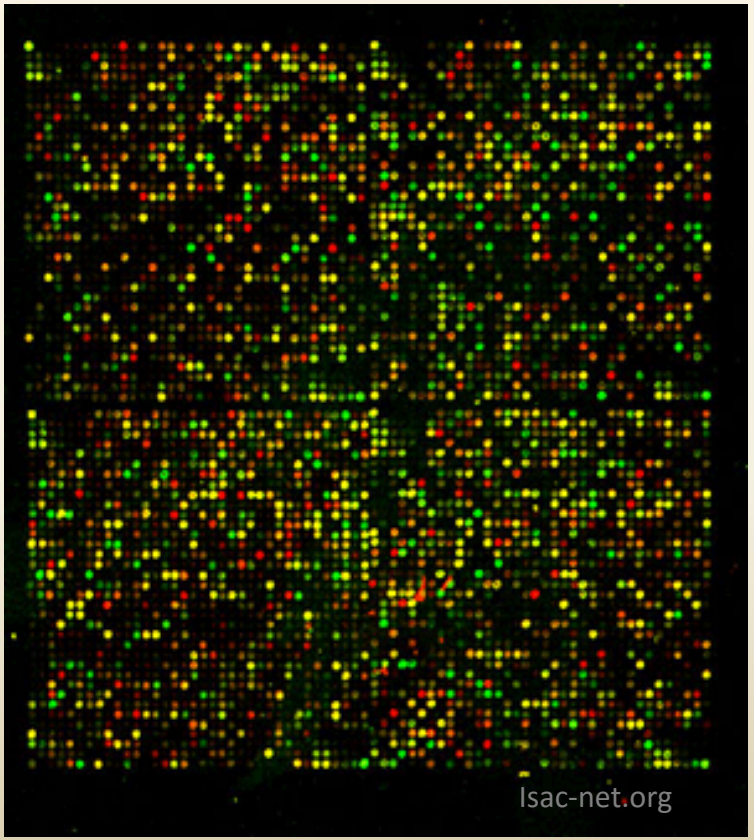
- qRT-PCR requires knowledge of gene sequence
- Hard manual work
- Low throughput
- Expression level relative to control (house-keeping gene)



Obtaining expression II

Microarray - expression determination

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

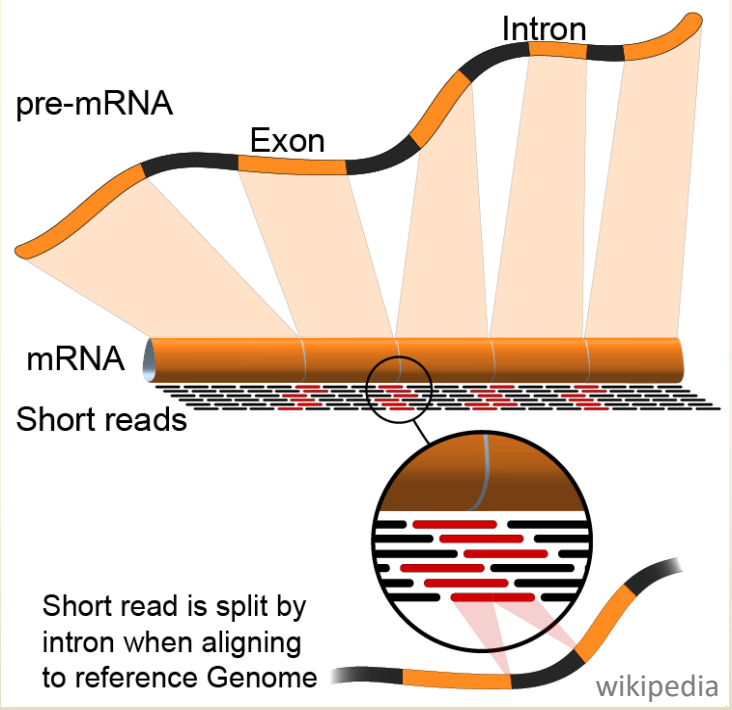


Next generation transcriptomics

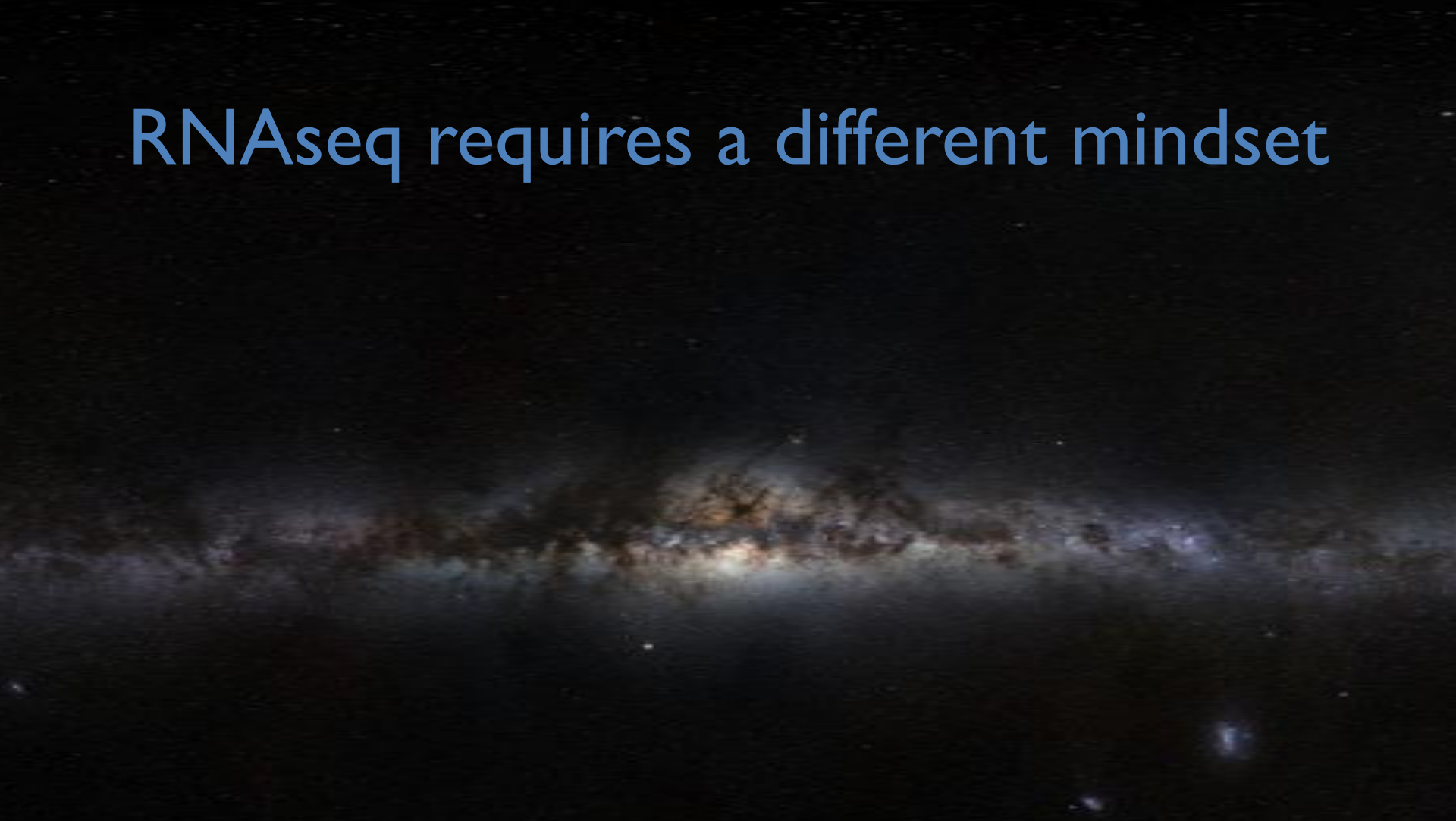
Next generation transcriptomics

RNA sequencing

- 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



RNAseq requires a different mindset

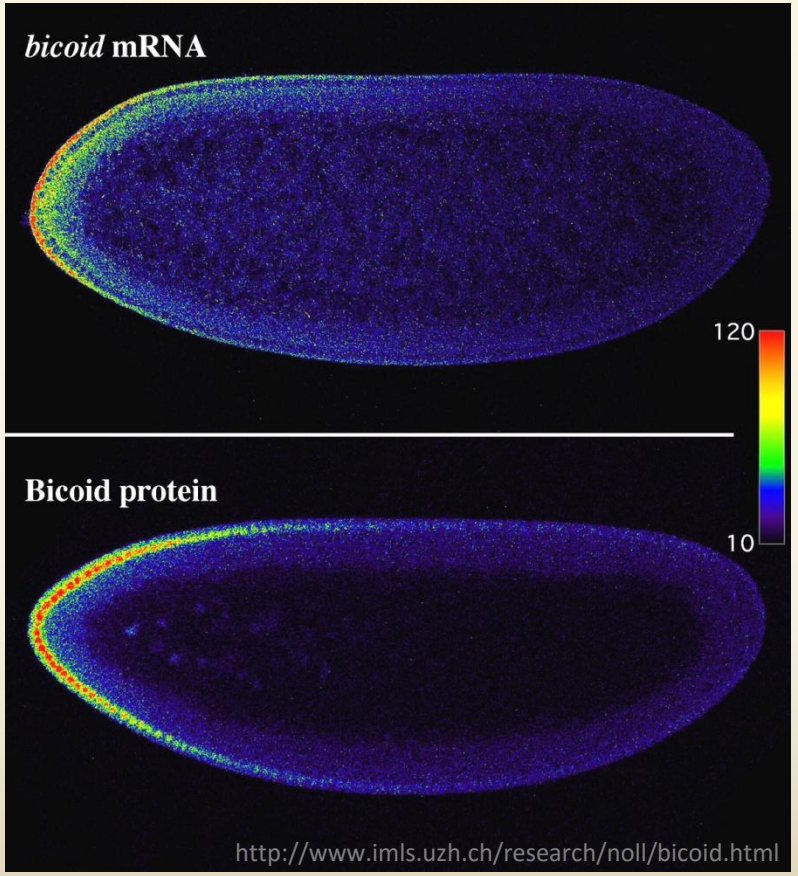


It's like watching a picture of the milky way when you are used to watching a picture of the sun…

We tend to think…

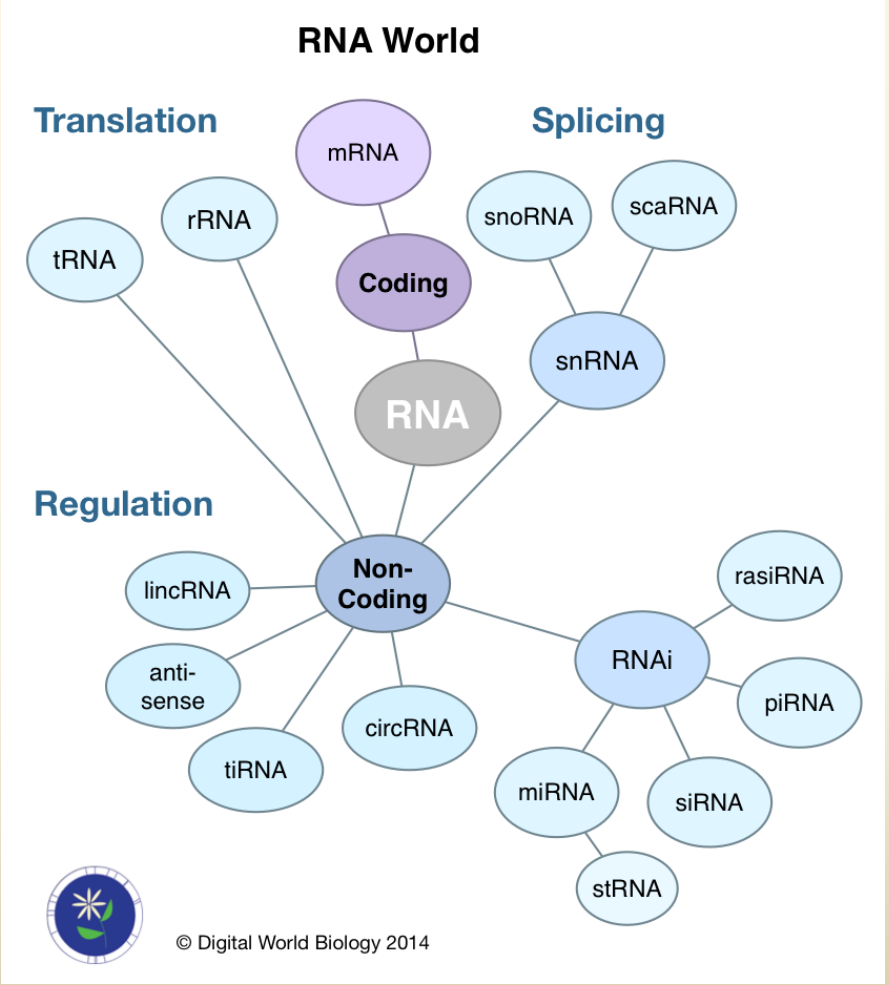
- Transcriptome = mRNA
- mRNA = Protein
- Protein = Biological relevance

- Things are seldom as simple as clear cut…



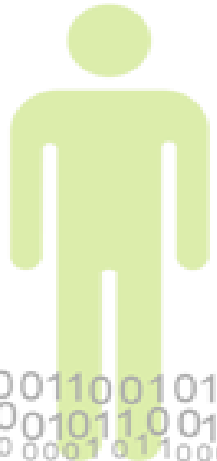
Before interpreting function

- Remember:
  - RNA decay
  - RNA editing
  - RNA splicing
  - Translation regulation
  - RNA interference
  - …



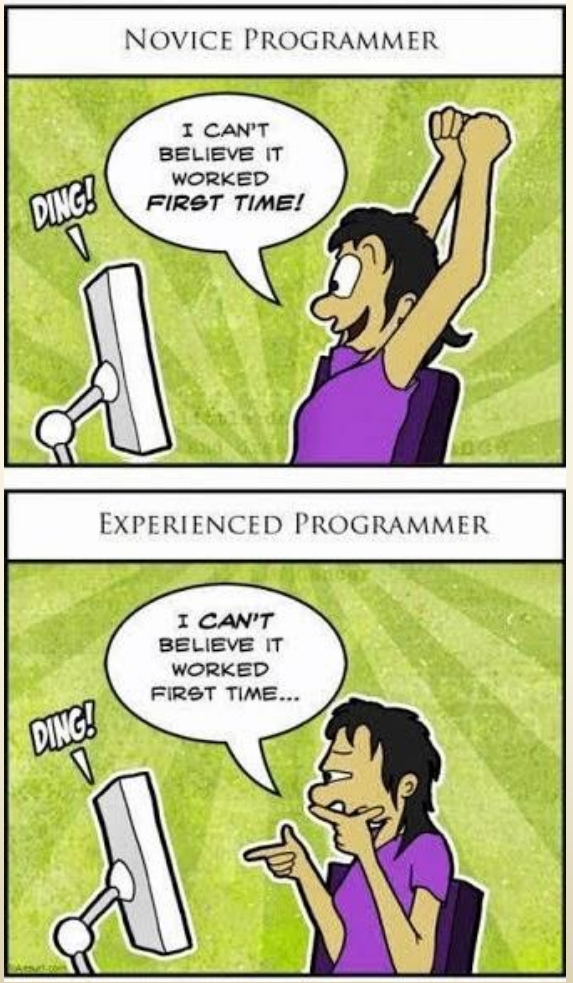
Cons

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



Pros

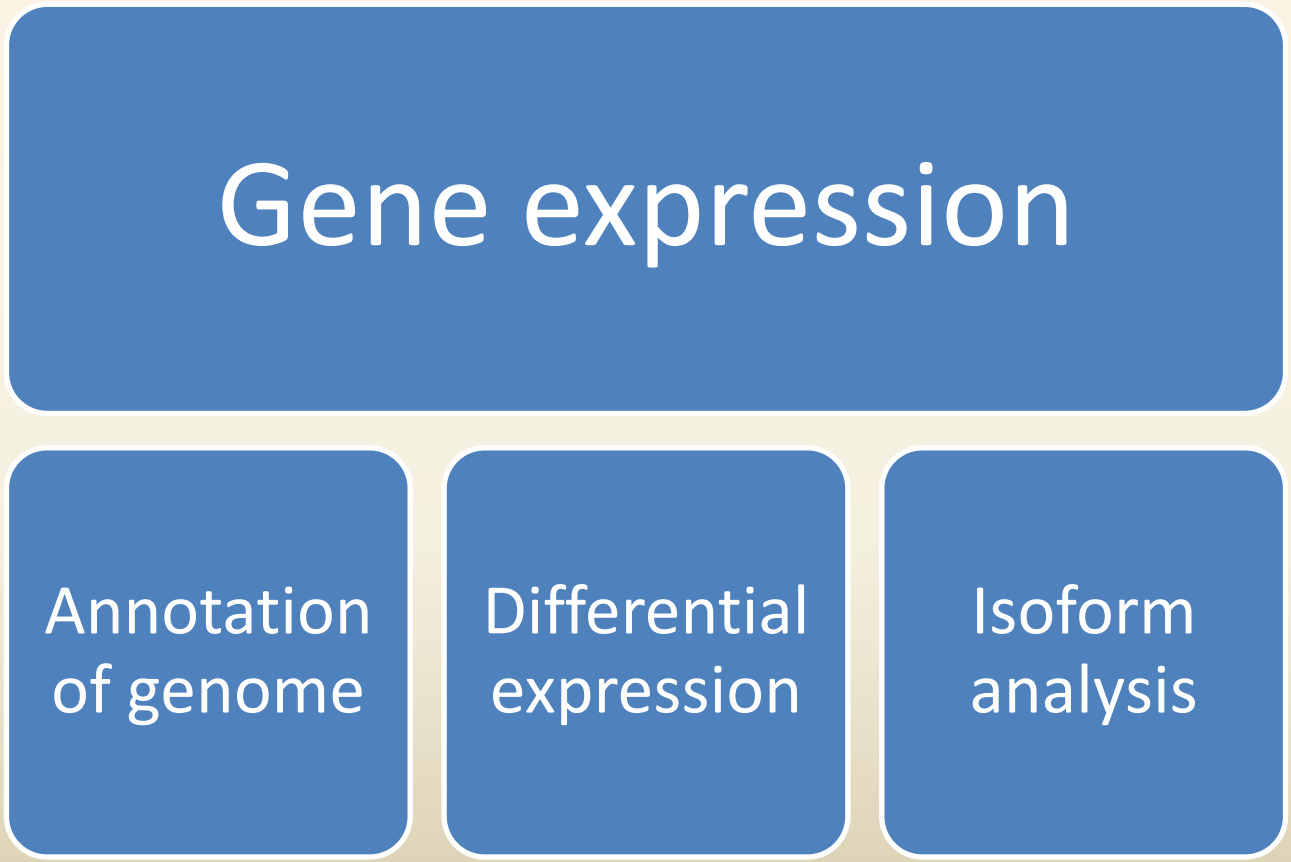
- Others have traversed the path you now set upon
- There are pipelines to help you manage the data
- Careful design will highlight your hypothesis beautifully
- The data you possess when you are finished are really cool and a great stepping stone for functional experiment

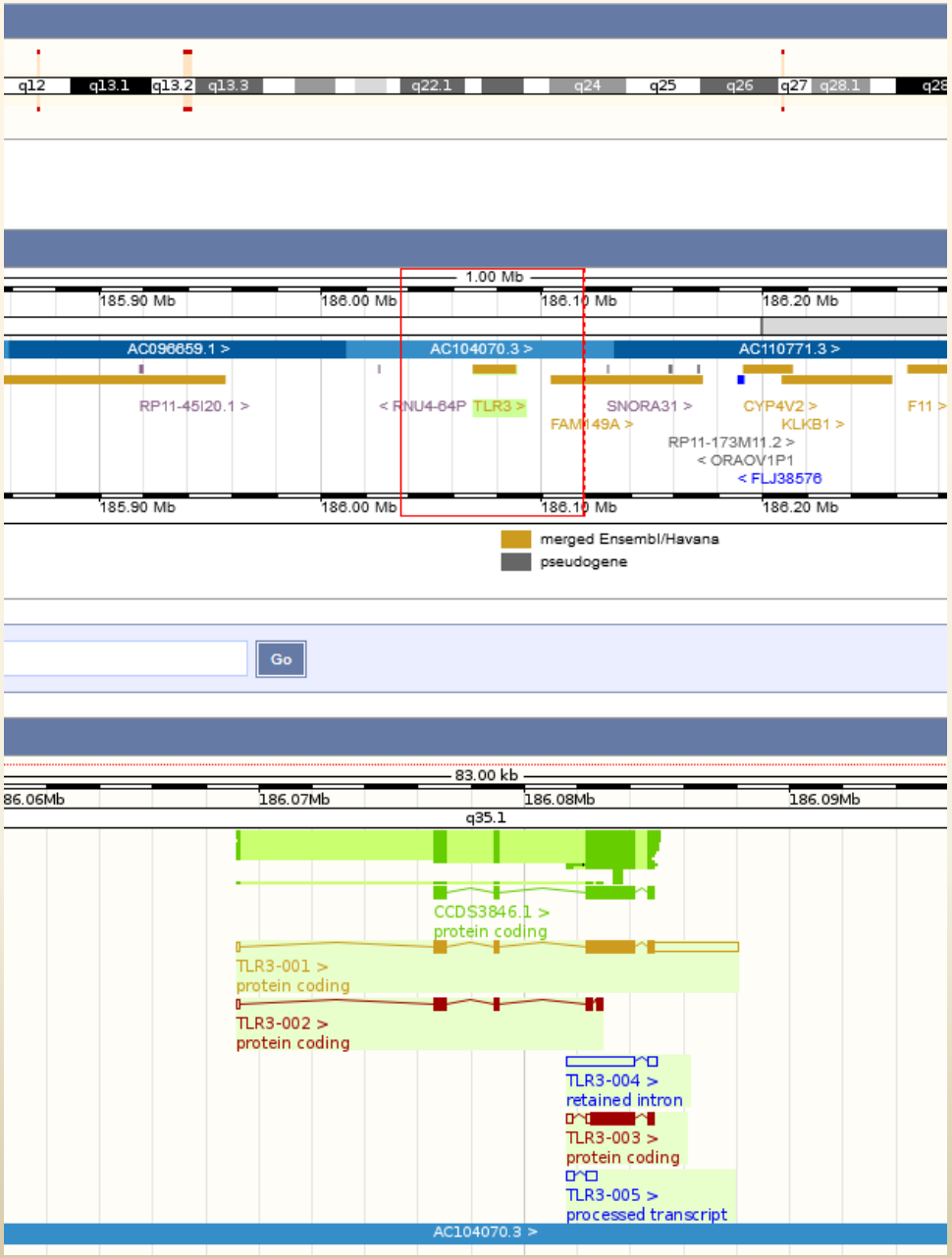


http://erika-vilches.com/

Uses of RNA data

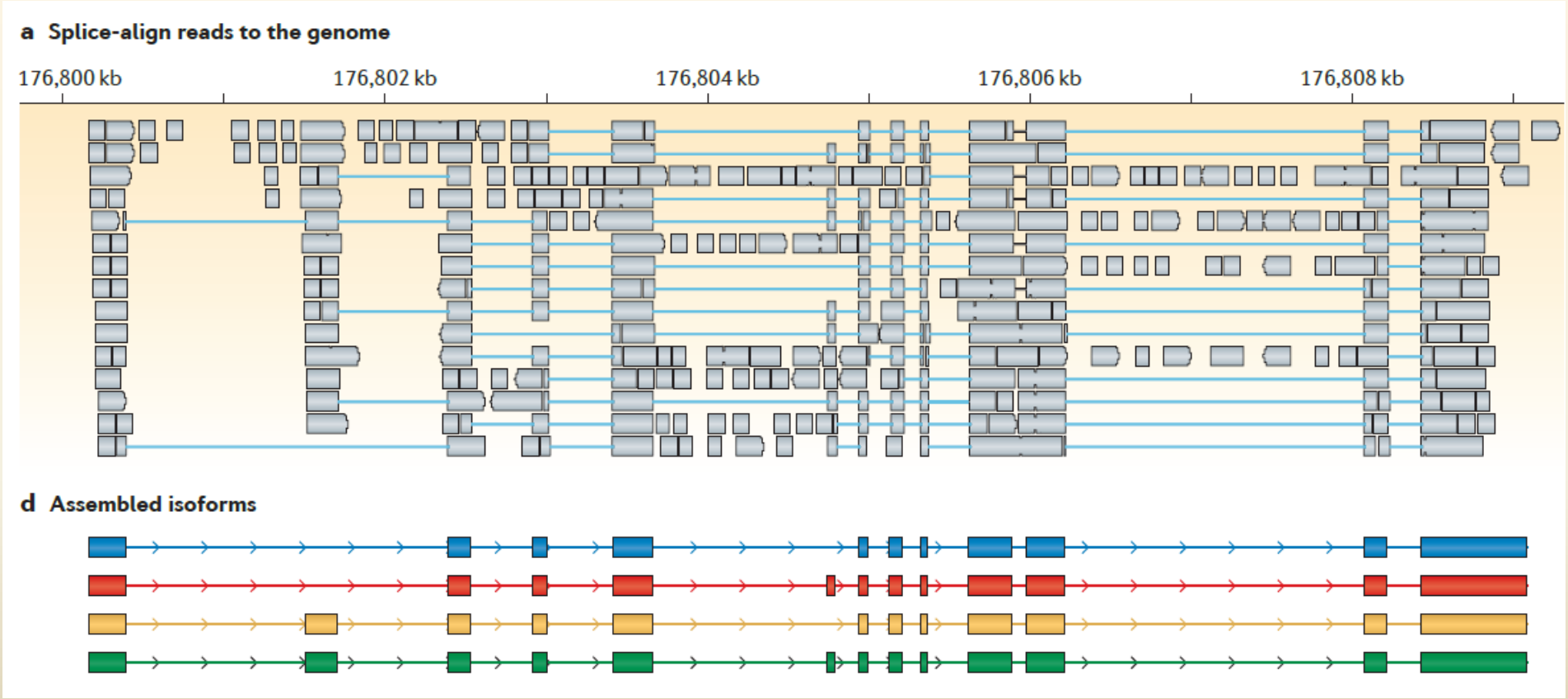
Uses of RNAseq





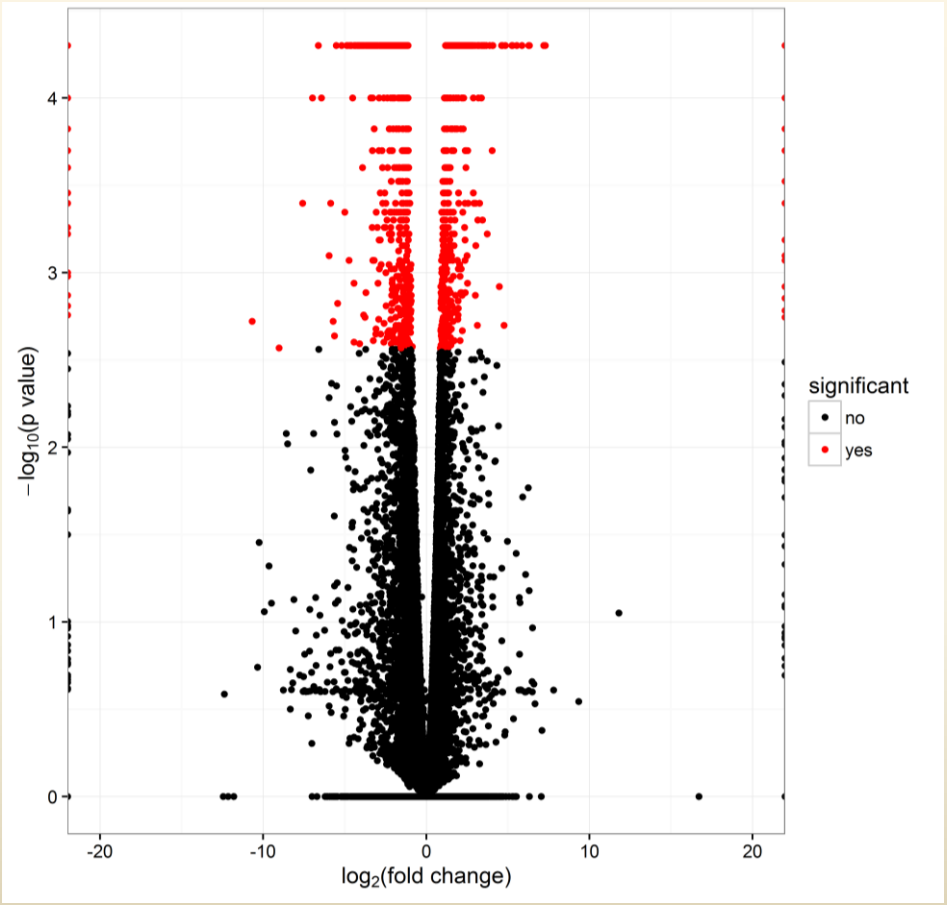
Uses of RNAseq

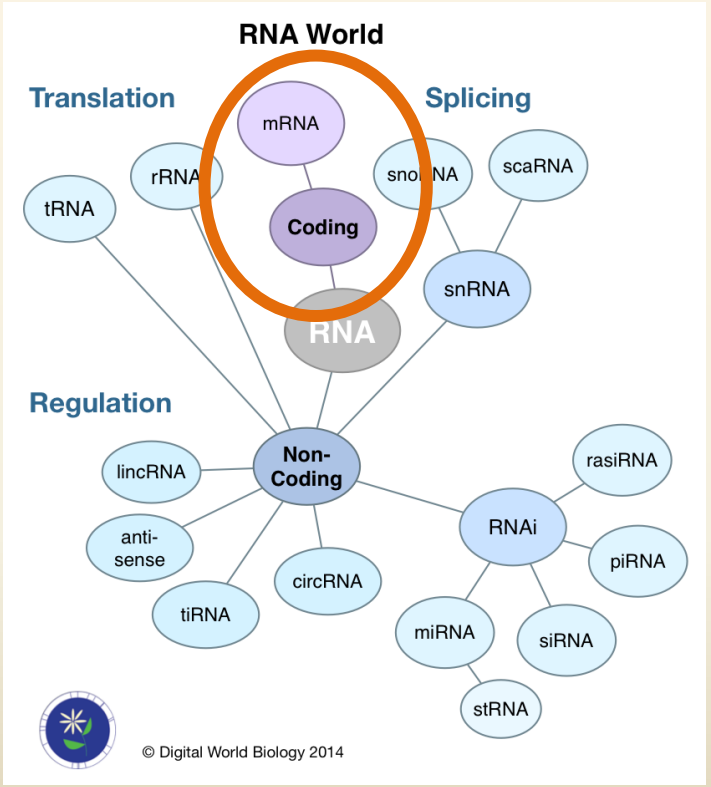
Uses of RNAseq



http://www.nature.com/nrg/journal/v12/n10/full/nrg3068.html

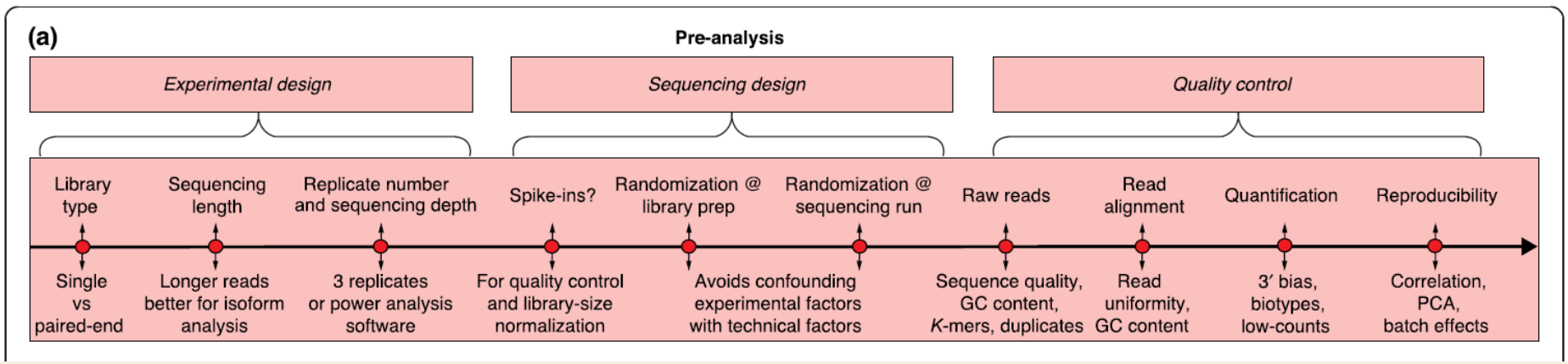
Uses of RNAseq





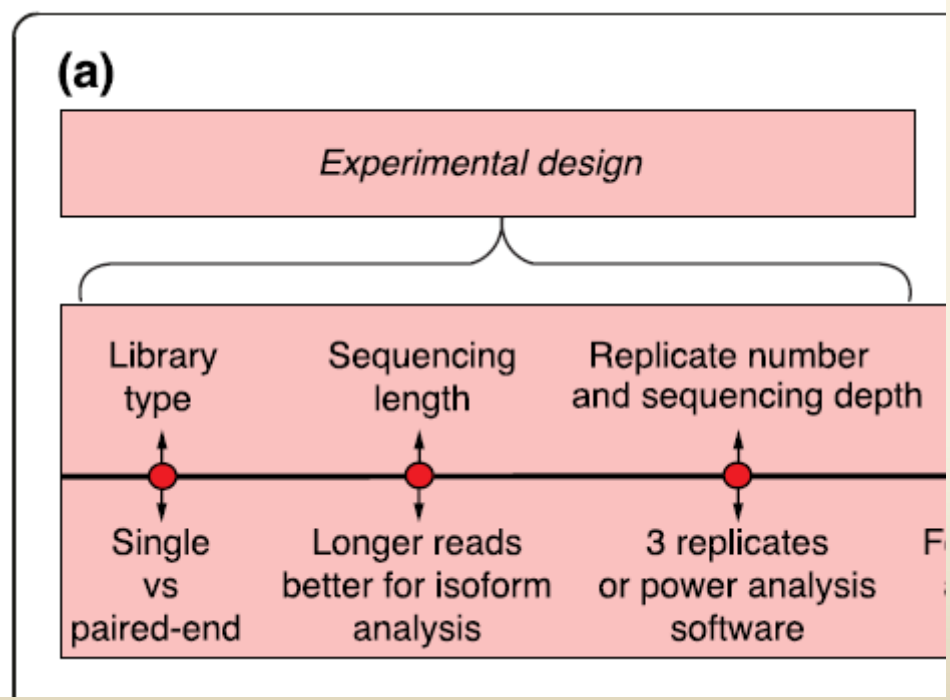
Overall RNAseq pipeline

Where to start



https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0881-8

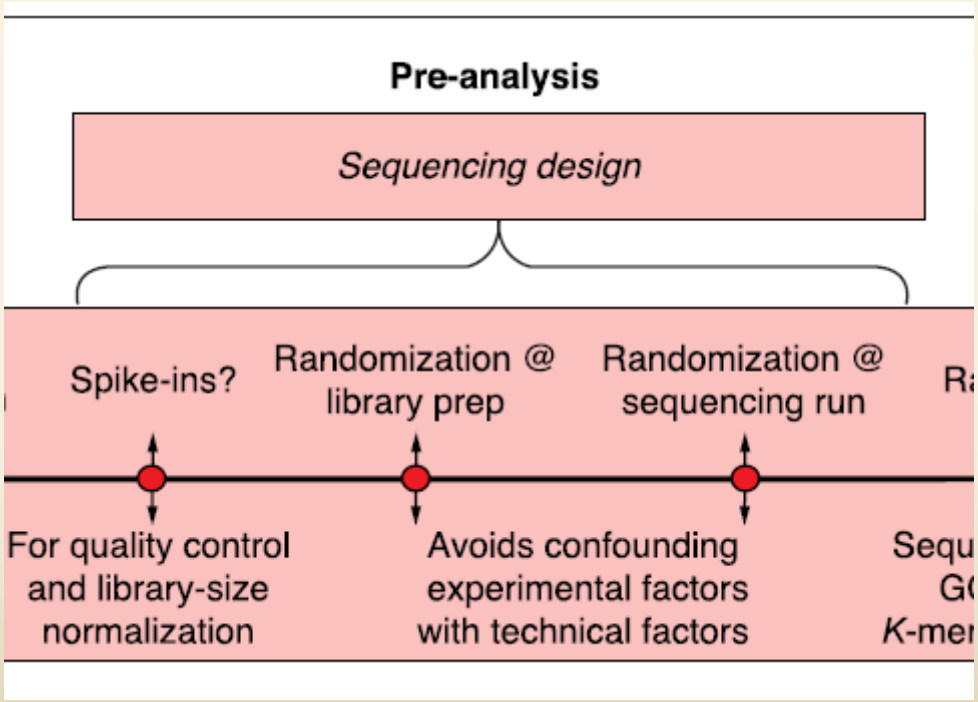
Where to start



Very important - You cannot answer your biological question if the design is off!

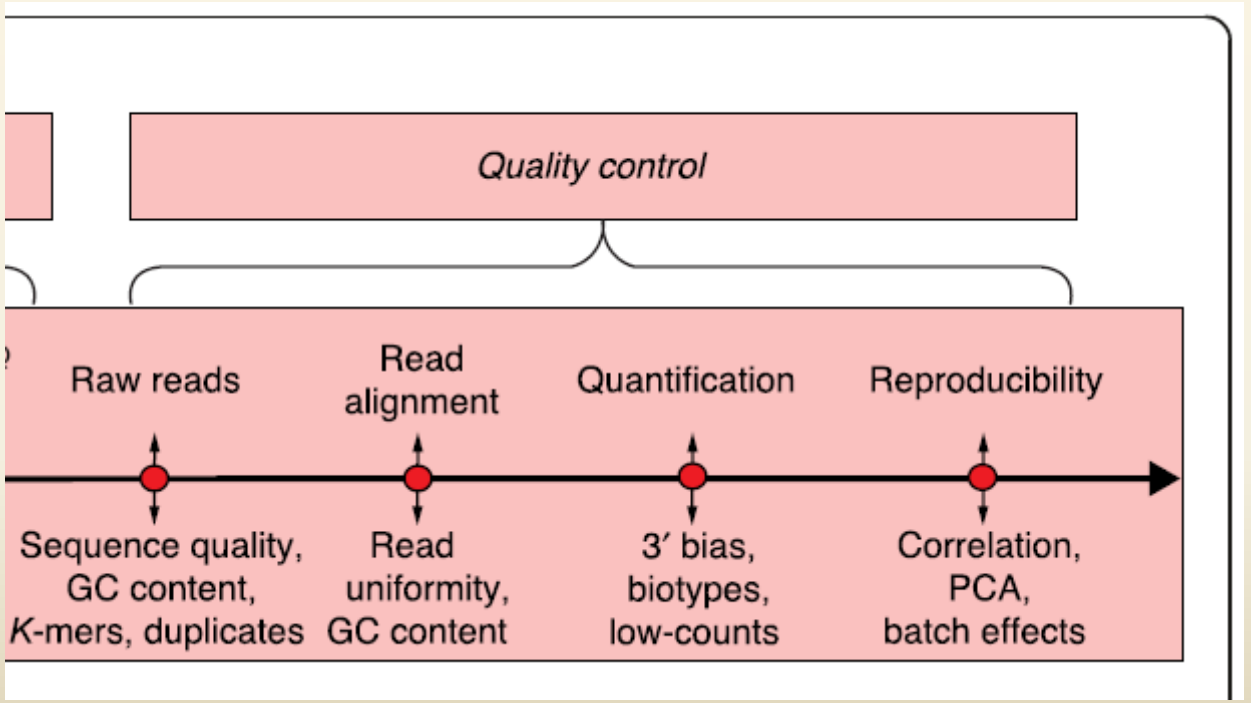
https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0881-8

Where to start



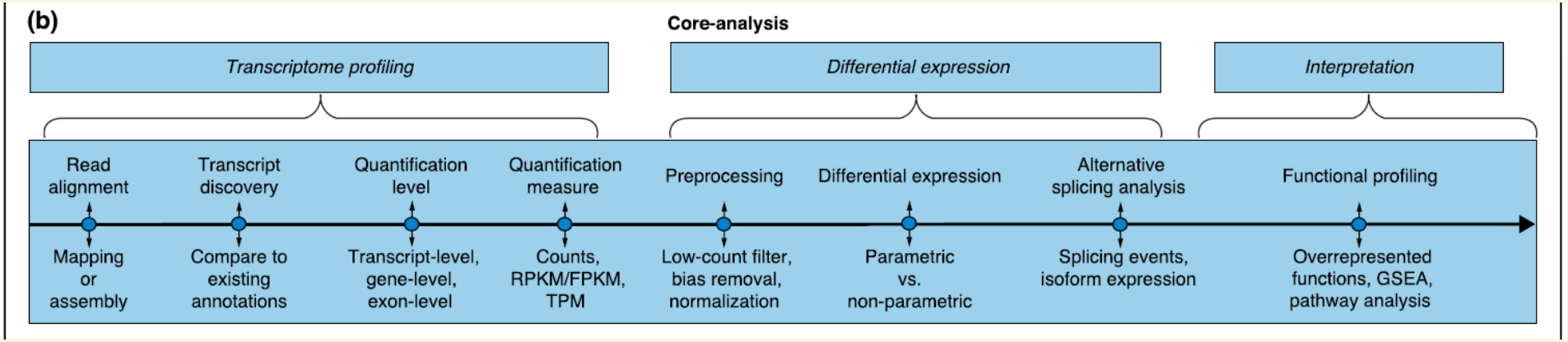
https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0881-8

Where to start



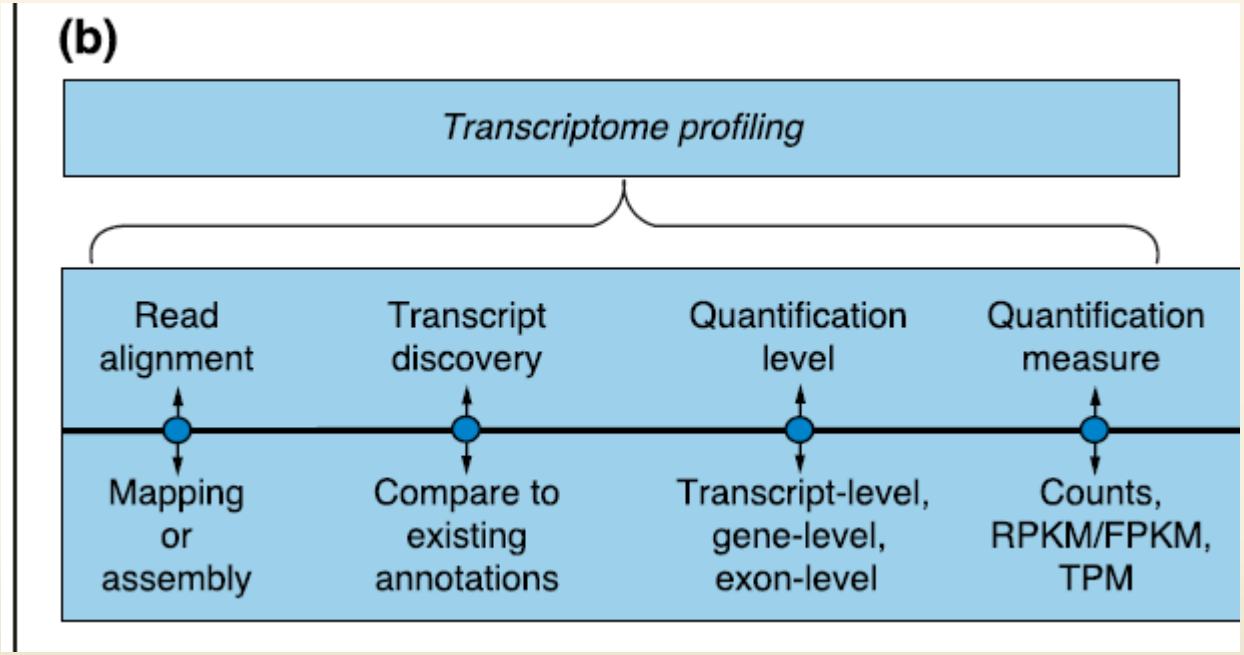
https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0881-8

How to continue



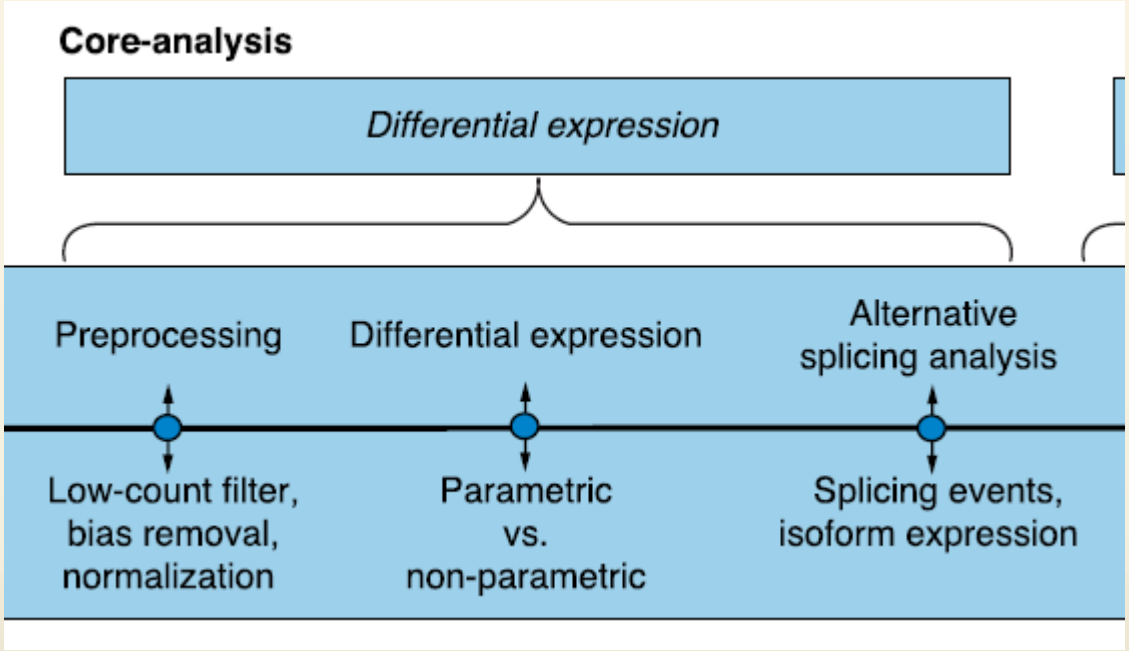
https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0881-8

How to continue



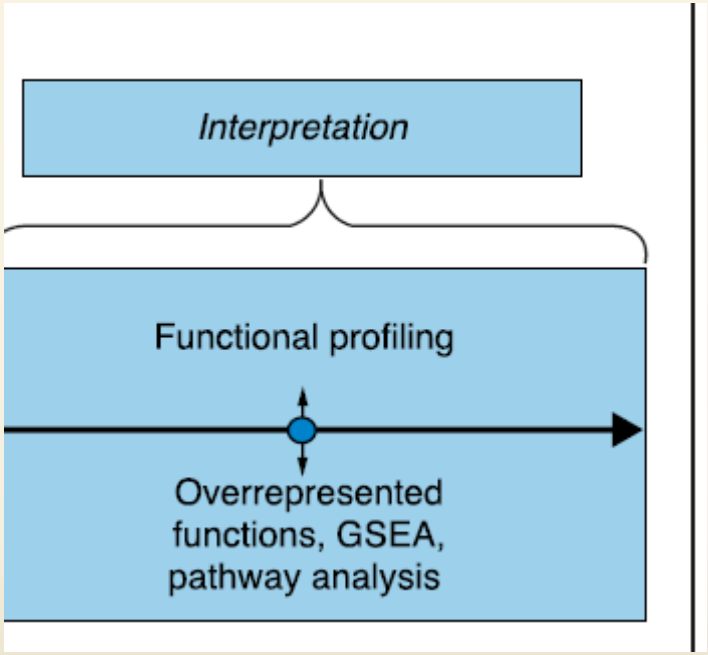
https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0881-8

How to continue



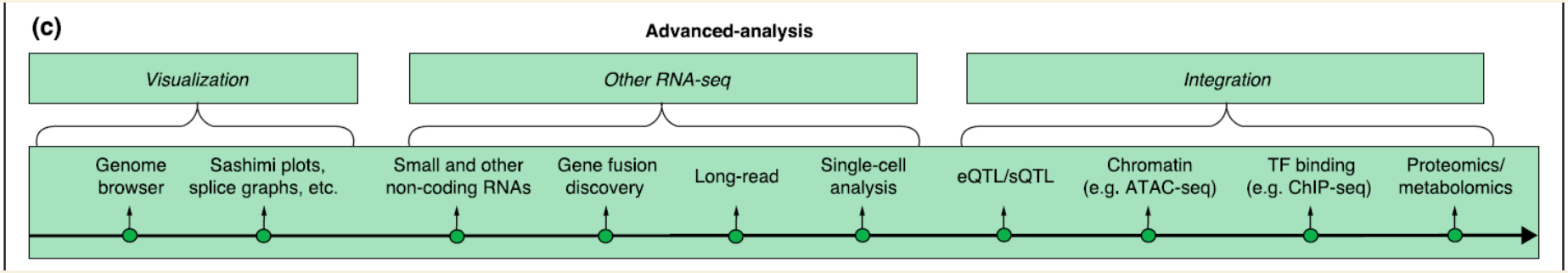
https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0881-8

How to continue



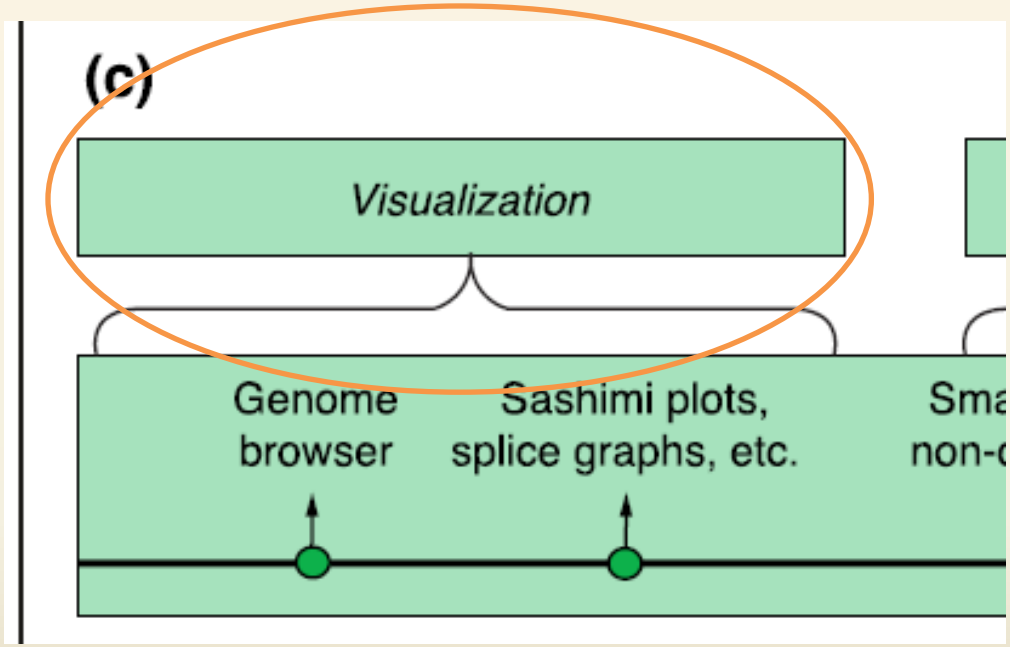
https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0881-8

Downstream options



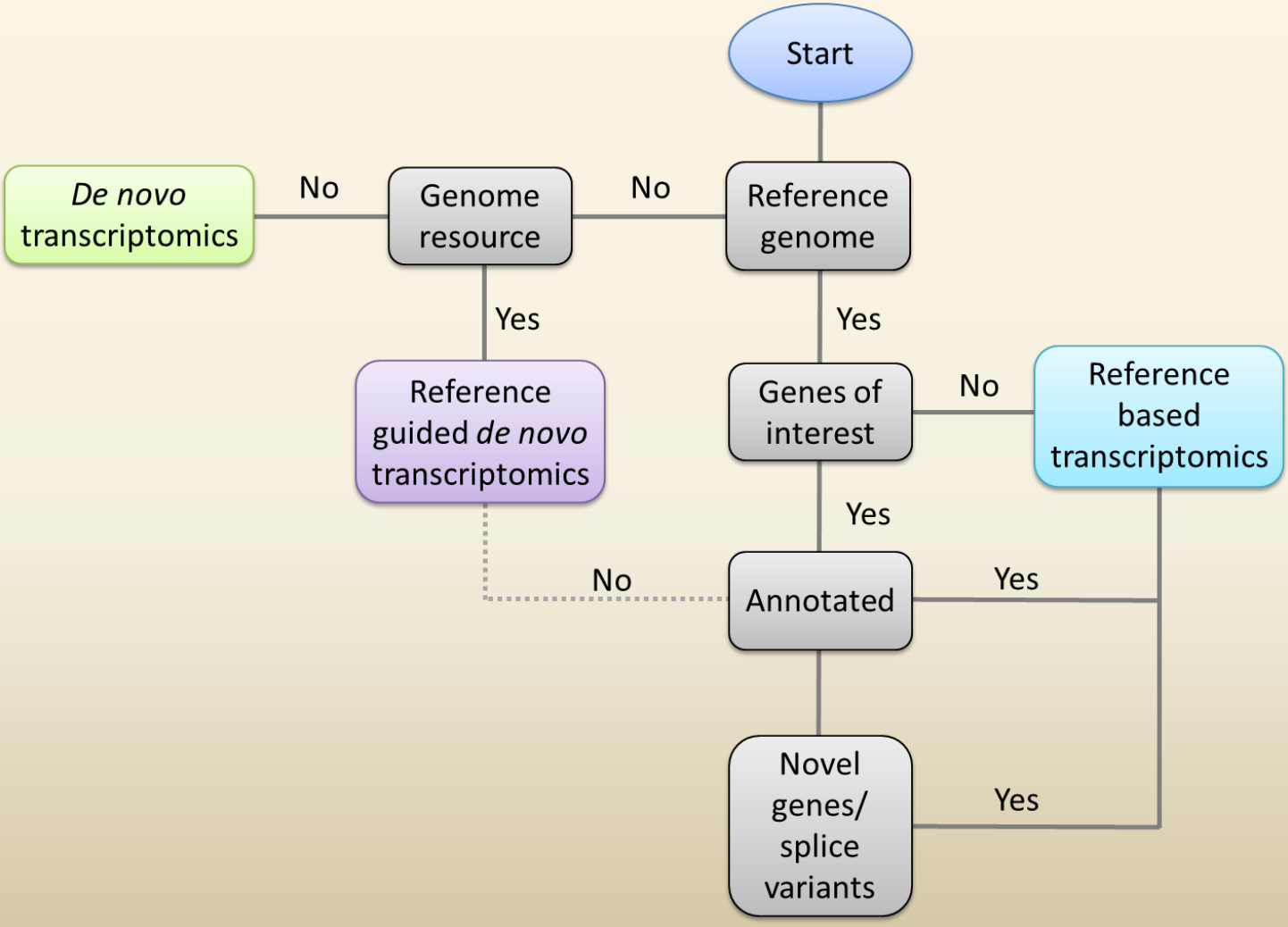
https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0881-8

Downstream options

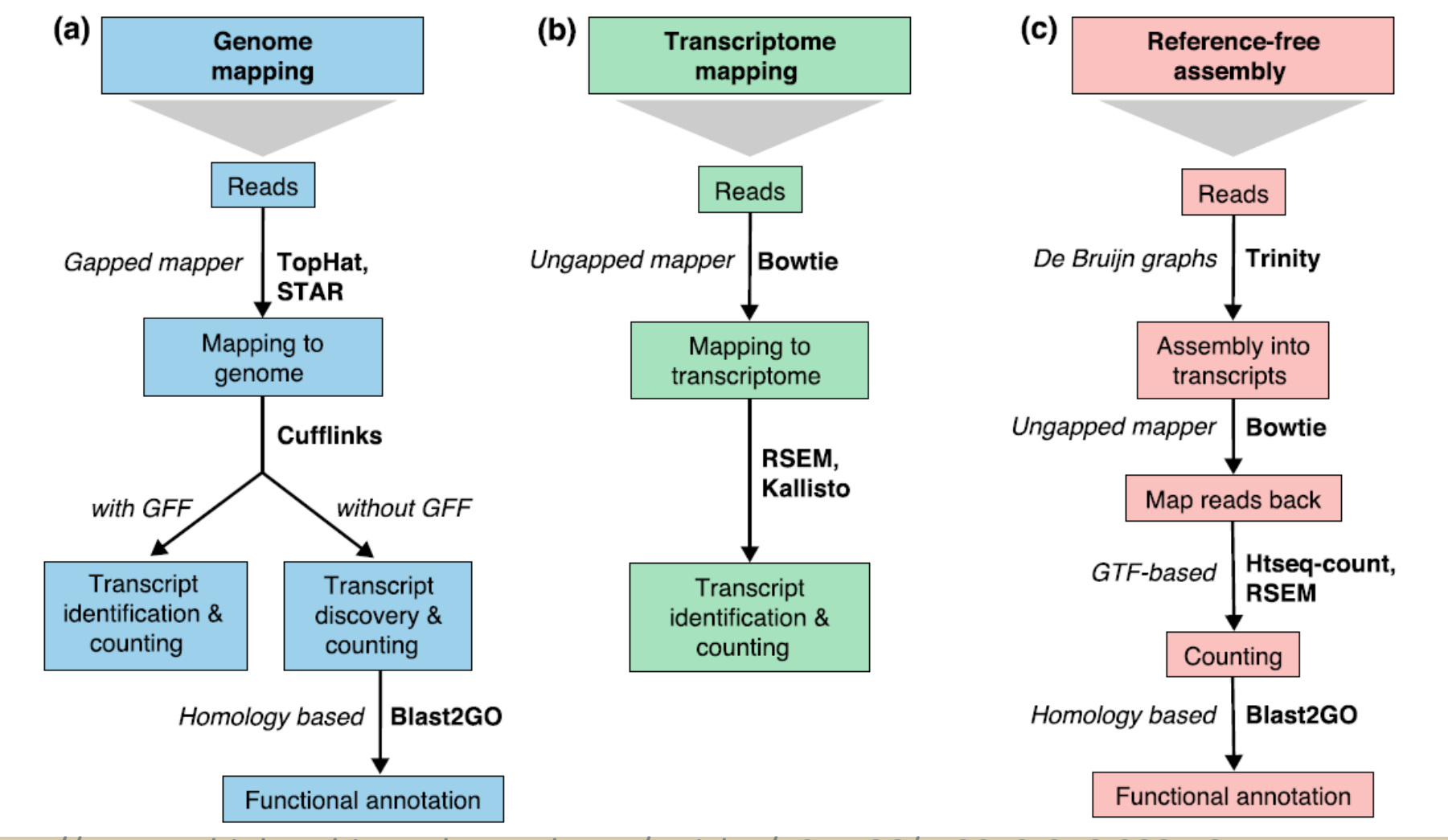


https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0881-8

Three ways to «Rome» – Differential gene expression

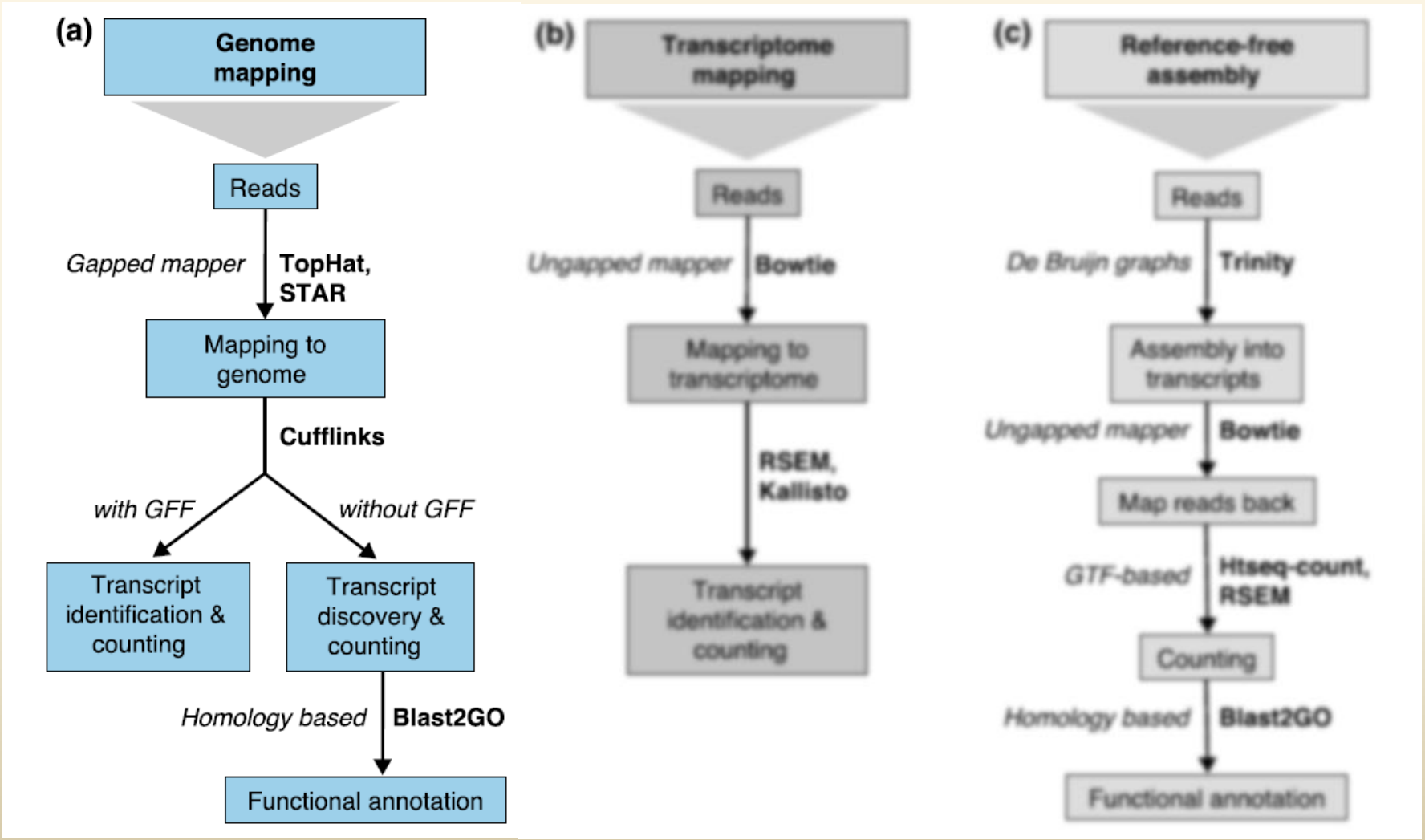


Three ways to «Rome» – Differential gene expression



https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0881-8

Three ways to «Rome» – Differential gene expression

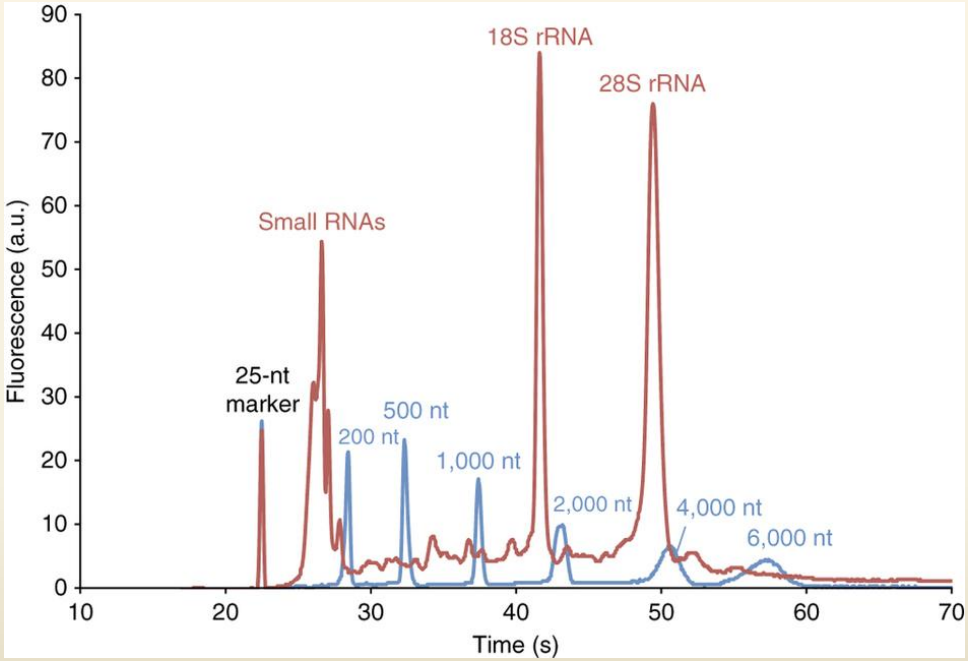


https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0881-8

Sample preparation

RNA isolation

- Aim for high quality RNA with good integrity and concentration
- Column based isolation loses all small RNAs
- Chloroform left-overs may interfere with sequencing reaction



Library preparation

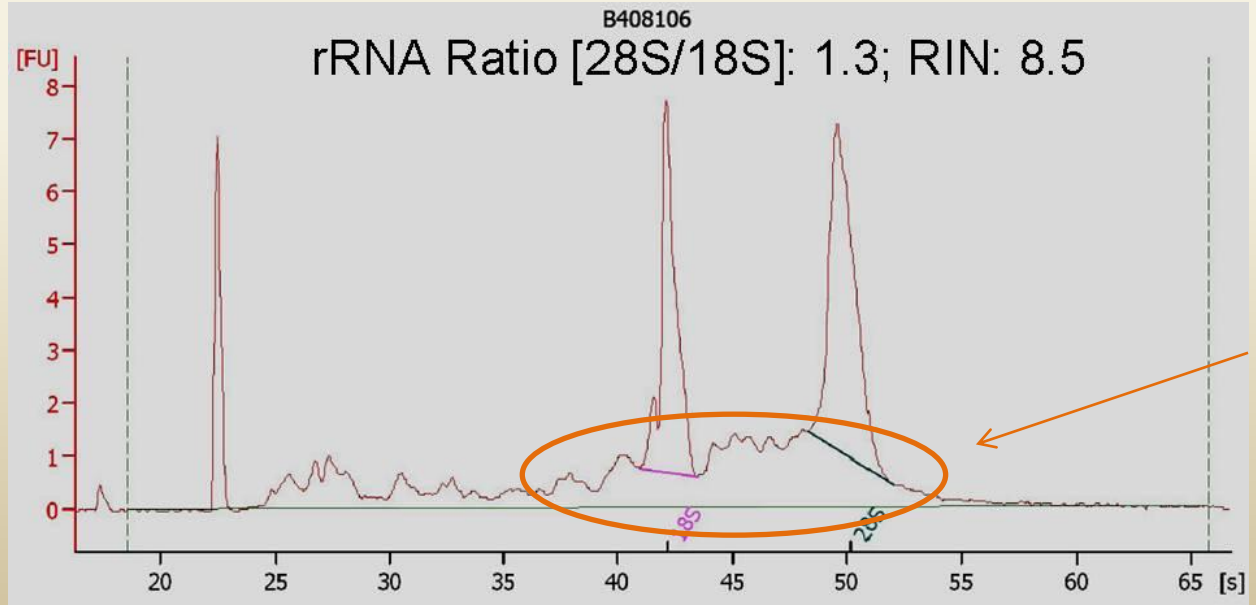
- TotalRNA works for most applications
- Depending on sequencing instrument physical or enzymatic shearing might be needed (affects needed RNA input amount)
- More than 24 samples – consider robot preparation

Library preparation II

- Depending on focus you may perform:
  - rRNA depletion
  - mRNA selection
  - Abundant transcript removal
  - smallRNA conservation
  - Skip library amplification
  - Strand specific library preparation

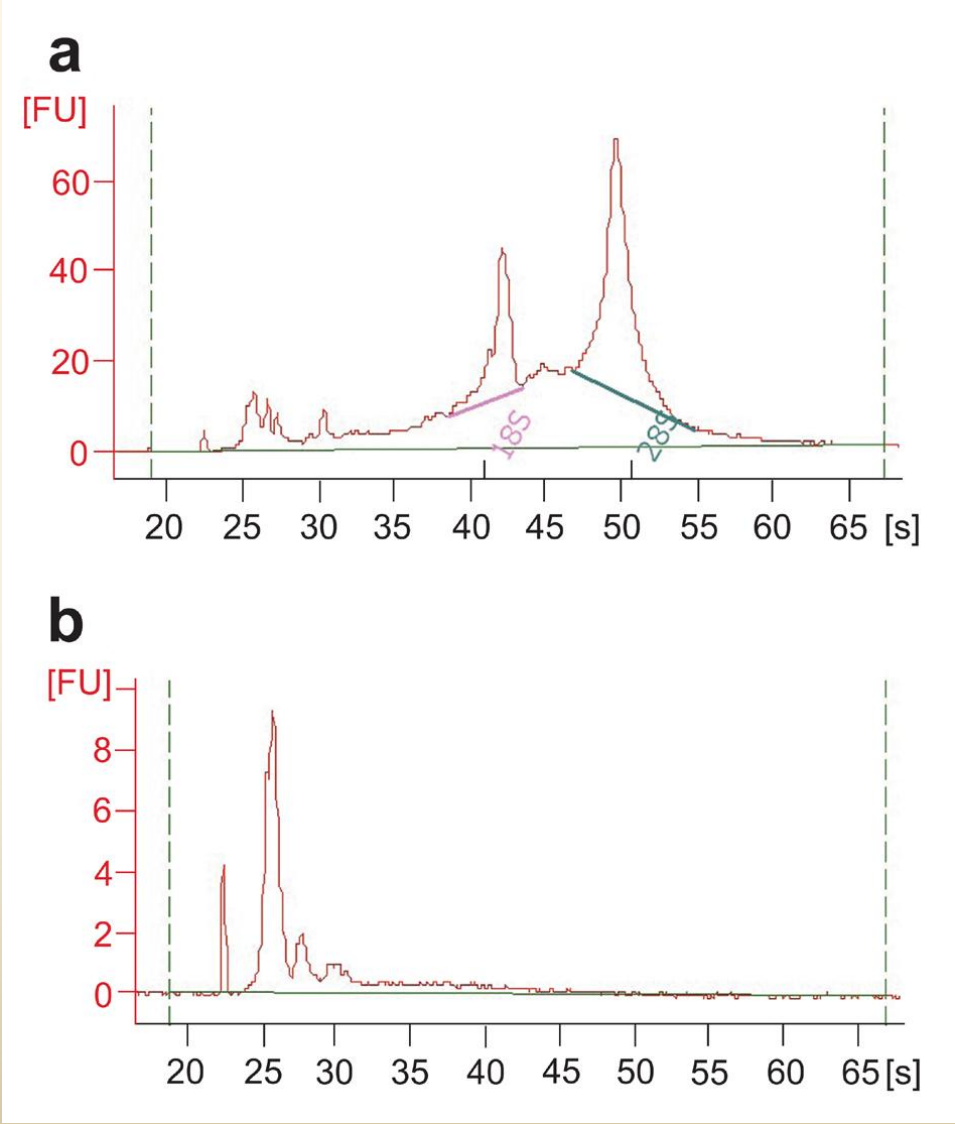
RNA enrichment / depletion

- Low conc. input RNA: deplete rRNA
- High conc. input RNA: enrich mRNA
- No polyA tail: deplete rRNA
- Enrich small RNA



mRNA

RNA enrichment / depletion

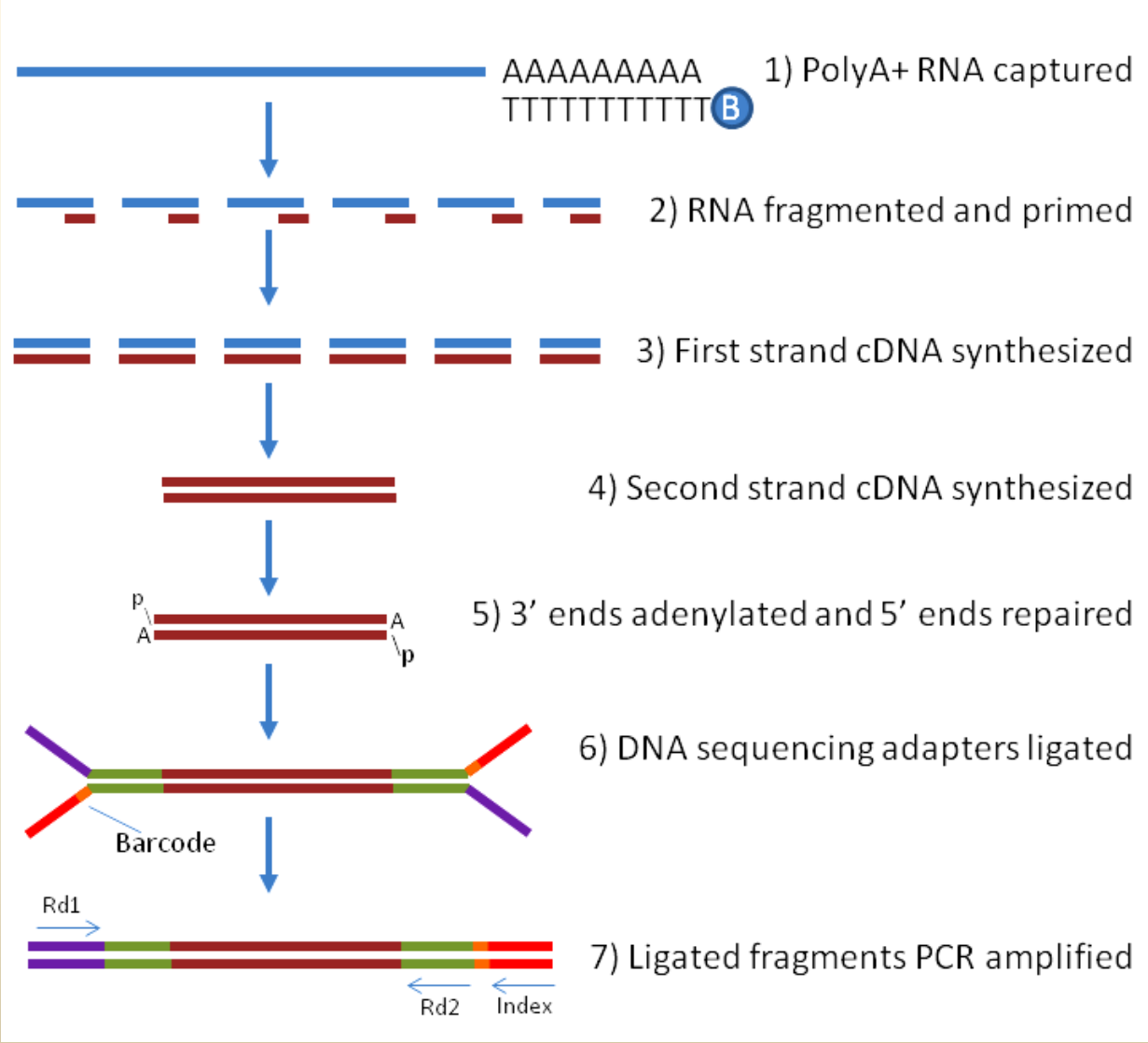


Ribo-Zero Gold Kit: improved RNA-seq results after removal of cytoplasmic and mitochondrial ribosomal RNA

Vladimir Benes, Jonathon Blake & Ken Doyle

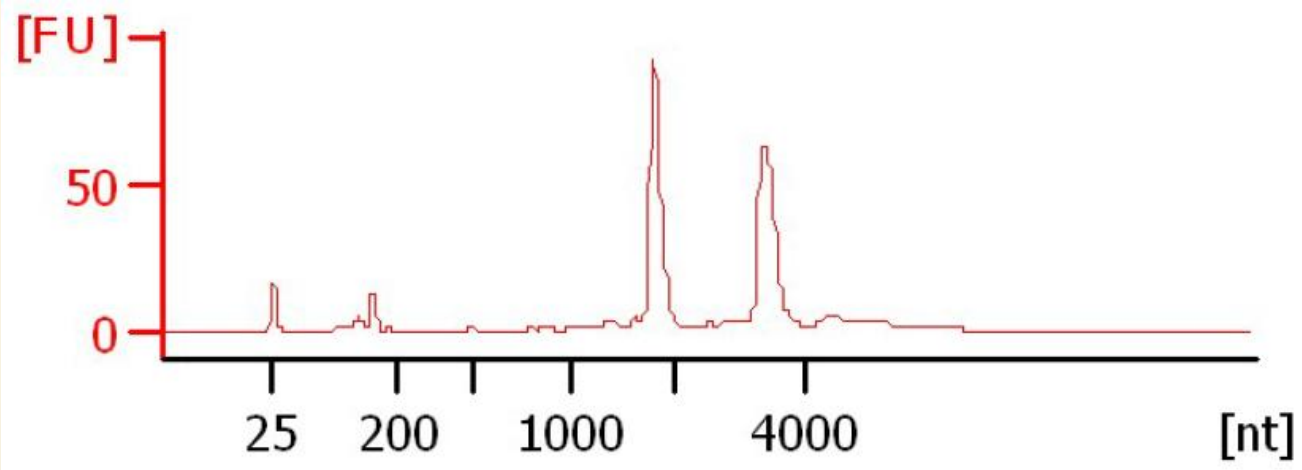
Nature Methods 8 (2011)

Library preparation – mRNA Illumina

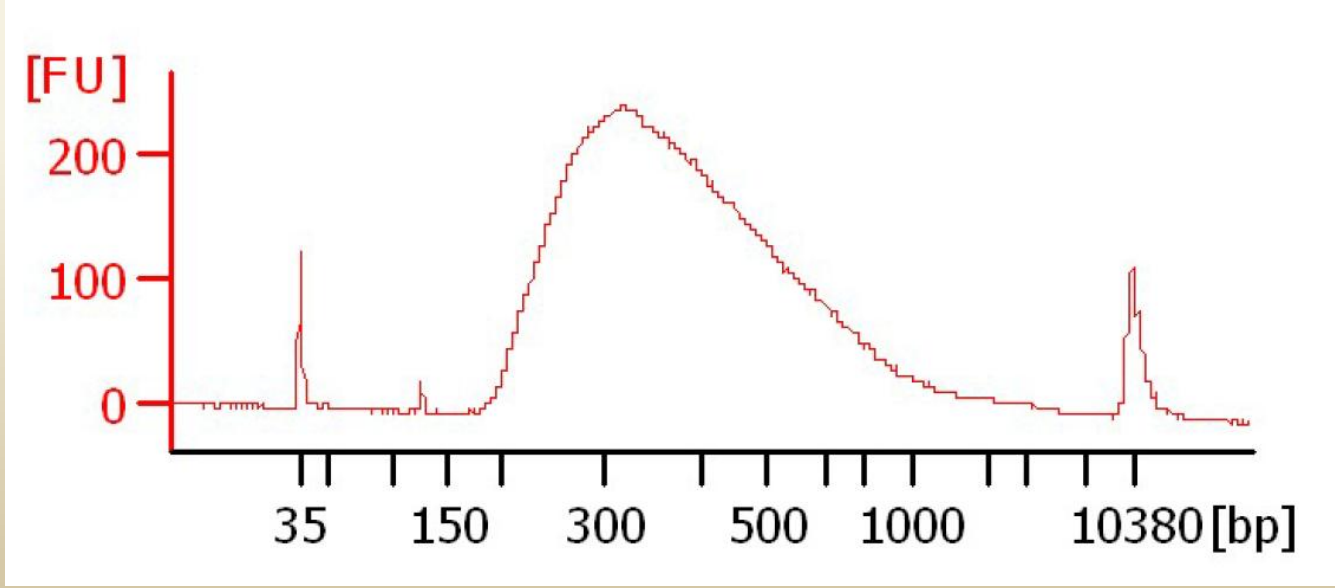


Final library – mRNA Illumina

Gone from this:
totalRNA with ribosomal peaks



To this:
mRNA selected library with ~350 bp fragment size



RNAseq technologies

Choose your sequencing technology

Model species

Differential expression

Isoforms

Read length

Paired-end

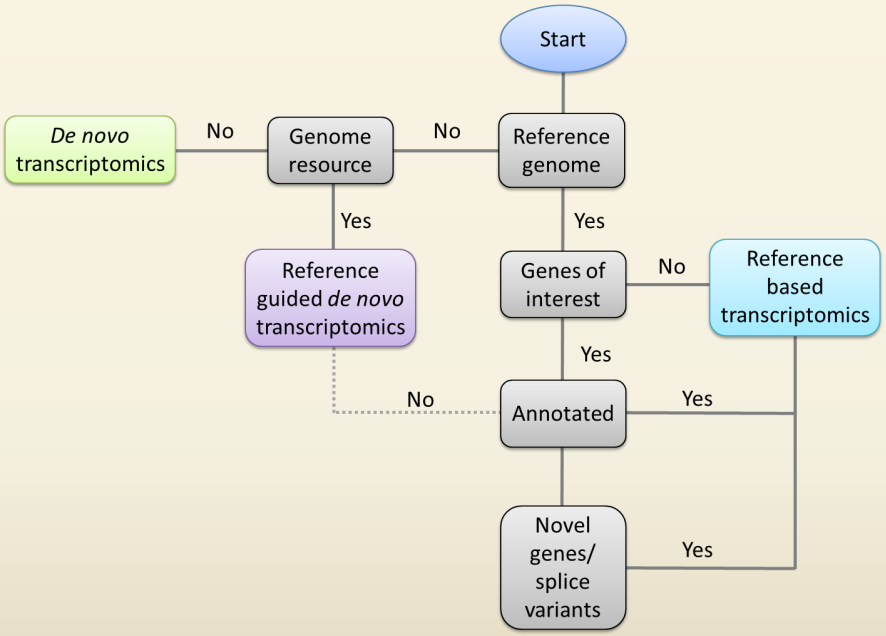
Genome resource

Money

Novel transcripts

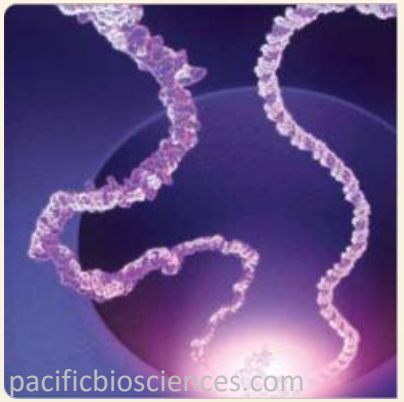
Resources

- Computing power
  - Dedicated computer or access to computing cluster
- Genome resources
  - Reference or draft genome
    - Large toolkit available
- No genome
    - Small toolkit available



PacBio

- Long read sequencing technology
- 16 SMRT Cells
- Sequences entire RNAs up to 10 kb
- Reconstruction of isoforms
- Detection of novel transcripts
- Expression analysis
- Great for reference transcriptomes



PacBio sequencing well

Prof. Kjetill Jakobsen in fron t of the NSC' PacBio

Illumina

- Short read paired-end technology
- 2 flowcells – 8 lanes each
- ~150 bp PE reads
- Reasonable reconstruction of isoforms
- Reasonable detection of novel transcripts
- Expression analysis
- Makes decent reference transcriptomes

www.illumina.com

Illumina HiSeq 4000

Sequencing output

Differential expression

- Model organism
  - Illumina >= 10 mill PE / sample
  - More for rare transcripts
- Non-model organism
  - Illumina >= 20 mill PE / sample
  - More for rare transcripts

Transcriptome assembly

- Depends on species
  - Illumina 100-150 mill PE reads minimum for vertebrates
  - For comparison yeast is sufficient with 4 mill PE stranded reads
  - PacBio vertebrate example:
    - ~25 000 full-length cDNAs
    - SMRT cells 1-2kb, 2-3kb and 3-6 kb

Experimental design