UiO IN-BIOS5000/9000

RNA-seq

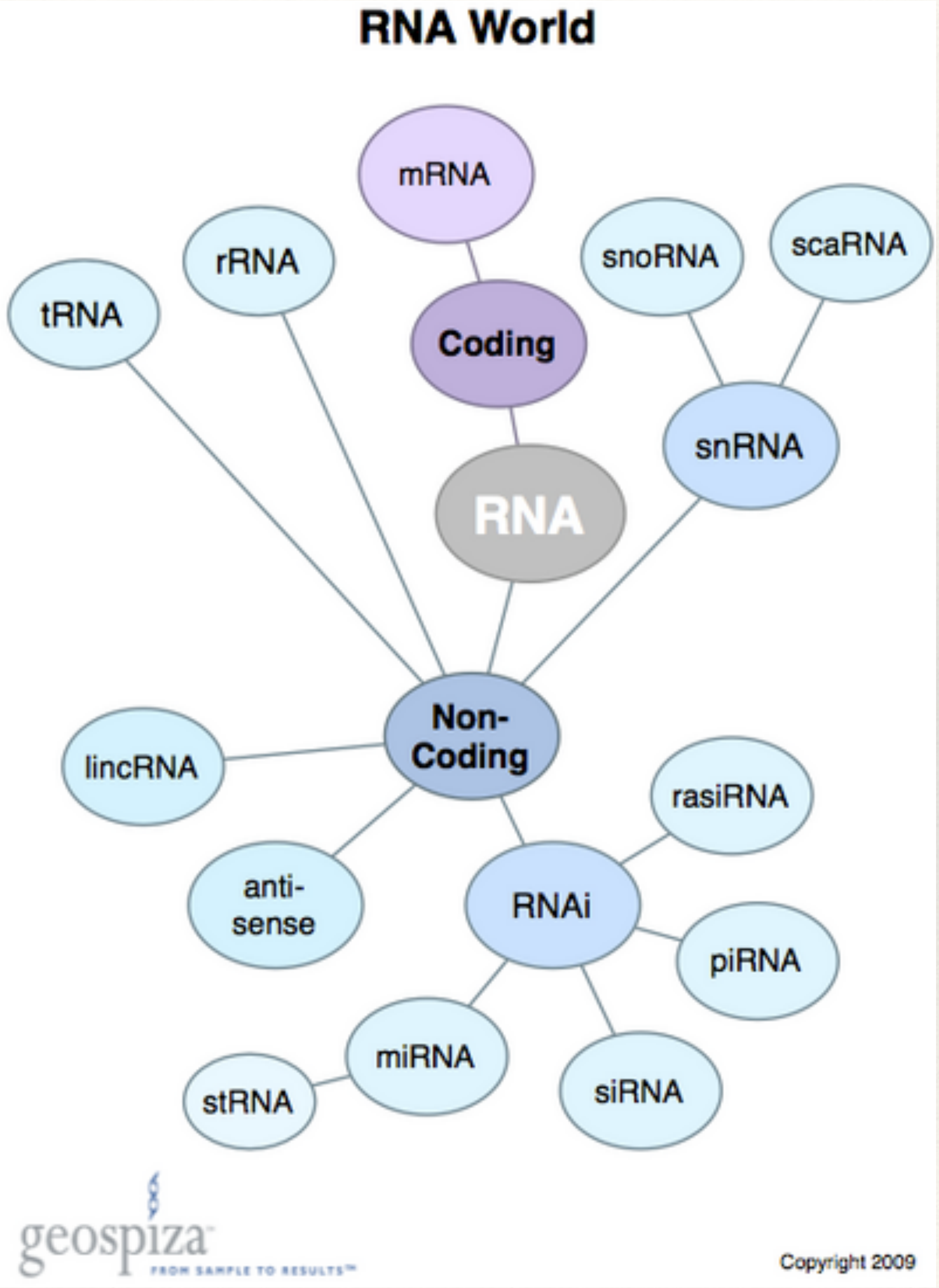
differential expression analysis

Arvind Sundaram
Nov 07, 2024

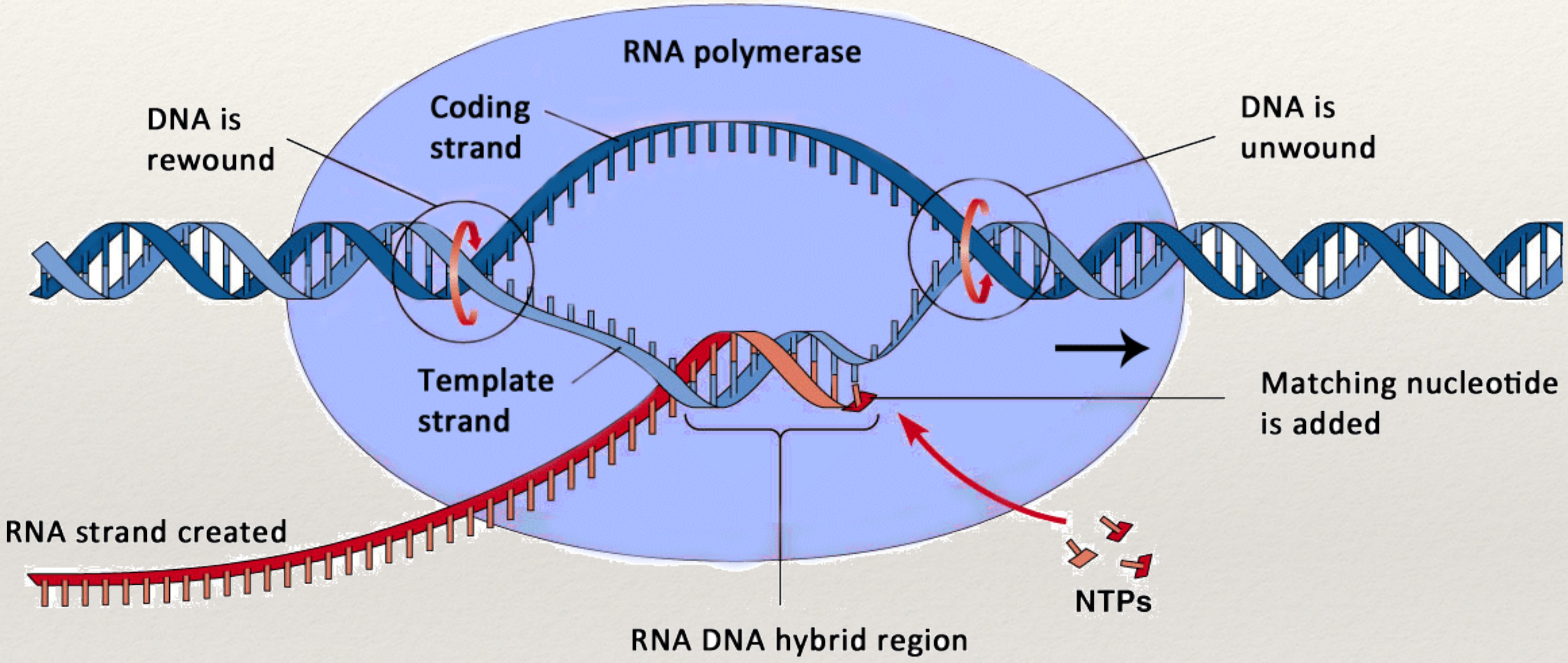
Norwegian Sequencing Centre
OUS, Ullevål, Oslo

Transcriptome

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



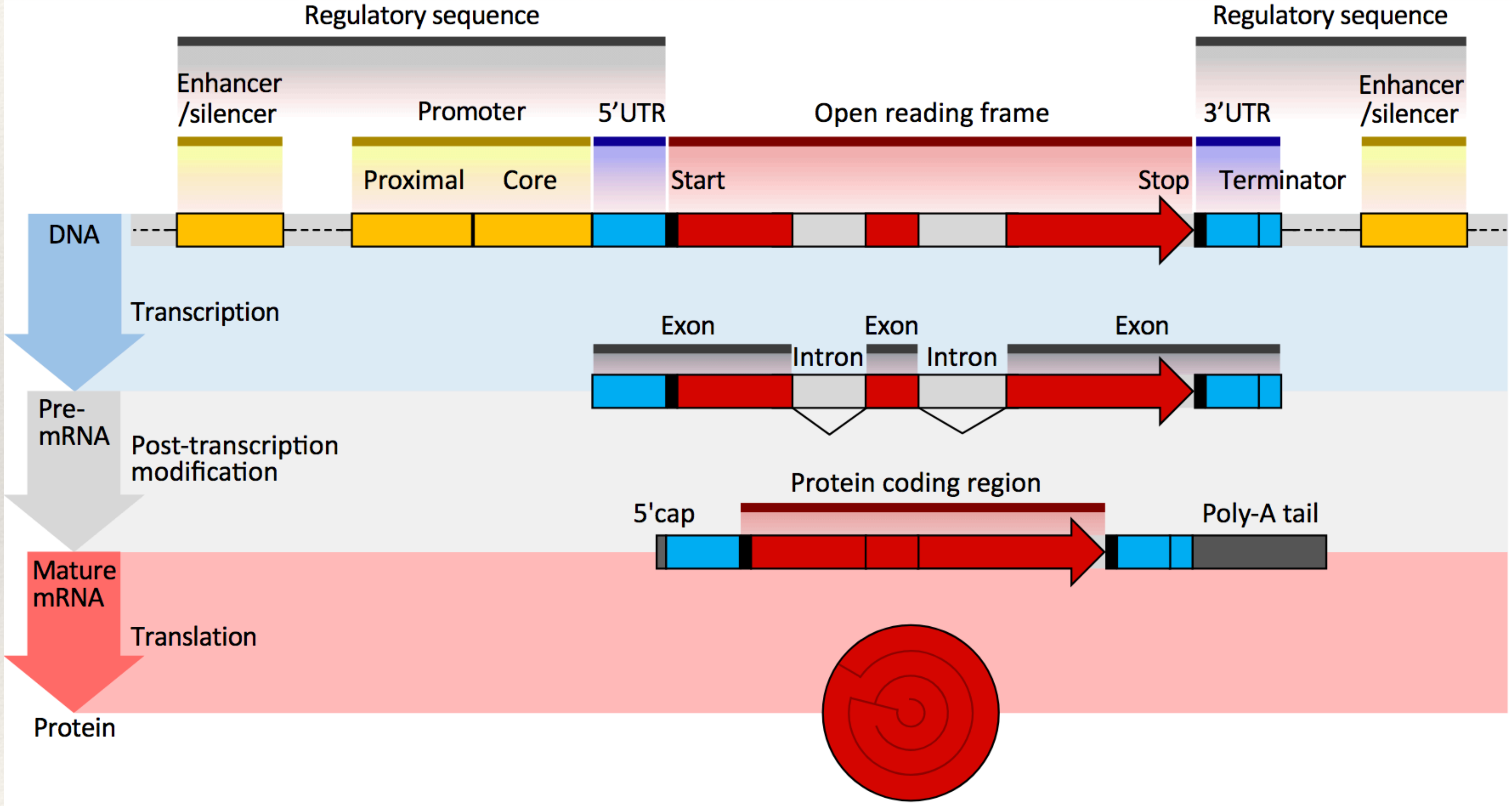
Transcription



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

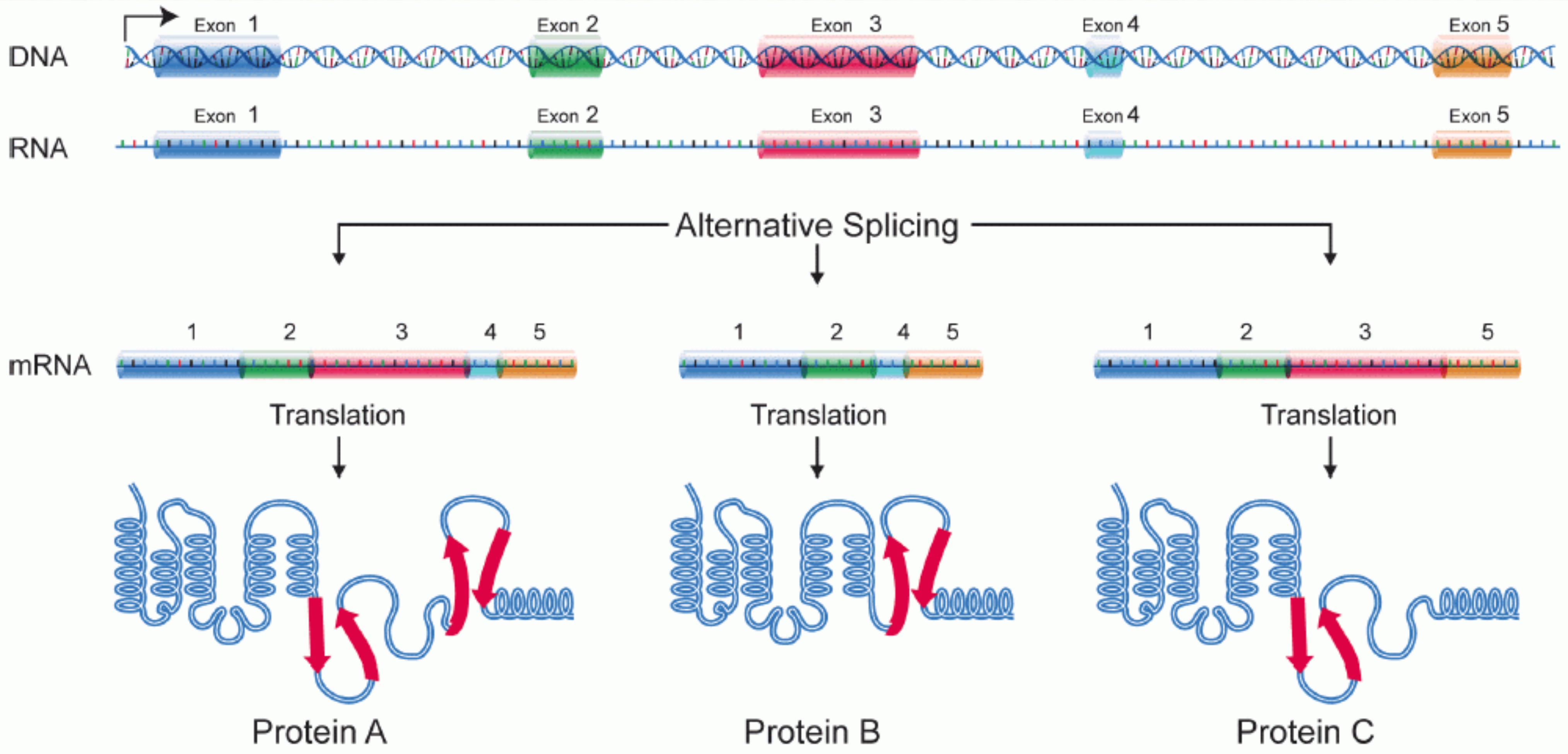
https://www.thinglink.com/scene/618815759912534018

Eukaryotic mRNA processing



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

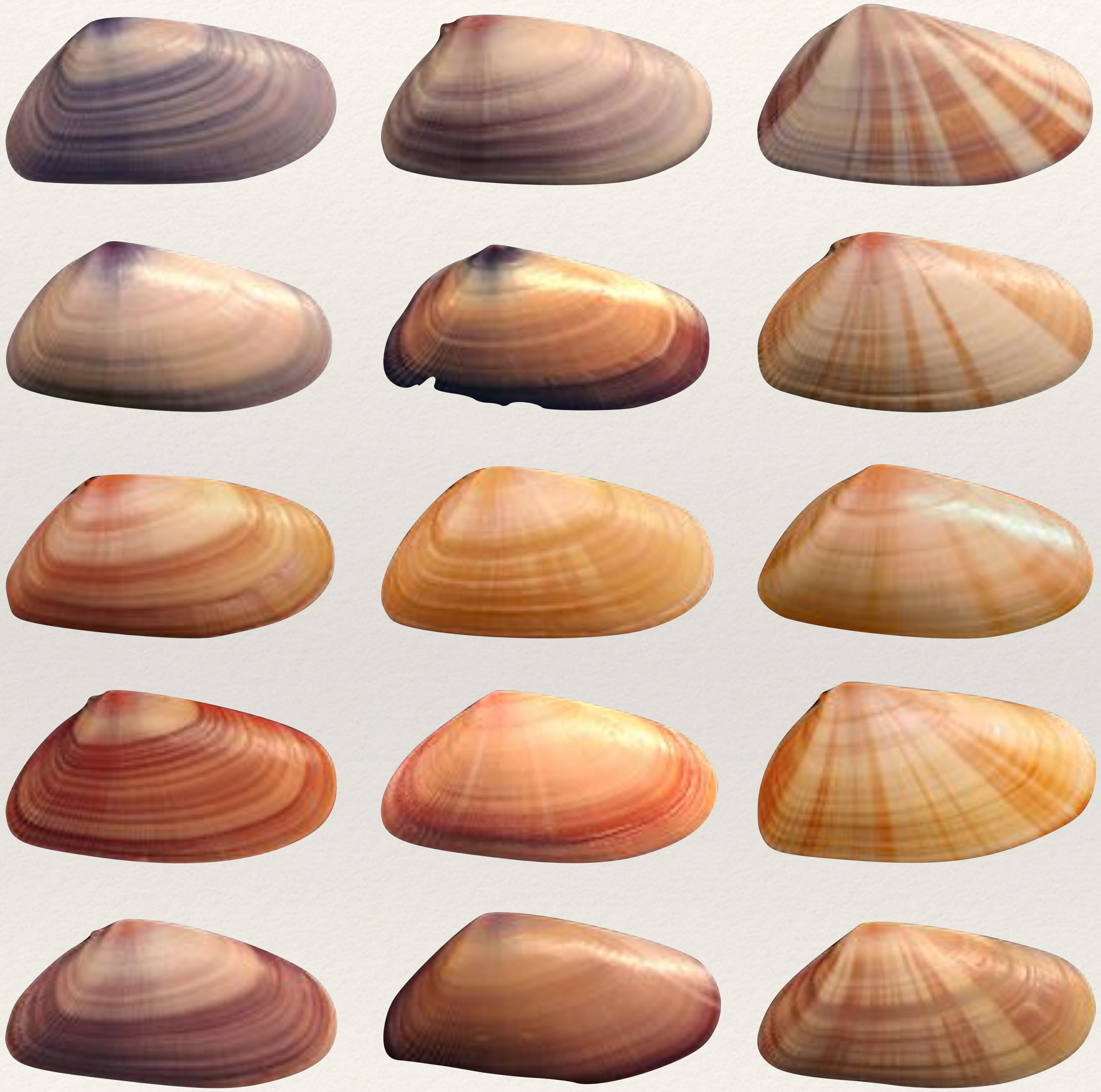
Eukaryotic mRNA processing



Splicing

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

Transcriptome

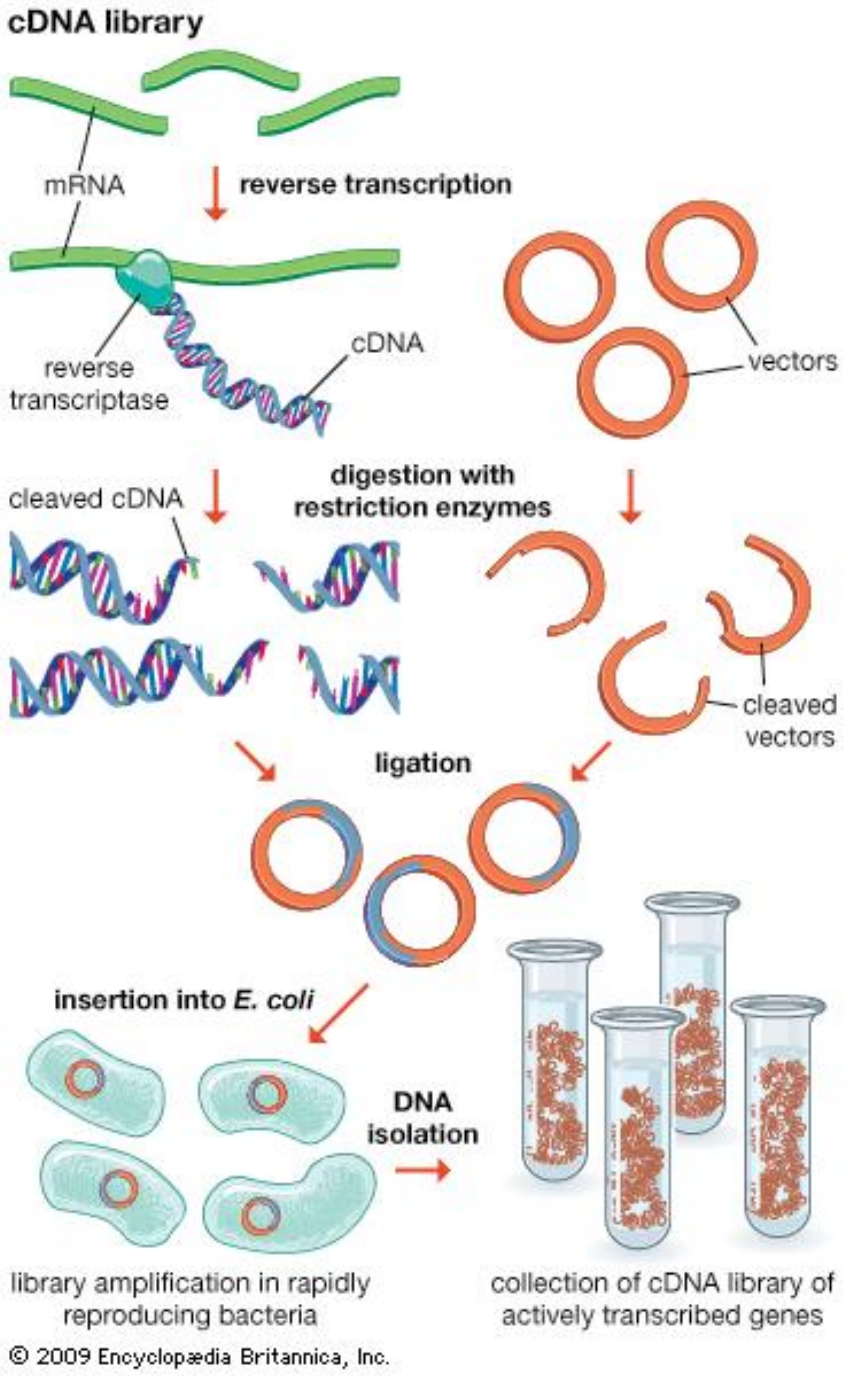


Study individual variation

By Debivort

Obtaining transcriptome

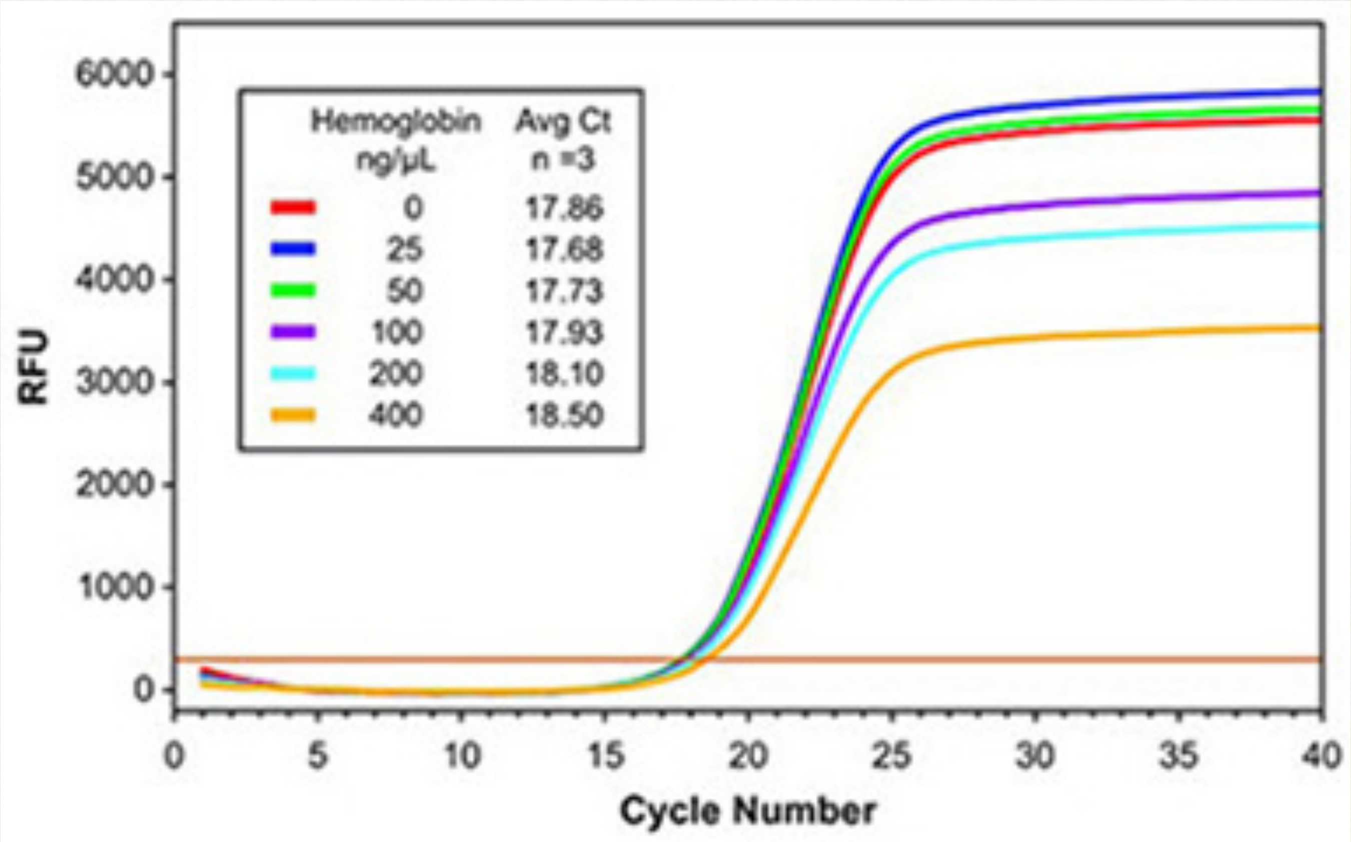
- Sanger sequencing
  - mRNA converted to the more stable cDNA
  - cDNA cleaved and ligated into vectors
  - Vectors amplified (cloned) in *E. coli*
  - DNA isolated = cDNA library
  - Sequenced on Sanger
  - Low throughput
  - High accuracy


cDNA library
mRNA
reverse transcription
reverse transcriptase
cDNA
vectors
cleaved cDNA
digestion with restriction enzymes
cleaved vectors
ligation
insertion into *E. coli*
DNA isolation
library amplification in rapidly reproducing bacteria
collection of cDNA library of actively transcribed genes

© 2009 Encyclopædia Britannica, Inc.

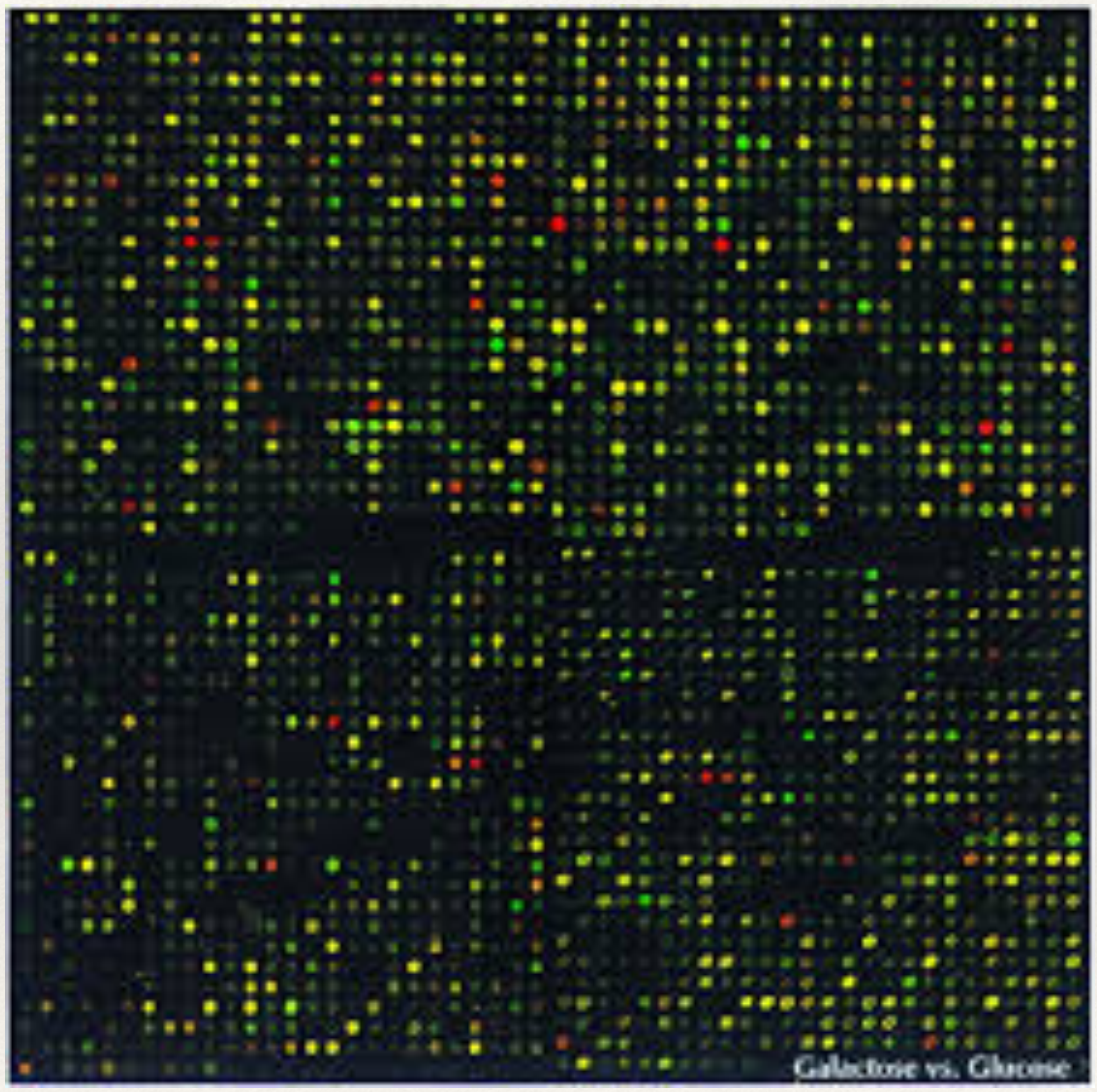
Quantifying expression

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



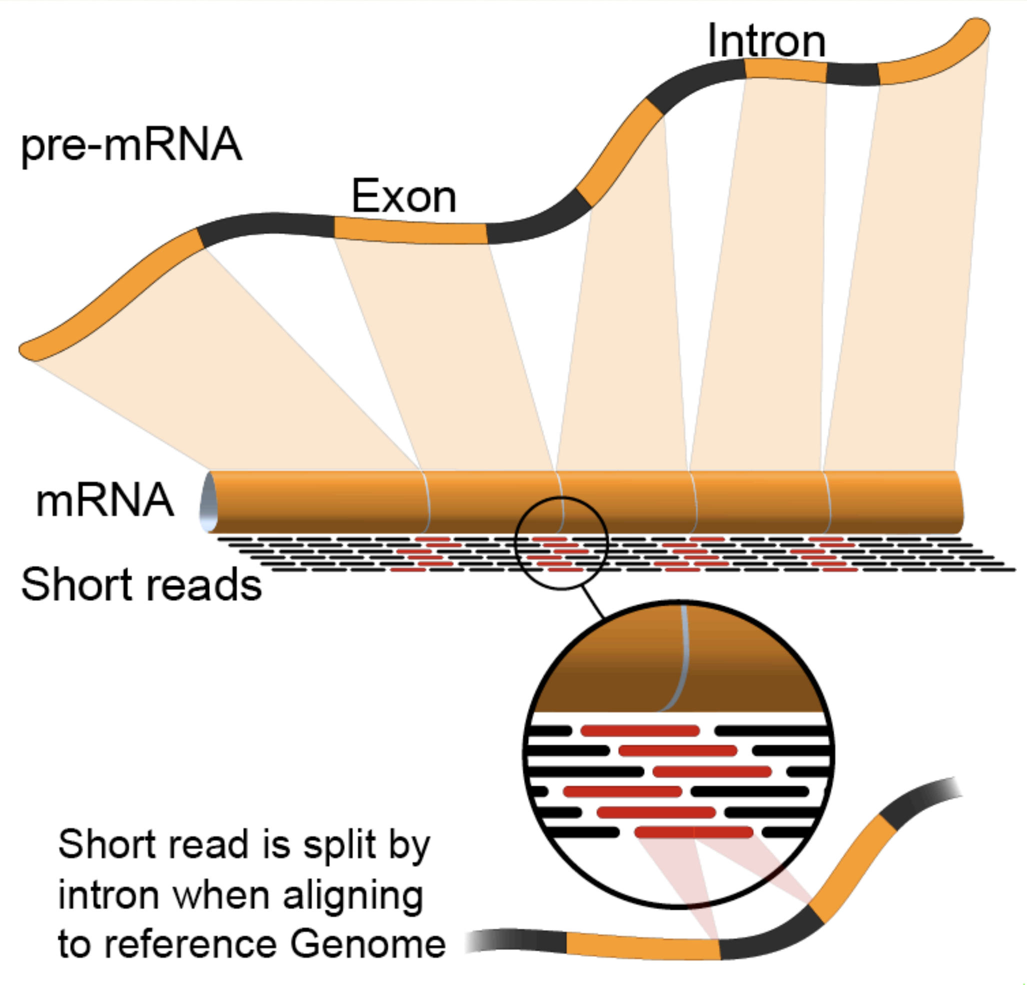
Quantifying expression

- 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


Intron
pre-mRNA
Exon
mRNA
Short reads
Short read is split by intron when aligning to reference Genome

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

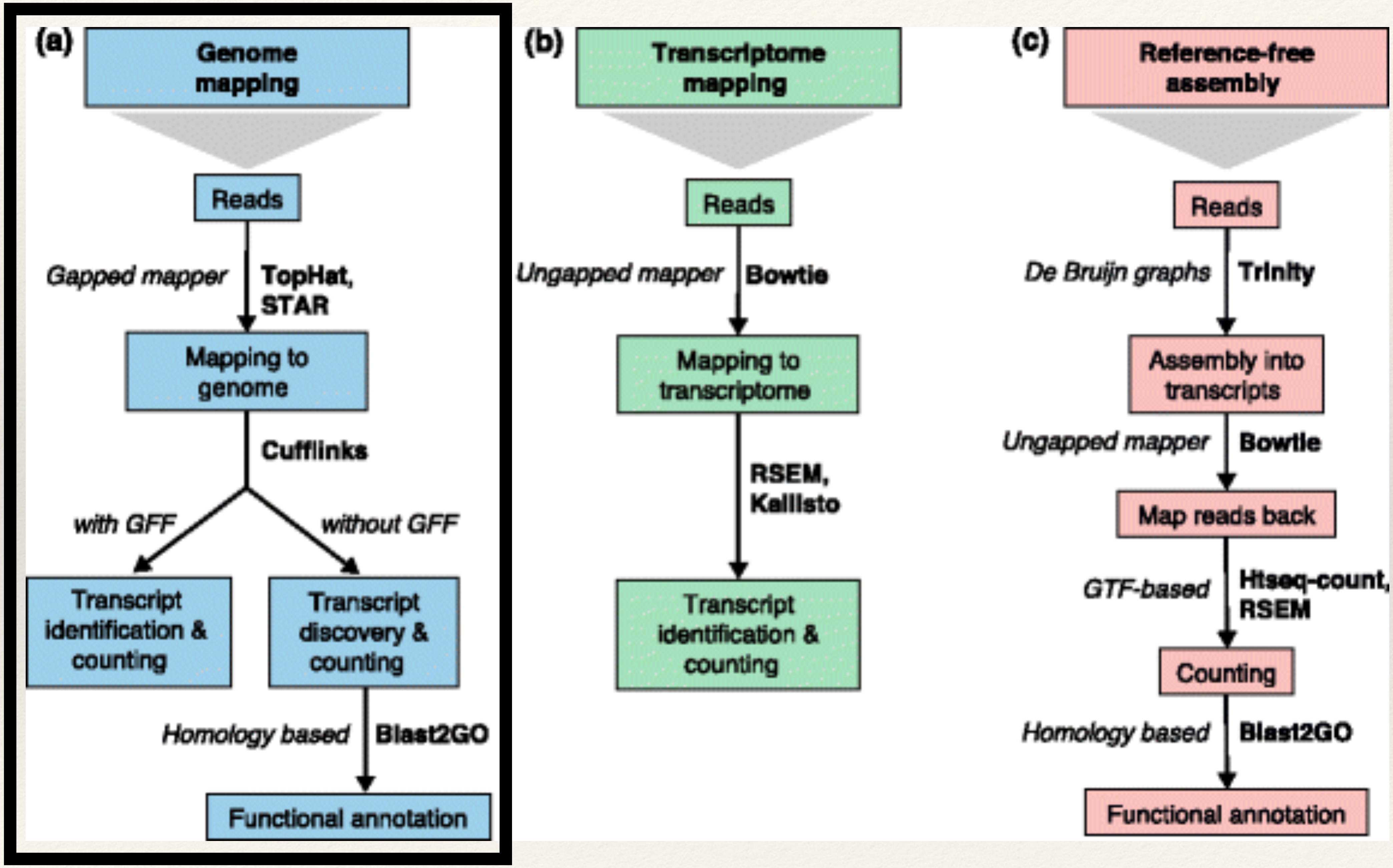
Sequence data analysis

- Is genome available?
- Well annotated?

- *De novo* approach
- Reference based approach
- Transcriptome
- Genome+Transcriptome
- Mixed approach??

- Short reads (Illumina) + Long reads (PacBio, ONT)

Mapping sequence data



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

Library prep (Illumina)

- RNA sequencing
  - Total RNA
  - mRNA
  - small RNA
  - Ribosome profiling

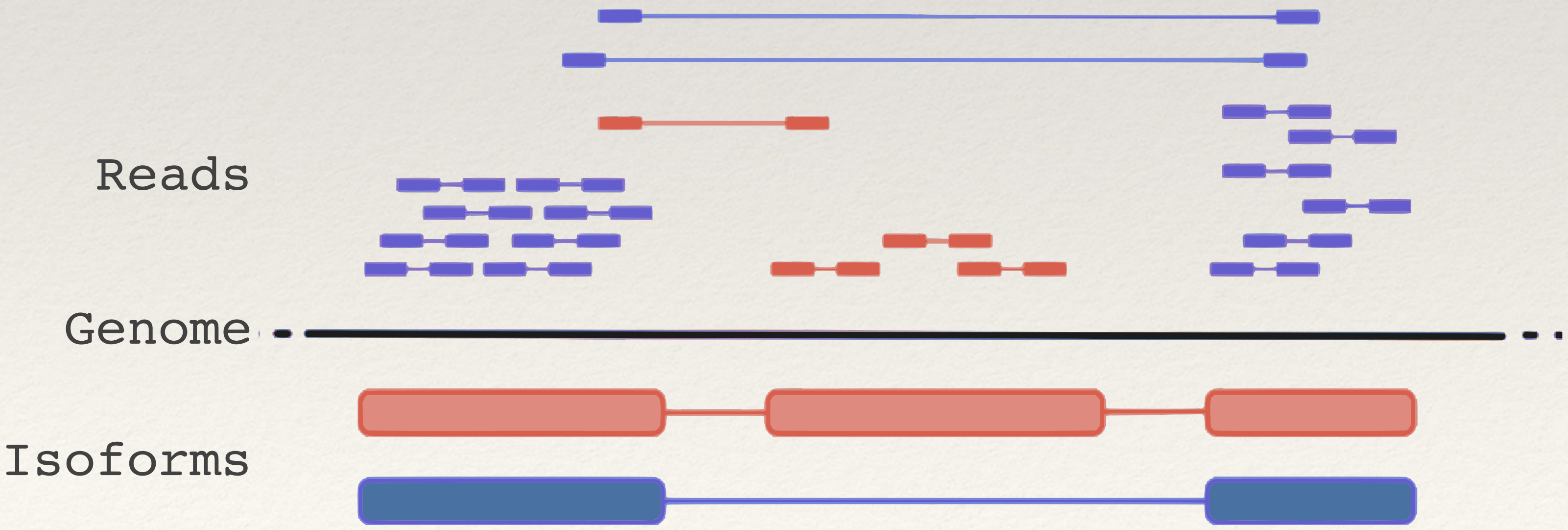
- TruSeq **Stranded** Total RNA kit
- TruSeq **Stranded** mRNA kit
- TruSeq small RNA kit

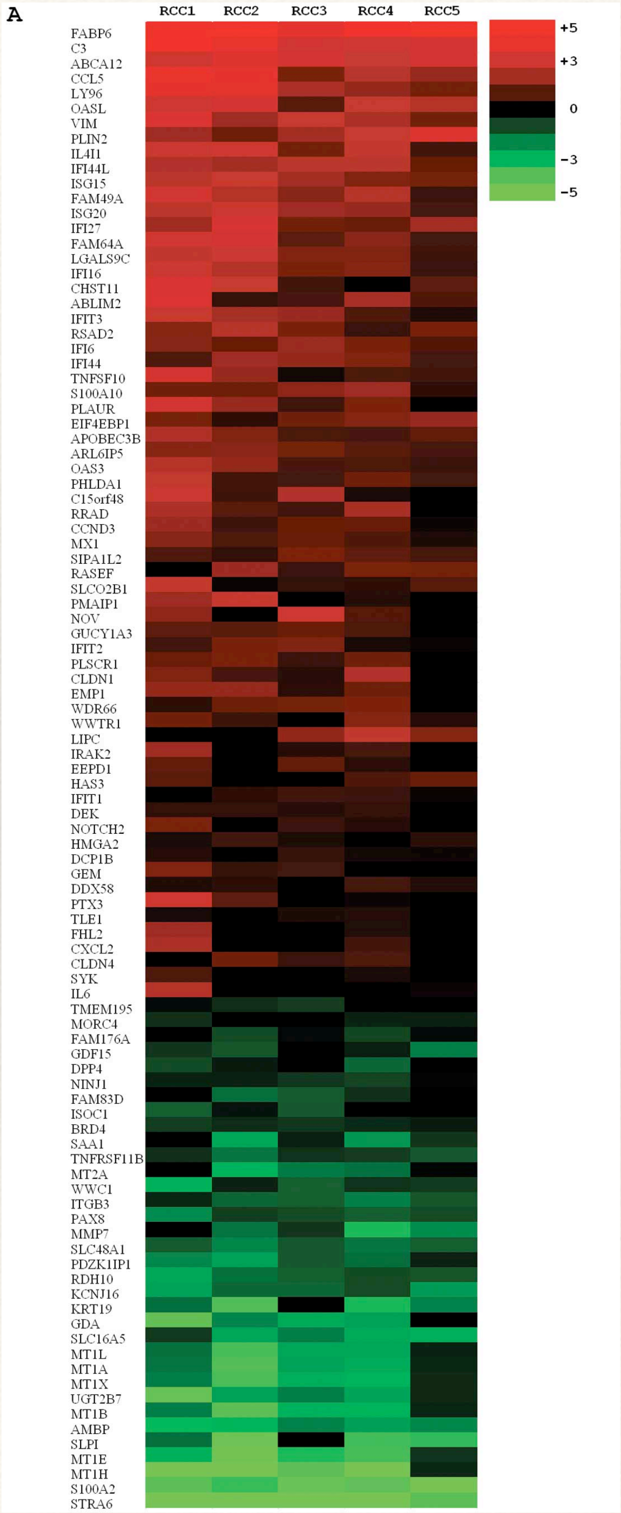
- High quality and quantity of RNA
- Do you want to sequence rRNA??

Depth

- RNA sequencing
  - Highly expressed known transcripts
  - Novel isoforms
  - Low expressed / rare transcripts

More depth





Counting

- Feature - genes, transcript or exon
- How many reads aligned to each feature of interest?
- What is the length of the feature?

- Raw count calculated from BAM files using featureCounts, HTSeq, etc
- Most (all) DE tools would require raw count file and not (pre) scaled data.

Differential expression

- Genes
- Transcripts (Isoforms)
- Allele specific expression
- Exon level expression

Normalisation

- Normalisation within and across samples
- Count gets converted to RPKM, FPKM or TPM
- RPKM (Reads Per Kilobase Million)

  Scaling factor = Total number reads / 1,000,000

  RPM = Read count per feature / scaling factor

  RPKM = RPM / Feature length in kilo bases
- FPKM (Fragments Per Kilobase Million)

  FPM = Fragment count per feature / scaling factor

  FPKM = FPM / Feature length in kilo bases
- TPM (Transcripts Per Kilobase Million)

  RPK = Read count per feature / Feature length in kilo bases

  Scaling factor = sum of RPK / 1,000,000

  TPM = RPK / Scaling factor

DESeq2 (or edgeR) is different!!

https://www.youtube.com/watch?v=UFB993xufUU

https://www.youtube.com/watch?v=TTUrtCY2k-w

DESeq2

- Generalised linear model fit
    - Using negative binomial distortion (aka gamma-Poisson distribution)
- Empirical Bayes shrinkage
    - for within-group variability, i.e., variability between replicates
- Fold change estimation
- Not just pair-wise comparison. Allows for complicated nested designs to be compared

Multiple hypothesis testing and FDR

Multiple hypothesis testing

- Thousands of genes = thousands of hypothesis tests (simultaneously)
- Increased chance of false positives! (Type I error)
  - e.g. you test for differential expression in 1000 genes that are not differentially expressed
  - You would expect 1000 x 0.05 = 50 of them to have a *P*-value < 0.05
- Individual *P*-values not useful : Need multiple testing statistic instead

False Discovery date (Benjamini & Hochberg 1995)

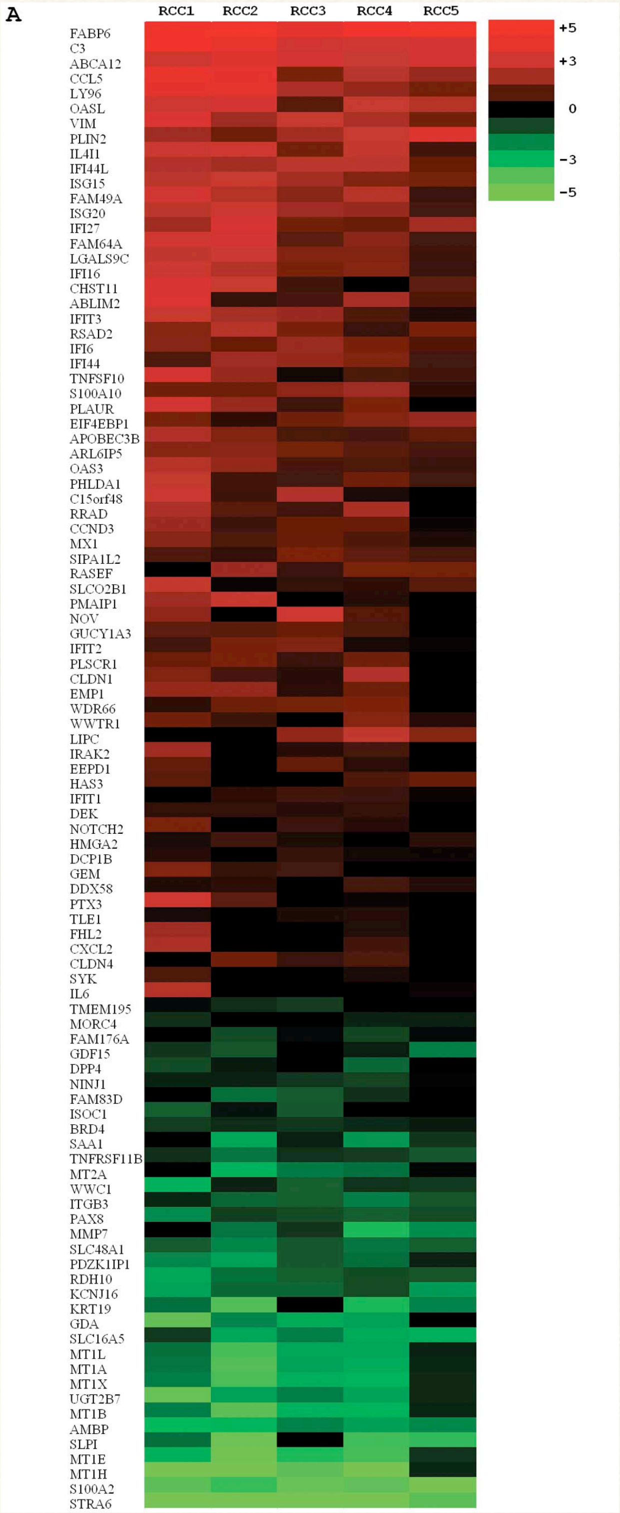
- The expected proportion of Type I errors among the rejected hypotheses
  - i.e. the proportion of false positives
  - Tends to be conservative if many genes are DE
    - FDR = 0.05 common for exploratory/broad scope studies
    - FDR < 0.05 common for medical applications and hunts for candidate genes

DESeq2

- plotDispEsts(): To look at the dispersion plots
- plotPCA(): To find outliers
- plotMA(): Exploring DE results

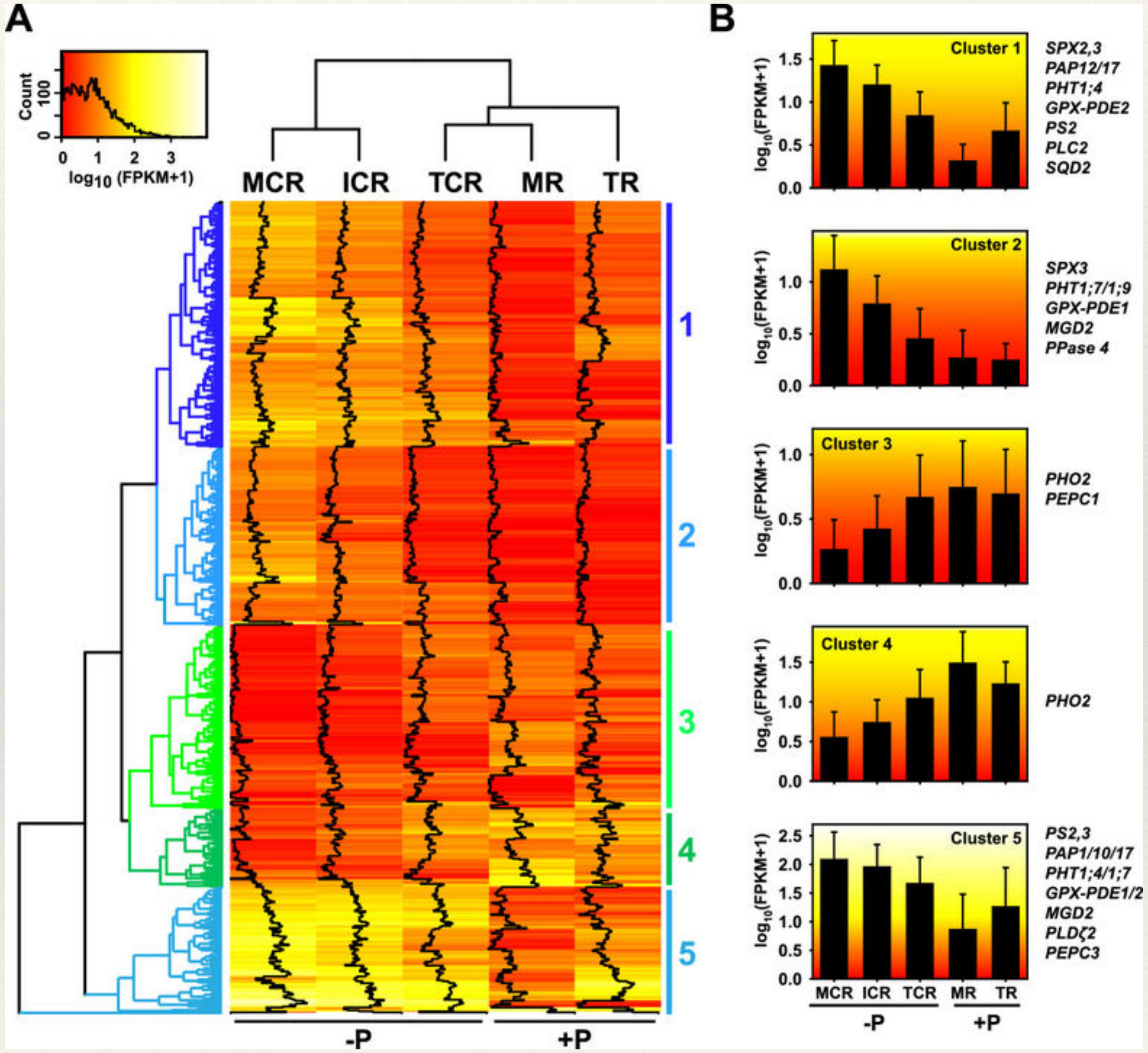
DE list. What next?

- Heatmap
  - Possible to make heatmaps using cummeRbund, DESeq2 and edgeR or in R
  - Tool: MeV TM4
  - Using normalised count information



DE list. What next?

- Clustering
  - Gene (feature level)
  - Sample level
  - Hierarchical
  - CAST: Clustering Affinity Search Technique
- Personal favorite - MeV TM4
  - Possible in R



https://doi.org/10.1186/1471-2164-15-230

DE list. What next?

- Functional profiling of gene lists
  - Gene Ontology (GO) enrichment analysis
    - Biological process
    - Cellular components
    - Molecular function
  - KEGG pathway enrichment analysis
- Tools
  - GOrilla (http://cbl-gorilla.cs.technion.ac.il/)
  - Comprehensive tool - g:Profiler (http://biit.cs.ut.ee/gprofiler/)