University of Zurich UZH

BIO634 - Day 2: RNA sequencing technologies

Carla Bello, carla.bello@ieu.uzh.ch

December 3-4th, 2019

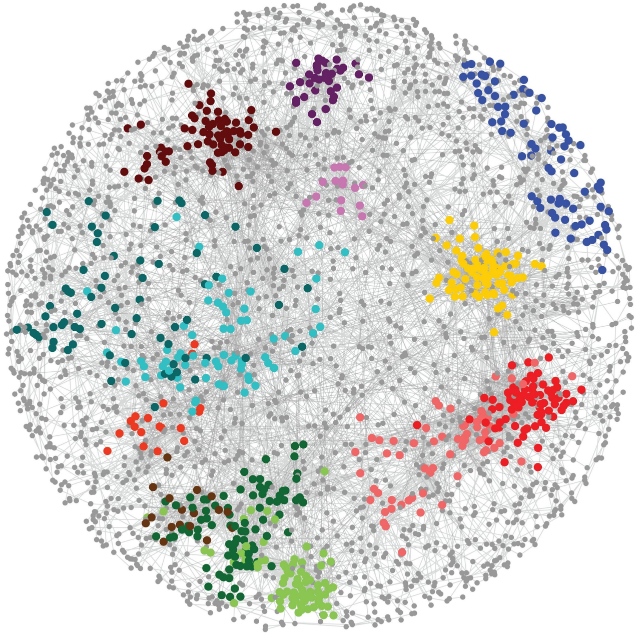
Zürich, CH

Overview

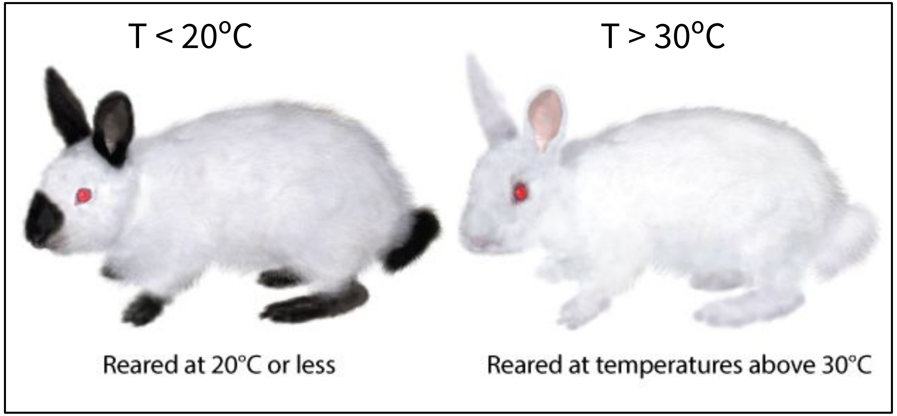
- **Introduction** to RNA sequencing
- **RNA sequencing workflow**: steps to analyze the data
- **Important considerations**: technical, biological replicates, etc
- **Abundance quantification:** Gene, exons, transcripts
- **Differential expression analyses:** DEseq, edgeR, etc

Gene expression and phenotype

Image from Wikipedia



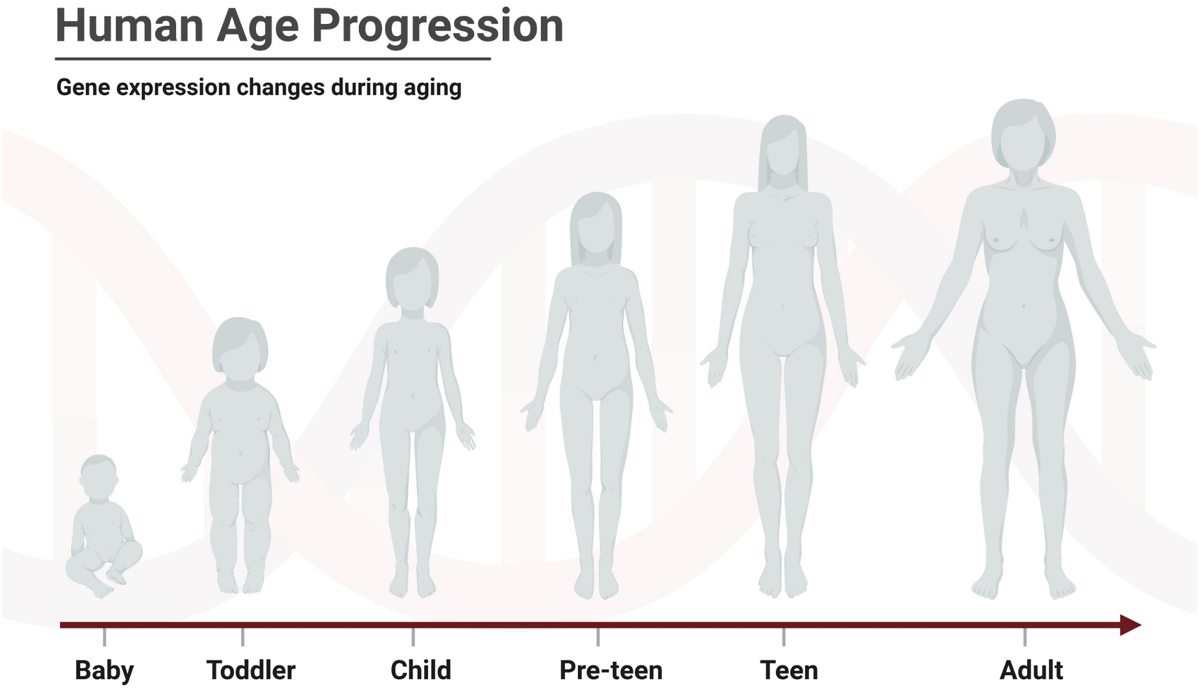
Himalayan rabbit



Example: A pigment gene is influenced by temperature. When the temperature is **< 20°C the gene is inactive**

Nothing in the genome makes sense except in the light of the transcriptome

RNA sequencing of whole genomes

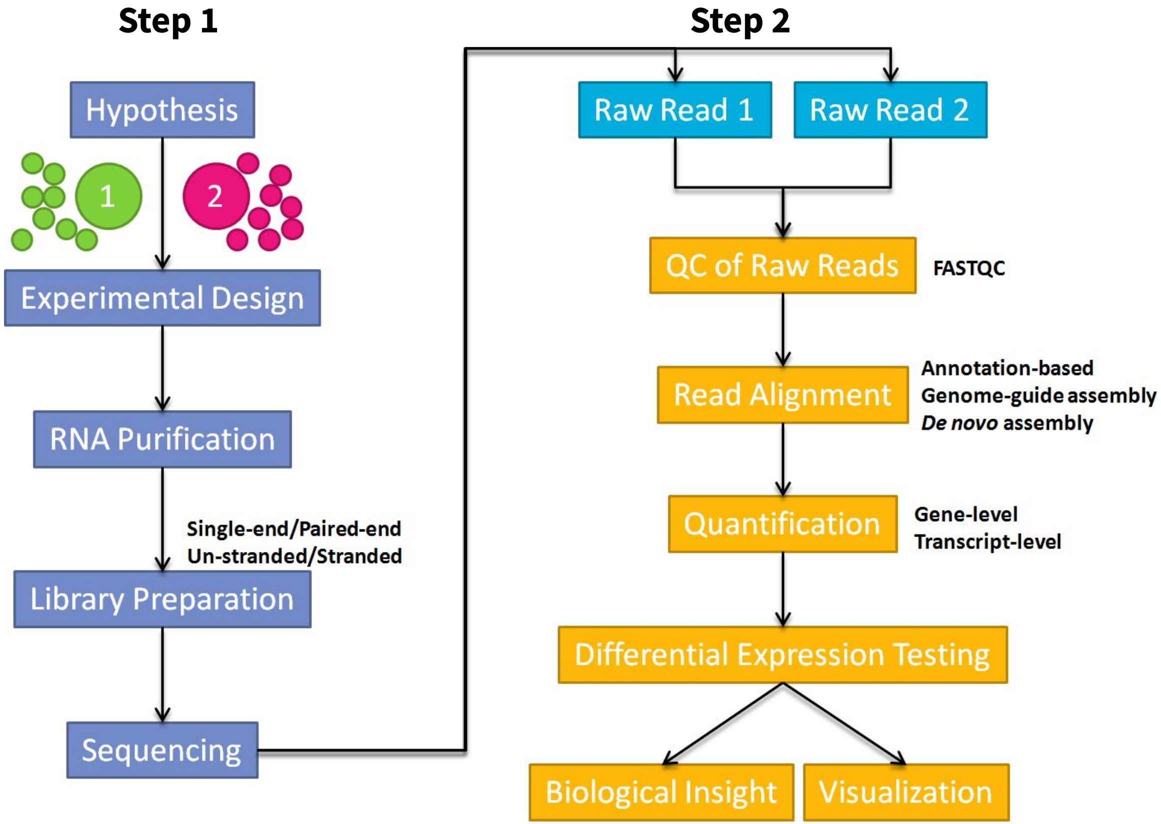


Expression of all the genes in the genome at different stages of aging

Applications of RNA sequencing

- **Gene** expression/**differential** gene expression
- **Detecting novel** or **alternative** transcripts
- **De-novo transcriptome** assembly
- **SNP analysis,** e.g disease association studies
- **Allele-specific expression**
- **RNA studies**: miRNA, tRNAs, snRNA, lncRNAs, etc.

RNA sequencing workflow



Abundance quantification

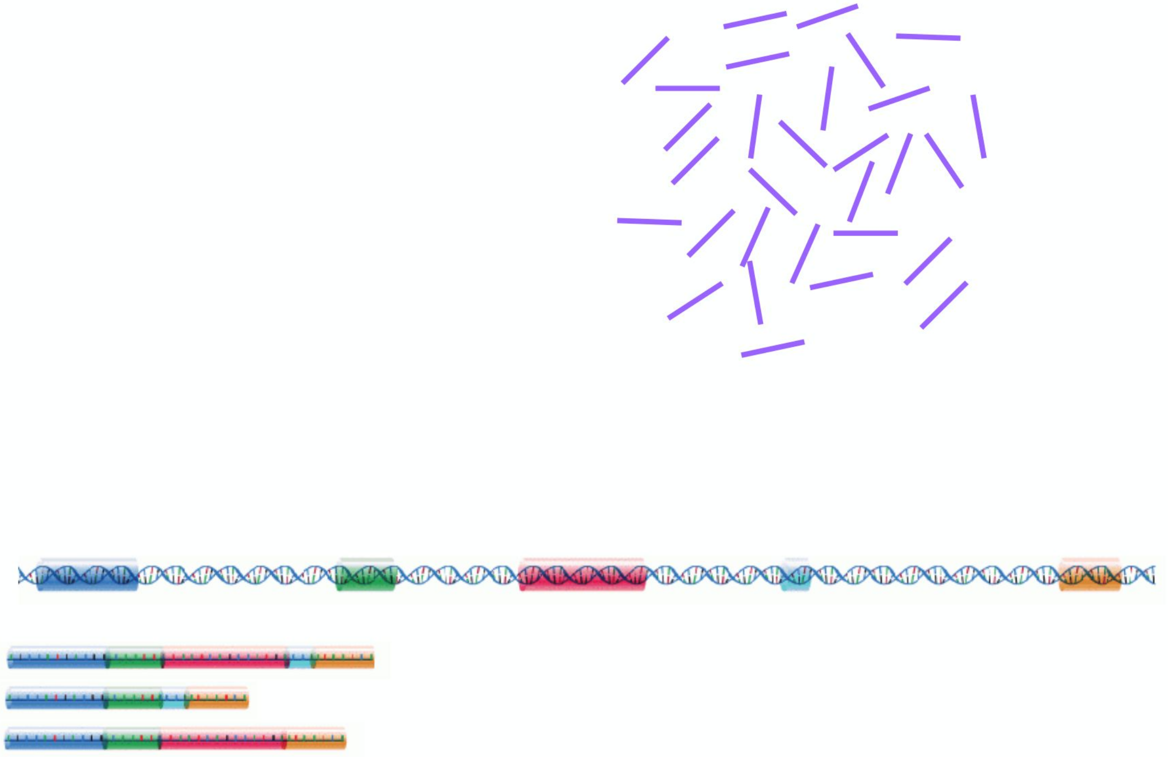


Image from Dr. Charlotte Soneson

Abundance quantification

Genome alignment of RNA-seq requires a splice-aware aligner (STAR, HISAT2)

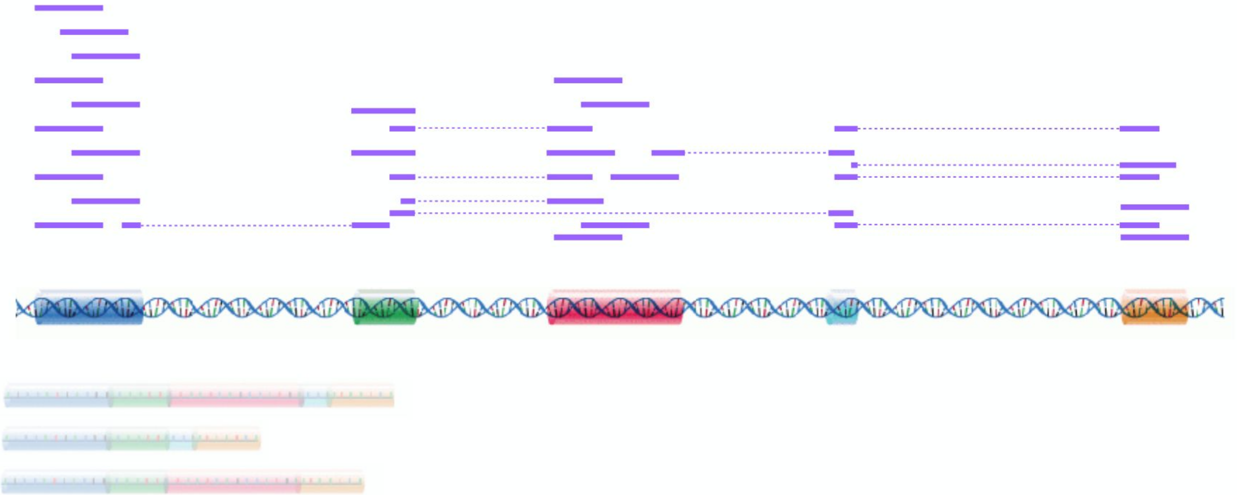
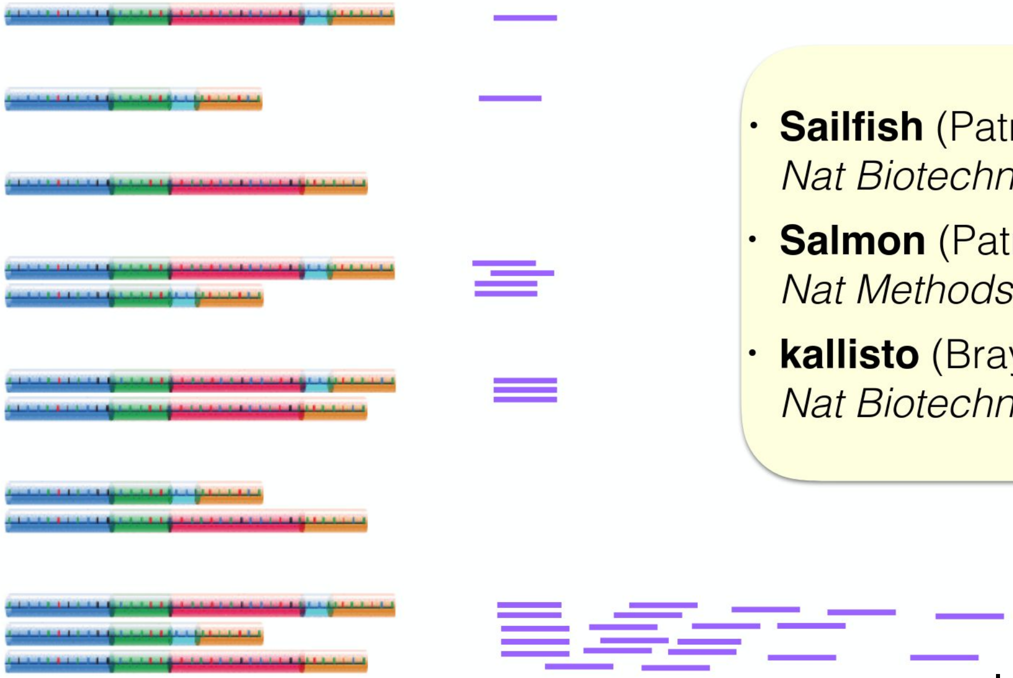


Image from Dr. Charlotte Soneson

Abundance quantification

Transcript-level counts, e.g. obtained by "alignment-free" estimation methods

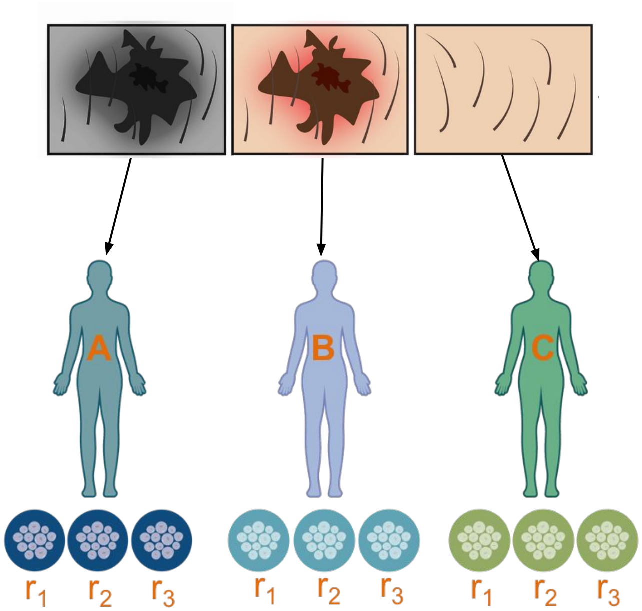


- **Sailfish** (Patro et al, *Nat Biotechnol* 2014)
- **Salmon** (Patro et al, *Nat Methods* 2017)
- **kallisto** (Bray et al, *Nat Biotechnol* 2016)

Image from Dr. Charlotte Soneson

Important considerations before sequencing

3 different samples
3 different conditions



3-12 replicates per sample/condition

Technical and biological replicates are important and should be taken into consideration **when planning the experiments**

Technical replicates

1. <u>Technical replicates:</u> **Biological material is the same** but the **technical steps used to measure gene expression are repeated.**

In particular **RNA-seq library preparation** (RNA fragmentation, cDNA synthesis and PCR amplification) **may introduce biases in the data**.

Biological replicates and statistical power

2. **Biological replicates**: Are different biological samples that are processed separately. They are **required if inference on the population is to be made**, with **three** biological replicates **being the minimum for any inferential analysis.**

3. **Desired statistical power**, that is the capacity for detecting statistically significant differences in gene expression between experimental groups.

Different methods to analyze RNA-seq data

- There are different packages for **differential expression analysis**, such as **edgeR** and **DESeq** based on negative binomial (NB) distributions or **baySeq** and **EBSeq** which are **Bayesian** approaches based on a **negative binomial model.**

- These packages work mostly by **estimating the variance mean dependance** in **count** data.

Factors to consider for RNAseq analyses

1) **Within sample**
   - Gene/transcript length
   - Relative expression (a few highly expressed genes)

2) **Between samples**
   - Sequencing depth (library size)
   - Sequencing biases

3) Raw read counts **are NOT directly** comparable **between** samples:
   Solution: Normalize read counts

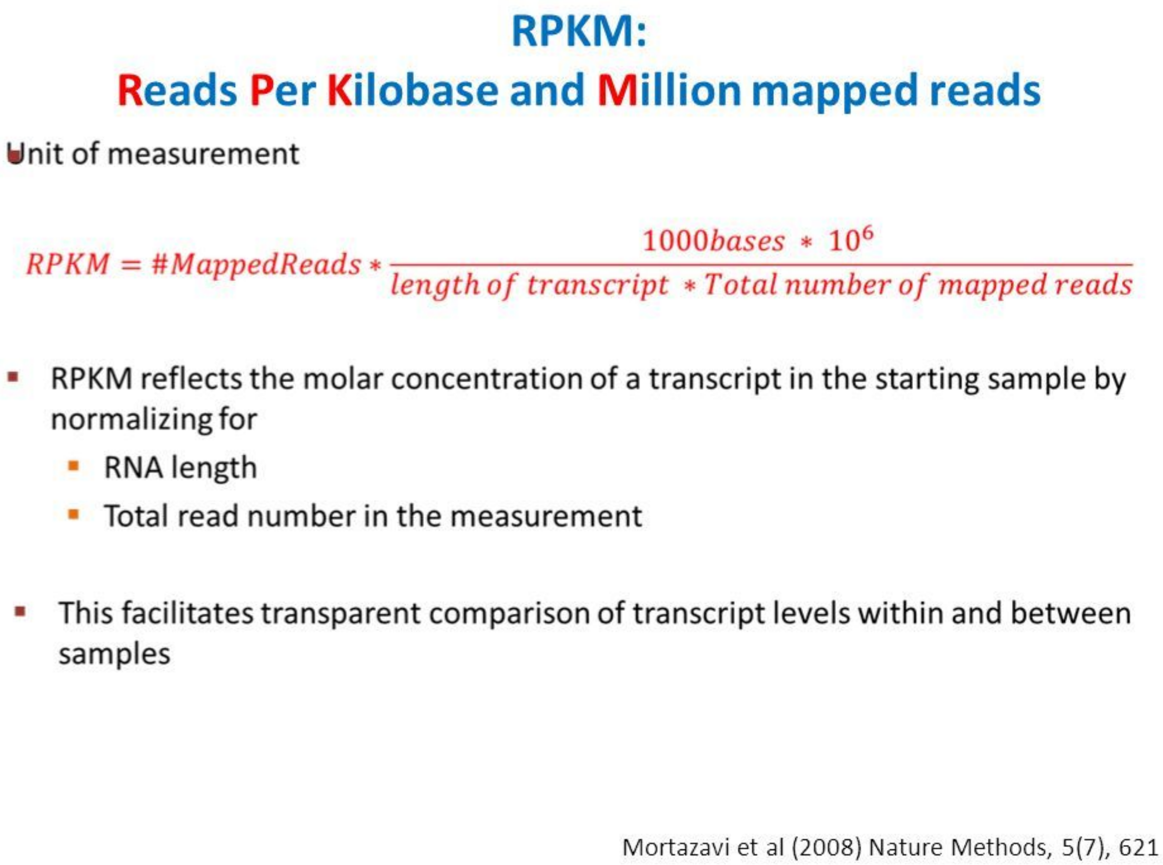
RNA-Seq Read Count Normalization

- **RPKM: Reads per kilobase of transcript** per million reads of library

- **FPKM: Fragments per kilobase of transcripts** per million reads of library

- **TPM**: **Transcripts per million** reads of library

RNA-Seq Read Count Normalization



RPKM:

Reads Per Kilobase and Million mapped reads

Unit of measurement

$$RPKM = \#MappedReads * \frac{1000bases\ *\ 10^6}{length\,of\ transcript\ *Total\ number\ of\ mapped\ reads}$$

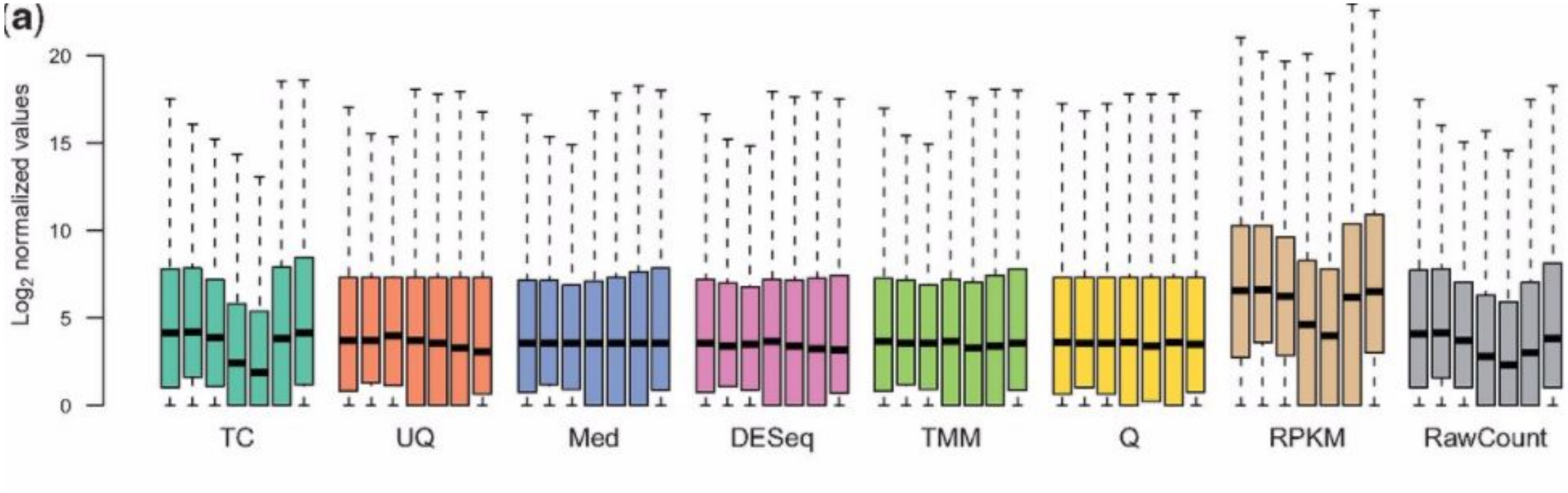
- RPKM reflects the molar concentration of a transcript in the starting sample by normalizing for
  - RNA length
  - Total read number in the measurement

- This facilitates transparent comparison of transcript levels within and between samples

Mortazavi et al (2008) Nature Methods, 5(7), 621

RNA-Seq Read count normalization

RPKM/FPKM are normalized counts. **DESeq/edgeR** requires **raw counts** as **input** as they have their **own normalisation methods**



Differential expression analyses

- **Many statistical methods available**
  - T-test
  - Poisson Distribution
  - Negative binomial

- **No clear consensus yet.**

- **Tools shown to perform well (under certain circumstances):**
  - LIMMA (TMM)
  - DESeq (RLE)
  - edgeR (TMM)
  - Cuffdiff (FPKM)
  - RSEM (EM)
  - Trinity

Identify genes that show differences in expression level between conditions (samples)

Differential expression analyses

1. Analyzing RNA sequencing data with **Salmon**
2. Exploration of ***airway*** **library**: airway smooth muscle
3. Differential analysis: comparison between **DEseq** and **edgeR**

Later in the afternoon

1. List of differentially expressed genes
2. Biological context
3. Pathway Analysis (differentially expressed biological pathways)
4. Gene Set Enrichment Analysis (GSEA) (functional enrichment between two biological groups)
5. Co-expression analysis

Hands-on session - Part II: RNA-seq

Please, go here and follow the instructions:

https://github.com/carlalbc

https://github.com/carlalbc/BIO634_2020/