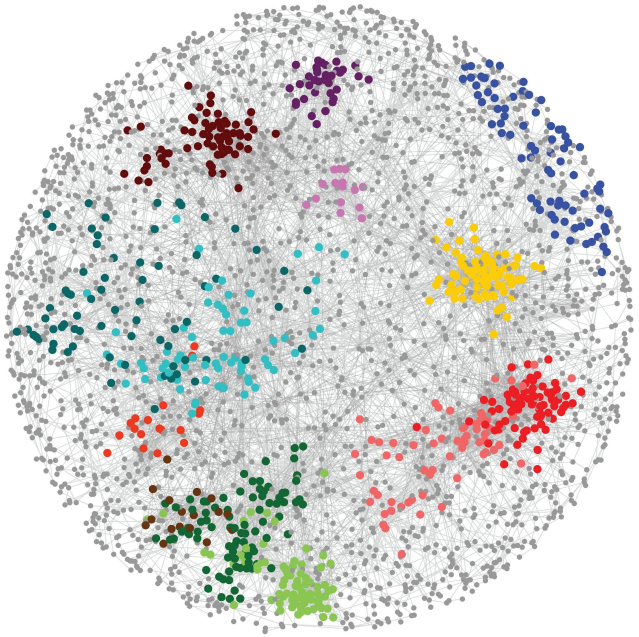
BIO634 - Day 2: RNA-sequencing technologies

September 17-18th, 2018

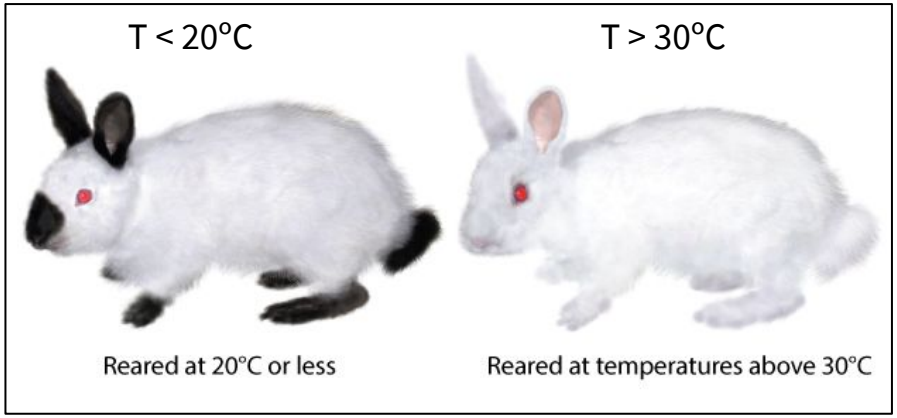
Carla Bello, carla.bello@ieu.uzh.ch

University of Zurich[UZH]

Gene expression and phenotype



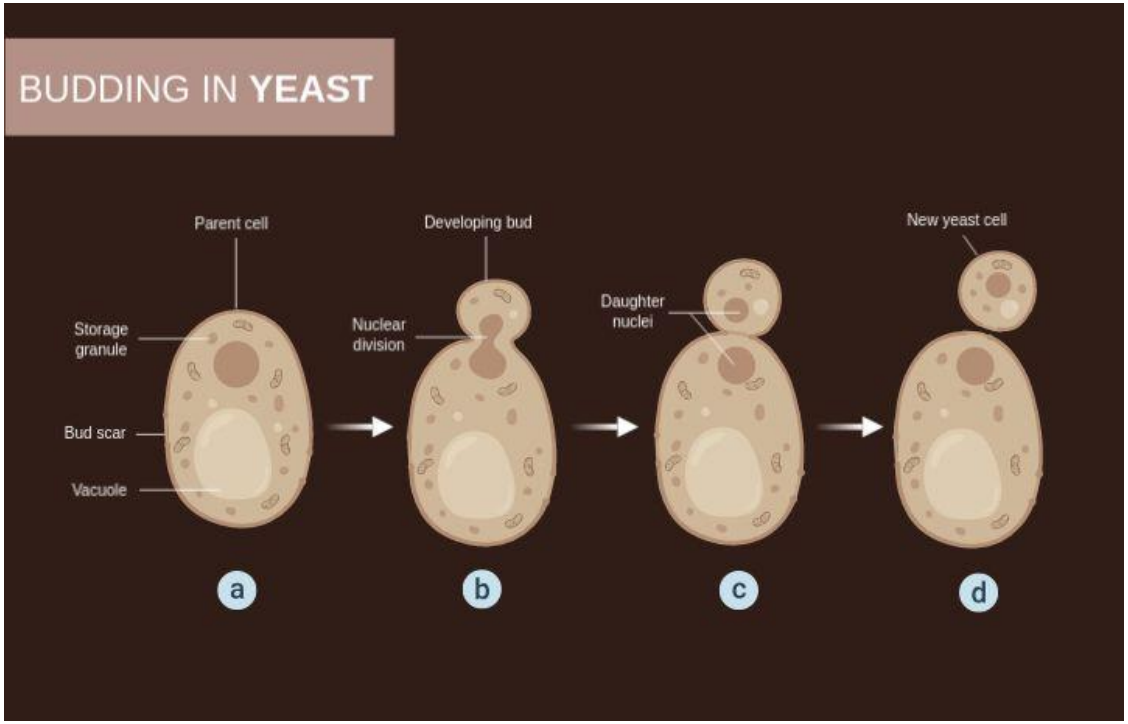
Himalayan rabbit



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

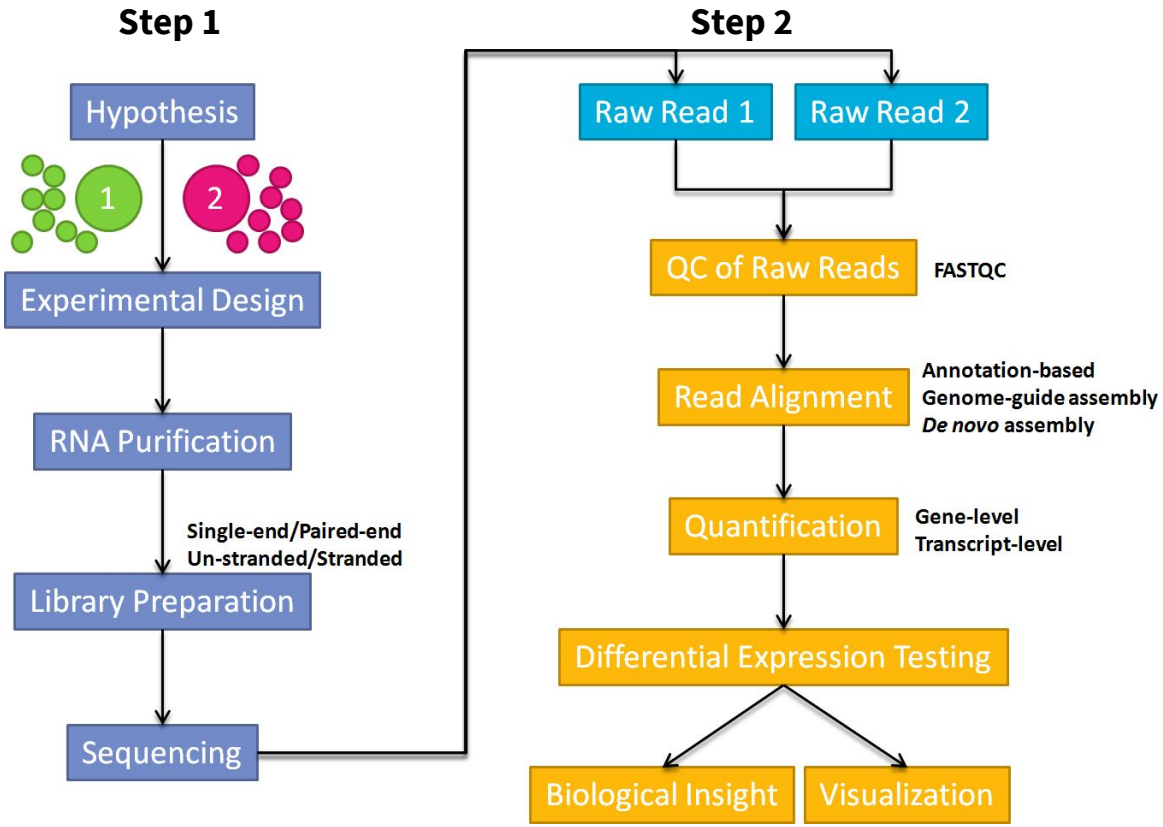


Expression of all the genes in the genome at different budding times in yeast

Applications of RNA sequencing

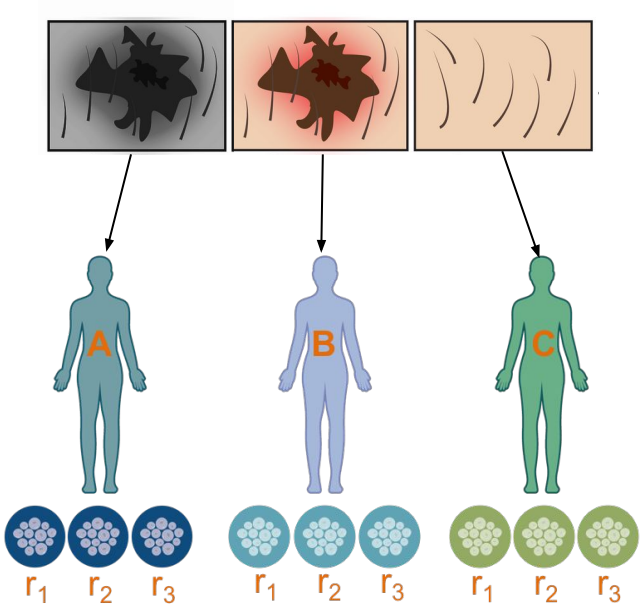
1. Gene expression/differential expression
2. Detecting novel/alternative transcripts
3. De-novo transcriptome assembly
4. SNP analysis (disease association)
5. Allele specific expression
6. RNAs (miRNA, snRNA, lncRNAs)

RNA sequencing workflow



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 and biological 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**.

Technical and biological replicates

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.

Methods to analyze RNA-seq data

- There are different methods 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.

Methods to analyze RNA-seq data

Factors to consider:

1) **Within sample**
   - Gene/transcript length
   - Relative expression (a few highly expressed genes)
2) **Between samples**
   - Sequencing depth (library size)
   - Sequencing biases

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

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

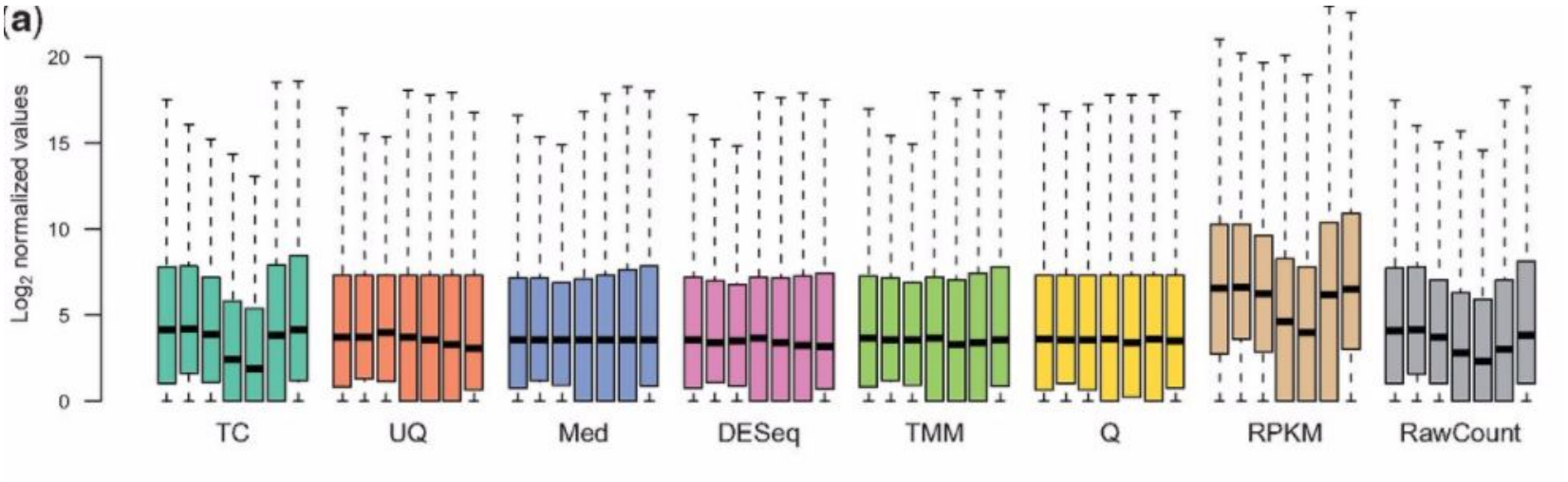
- RPKM reflects the molar concentration of a transcript in the starting sample by normalizing for
  - RNA length
  - Total read number in the measurement

- 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

I. Aligning transcriptomes with STAR

II. Exploration of *airway* library (airway smooth muscle)

III. 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_2018/