RNAseq: What to do?

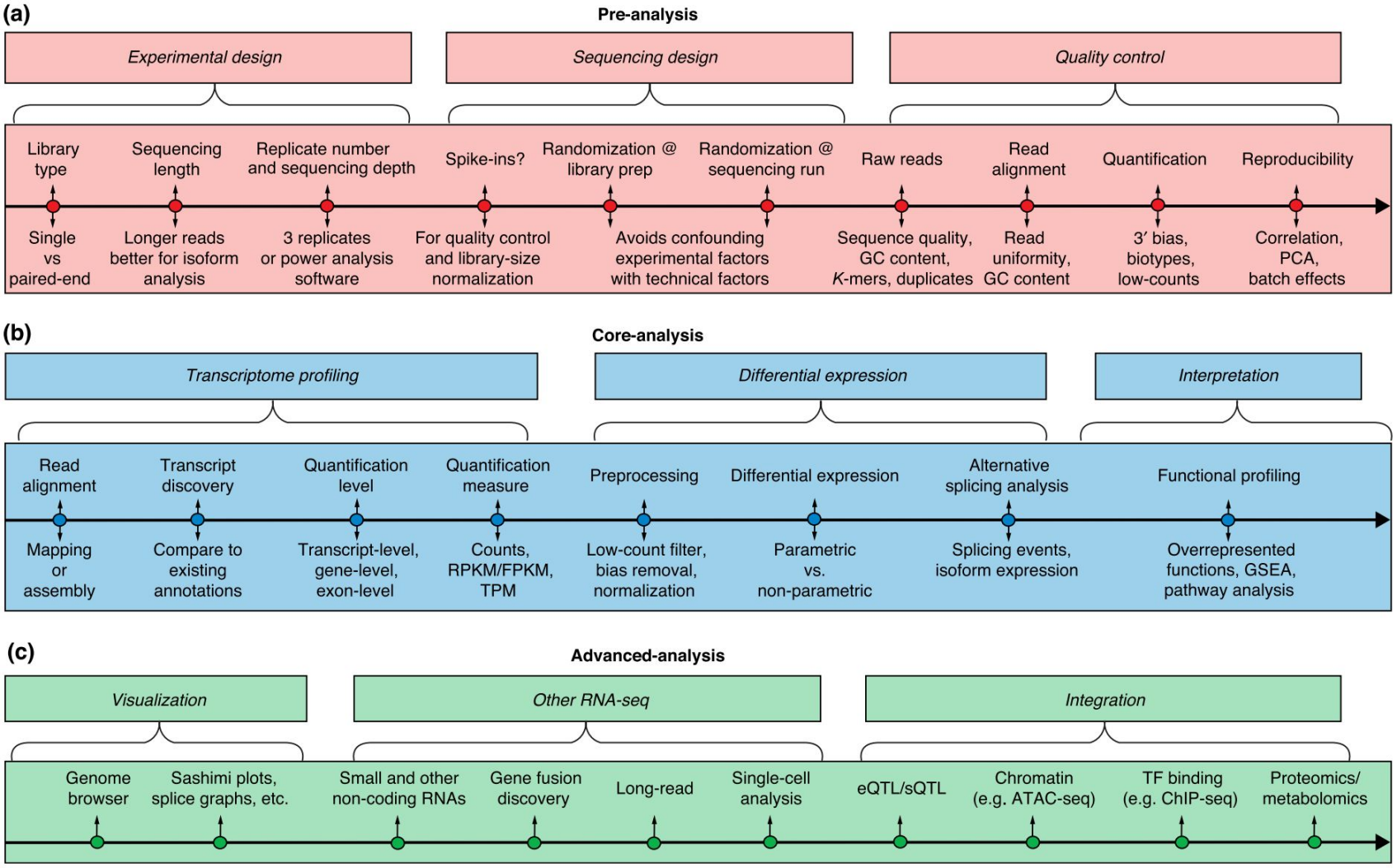
Guest Lecture: Juli Petereit, PhD
Bioinformatician and Data Scientist

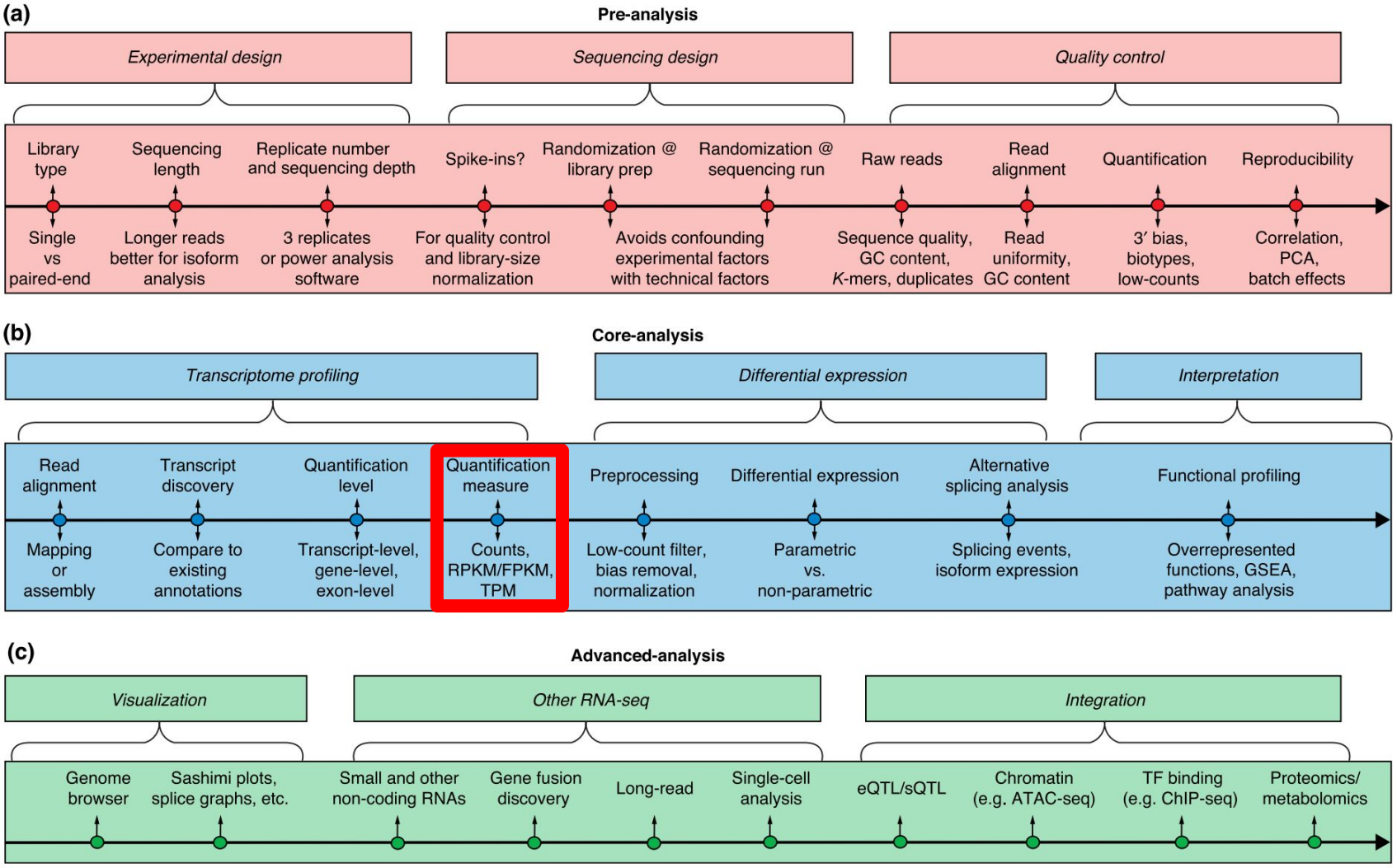
What is RNAseq?

“RNA sequencing (RNA-Seq, RNAseq), also called whole transcriptome shotgun sequencing, uses next-generation sequencing (NGS) to reveal the presence and quantity of RNA in a biological sample at a given moment in time.” https://en.wikipedia.org/wiki/RNA-Seq

Most common application of RNAseq is to quantify gene and transcript expression.

So how do we go about it?


(a)
Pre-analysis
Experimental design
Sequencing design
Quality control
Library type
Sequencing length
Replicate number and sequencing depth
Spike-ins?
Randomization @ library prep
Randomization @ sequencing run
Raw reads
Read alignment
Quantification
Reproducibility
Single vs paired-end
Longer reads better for isoform analysis
3 replicates or power analysis software
For quality control and library-size normalization
Avoids confounding experimental factors with technical factors
Sequence quality, GC content, K-mers, duplicates
Read uniformity, GC content
3′ bias, biotypes, low-counts
Correlation, PCA, batch effects
(b)
Core-analysis
Transcriptome profiling
Differential expression
Interpretation
Read alignment
Transcript discovery
Quantification level
Quantification measure
Preprocessing
Differential expression
Alternative splicing analysis
Functional profiling
Mapping or assembly
Compare to existing annotations
Transcript-level, gene-level, exon-level
Counts, RPKM/FPKM, TPM
Low-count filter, bias removal, normalization
Parametric vs. non-parametric
Splicing events, isoform expression
Overrepresented functions, GSEA, pathway analysis
(c)
Advanced-analysis
Visualization
Other RNA-seq
Integration
Genome browser
Sashimi plots, splice graphs, etc.
Small and other non-coding RNAs
Gene fusion discovery
Long-read
Single-cell analysis
eQTL/sQTL
Chromatin (e.g. ATAC-seq)
TF binding (e.g. ChIP-seq)
Proteomics/ metabolomics


(a)
Pre-analysis
Experimental design
Sequencing design
Quality control
Library type
Sequencing length
Replicate number and sequencing depth
Spike-ins?
Randomization @ library prep
Randomization @ sequencing run
Raw reads
Read alignment
Quantification
Reproducibility
Single vs paired-end
Longer reads better for isoform analysis
3 replicates or power analysis software
For quality control and library-size normalization
Avoids confounding experimental factors with technical factors
Sequence quality, GC content, K-mers, duplicates
Read uniformity, GC content
3′ bias, biotypes, low-counts
Correlation, PCA, batch effects
(b)
Core-analysis
Transcriptome profiling
Differential expression
Interpretation
Read alignment
Transcript discovery
Quantification level
Quantification measure
Preprocessing
Differential expression
Alternative splicing analysis
Functional profiling
Mapping or assembly
Compare to existing annotations
Transcript-level, gene-level, exon-level
Counts, RPKM/FPKM, TPM
Low-count filter, bias removal, normalization
Parametric vs. non-parametric
Splicing events, isoform expression
Overrepresented functions, GSEA, pathway analysis
(c)
Advanced-analysis
Visualization
Other RNA-seq
Integration
Genome browser
Sashimi plots, splice graphs, etc.
Small and other non-coding RNAs
Gene fusion discovery
Long-read
Single-cell analysis
eQTL/sQTL
Chromatin (e.g. ATAC-seq)
TF binding (e.g. ChIP-seq)
Proteomics/ metabolomics

RNAseq, RNA-Seq, nextGen - which one is correct?

Research community does not even agree on a common abbreviation: RNAseq, RNA-seq, RNA-Seq.

Thus it is not surprising that there are many variations of RNAseq protocols and analyses as well.

This is a challenge for new and newer users to appreciate all of the steps necessary to properly conduct an RNA-seq study.

I recommend to consult with as many people as possible, especially your bioinformatician, and don’t try to tackle the problem alone.

So what do I have to do? Which analysis procedure?

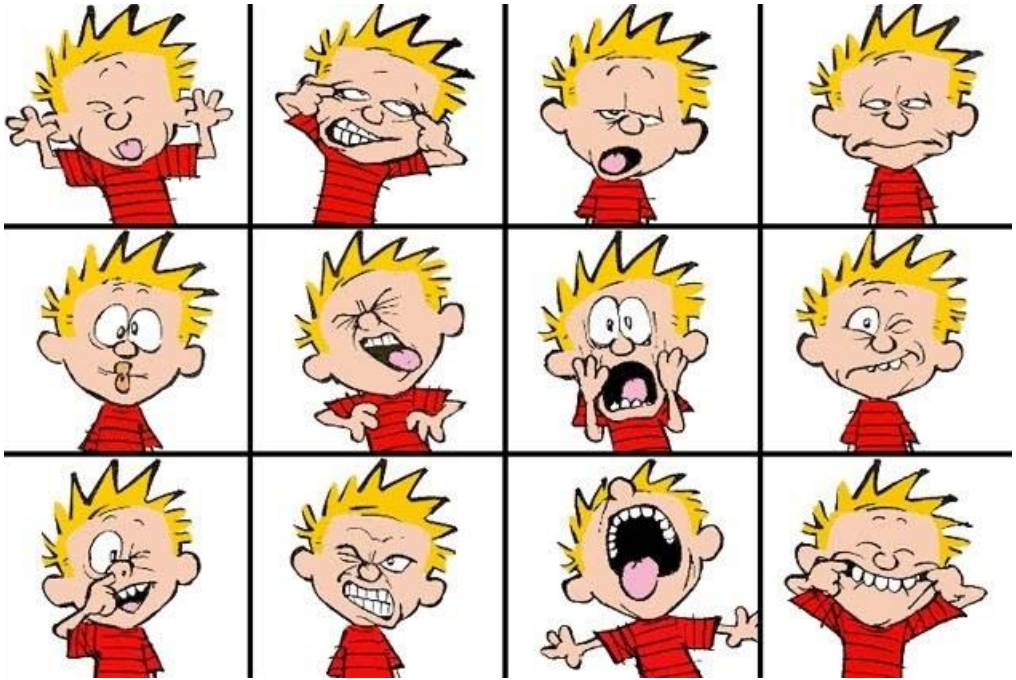
There is not one golden pipeline, -rule or -standard.

RNAseq analysis must be adjusted according to the scientific question/objective and organism being studied.

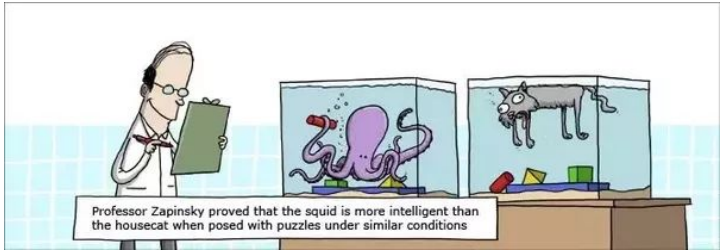
Individual experimental scenarios potentially require adjusted/different methods for transcript quantification, normalization, and other downstream analyses (e.g., differential gene expression)

What is the appropriate experimental setup and analysis strategy?

This is How I feel most days =)



Experimental Design



Define a good experimental design

- Define a precise research question.
- What is the underlying hypothesis?
- Which library type should be utilized?
- What sequencing depth and number of replicates are appropriate for the biological system under study?
- Does my protocol allow for adequate execution of the sequencing experiment itself?
- Are there any steps that possibly introduce unnecessary biases?

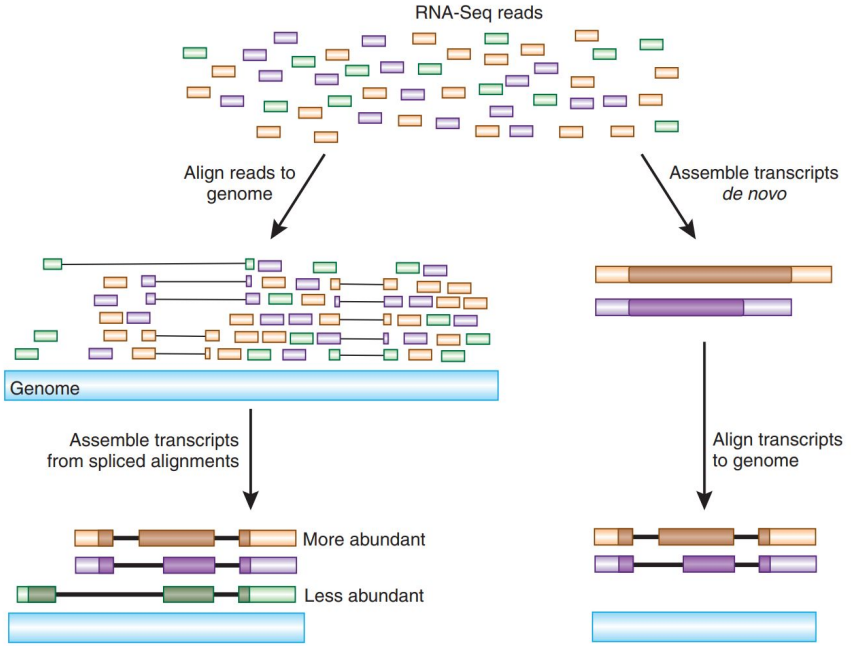
Assembly versu Alignment

For non-model organism without reference genome

1) Assembling reads de novo into contigs
2) Mapping these contigs onto the transcriptome

For model organism

Identify transcripts by mapping reads onto the genome



Nature Biotech (2010) 28, 421-423

Replicates

Often the number of replicates is simply governed by availability of samples.

The number of samples should depend on:

- technical variability in the RNA-seq procedures
- biological variability of the system/organism under study
- desired statistical power (power analysis)

Types of sequencing

Sequencing option depends on analysis goals

Single-end (SE) reads: cheaper, often sufficient for well-annotated organisms

Paired-end (PE) reads: preferable for de novo transcript discovery or isoform expression analysis

Short reads: could be good enough

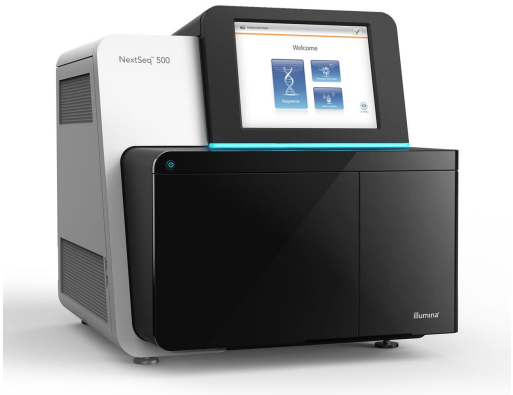
Long reads: improve mappability and transcript identification, preferable for poorly annotated transcriptomes, faster alignment

For more: http://www.anthonybaldor.com/sequencing-run-types/

Sequencers

Each technology has its own proprietary software, preparation kits, read lengths, quirks, etc. and therefore require different analyses procedures.

Illumina



PacBio

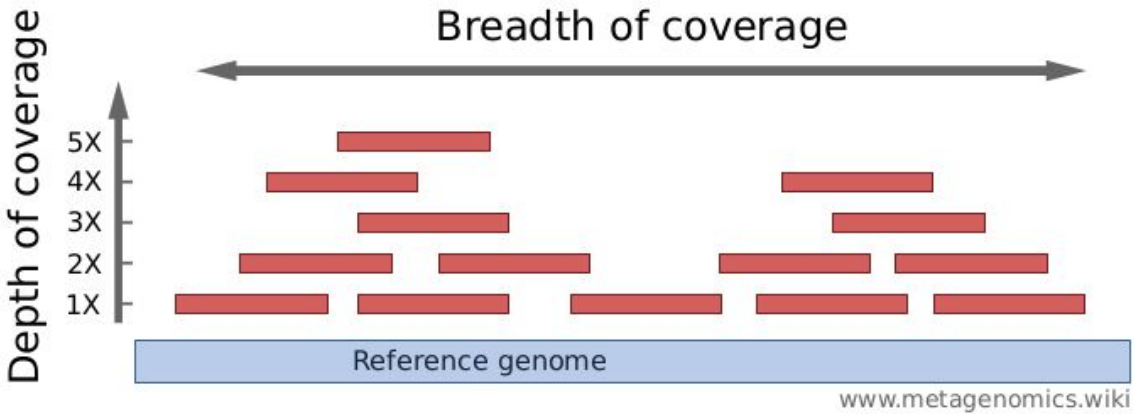
Oxford Nanopore

Ion Torrent

Sequencing Coverage

Sequencing depth or library size is the number of sequenced reads for a given sample.

Deep sequencing improves quantification and identification, but not always is more necessarily better.



For more: Sims, D., *et al*. “Sequencing depth and coverage: key considerations in genomic analyses”, Nature Reviews - Genetics(2014)

Quality Control, Followed by Quality Control

To present a high quality outcome, the results have to be reproducible and reliable.

Apply quality control checks at different stages of the analysis to ensure both:

- Raw reads
- Read Alignment
- Quantification
- Reproducibility

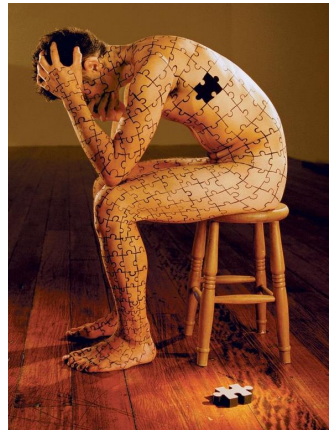
Analysis of RNAseq data

RNA-seq data undergoes multiple steps and thus several quality-control checkpoints need to be considered.

Raw read -> read alignment -> quantification -> counts -> expression measures

At each of these steps, specific checks should be applied to monitor the quality of the data.

Read alignment



An important mapping quality parameter is the percentage of mapped reads, which is a global indicator of the overall sequencing accuracy.

Other important parameters are the uniformity of read coverage on exons and the mapped strand.

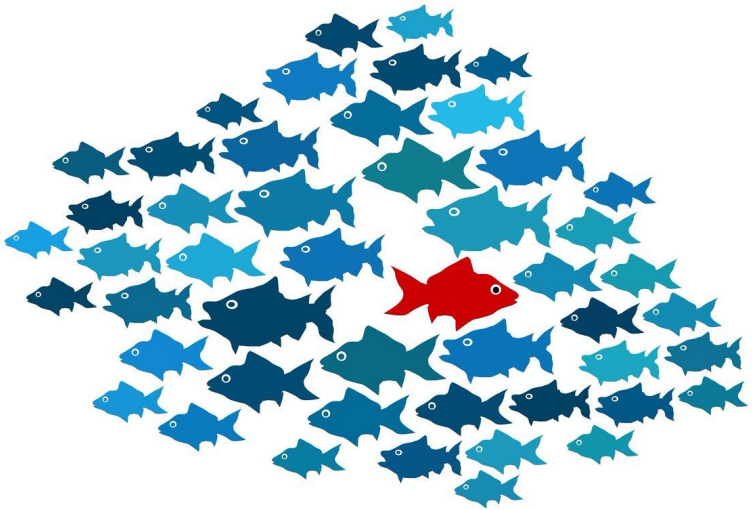
Quantification Checks

- GC content
- Gene length biases (relevant for normalization)
- Biotype composition (quality of the RNA purification step)
- Sample similarity and variance

R packages (e.g., **NOISeq**, **EDA-Seq**) provide useful plots for quality control of count data

HTSeq-count or **featureCounts** combine raw counts of mapped reads and quantify counts at the gene level.

Control on entire dataset



Check for global quality

Are the (technical and biological) replicates reproducible?

Is the distance relatively small an correlation relatively high among replicates?

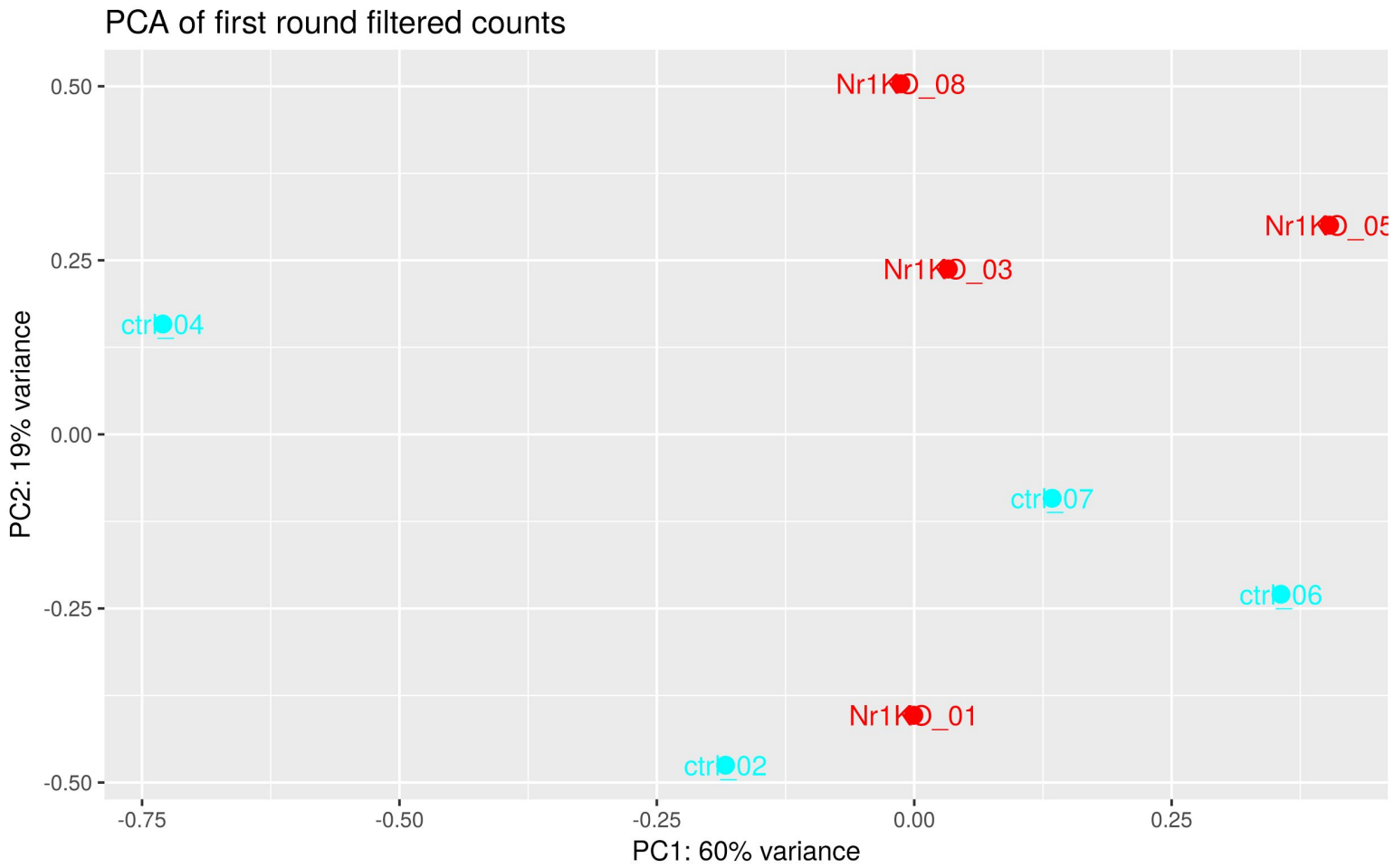
Are there outlying replicates? Why are they outliers?

Is there a separation between experimental samples?

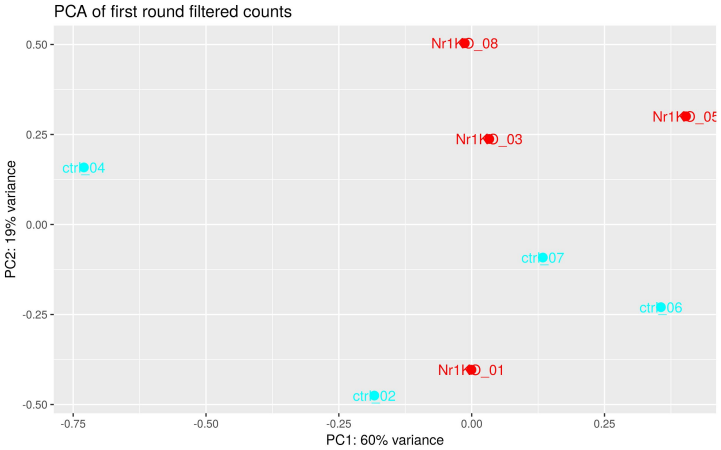
Control on entire dataset

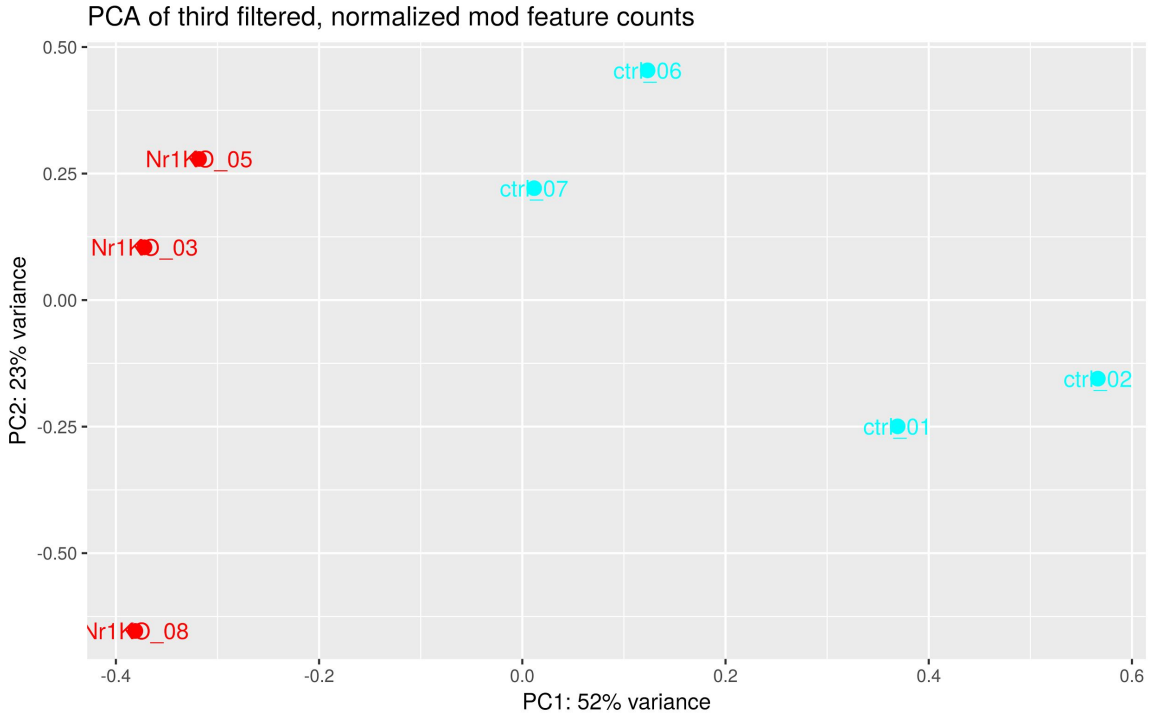
There is no clear standard to control for variation with replicates. The amount of variation depends on the heterogeneity of the experimental system, organism, and experimental design.

Principal component analysis (PCA) is a statistical procedure to investigate the relationships among samples.

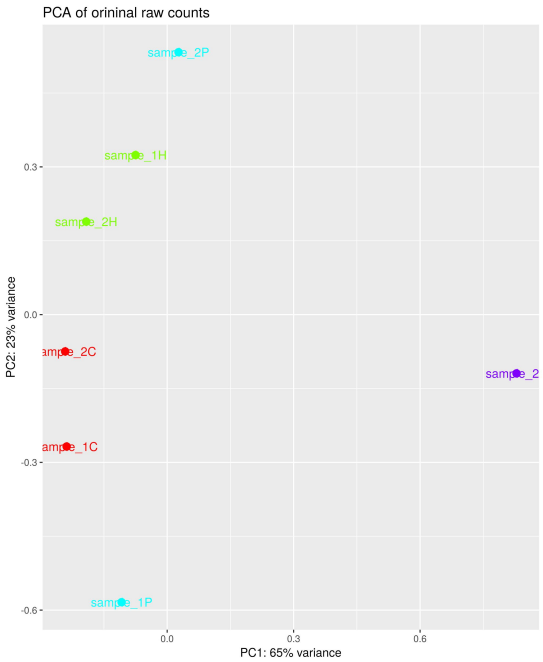

PCA of first round filtered counts
0.50
0.25
0.00
-0.25
-0.50
PC2: 19% variance
-0.75
-0.50
-0.25
0.00
0.25
PC1: 60% variance
Nr1KO_08
Nr1KO_05
Nr1KO_03
ctrl_04
ctrl_07
ctrl_06
Nr1KO_01
ctrl_02

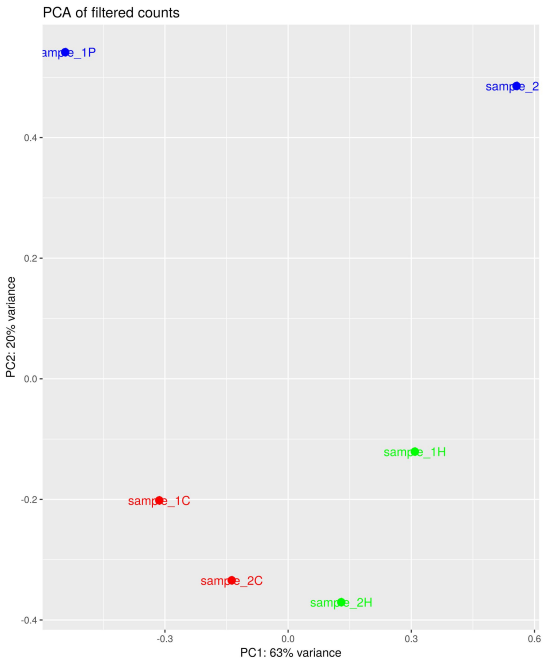
Hours Later . . .

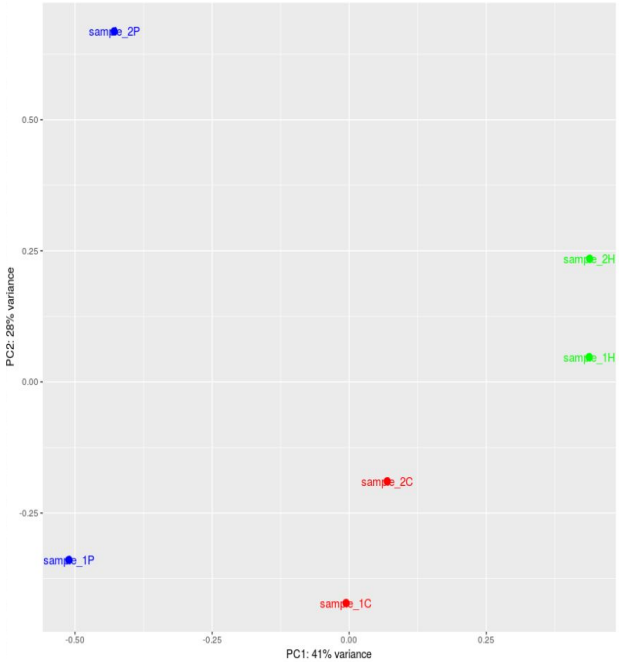

PCA of first round filtered counts
Nr1KO_08
Nr1KO_05
Nr1KO_03
ctrl_04
ctrl_07
ctrl_06
Nr1KO_01
ctrl_02
0.50
0.25
0.00
-0.25
-0.50
PC2: 19% variance
-0.75
-0.50
-0.25
0.00
0.25
PC1: 60% variance

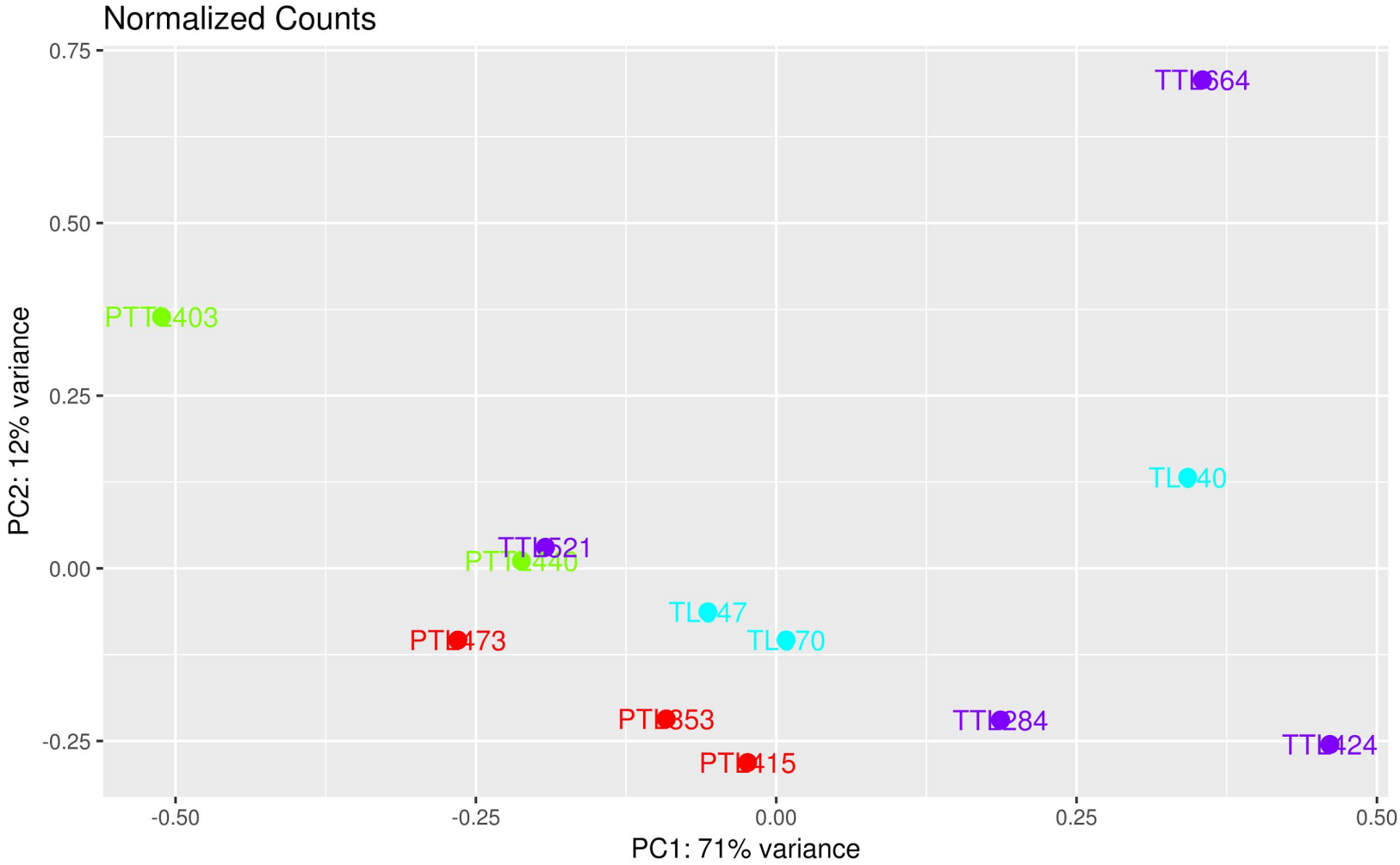

PCA of third filtered, normalized mod feature counts
ctrl_06
Nr1KO_05
ctrl_07
Nr1KO_03
ctrl_02
ctrl_01
Nr1KO_08
0.50
0.25
0.00
-0.25
-0.50
PC2: 23% variance
-0.4
-0.2
0.0
0.2
0.4
0.6
PC1: 52% variance

Preprocessing







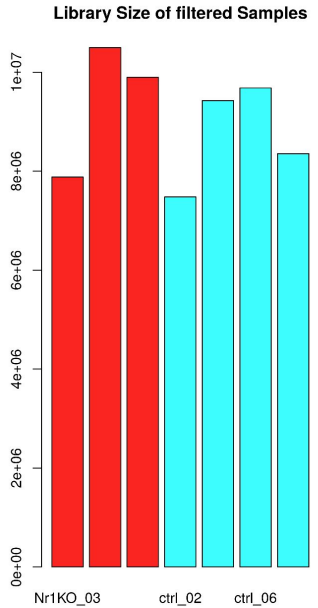

Normalized Counts
0.75
0.50
0.25
0.00
-0.25
PC2: 12% variance
-0.50
-0.25
0.00
0.25
0.50
PC1: 71% variance

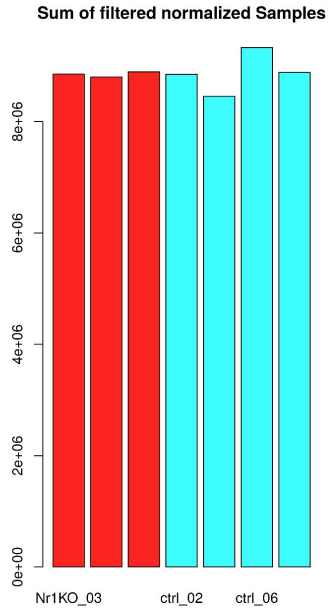
RNAseq data Normalization

Many different normalization methods: RPKM, FPKM, TPM, CPM, TMM, ...

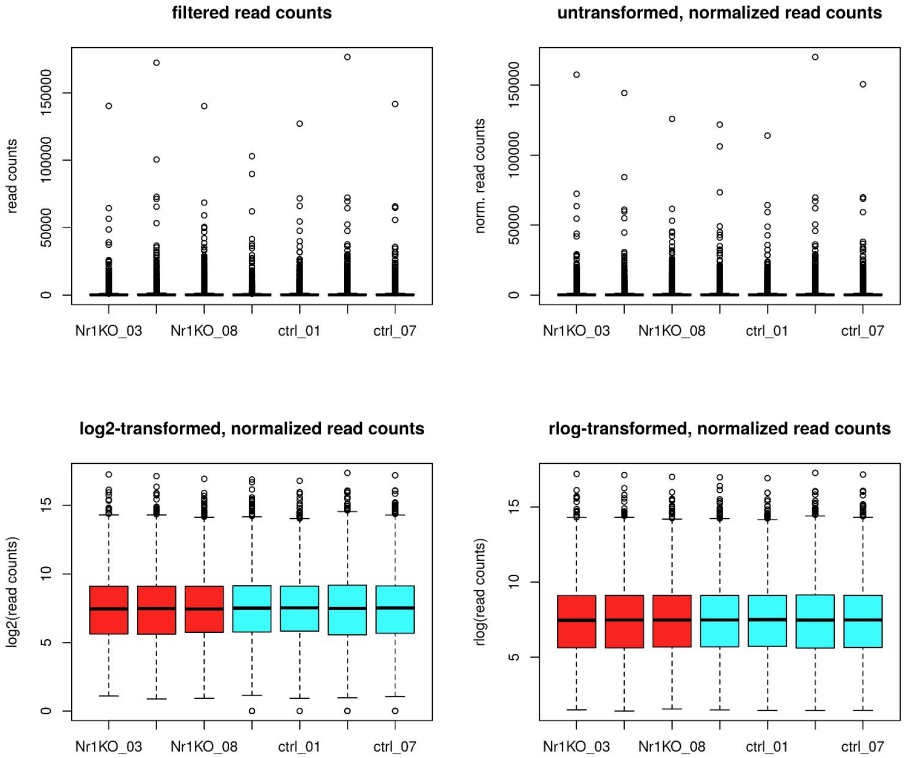
TPMs (Transcripts Per Kilobase Million) are normalized for gene length and library size and denotes the proportion of counts within that sample of a gene.

TPM values are used for clustering, heatmaps, and manuscripts, but NOT for differential gene expression analysis.





RNAseq data Normalization



Not all variation is bad.

Be careful transforming data, as biological variation might be massaged away!!

Yay we are Finally ready to analyze the data

The most common application of RNAseq is to estimate gene and transcript expression and to quantify differential expression between tissues/conditions at the genome level.

Downstream analyses are somewhat similar to microarray data, but don’t forget about the data distribution.

Microarray data often transform into a normal distribution, whereas RNAseq has a negative binomial distribution.

Differential Gene Expression Analysis

The most popular R packages are: **edgeR** and **DESeq2**

Both applications require the raw filtered counts as their input.

Correcting for gene length is not necessary when comparing changes in gene expression within the same gene across samples.

Recommendation: At least three biological replicates.

The **NOISeq** R package includes diagnostic plots to identify origins of biases in RNA-seq data.

Whole System networks

Very current topic in life sciences.

Network analysis is a versatile tool and can present a variety of interactions and organizations at a systems level.

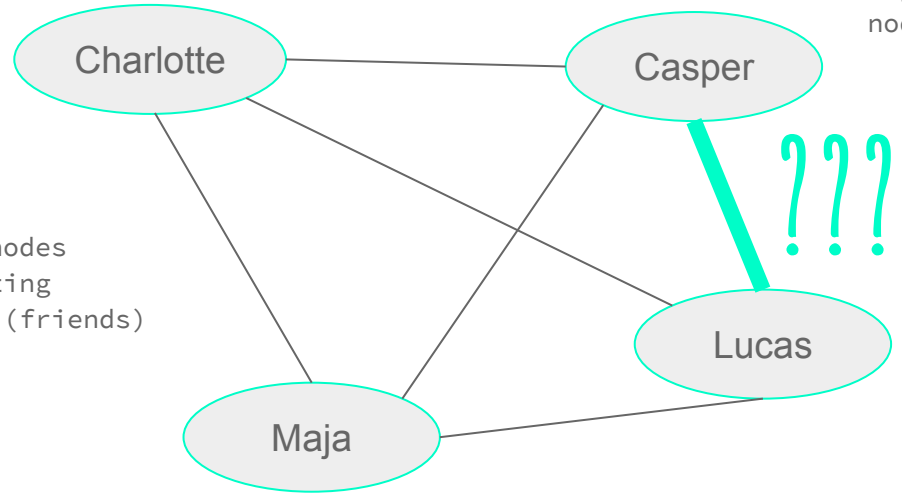
The study of node-edge graphs:

- Node = elements in the system
- Edges = a type of relationship/association/interaction between elements

Conceptual Example

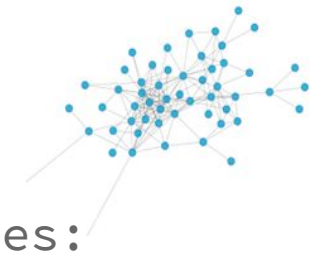
How does Facebook recommend people you might know?

Members of Facebook are represented by network nodes



Edges between nodes represent existing Facebook links (friends)

Biological Networks - a node-edge graph



Nodes are elements in the system, molecular entities:

- genes, DNA, RNA, peptides, proteins, metabolites, transcription factors, other biomolecules

Edges are relationships/associations between elements:

- interaction, regulation, transformation, activation, similar in expression behavior or value, sequence similarity

Can reveal organizational principles and structure of biological systems at the cellular level.

Networks in Biology

1970, Kauffman was the first to use Boolean networks as a biological network modeling paradigm.

Use in biology has increased exponentially.

Established role in research: Systems Biology, Network Biology, Network Medicine.

Have been used for structural and functional studies of genes, proteins, and metabolites.

Network model construction is specific to research question and dataset.

Biological Network Models

Protein-Protein Interaction (PPI) Networks

Co-expression Networks

Metabolic Networks

Regulatory Networks

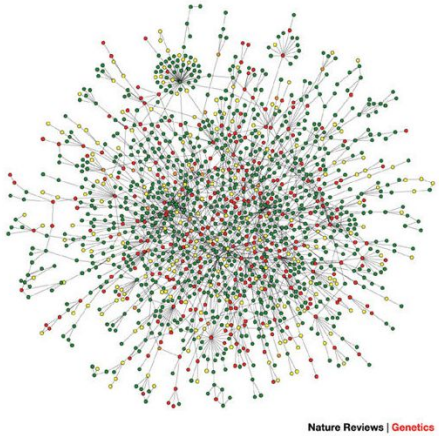
Signaling Network

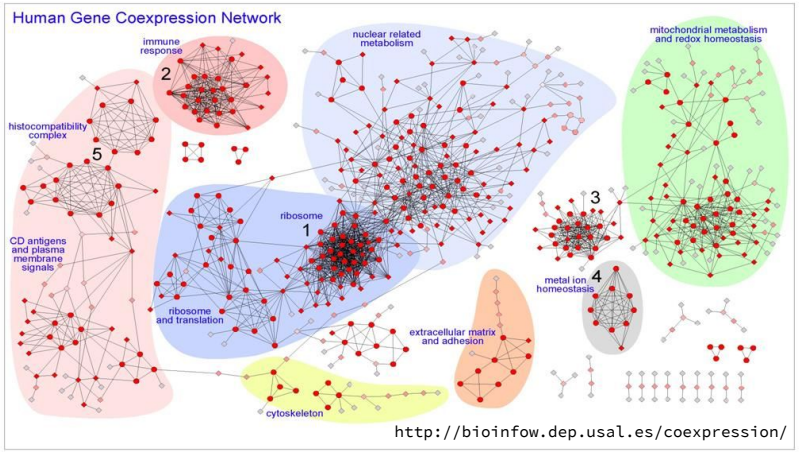
Evolutionary Tree of Life

...

Figure 2 : Yeast protein interaction network.

From: Network biology: understanding the cell's functional organization

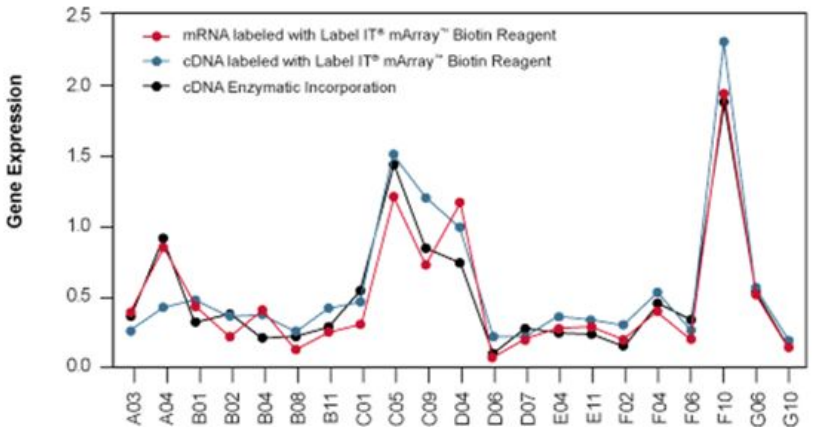




Co-expression Networks

Measurement of expression levels of individual genes (or proteins, etc.) over a time-series or/and multiple conditions.

Genes with similar expression are hypothesized to have common functions.

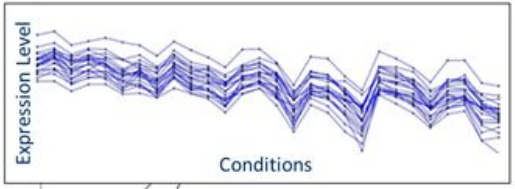


petal

Expression Data

Network Model

Constructing gene co-expression networks



Group of tight expression profiles

Groupings via Cliques

Tutorial

```
source("https://bioconductor.org/biocLite.R")
biocLite("DESeq2")

install.packages(""ggplot2")

install.packages("ggfortify")
```

See file

References

Conesa, A, *et al.* “A survey of best practices for RNA-seq data analysis”, Genome Biology(2016)

Barabási, A-L and Oltvai, ZN. “Network biology: understanding the cell's functional organization”, Nature Reviews - Genetics(2004)

Prieto, C, *et al.* “Human Gene Coexpression Landscape: Confident Network Derived from Tissue Transcriptomic Profiles” PLoS One(2008)

Petereit, J, *et al.* “petal: Co-expression network modelling in R”, BMC Systems Biology(1016)