Leids Universitair Medisch Centrum

Introduction to Transcriptomics

Molecular Data Science: from disease mechanisms to personalized medicine

Rodrigo C de Almeida

Biomedical Data Sciences,
Molecular Epidemiology

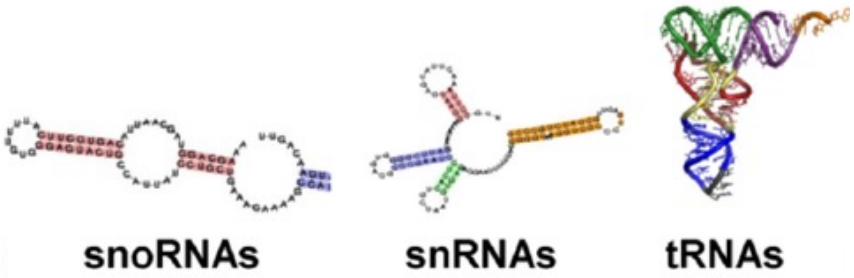
r.coutinho_de_almeida@lumc.nl

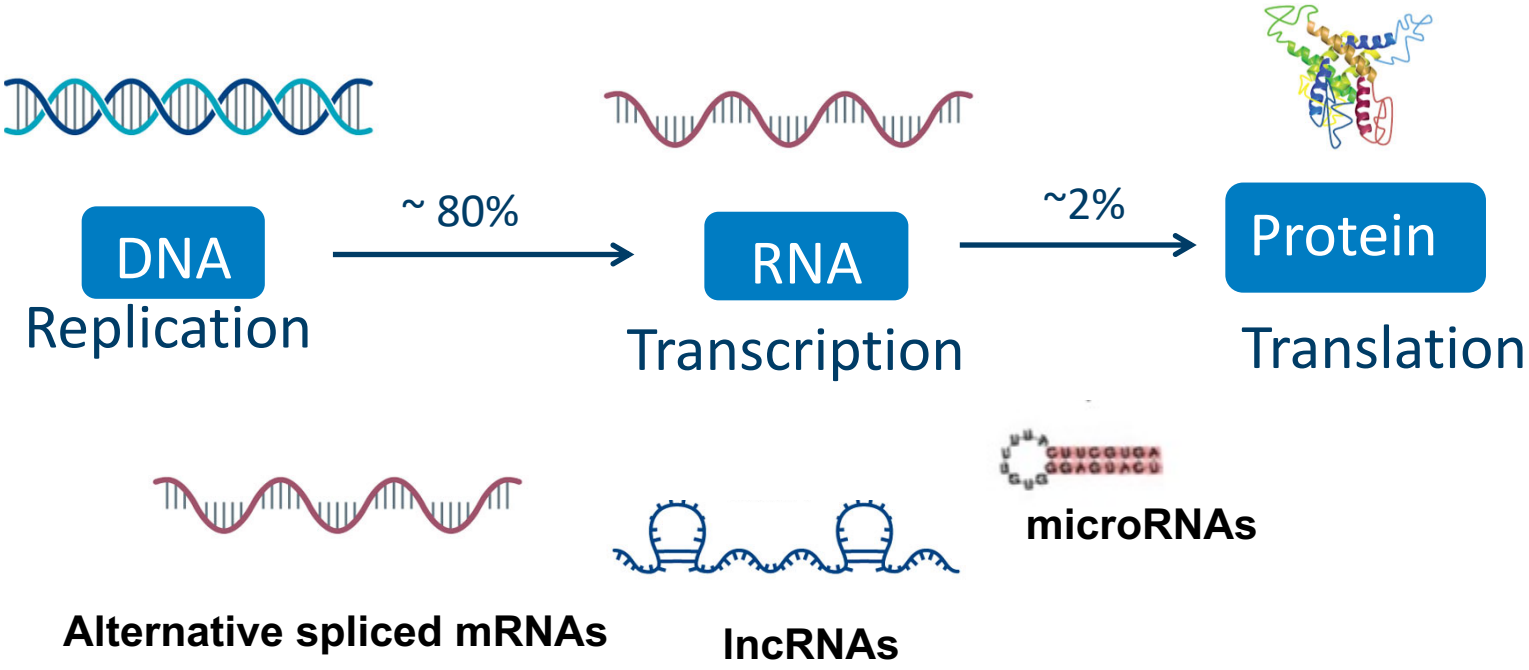
@rodcoutalmeida

Outline

- Transcriptome;
- Methods to study the transcriptome;
- RNA-seq;
- Differential expression analysis;

The Central Dogma of Molecular Biology





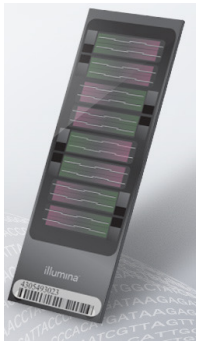
Transcriptomics

The **transcriptome** is the complete set of transcripts (mRNA, rRNA, tRNA, and non-coding RNA) in a cell, and their quantity, for a specific developmental stage or physiological condition. Wang et al., Nat Rev 2011

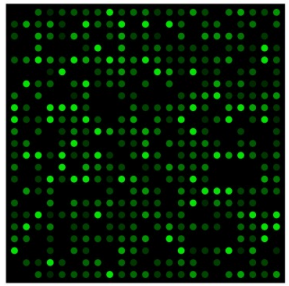
What can the transcriptome tell us?

- Where and when each gene is expressed in the cells and tissues of an organism;
- Changes in the normal level of gene activity in the transcriptome may reflect of contribute to disease;
- Researchers can get a genome-wide picture on what genes are active in a tissue;

Two major technologies to study the transcriptome

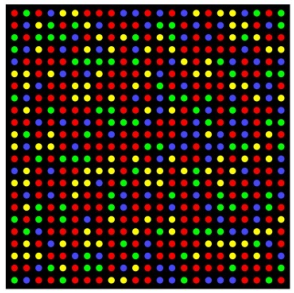


Microarray chip



Microarray

Sequencing flow cell



RNA-seq

Microarray vs RNA-seq

Microarray

- Based in prior knowledge (probes);
- Higher througput;
- Data analysis more user friendly compare to RNA-seq;

RNA-seq

- Not limited by prior knowledge of the genome (full transcriptome);
- Higher dynamic range of expression levels over which transcripts can be detected (> 8000-fold range);
- Higher sensitivity for genes expressed either at low level;
- Lower technical variation and higher levels of reproducibility;
- Gives single base resolution about transcriptional features (alternative splicing and allele-specific expression);

Applications of RNA-seq

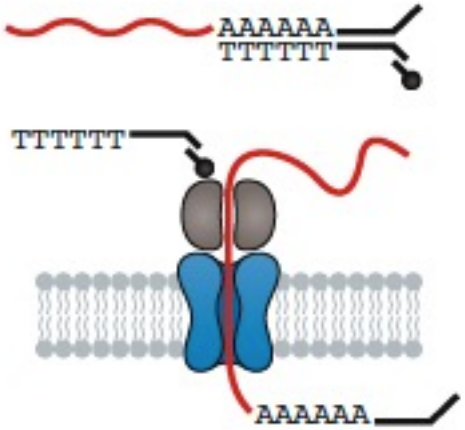
- Gene expression profiling between samples;
- Diagnostics through expression profiling;
- Identify alternative splicing events;
- Allele-specific expression, SNPs and gene fusions;
- Exon dosage (quantification);
- Identify non-coding RNAs (eg. microRNAs);
- Identification of human pathogens;

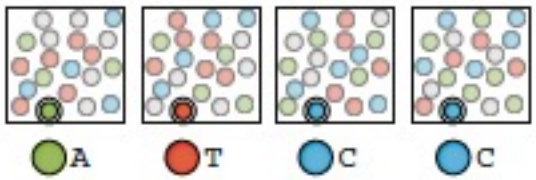
Main sequencing technologies for RNA-seq

Illumina

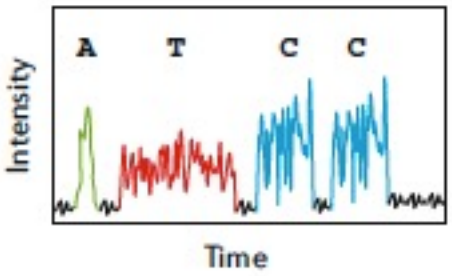
Pacific Biosciences

Oxford Nanopore

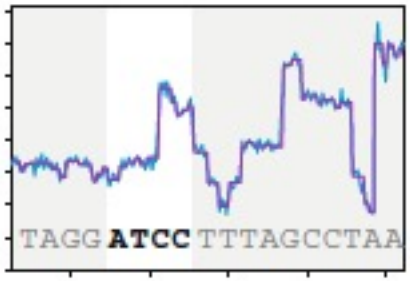




Short-read



Long-read

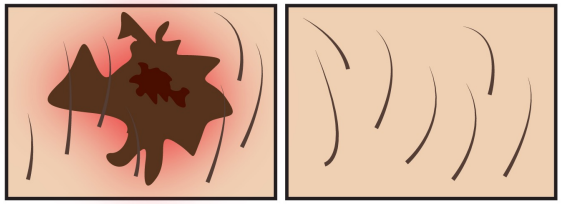


Direct

Stark et al., Nat Rev Gen 2019

Typical RNA-seq experiments (Short-read)

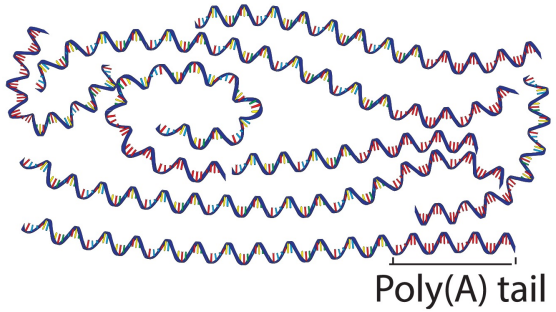
Samples of interest



Condition 1 (e.g. tumor)

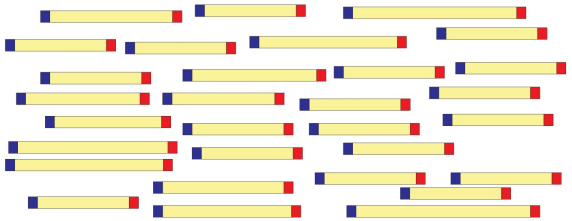
Condition 2 (e.g. normal)

Isolate RNAs

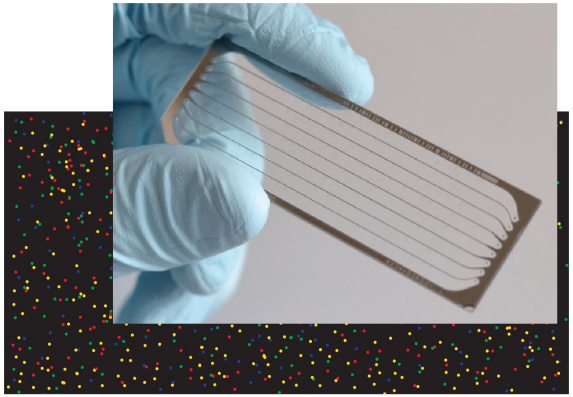


Poly(A) tail

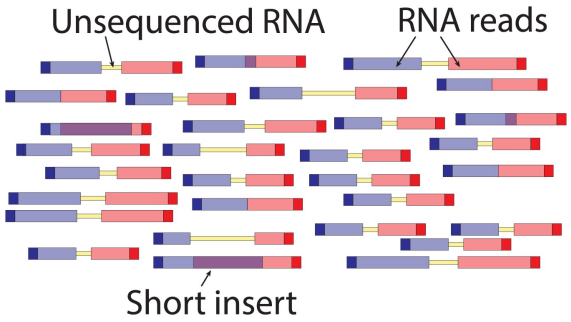
Generate cDNA, fragment, size select, add linkers



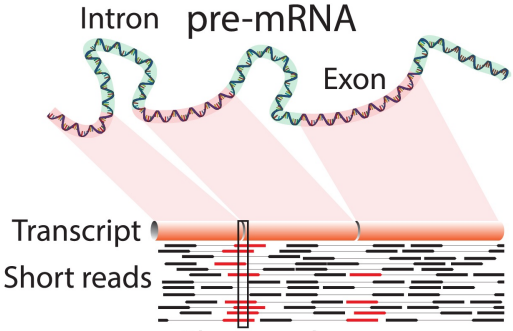
Sequence ends



100s of millions of paired reads
10s of billions bases of sequence



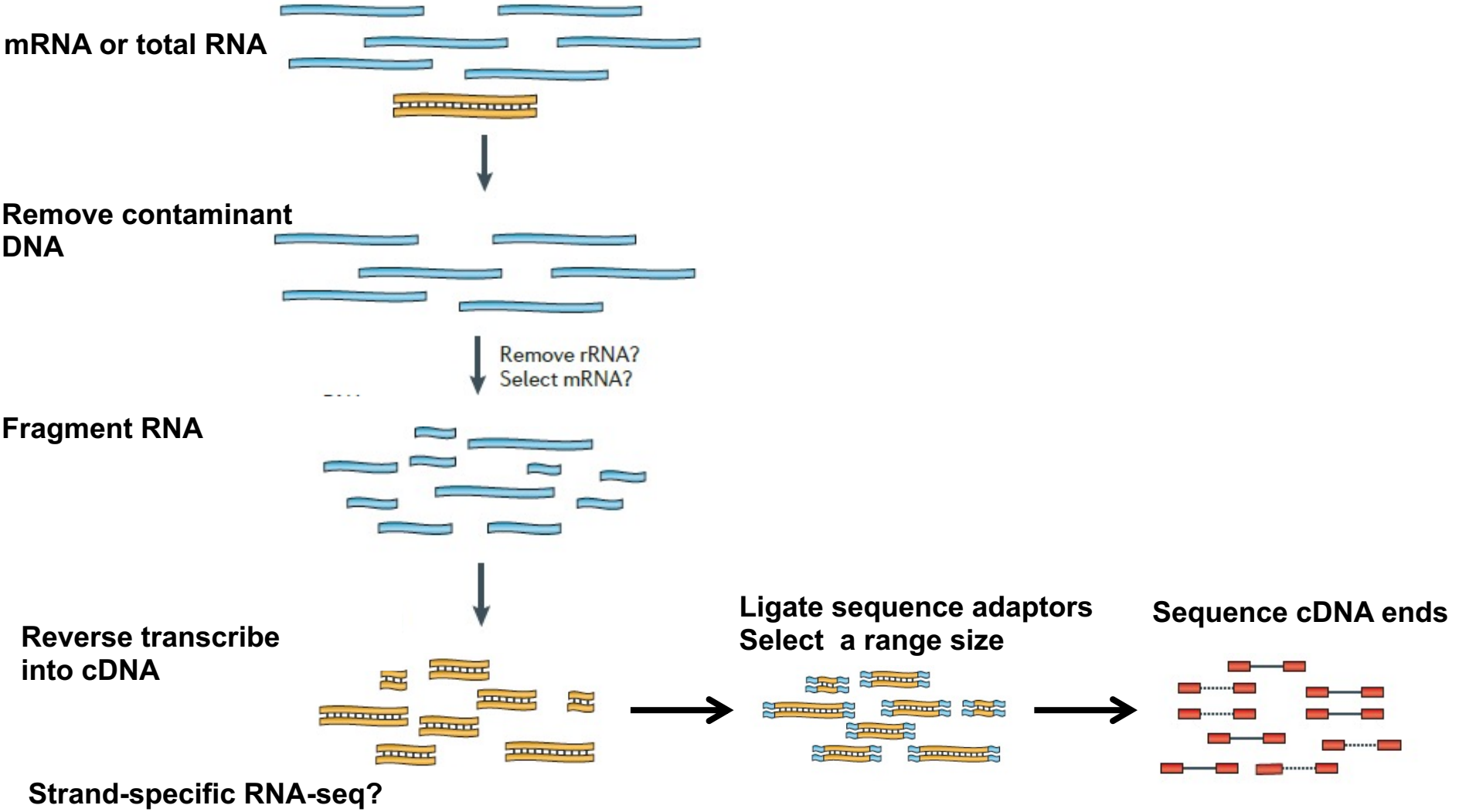
Map to genome, transcriptome, and predicted exon junctions



Downstream analysis

Source: Wikipedia

RNA-seq data generation



Adapted from Martin and Wang., Nat.Rev.Gen. 2011

~ 2Gb of expression data, and now?

FASTQ file

```
@22:16362385-16362561W:ENST00000440999:2:177:-40:244:S/2
CCAGCCCACCTGAGGCTTCTTTTTCCTTCCCAAGCCACATCACCATCCTGGTGGAACTCTCCTGTGAGGACAGCCA
+
GGFF<BB=>GBGIIIIIIIIIIIIIIEGEHGHHIIIIIIIIHFHBB2/:=??EGGGEGFHHIHHEDBD?@@DDHHD
@22:16362385-16362561W:ENST00000440999:3:177:-56:294:S/2
GCGTGAGCCACAGGGCCCAGCCCACCTGAGGCTTCTTTTTCCTTCCCAAGCCACATCACCATCCTGGTGGAACTCT
+
@=ABBBBIIIIIIIIHHGGGGIIDBDIIIIIIGIIIIHIIIIHFDD@BBDBGGFIDEE8DCC/29>BGFCGHHHGF
@22:16362385-16362561W:ENST00000440999:4:177:137:254:S/1
TCACCATCCTGGTGGAACTCTCCTGTGAGGACAGCCAAGGCCTGAACTACCTGCaGTGGGGAGCACCTCAGGGTTT
+
DDGBBCGGGIGGGBDDDHIIGGDGD77=BDIIIIIIIIFHHHHIIIHEFFHGGDD8A>DEGHHIFDDHH8@BEDDI
@22:16362385-16362561W:ENST00000440999:5:177:68:251:S/2
AGGGTTTGCCCAGGCAACCAGCCAGCCCTGGTCCAAGGCATCCTGGAGCGAGTTGTGGATGGCAAAAAGACNCGCC
+
HIGHIHFHEGE4111:.;8@?@HDIIIIIIIEGGIHHHIIGA?=:FIIIDD8.02506A8=AC##############
@22:16362385-16362561W:ENST00000440999:6:177:348:453:S/1
AAGGCCTGAACTACCTGCGGTGGGGAGCACCTCAGGGTTTGCCCAGGCAACCAGCCAGCCCTGGTCCAAGGCATCC
+
B9?@8=42:E@GDEDIIIIIGGHIIIFBEEAGIIDIIDHHGGHIIEGEIIIIIHIHFHFFEEFGGGGGB88>:DGH
@22:51205934-51222090C:ENST00000464740:132:612:223:359:S/2
GGAAGTATGATGCTGATGACAACGTGAAGATCATCTGCCTGGGAGACAGCGCAGTGGGCAAATCCAAACTCATGGA
+
IIEHHHHHIIIIIIIHGGDGHHEDDG8=;?==19;<<>D@@GGGIIHIIHGGDDHGBA=ABEG@@DFCCAA<:=>8
@22:51205934-51222090C:ENST00000464740:125:612:-1:185:S/1
TGGAGTGCGCTGCGGCGCGAGCTGGGCCGGCGGGCGTGGTTCGAGAGCGCGCAGAGTCCAGACTGGCGGCAGGGCC
+
HHIIIHIDGG@;=@GIIIIIDDGBBBEDB@8>5554,/':9B@@C?==@1:2@?=GG=;<HHHHGIIHHEC-;;3?
```

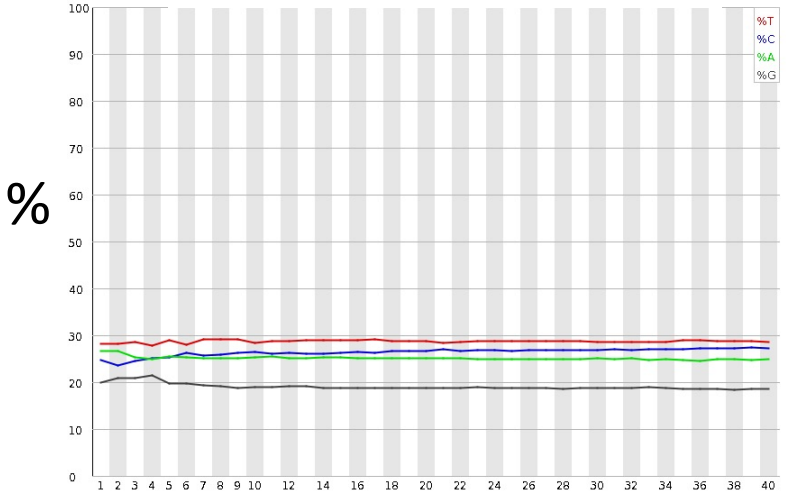
RNA-seq analysis

- Quality Control;
- Alignment and Quantification;
- Normalization;
- Differential expression;
- Pathway analysis

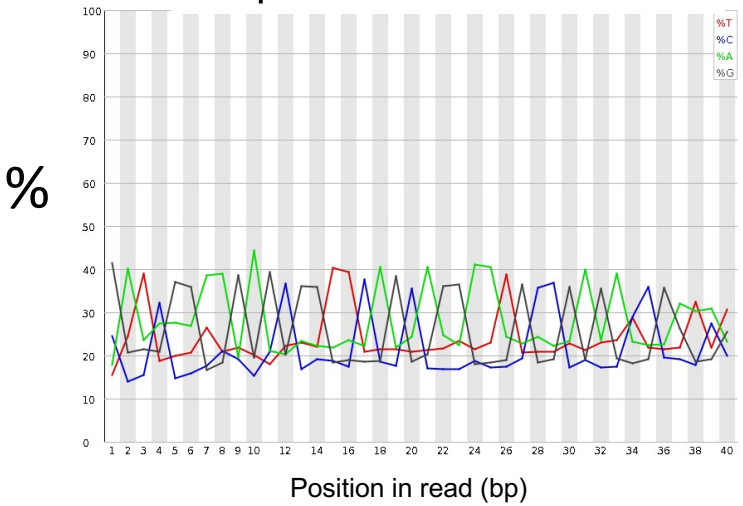
QC: Raw Data

Sequence bias

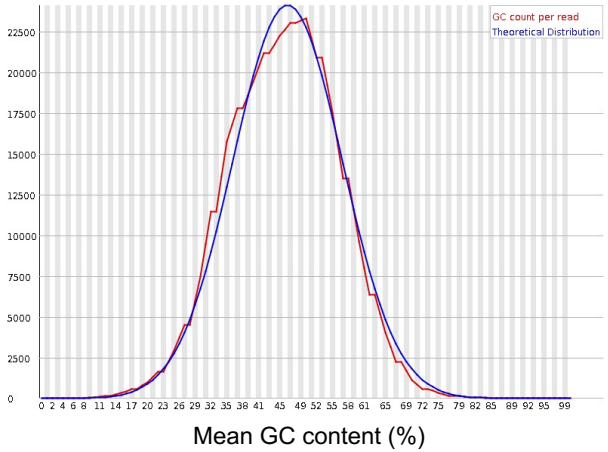
Sequence across all bases



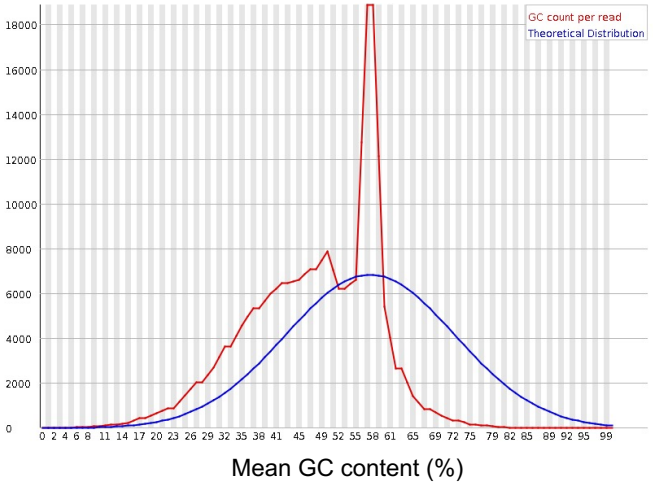
Sequence across all bases



GC distribution over all sequences



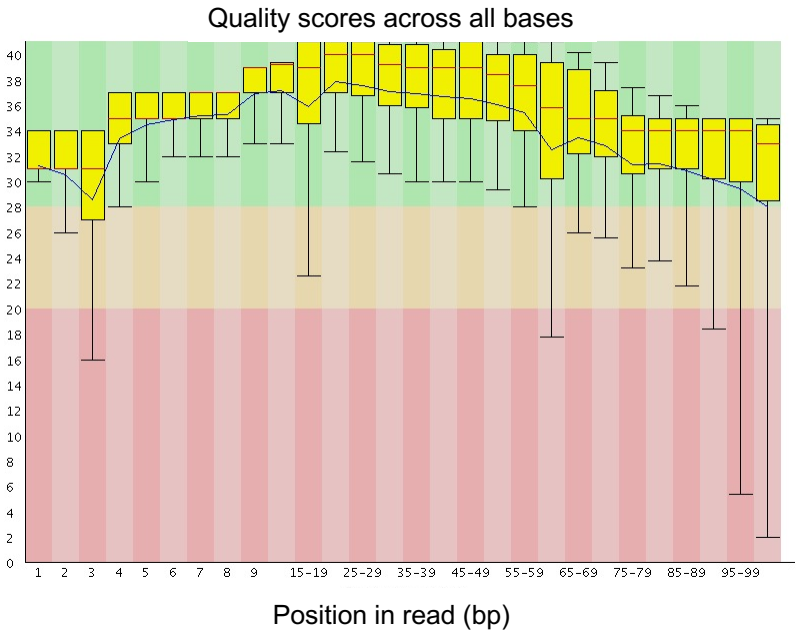
GC distribution over all sequences



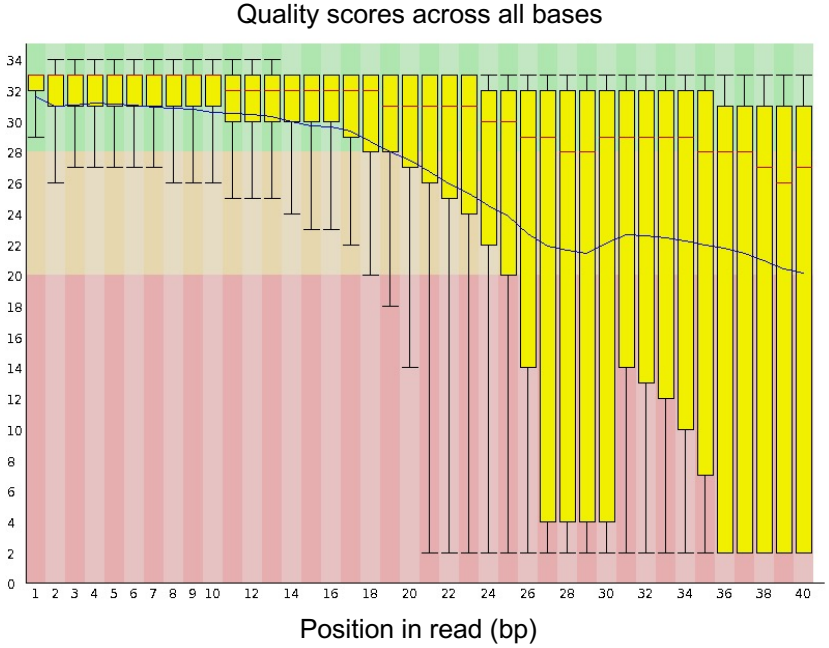
Quality Control (QC)

FastQC

Per base sequence quality



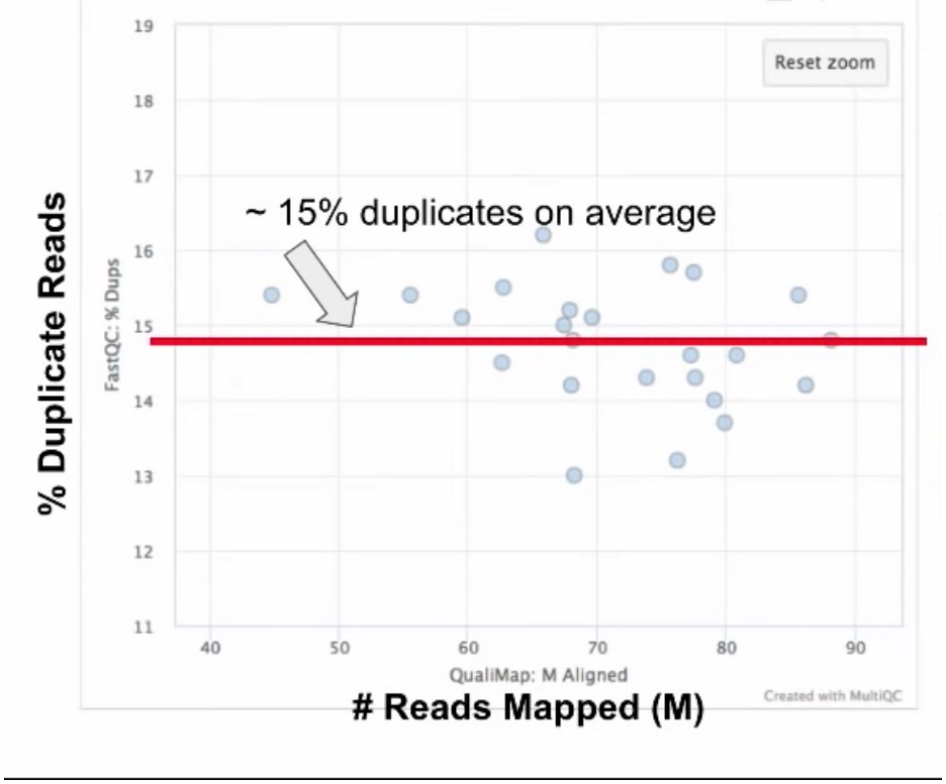
Good sample



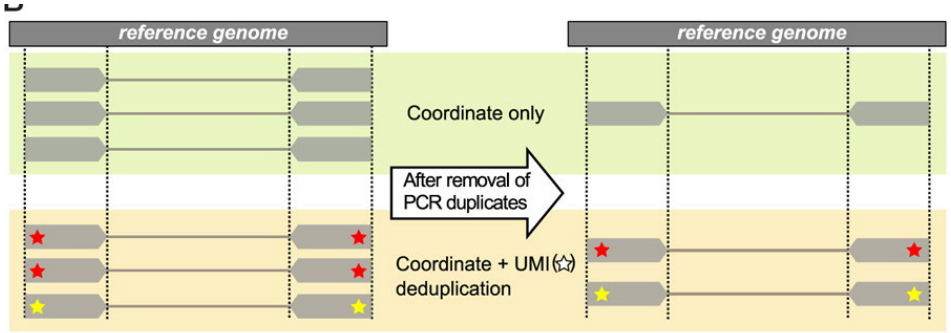
Bad sample

Duplicate reads

%PCR



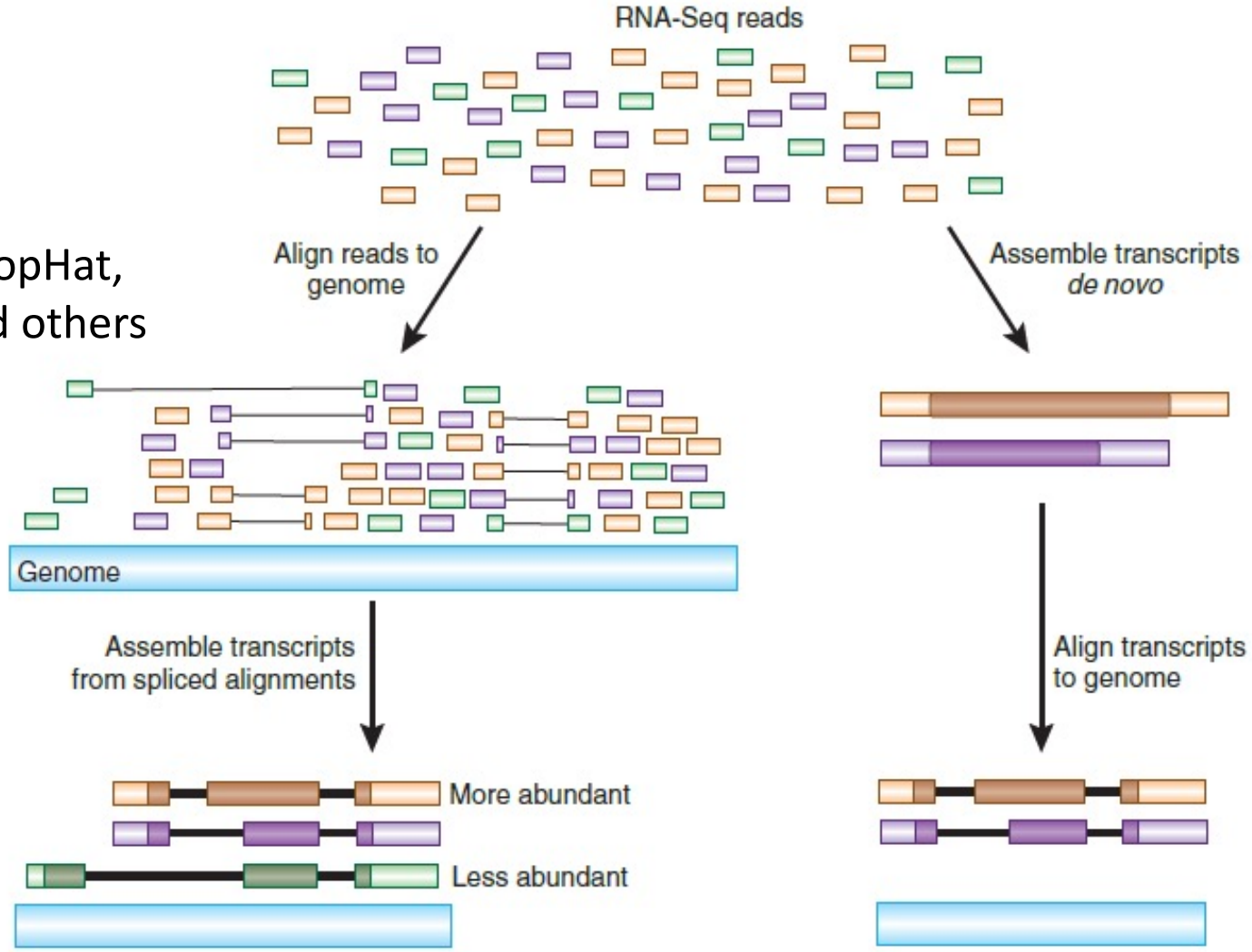
Partially solve using UMI



Hong et al., Biotechniques 2017

RNA-seq align and assemble

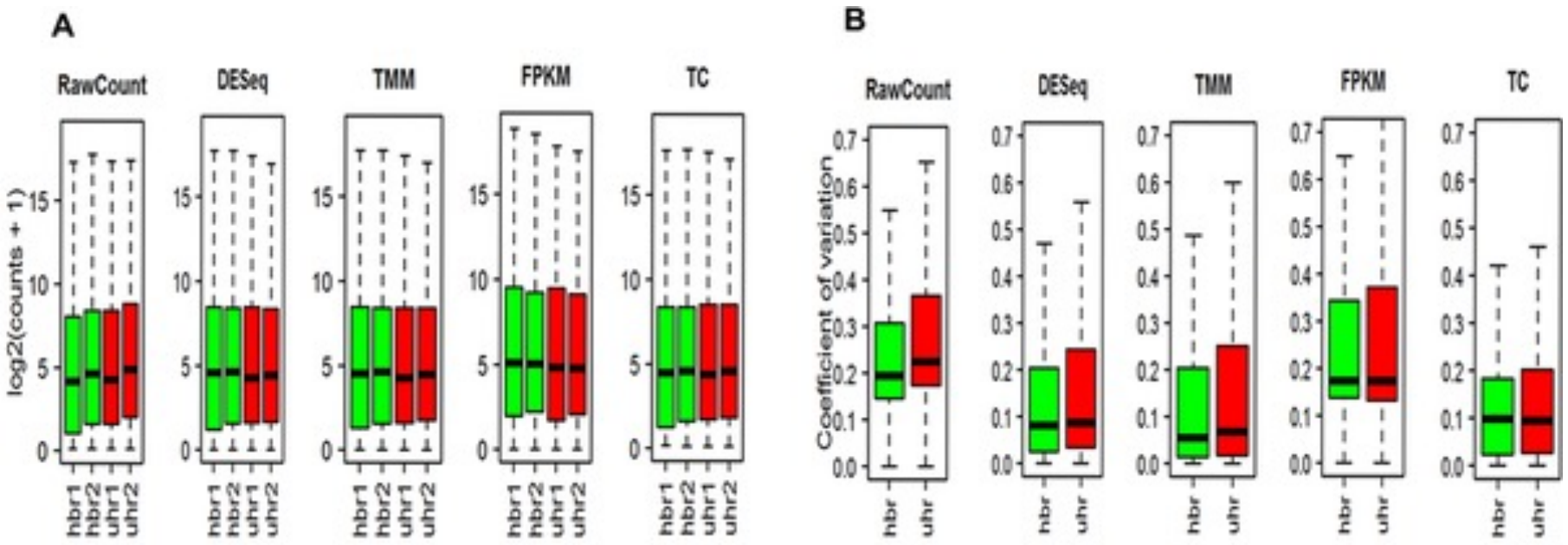
GSNAP, TopHat, STAR and others



Trinity, Kallisto and others

Haas BJ and Zody MC. , Nat.Biotech. 2010 .

Normalization required



Li X, , et al. (2017) PLOS ONE 12(5): e0176185

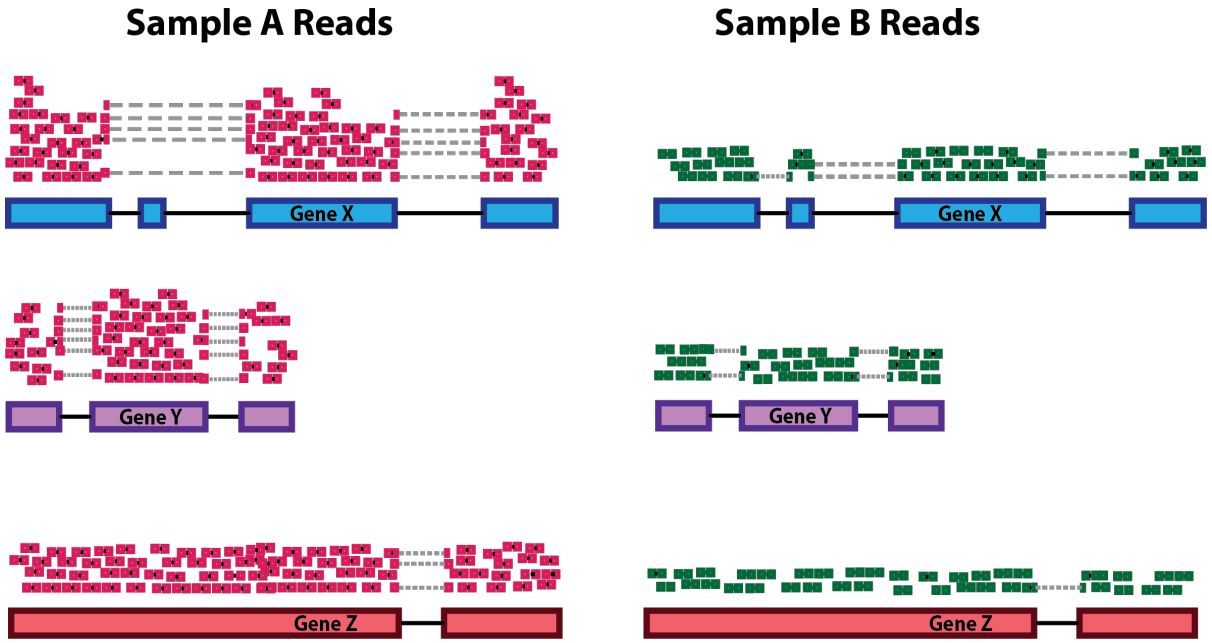
Normalization Methods

Necessary due to variable sequencing depth of RNA-Seq samples;

- Normalization for library size more important than gene length;
- Normalization for gene length only relevant for comparing expression across different genes/features;
- Simple size normalization can be skewed by highly overrepresented RNAs;

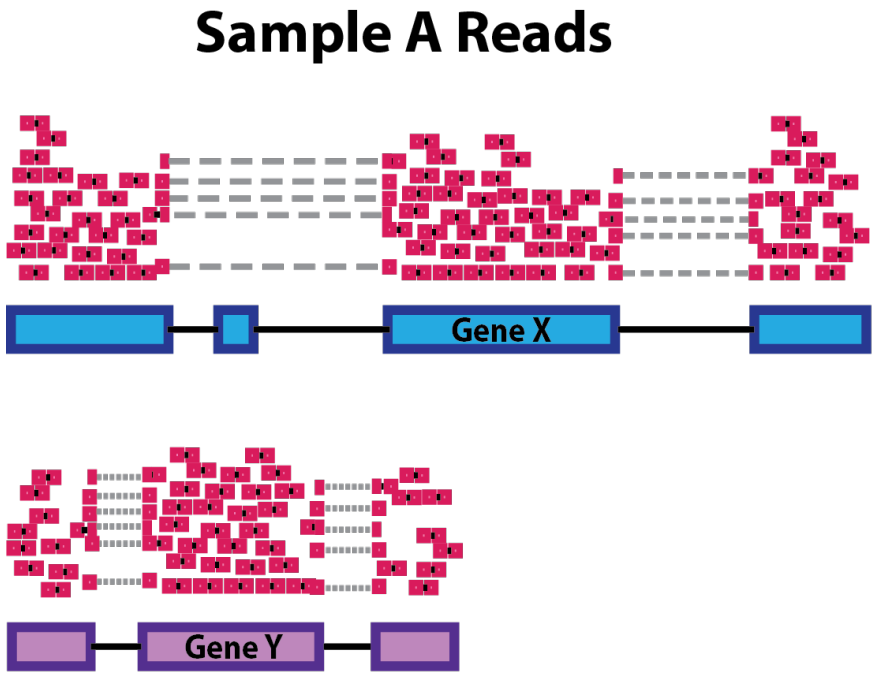
Sequencing depth

Accounting for sequencing depth is necessary for comparison of gene expression between samples. In the example below, each gene appears to have doubled in expression in Sample A relative to Sample B, however this is a consequence of Sample A having double the sequencing depth



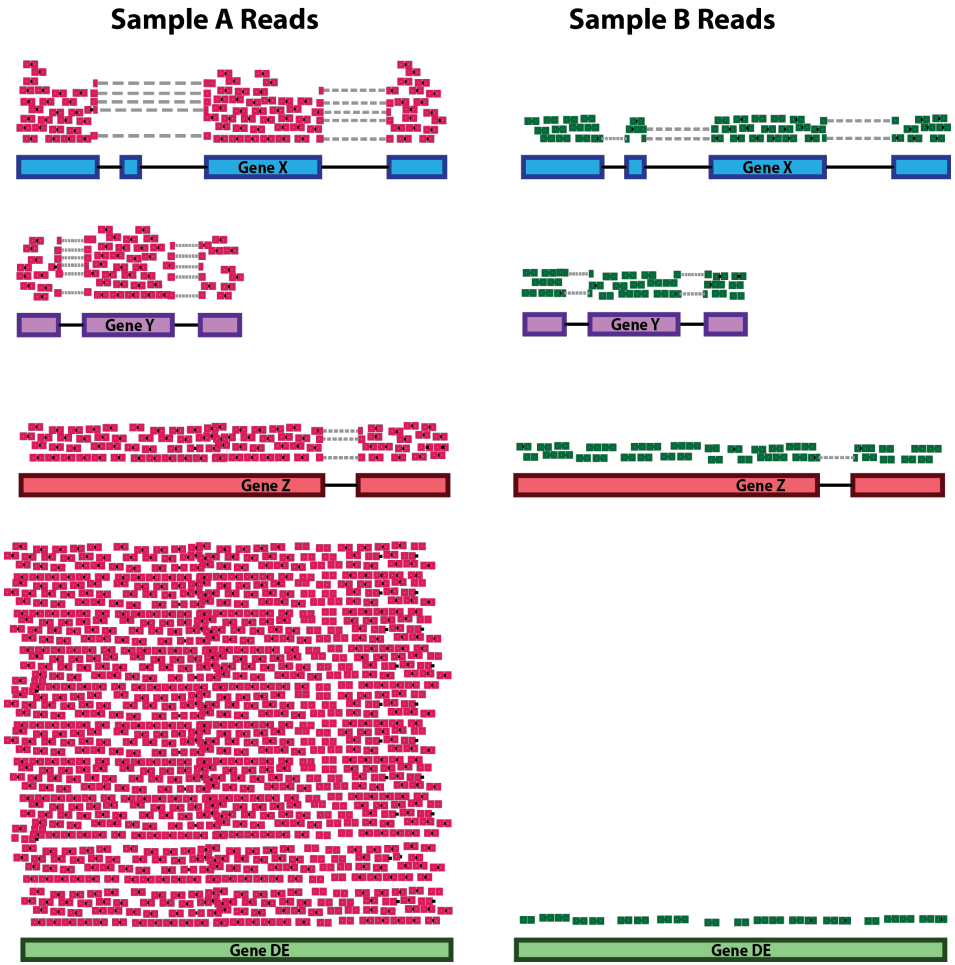
Gene length

Accounting for gene length is necessary for comparing expression between different genes within the same sample. In the example, Gene X and Gene Y have similar levels of expression, but the number of reads mapped to Gene X would be many more than the number mapped to Gene Y because Gene X is longer.



RNA composition

Accounting for RNA composition is recommended for accurate comparison of expression between samples and is particularly important when performing differential expression analyses.



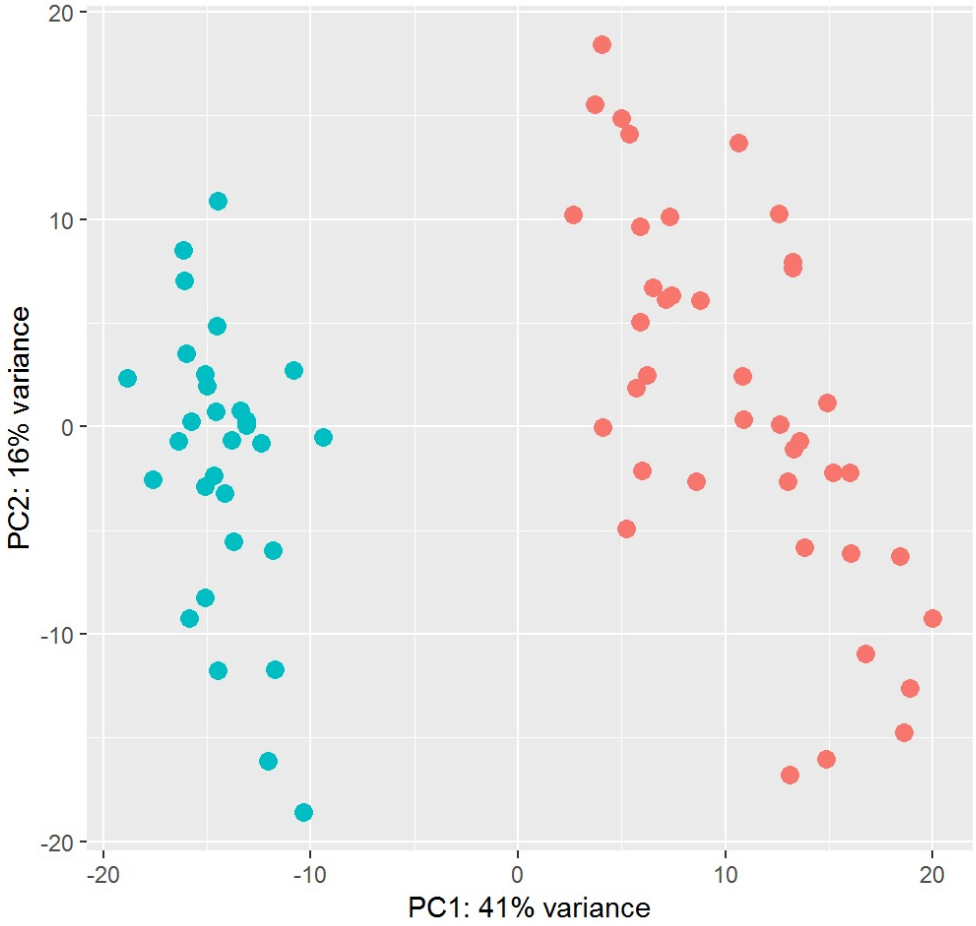
Examples of common normalization methods

- Log and relative log transformation;
- Variance-stabilizing transformation;

- RPKM (reads per kb per million mapped reads) - not for statistical testing;
- FPKM (fragment per kb per million mapped reads);
- CPM (counts per million reads);
- TMM (trimmed mean of M values);

- Median ratio method (size factor);
- Quantile normalization methods;

Batch effect



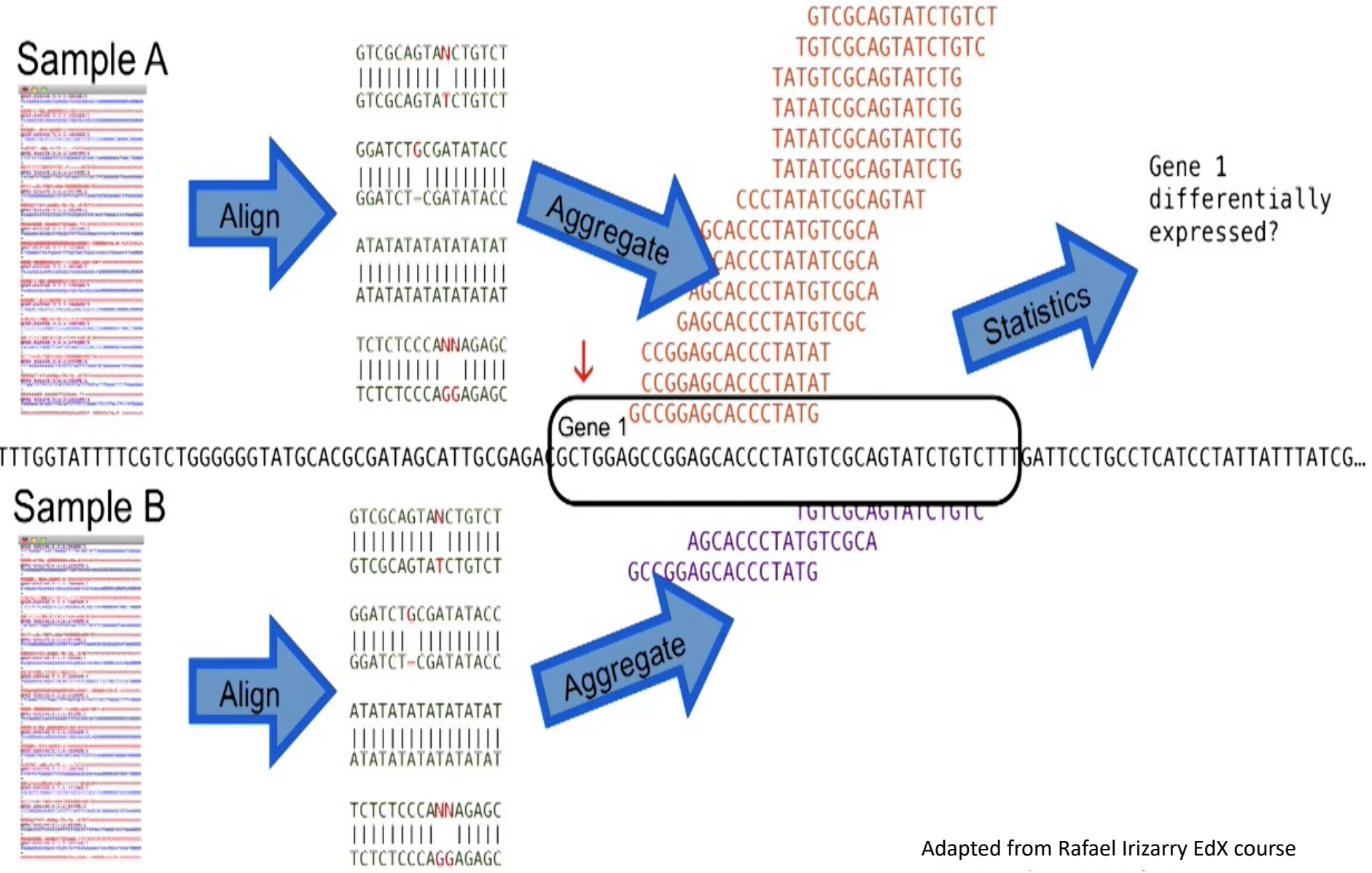
To remove:

- Include batches as covariant.
- Surrogate variables (hidden batch effects).
- Adjust for known variation (batches).

Sofwares:

Combat-Seq, limma, SVA, and others

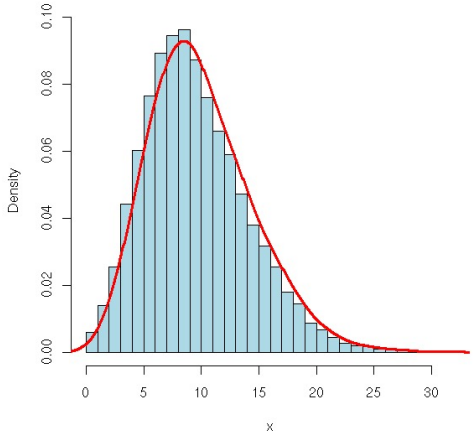
RNA-seq data analysis overview



Adapted from Rafael Irizarry EdX course

Statistical Testing in DEG Analysis

- Most statistical methods for RNA-Seq DEG analysis use negative binomial distribution (NB) or Poisson distribution along with modified statistical tests based on that;



- **The multiple testing issue**:
- False Discovery Rates (FDRs) using the Benjamini-Hochberg method;
- Bonferroni correction;
- **DESeq2**: NB with raw counts; Wald test, generalized linear model
- edgeR: NB with raw counts; empirical Bayes for estimating dispersion; generalized
- Linear model with likelihood ratio tests or quasi-likelihood F-tests

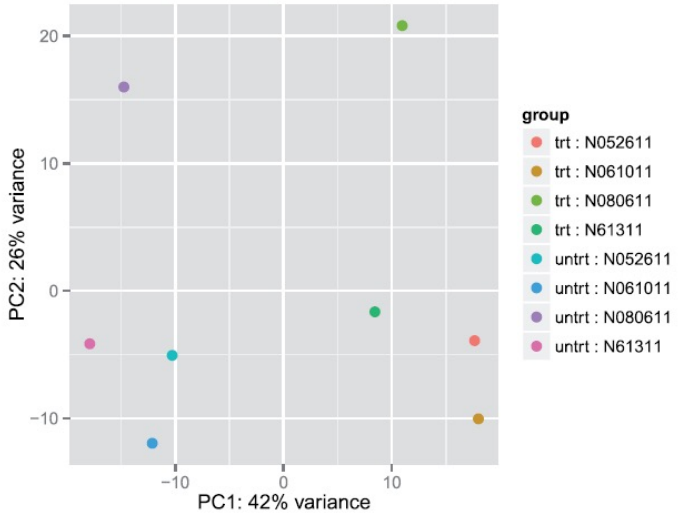
DESEQ2 Statistics

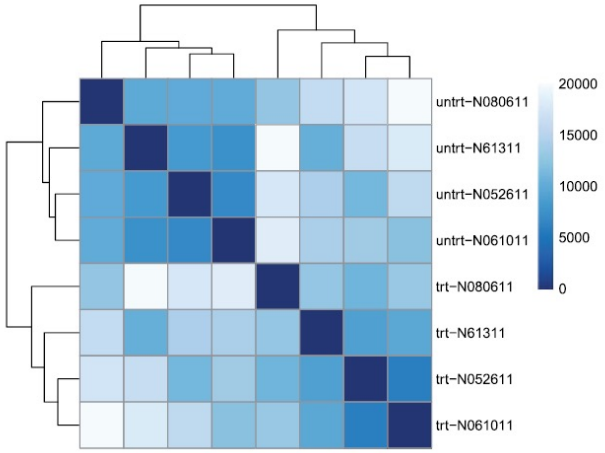
- Are the counts we see for gene A in condition 1 consistent with those for gene A in condition 2?

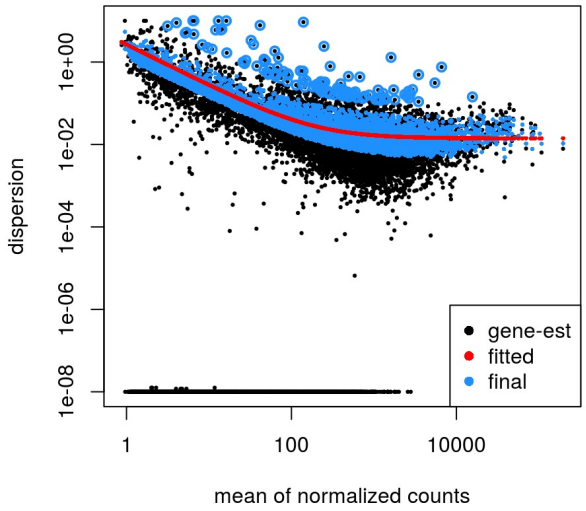
- Size factors
  - Estimator of library sampling depth
  - More stable measure than total coverage
  - Based on median ratio between conditions

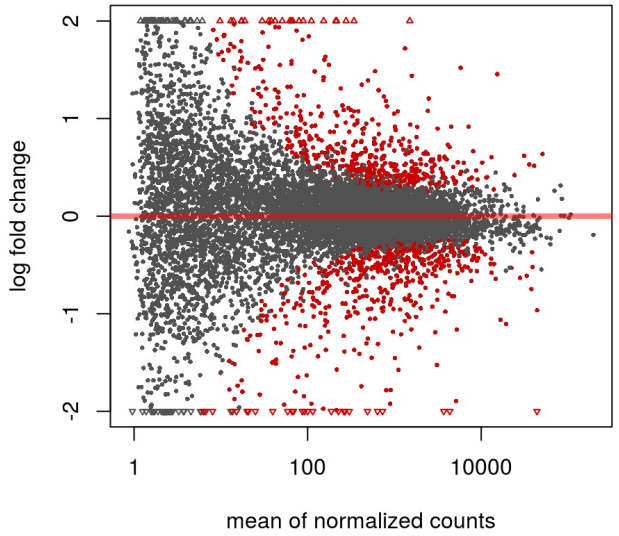
- Variance – required for NB distribution
  - Insufficient observations to allow direct measure
  - Custom variance distribution fitted to real data
  - Smooth distribution assumed to allow fitting

Exploratory DESeq2









Steps in DEG Analysis

Estimate variability - (common and genewise dispersion)

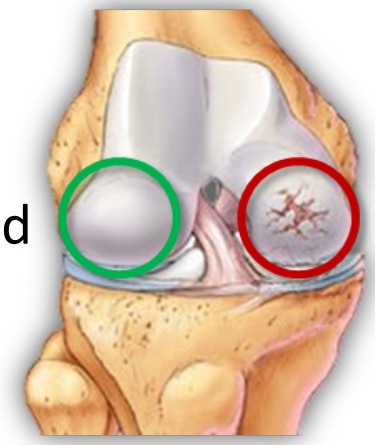
- Determine fold change between samples (e.g. treatment and control)
- Determine significance (p-value)
- Correct for multiple testing (corrected p-value, false discovery rate)

➢ Selection of DEG sets based on FDR (and possibly min/max fold-change)

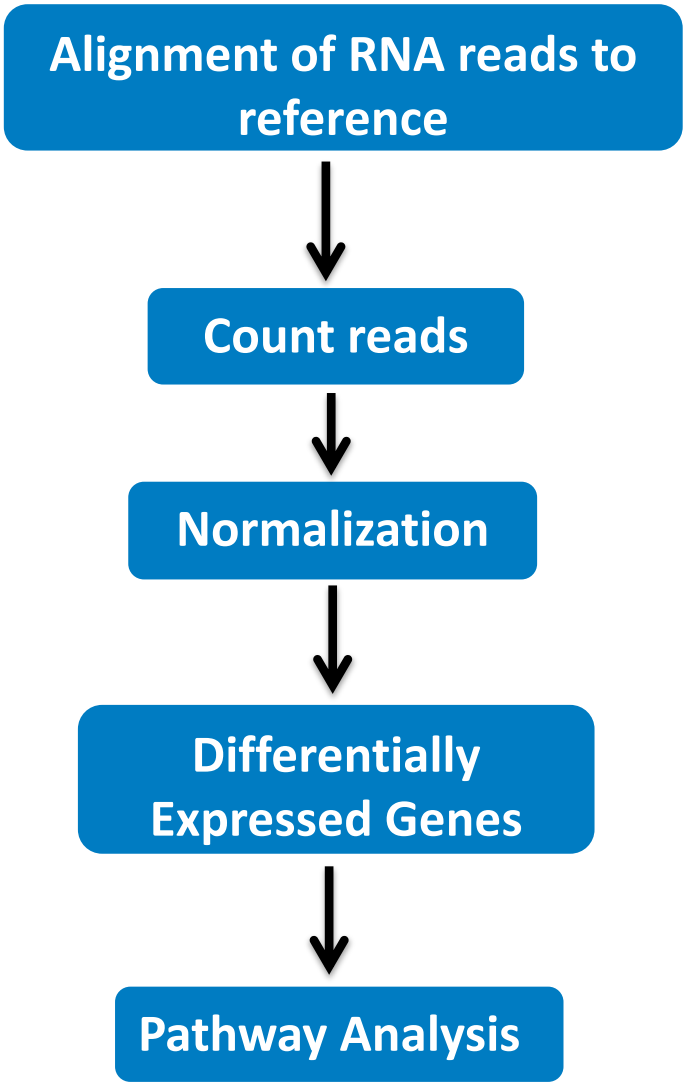
Complex Experimental Designs

Facilitated by generalized linear models (GLMs). Examples:

- Interaction effects
- Blocking
- **Paired samples**
- Batch effects
- ANOVA-like tests

Preserved  Lesioned

Typical workflow of RNA-Seq Gene Expression Data

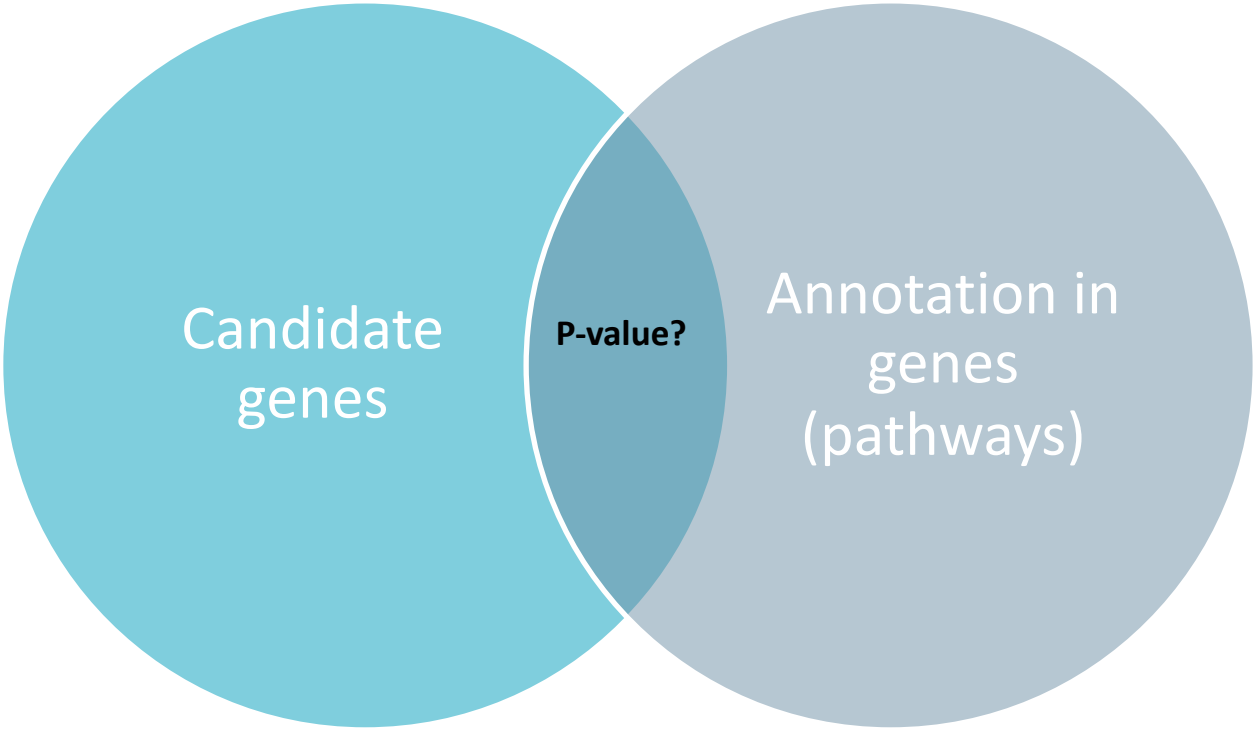


Pathways database

GENEONTOLOGY
Unifying Biology

Pathways analysis

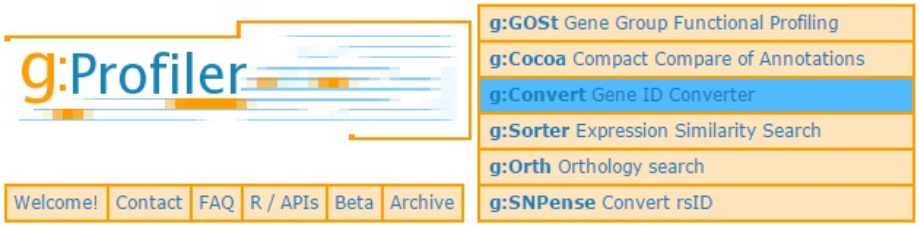
- Are there more annotations in a gene list than expected?

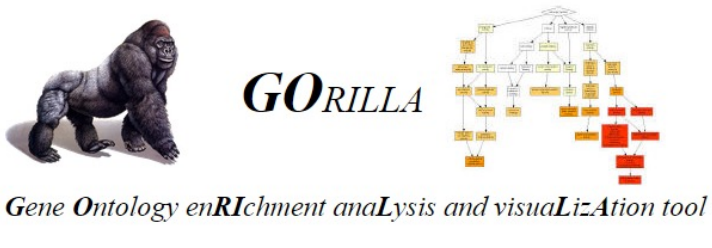


Tools for functional gene list analysis

There are many different tools available, both free and commercial

Popular tools include:





- Categorical Statistics;
- Biggest selection of gene sets;
- Simple interface, but limited options:
- No species information;
- No background list option;
- Simple interactive visualisation;
- Novel scoring scheme to rank hits;
- Implemented in R statistical language;

Questions & Practicals

QUESTIONS?

r.coutinho_de_almeida@lumc.nl

@rodcoutalmeida