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

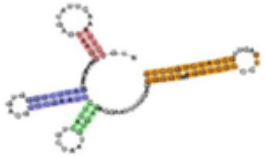
r.coutinho_de_almeida@lumc.nl

Outline

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

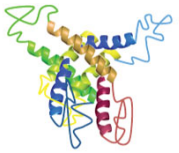
The Central Dogma of Molecular Biology

snoRNAs



snRNAs

tRNAs



DNA

Replication

~ 80%

RNA

Transcription

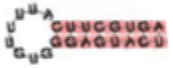
~2%

Protein

Translation

Alternative spliced mRNAs

lncRNAs



microRNAs

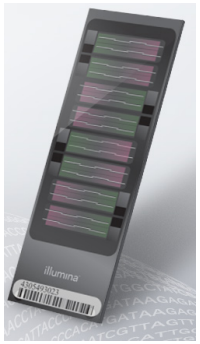
Transcriptomics

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

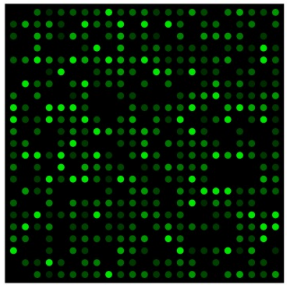
What can the transcriptome tell us?

- 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

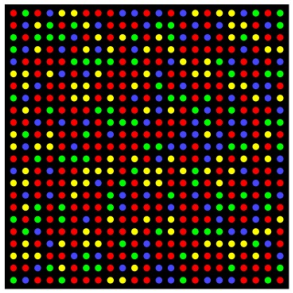


Microarray chip



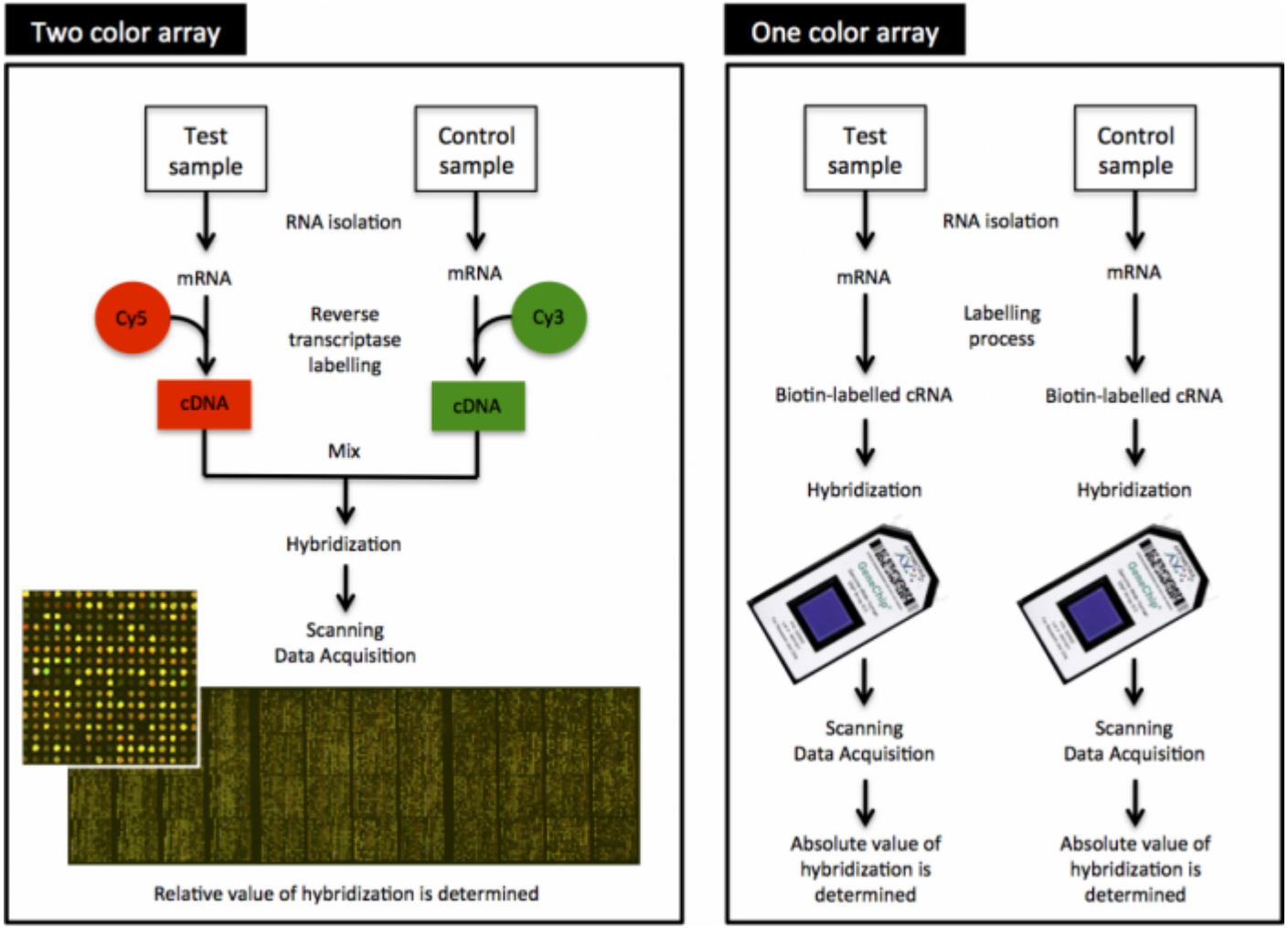
Microarray

Sequencing flow cell



RNA-seq

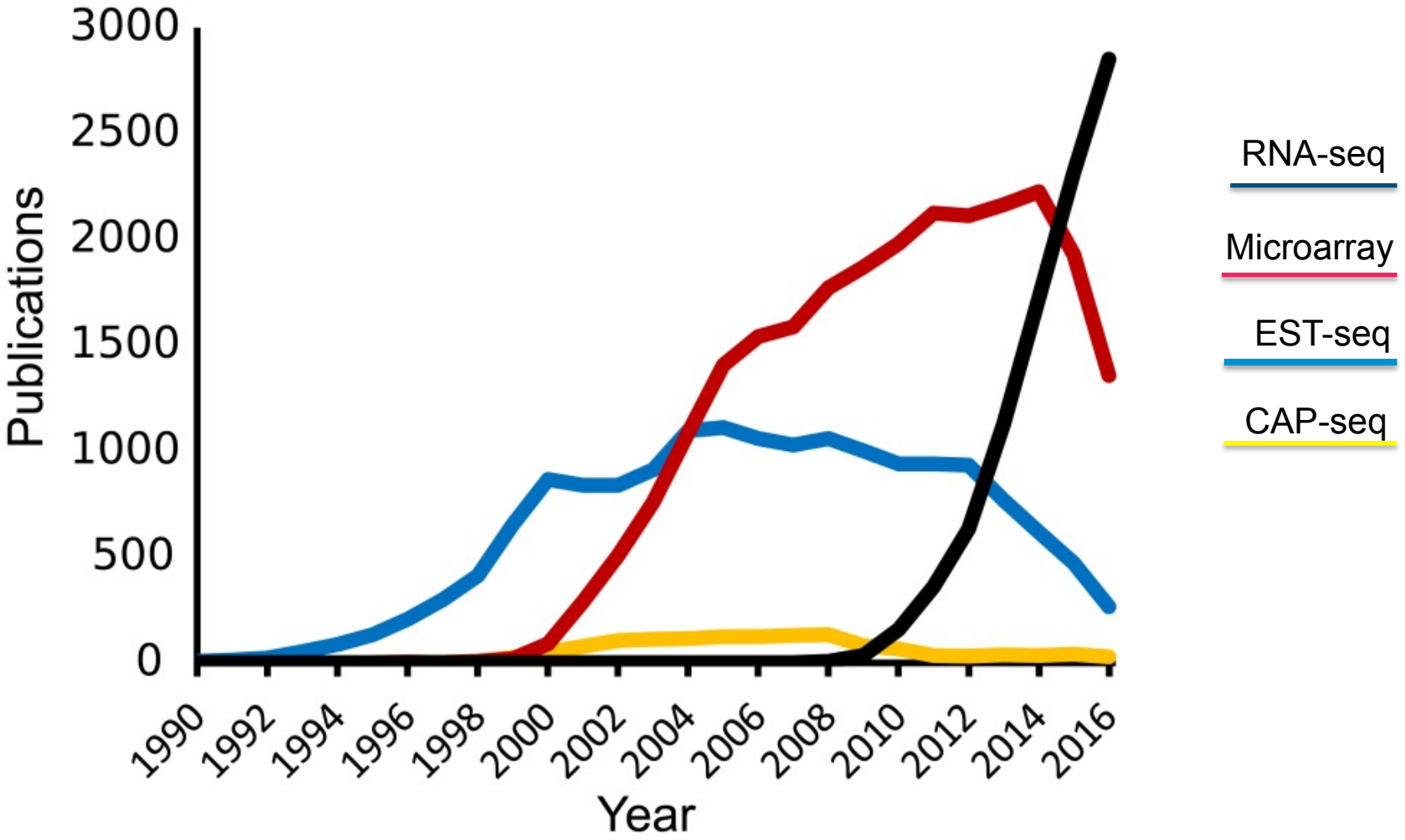
Microarray


Two color array
Test sample
Control sample
RNA isolation
mRNA
mRNA
Cy5
Cy3
Reverse transcriptase labelling
cDNA
cDNA
Mix
Hybridization
Scanning
Data Acquisition
Relative value of hybridization is determined
One color array
Test sample
Control sample
RNA isolation
mRNA
mRNA
Labelling process
Biotin-labelled cRNA
Biotin-labelled cRNA
Hybridization
Hybridization
Scanning
Data Acquisition
Scanning
Data Acquisition
Absolute value of hybridization is determined
Absolute value of hybridization is determined

Microarray and RNA-Seq Depositories

- NCBI GEO: http://www.ncbi.nlm.nih.gov/geo
- ArrayExpress: http://www.ebi.ac.uk/arrayexpress/
- recount2: https://jhubiostatistics.shinyapps.io/recount/

Transcriptomics method use over time



Lowe et al., PLoS Comput Biol, 2017

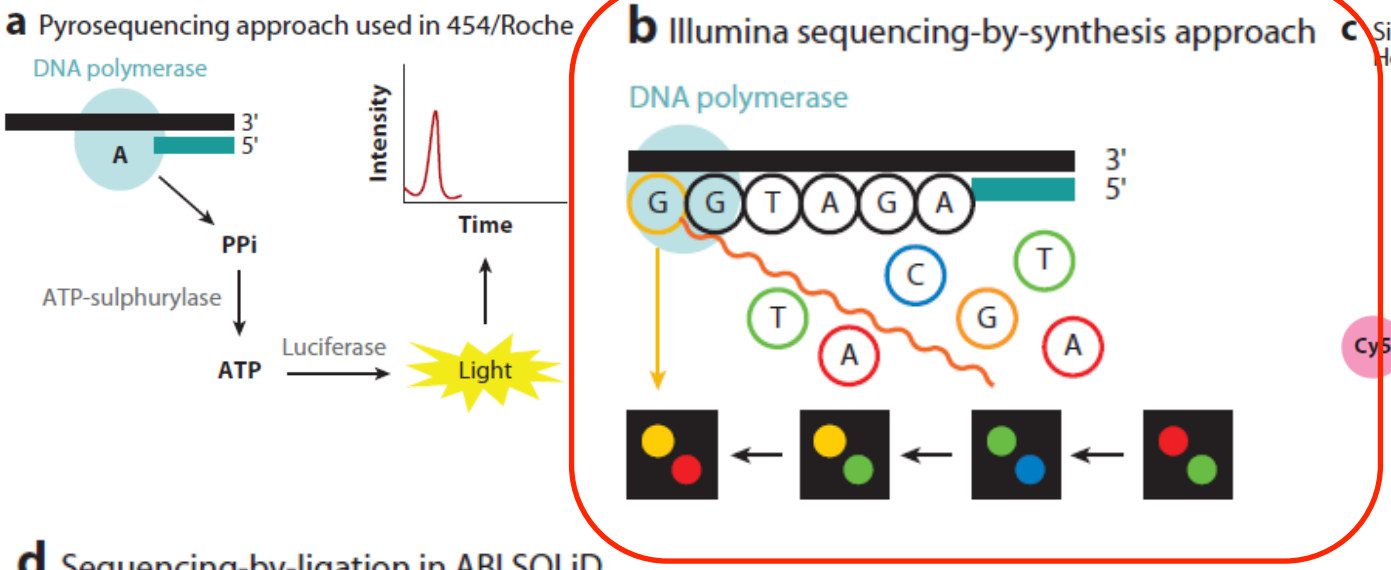
Advantages of RNA-seq over microarray approach

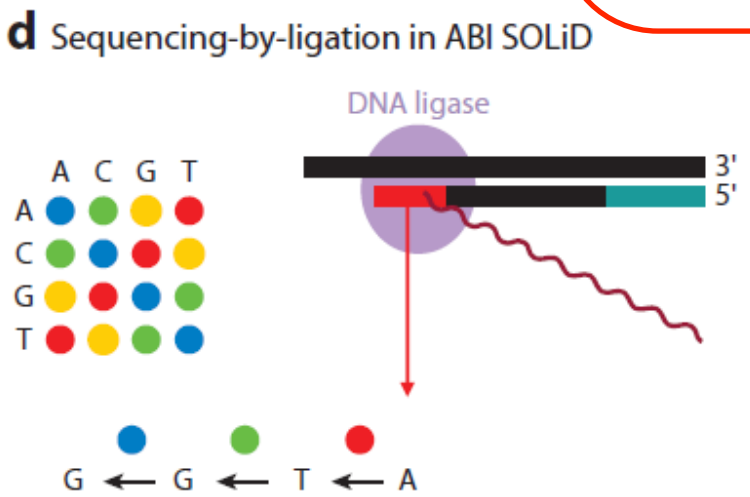
- Higher sensitivity for genes expressed either at low level;
- Higher dynamic range of expression levels over which transcripts can be detected (> 8000-fold range);
- Lower technical variation and higher levels of reproducibility;
- Not limited by prior knowledge of the genome of the organism;
- Gives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

Applications of RNA-seq

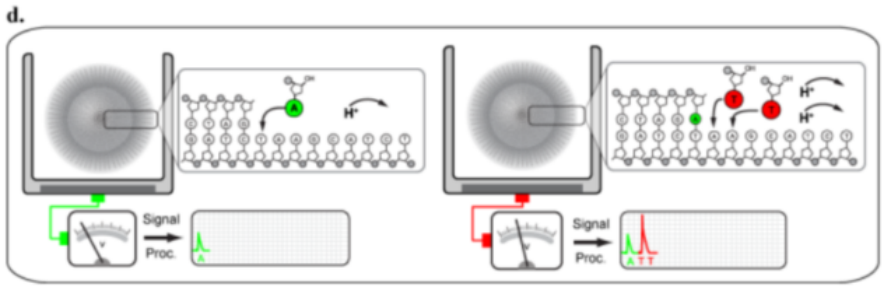
- Gene expression profiling between samples;
- Diagnostics through expression profiling;
- Identify alternative splicing events;
- Allele-specific expression, SNPs and gene fusions;
- Exon dosage (quantification);
- Identify non-coding RNAs (eg. microRNAs);
- Identification of human pathogens;

Types of RNA-seq methods





Ion semiconductor sequencing in Ion Proton



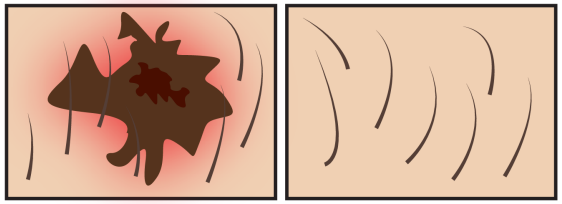
Morozova et al., Annu. Rev. Genomics Hum. Genet. 2009

Sequencing by synthesis

https://www.youtube.com/watch?v=fCd6B5HRaZ8

Typical RNA-seq experiments

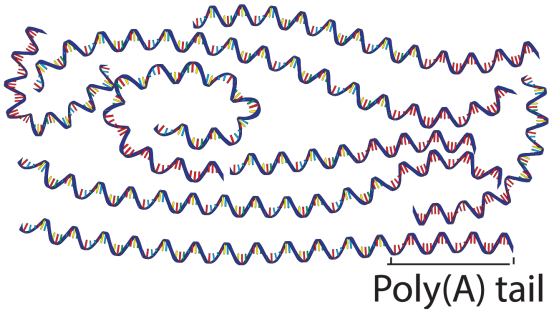
Samples of interest



Condition 1 (e.g. tumor)

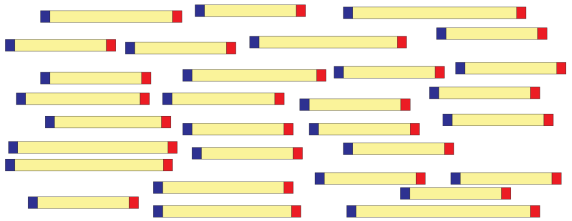
Condition 2 (e.g. normal)

Isolate RNAs

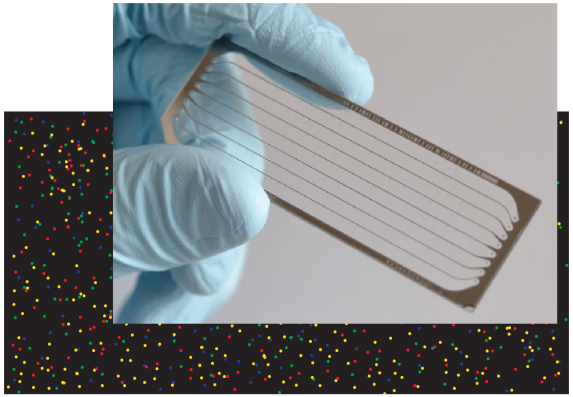


Poly(A) tail

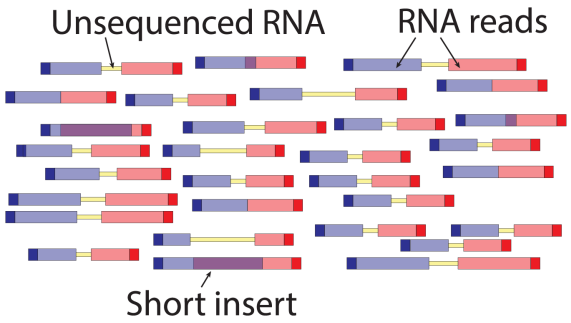
Generate cDNA, fragment, size select, add linkers



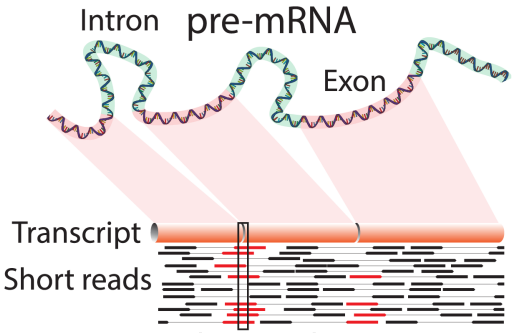
Sequence ends



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



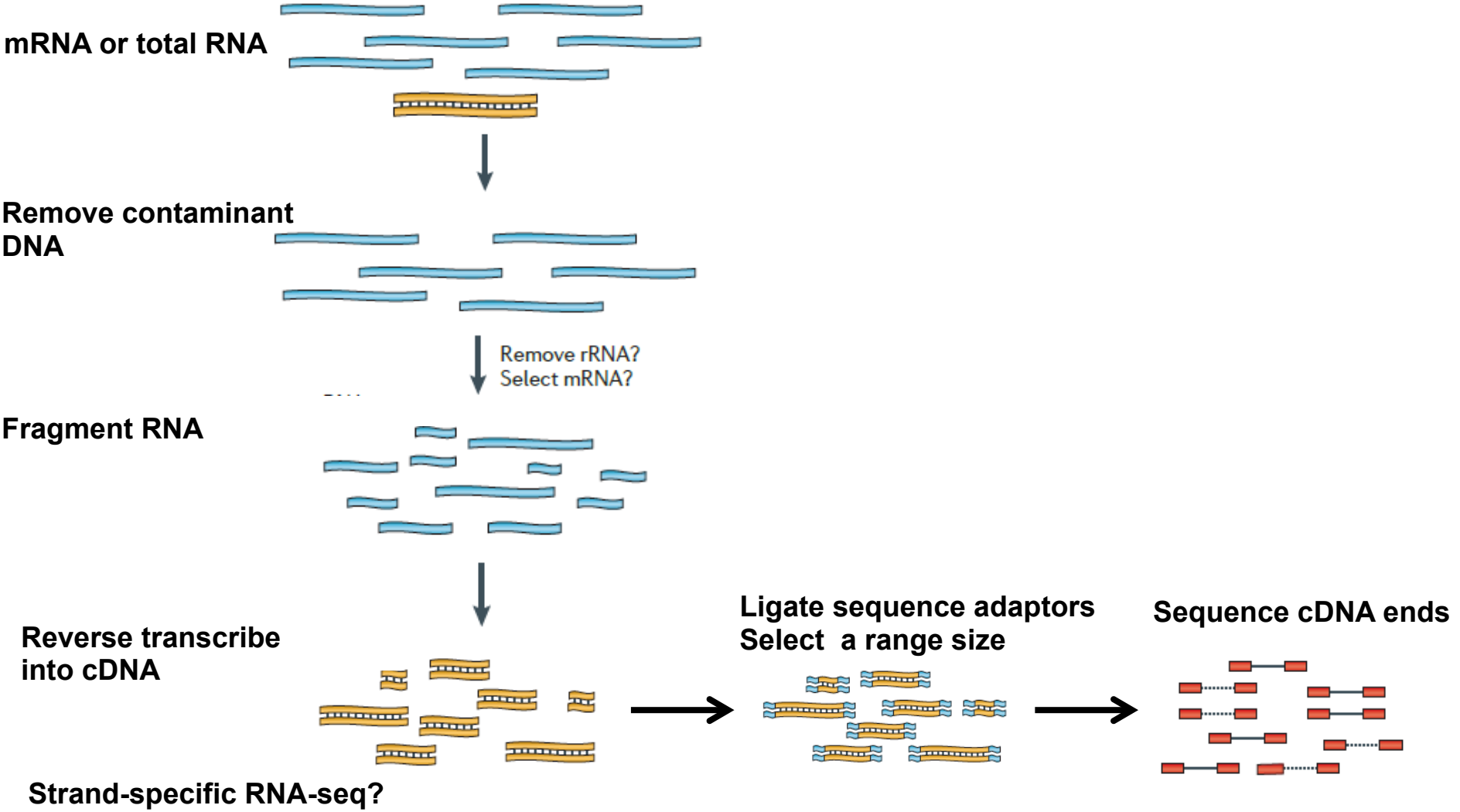
Map to genome, transcriptome, and predicted exon junctions



Downstream analysis

Source: Wikipedia

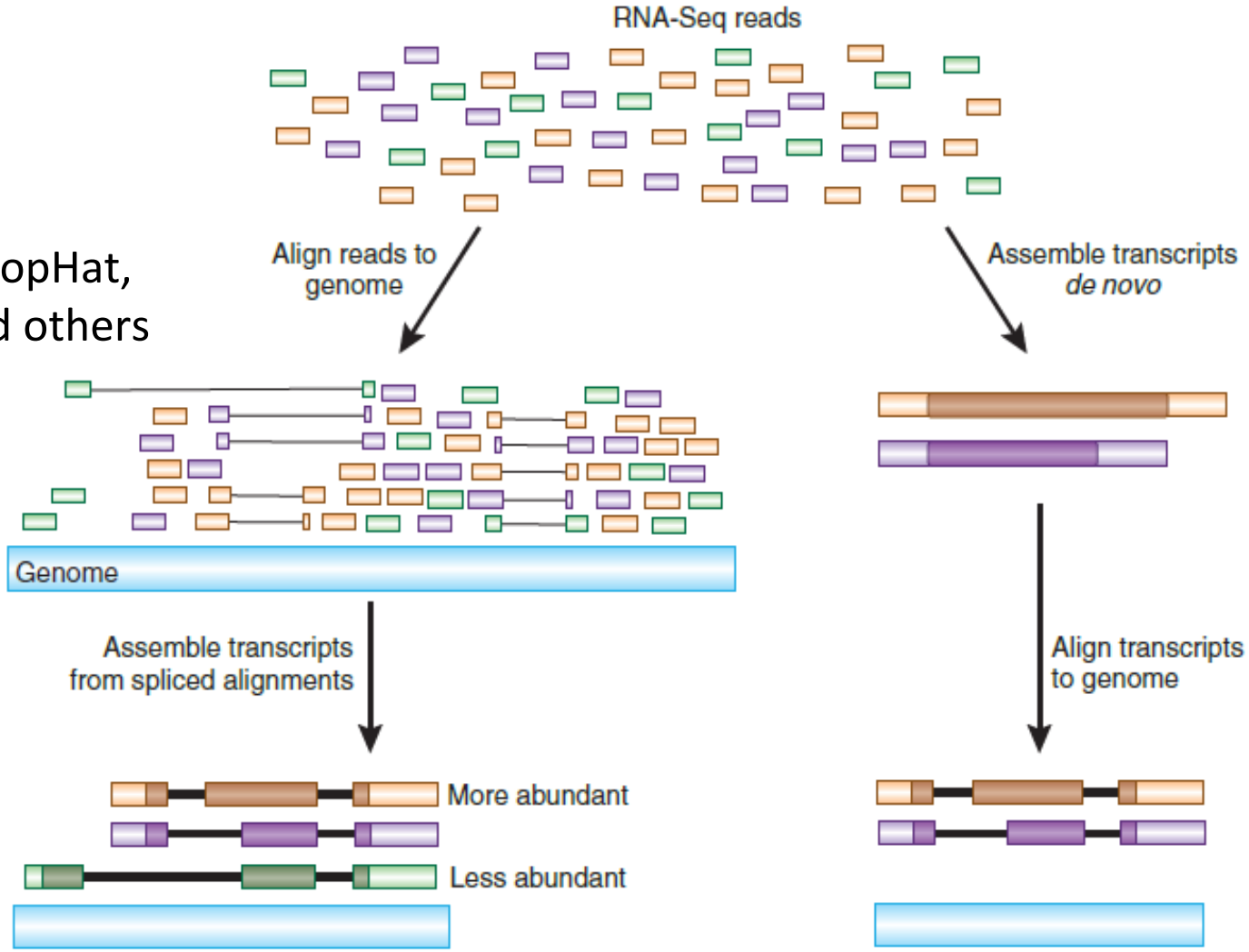
RNA-seq data generation



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

RNA-seq align and assemble

GSNAP, TopHat, STAR and others



Trinity, Kallisto and others

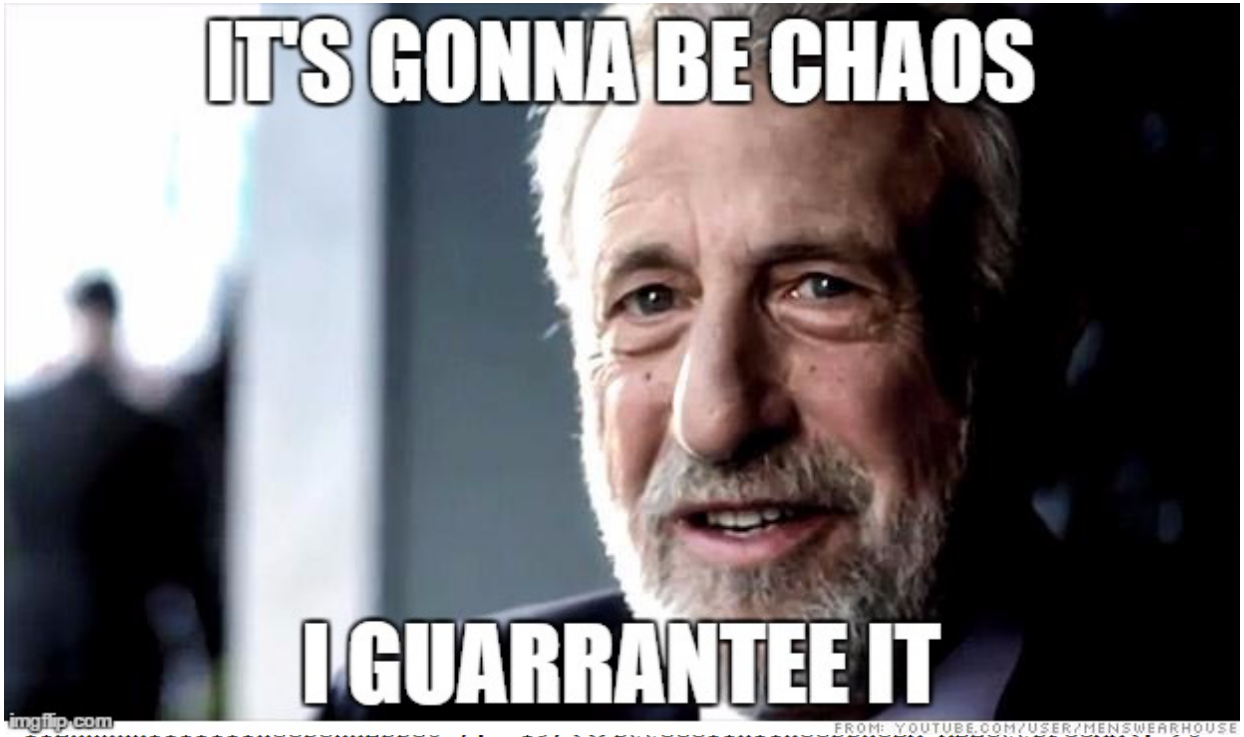
Haas BJ and Zody MC. , Nat.Biotech. 2010 .

RNA-seq analysis

- Quality Control;
- Normalization;
- Differential expression;
- Pathway analysis

~ 2Gb of expression data, and now?

FASTQ file

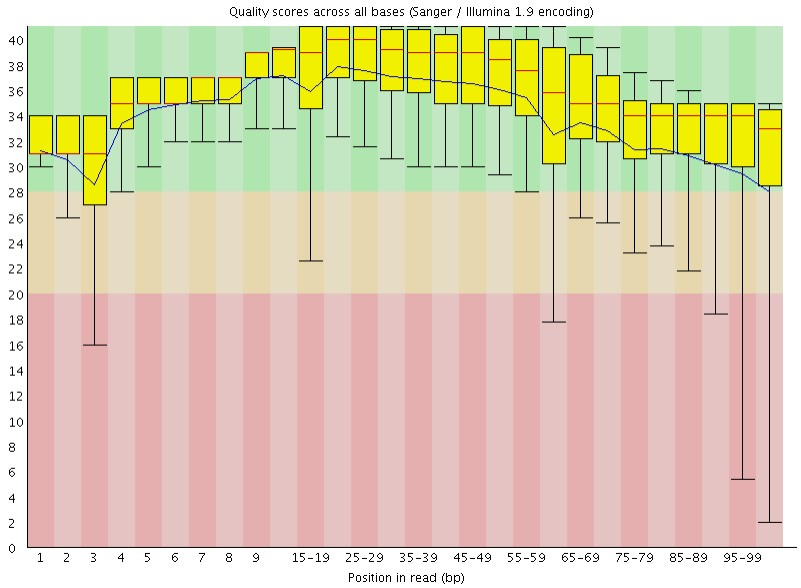



```
@22:51205934-51222090C:ENST00000464740:125:612:-1:185:S/1
TGGAGTGCGCTGCGGCGCGAGCTGGGCCGGCGGGCGTGGTTCGAGAGCGCGCAGAGTCCAGACTGGCGGCAGGGCC
+
HHIIIHIDGG@;=@GIIIIIDDGBBBEDB@8>5554,/':9B@@C?==@1:2@?=GG=;<HHHHGIIHHEC-;;3?
```

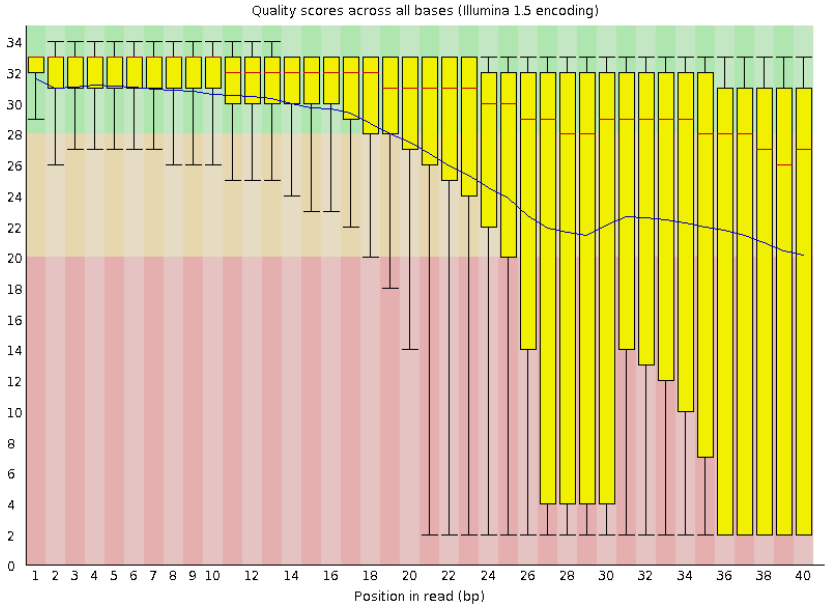
Quality Control (QC)

FastQC

Per base sequence quality



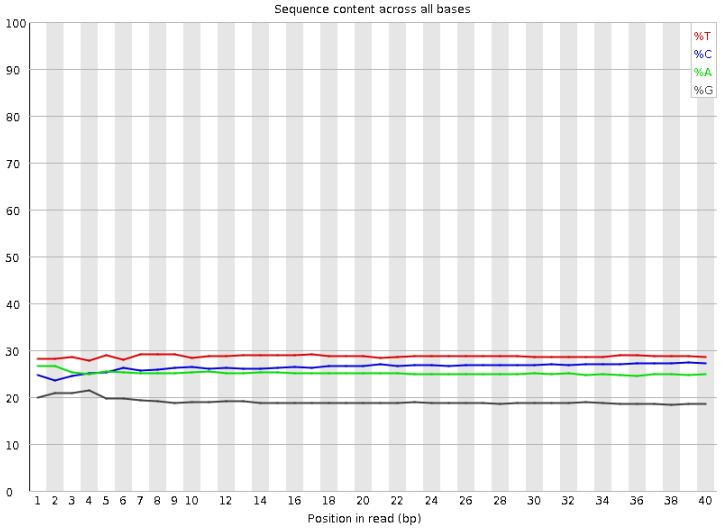
Good sample

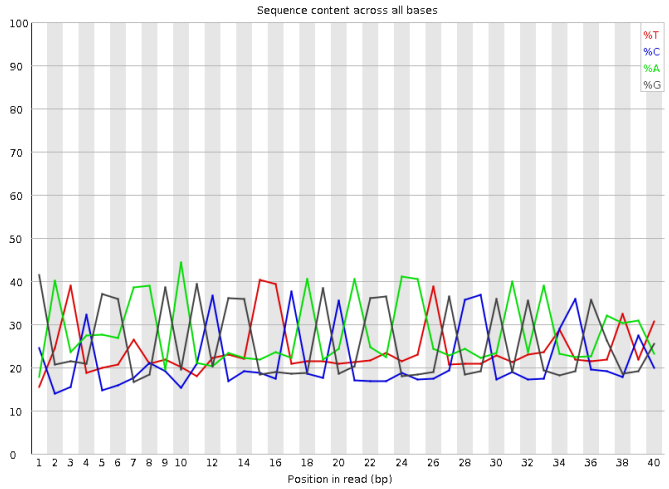


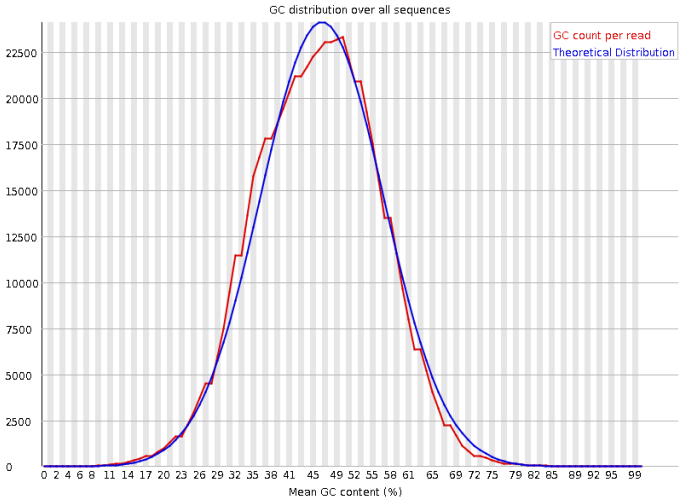
Bad sample

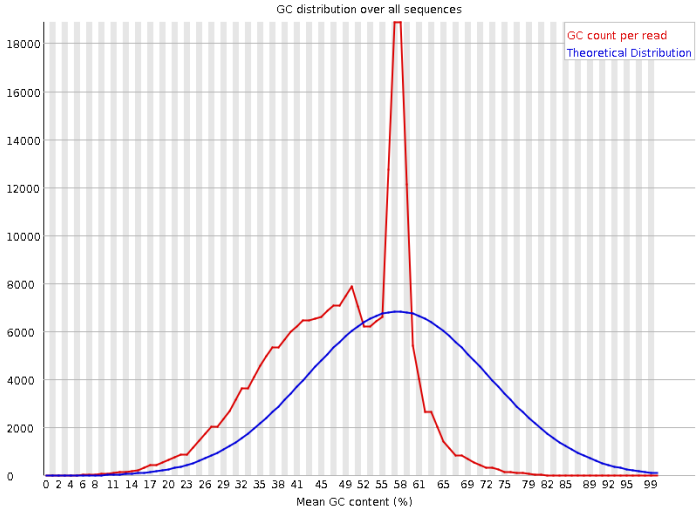
QC: Raw Data

Sequence bias

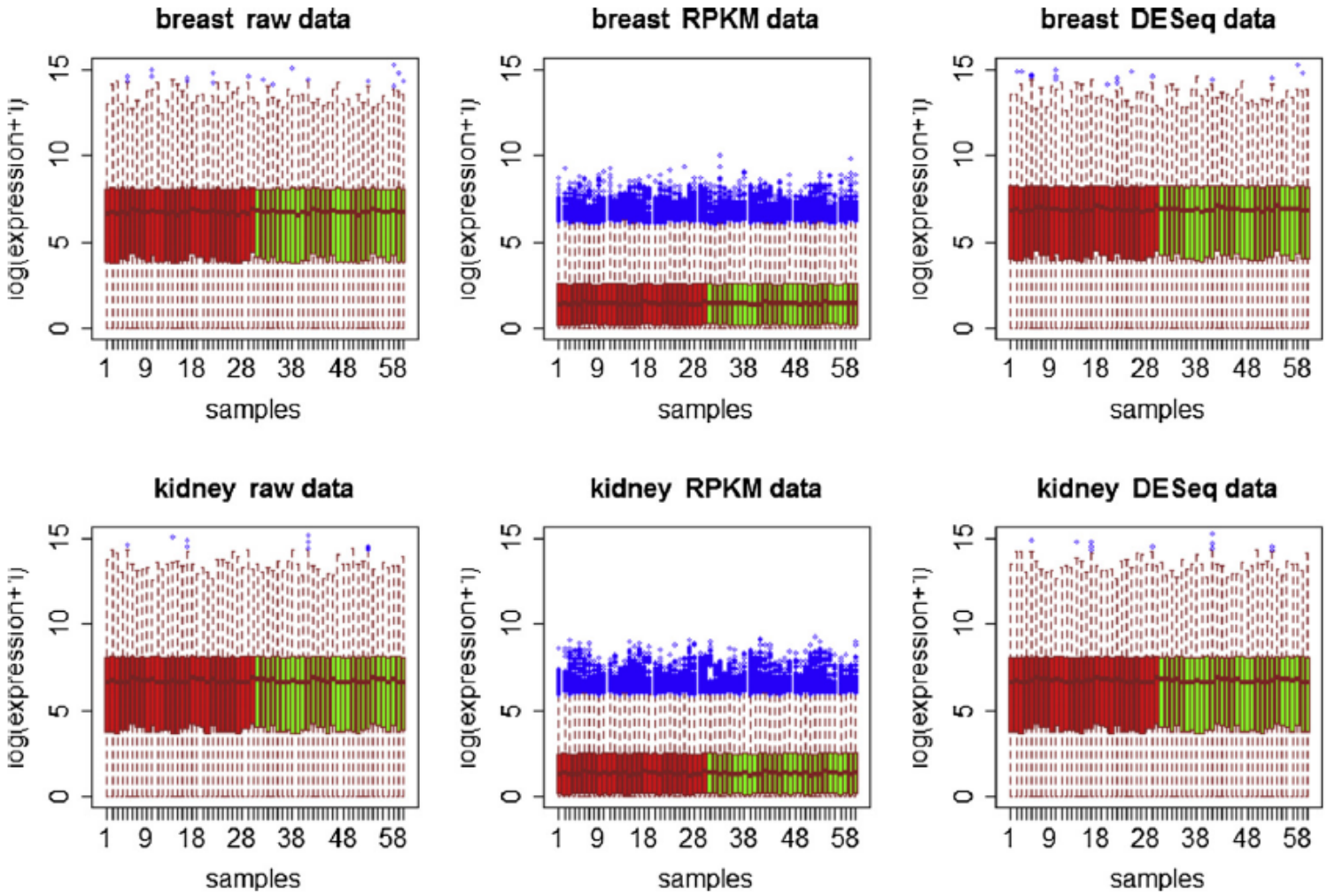








Normalization required



H. Han, K. Men, *J B Informatics 85 (2018) 80–92*

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;

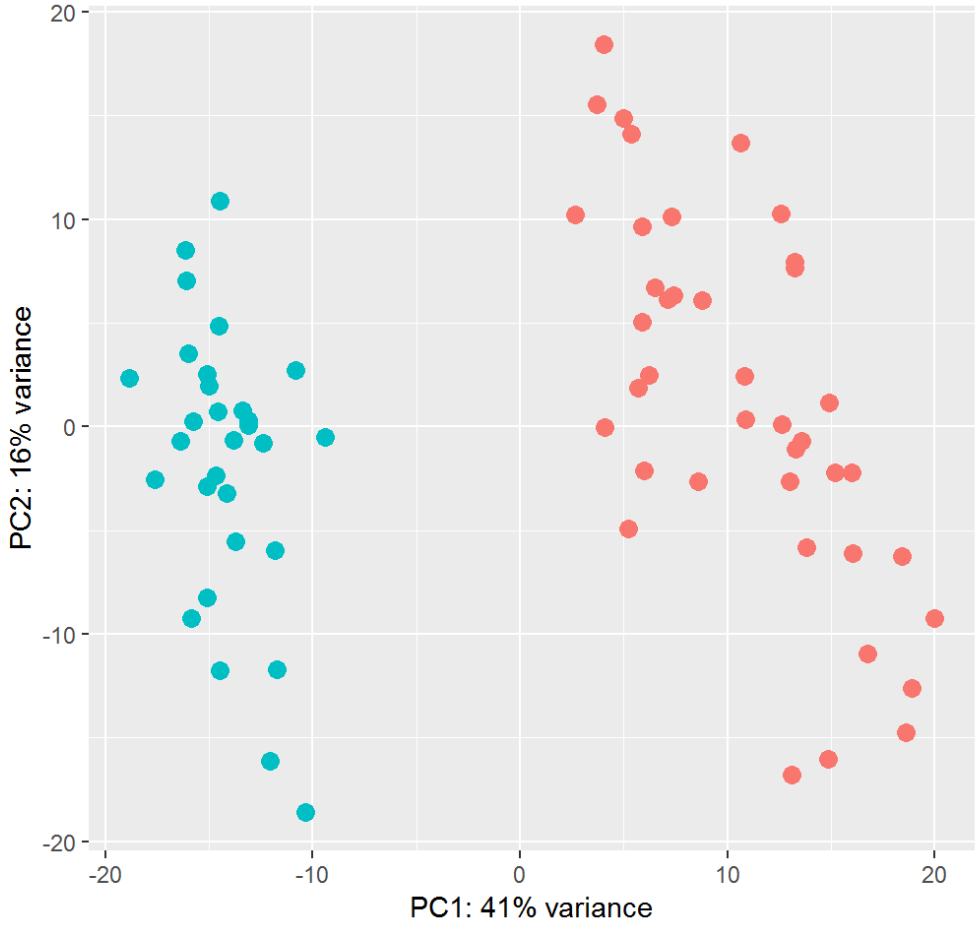
Examples of common normalization methods

- **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



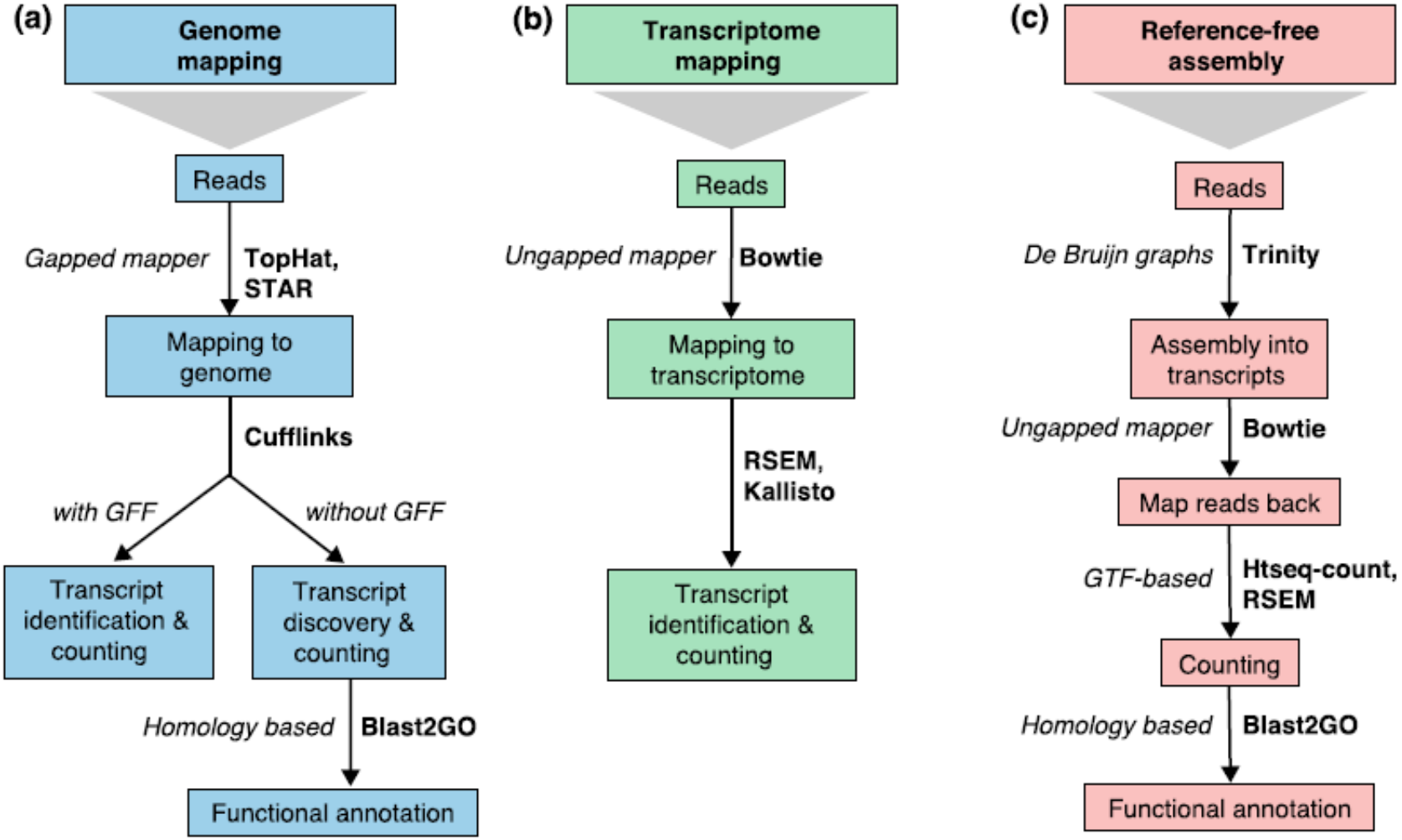
To remove:

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

Sofwares:

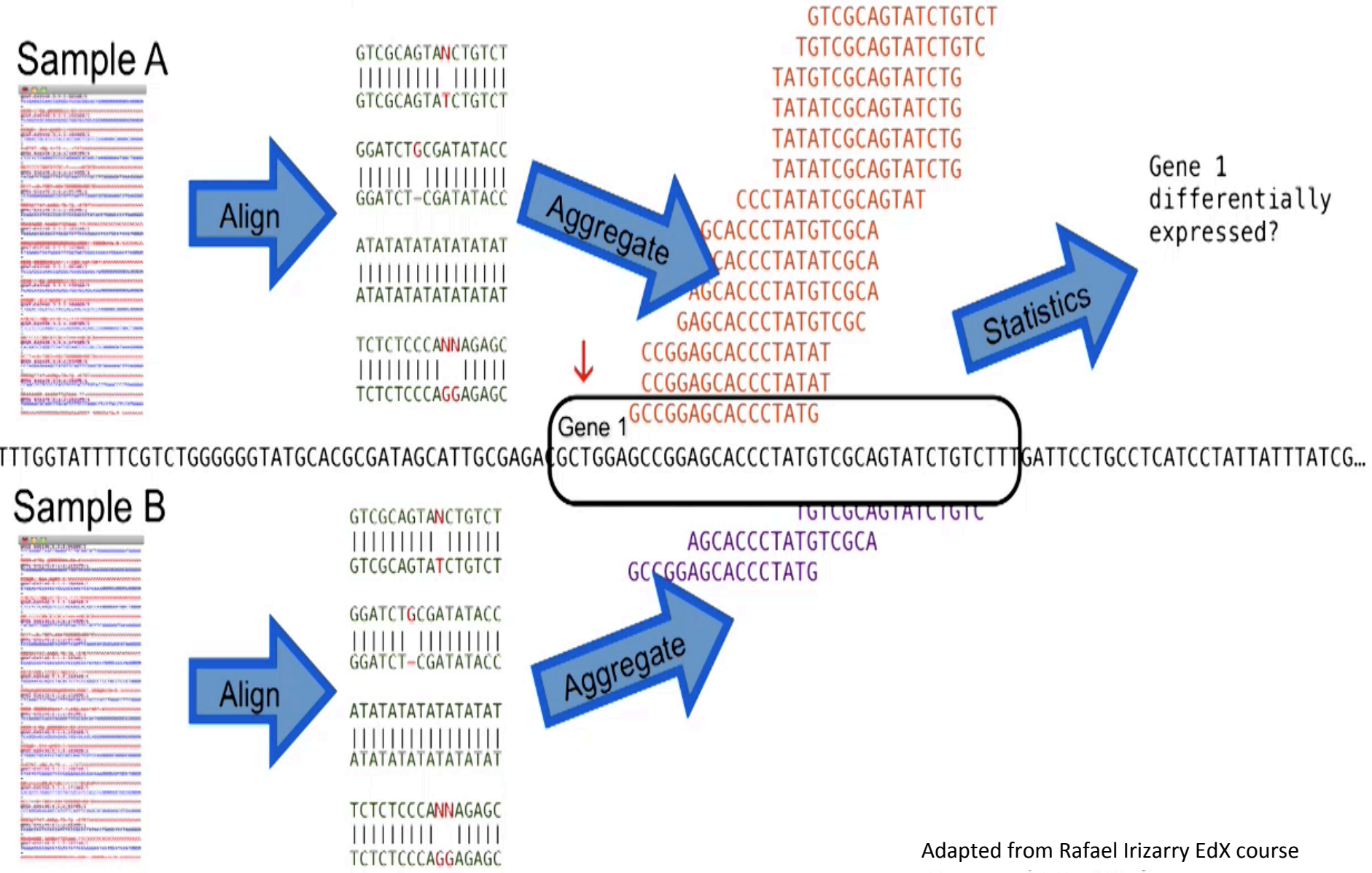
Combat, limma, sva, and others

RNA-seq options



Conesa et al. Genome Biology (2016)

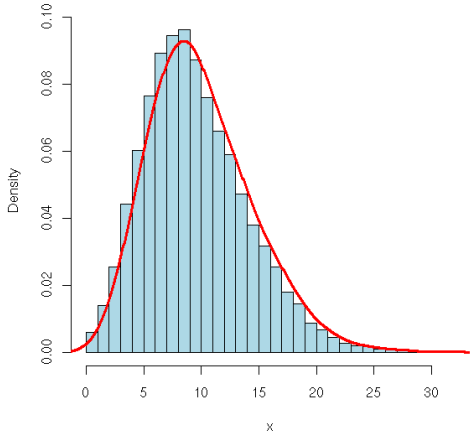
RNA-seq data analysis overview



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

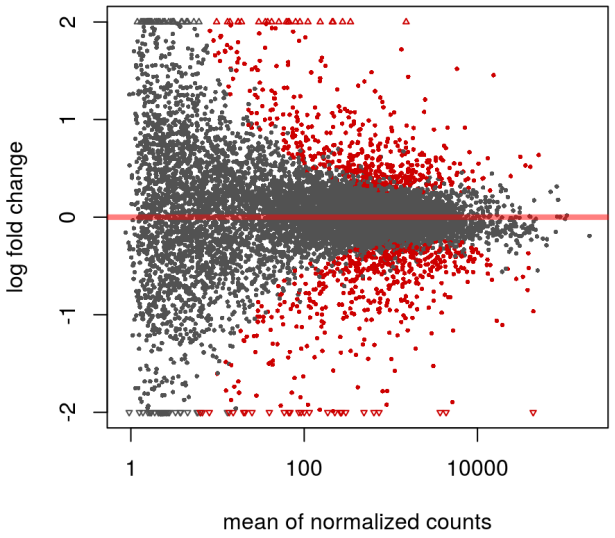
DESEQ2 Statistics

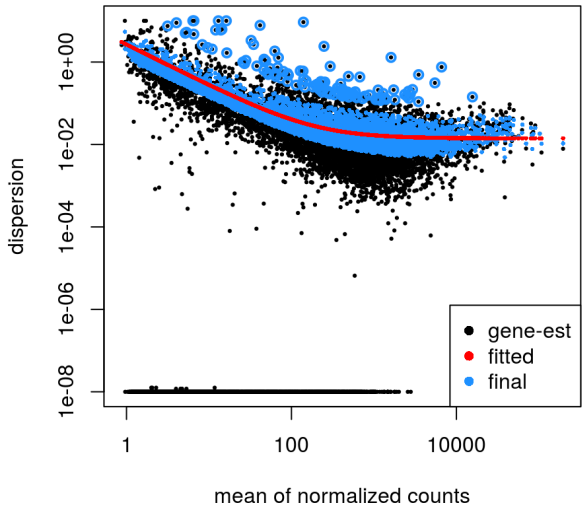
- Are the counts we see for gene A in condition 1 consistent with those for gene A in condition 2?

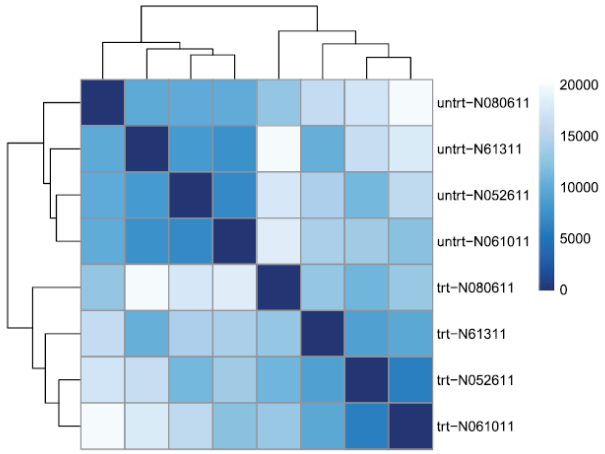
- Size factors
  - Estimator of library sampling depth
  - More stable measure than total coverage
  - Based on median ratio between conditions

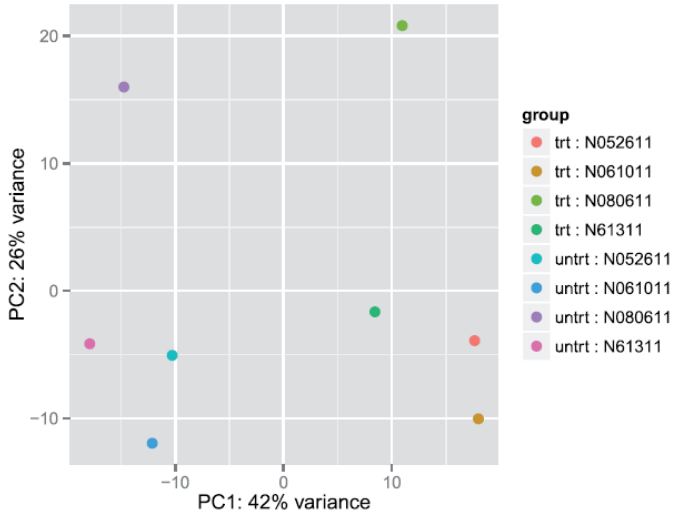
- Variance – required for NB distribution
  - Insufficient observations to allow direct measure
  - Custom variance distribution fitted to real data
  - Smooth distribution assumed to allow fitting

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)

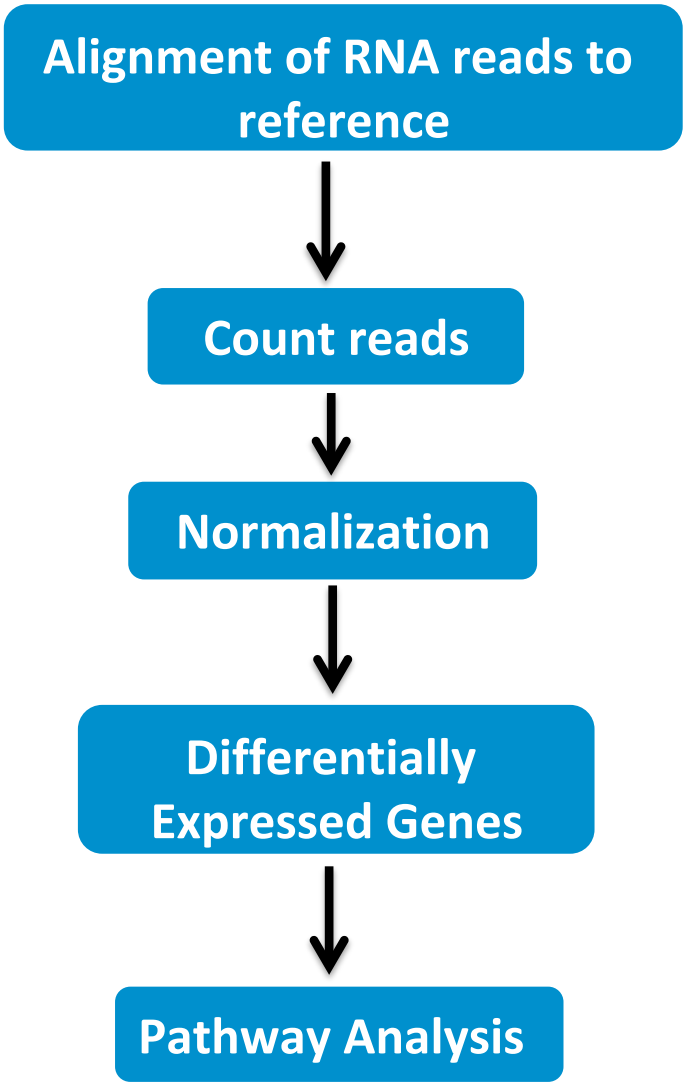
➢ Selection of DEG sets based on FDR (and possibly min/max fold-change)

Complex Experimental Designs

Facilitated by generalized linear models (GLMs). Examples:

- Interaction effects
- Blocking
- **Paired samples** (automatically adjustment for batch effects)
- Batch effects
- ANOVA-like tests

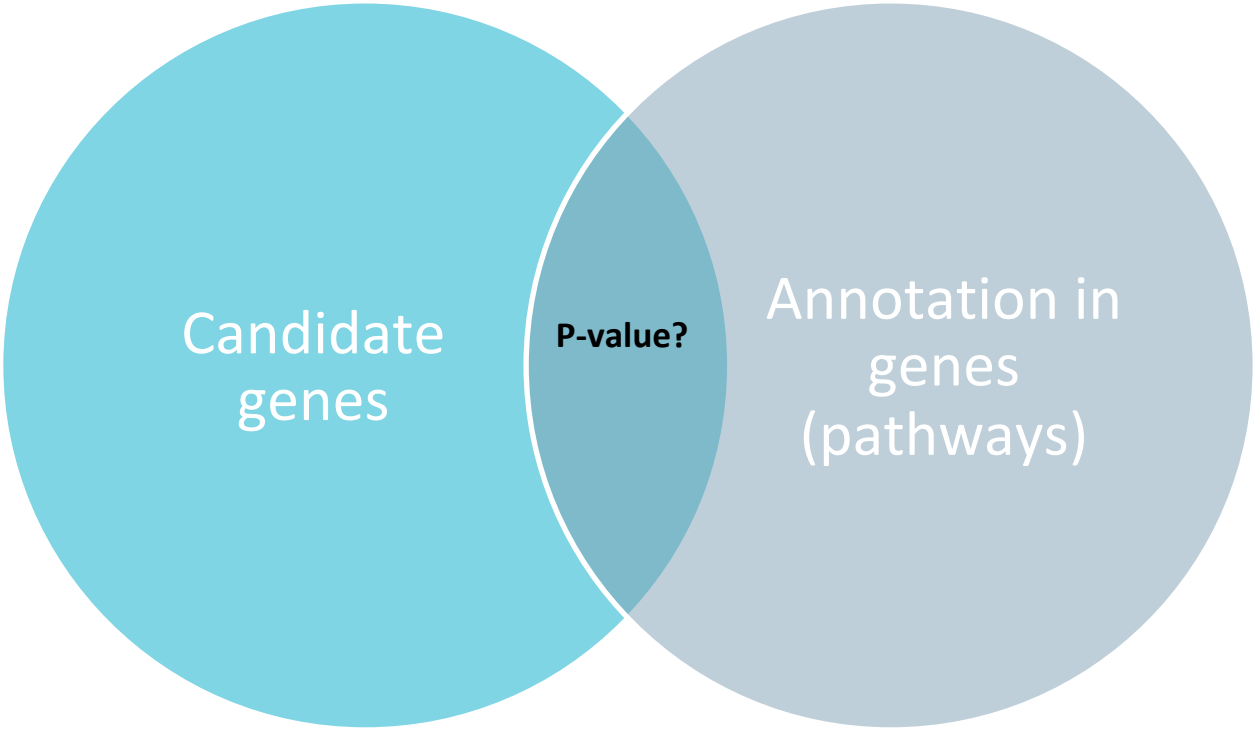
Typical workflow of RNA-Seq Gene Expression Data



Pathways database

Pathways analysis

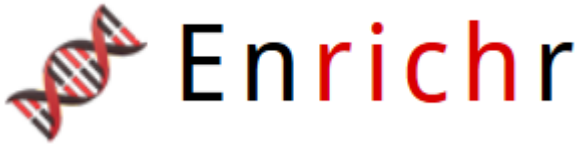
- Are there more annotations in a gene list than expected?

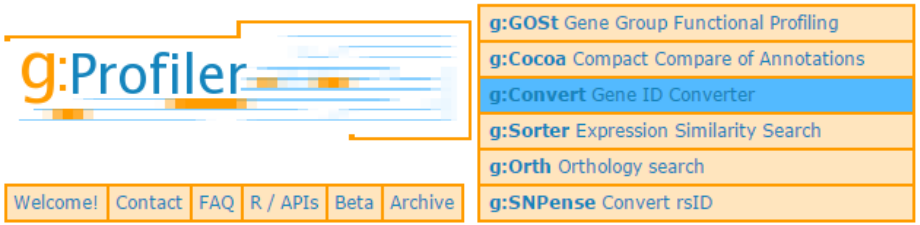


Tools for functional gene list analysis

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?