Leids Universitair Medisch Centrum

# Introduction to Transcriptomics

## Molecular Data Science: from disease mechanisms to personalized medicine

Rodrigo C de Almeida

Biomedical Data Sciences,

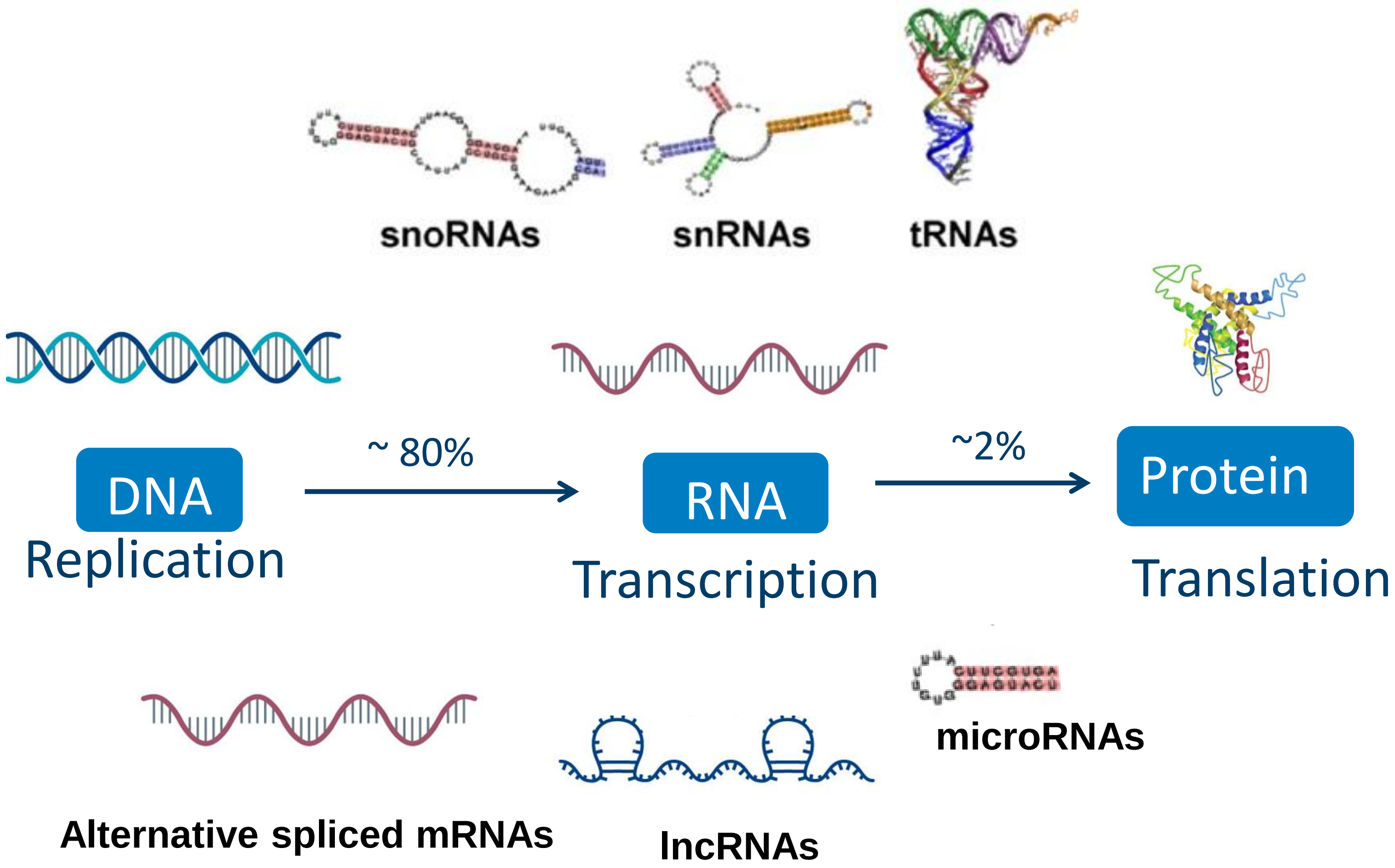
Molecular Epidemiology

r.coutinho_de_almeida@lumc.nl

# Outline

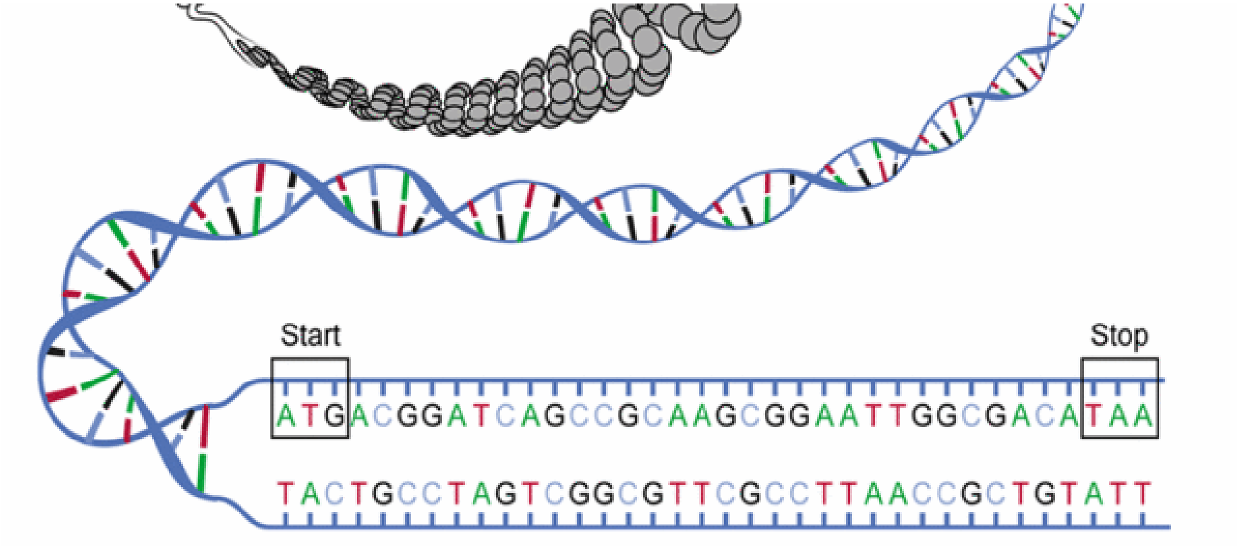
- Transcriptome;
- Methods to study the transcriptome;
- RNA-seq;
- Differential expression analysis;

# The Central Dogma of Molecular Biology



# 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



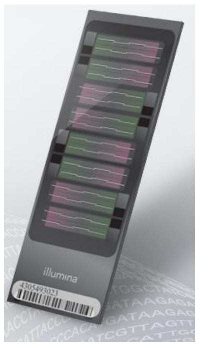
# What can the transcriptome tell us?

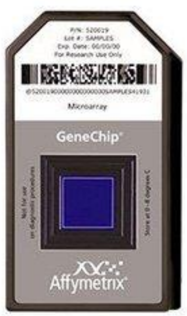
- Where and when each gene is expressed in the cells and tissues of an organism;

- Changes in the normal level of gene activity in the transcriptome may reflect of contribute to disease;

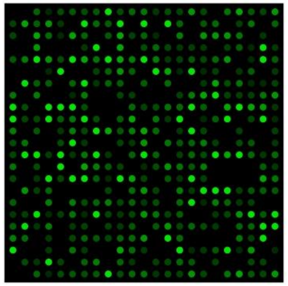
- Researchers can get a genome-wide picture on what genes are active in a tissue;

# Two major technologies to studies the transcriptome



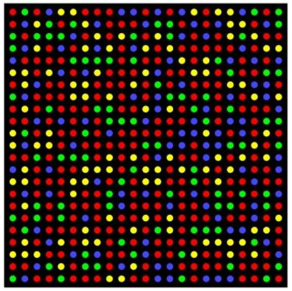




Microarray chip



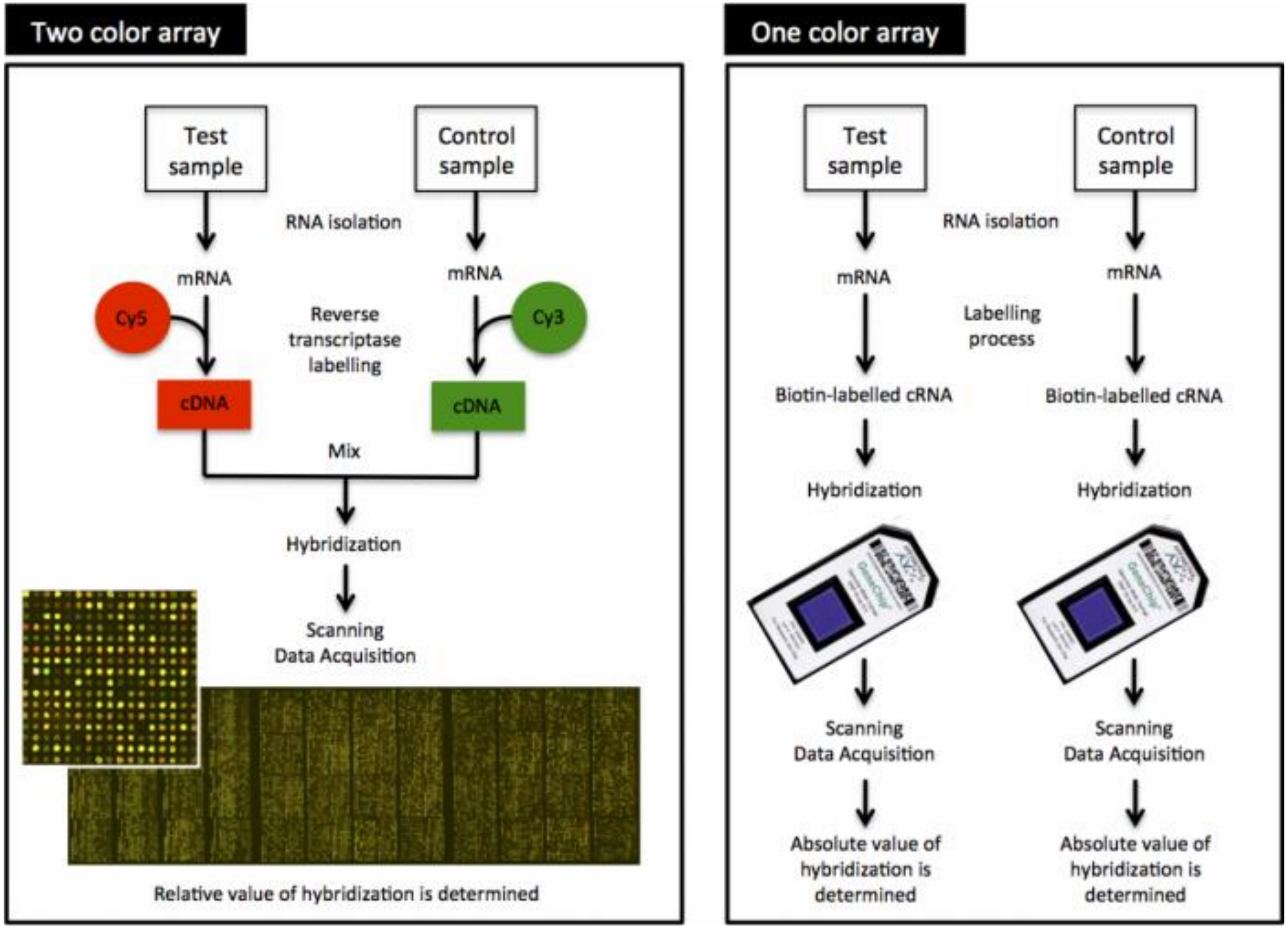
**Microarray**

Sequencing flow cell



**RNA-seq**

# Microarray

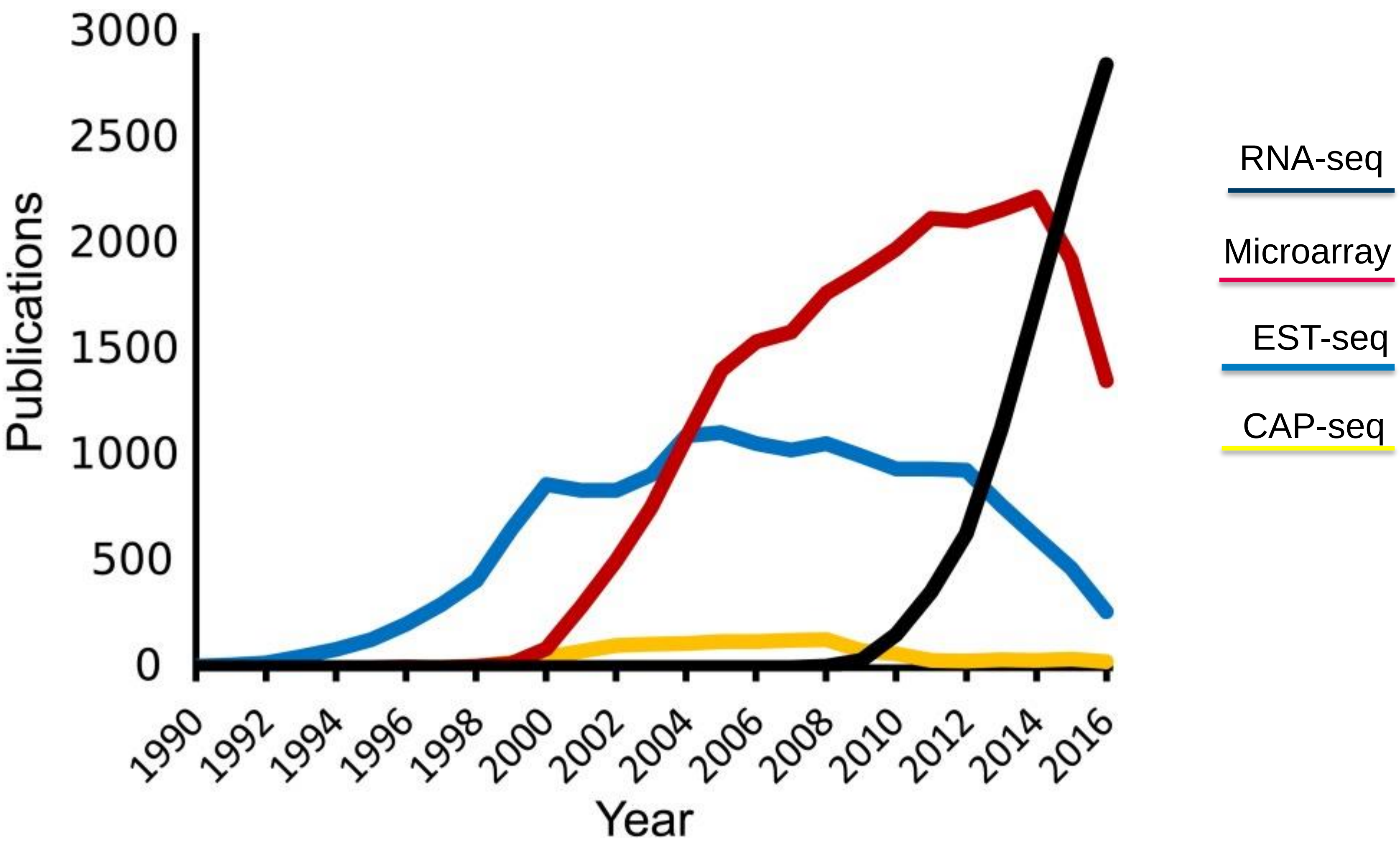


Two color array
Test sample
Control sample
RNA isolation
mRNA
mRNA
Cy5
Cy3
Reverse transcriptase labelling
cDNA
cDNA
Mix
Hybridization
Scanning
Data Acquisition
Relative value of hybridization is determined
One color array
Test sample
Control sample
RNA isolation
mRNA
mRNA
Labelling process
Biotin-labelled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

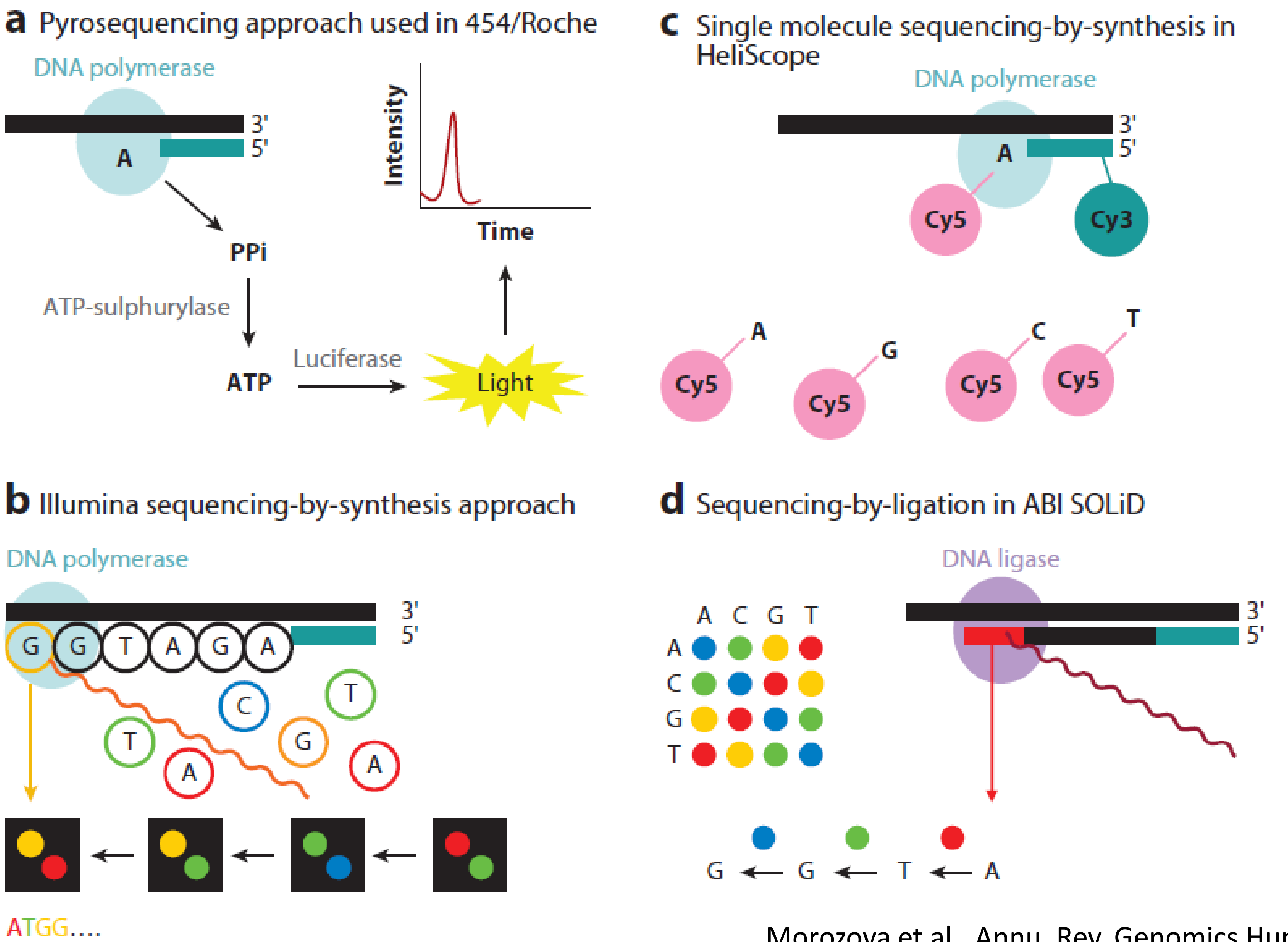
# Advantages of RNA-seq over microarray approach

- 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

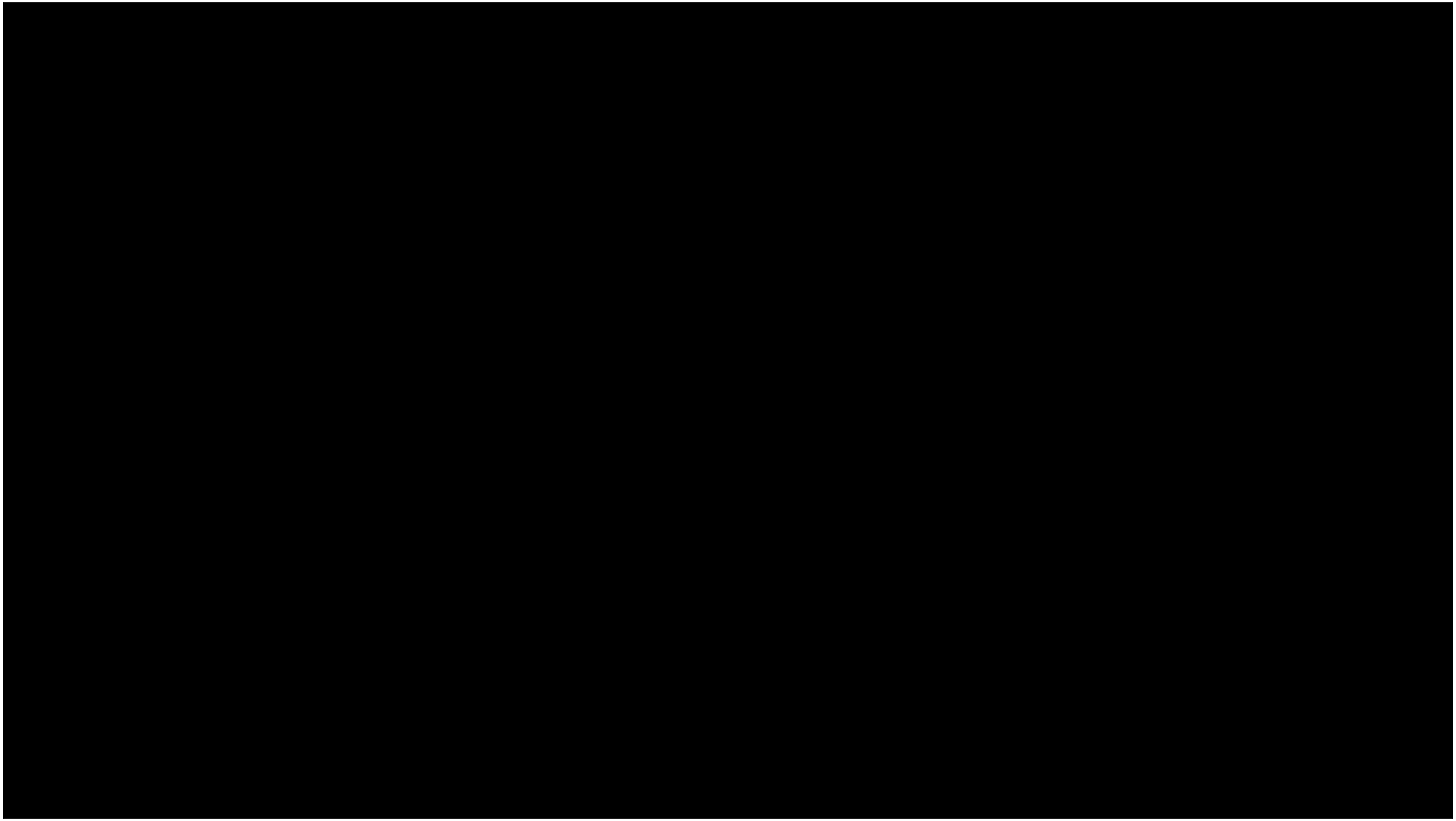
- Gene expression profiling between samples;
- Diagnostics through expression profiling;
- Identify alternative splicing events;
- Allele-specific expression, SNPs and gene fusions;
- Exon dosage (quantification);
- Identify non-coding RNAs (eg. microRNAs);
- Identification of human pathogens;

# Types of RNA-seq methods



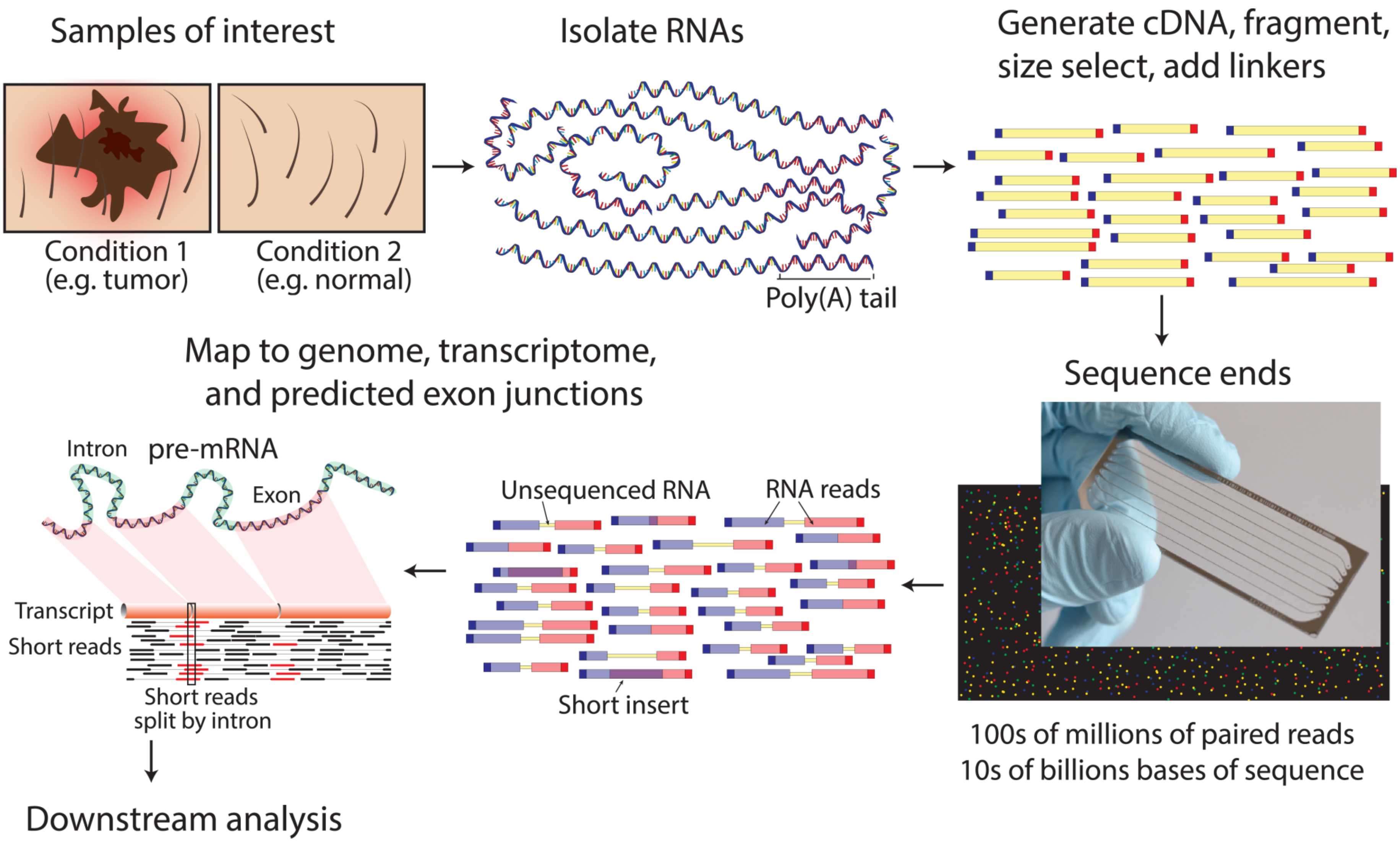

Morozova et al., Annu. Rev. Genomics Hum. Genet. 2009

# Sequencing by synthesis

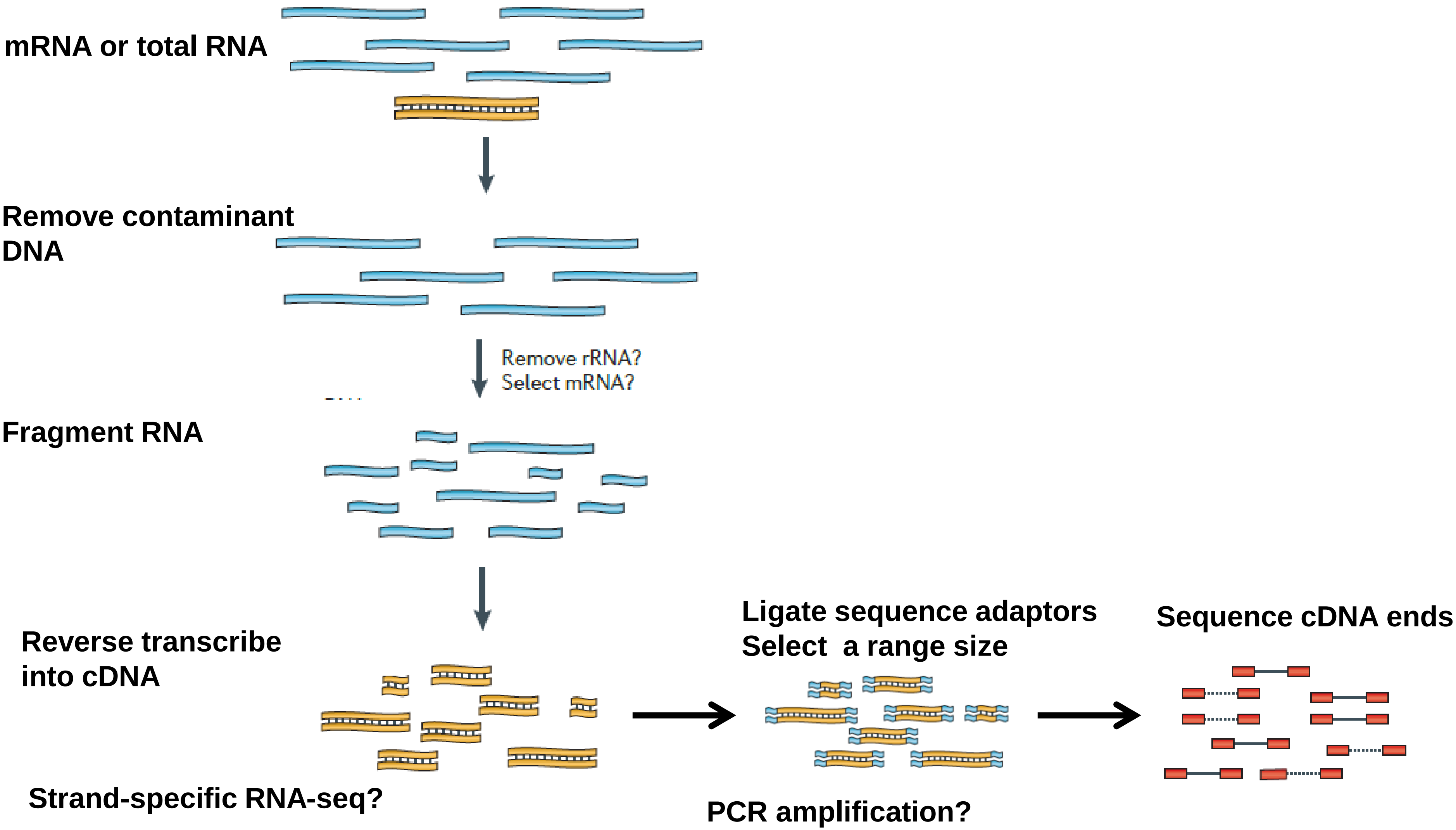


https://www.youtube.com/watch?v=fCd6B5HRaZ8

# Typical RNA-seq experiments



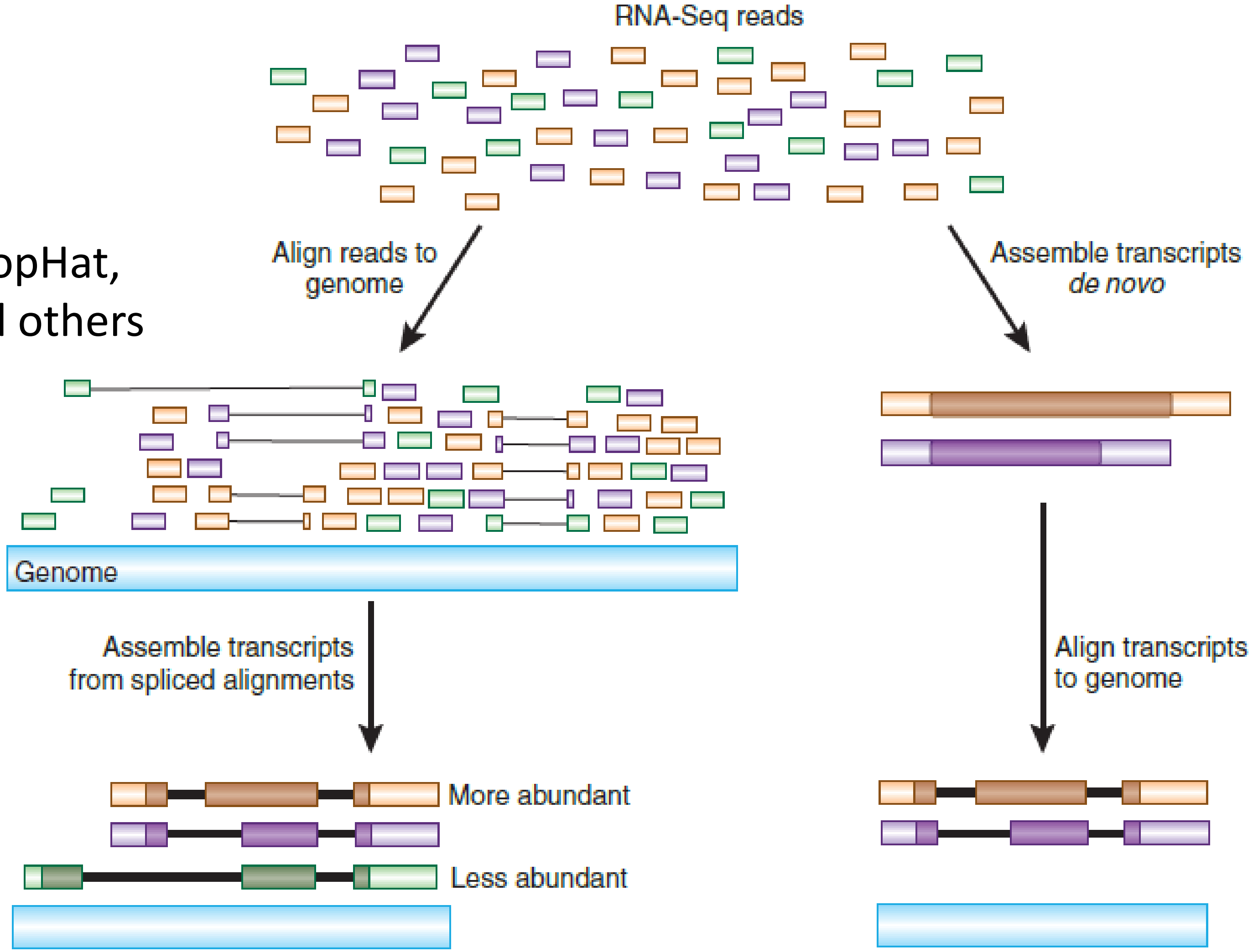
# RNA-seq data generation




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 .

# QUESTIONS?

# RNA-seq analysis

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

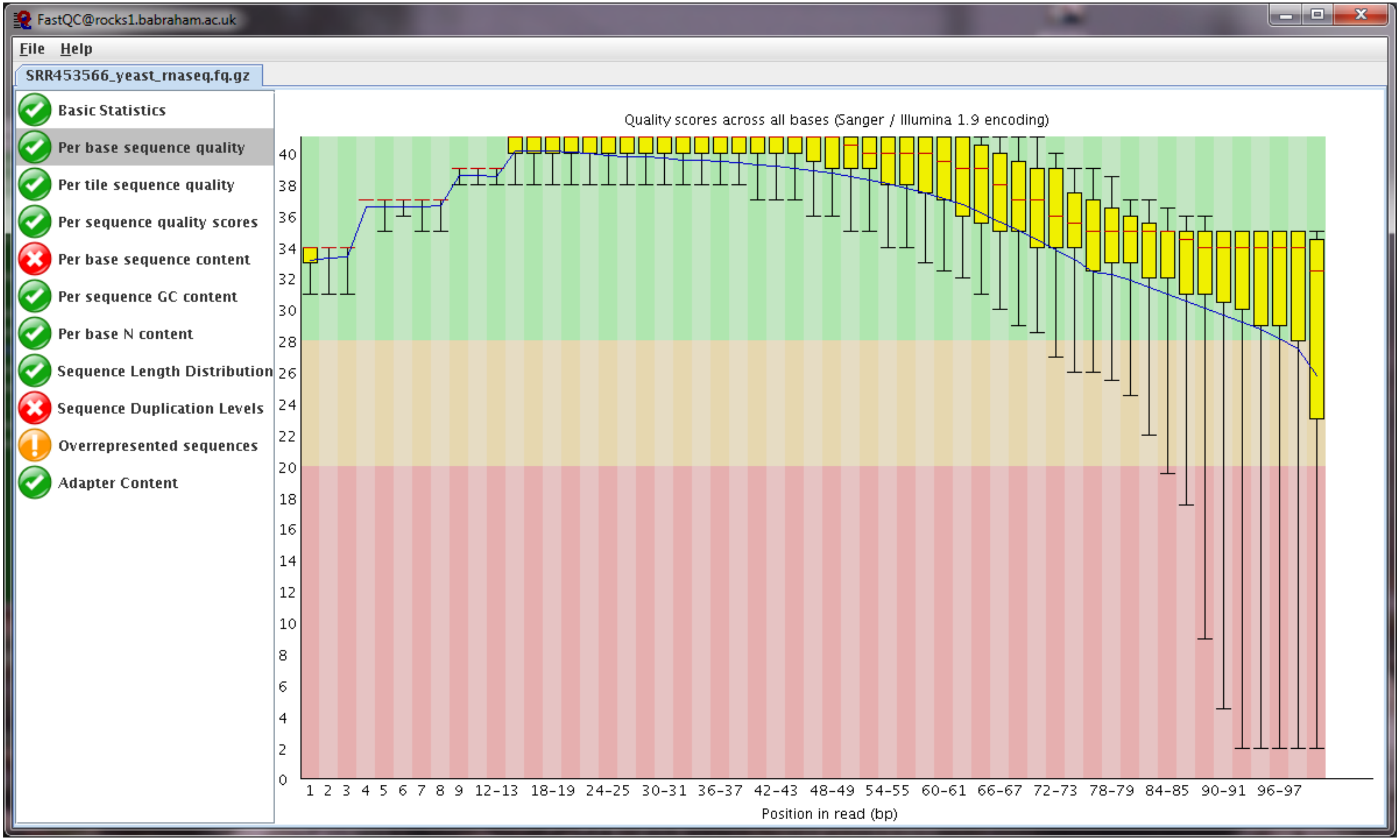
# ~ 3Gb of expression data, and now?

```
@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
+
@=ABBBBIIIIIIIIIHHGGGGIIDBDIIIIIIGIIIIHIIIIHFDD@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=;<HHHGIIHHEC-;;3?
```

FASTQ file

# Quality Control (QC)

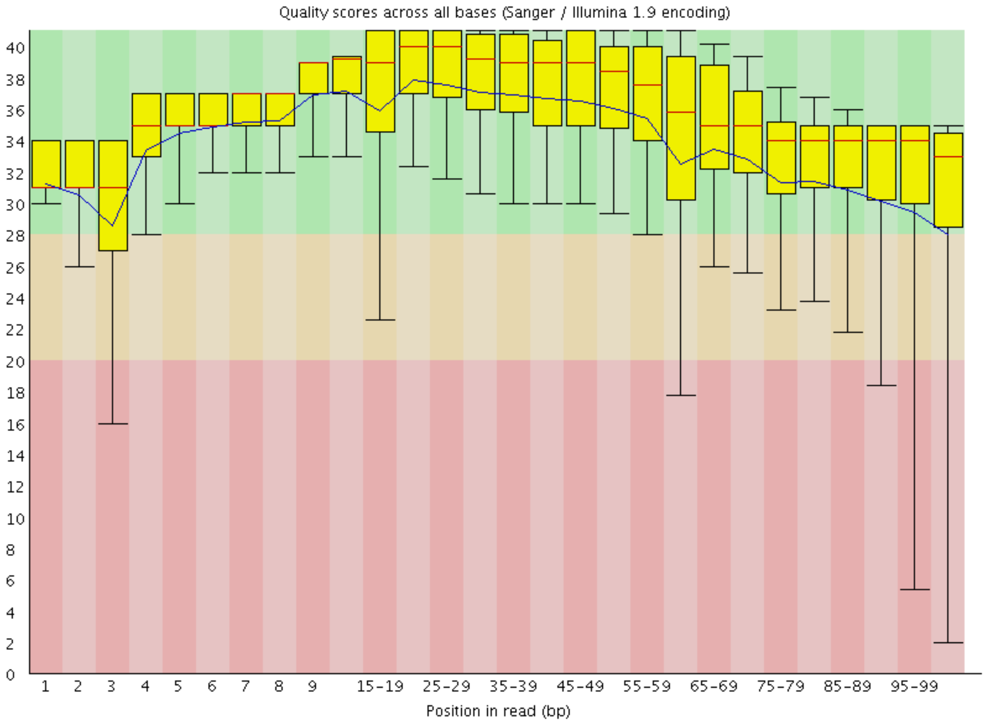
## FASTQC

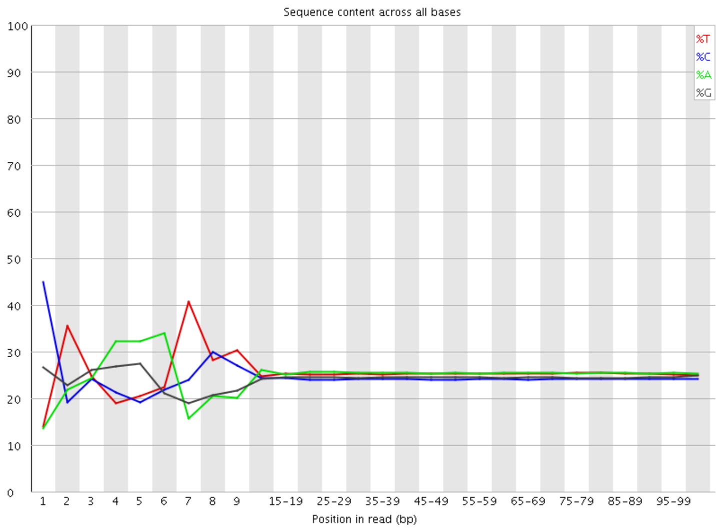


# QC: Raw Data

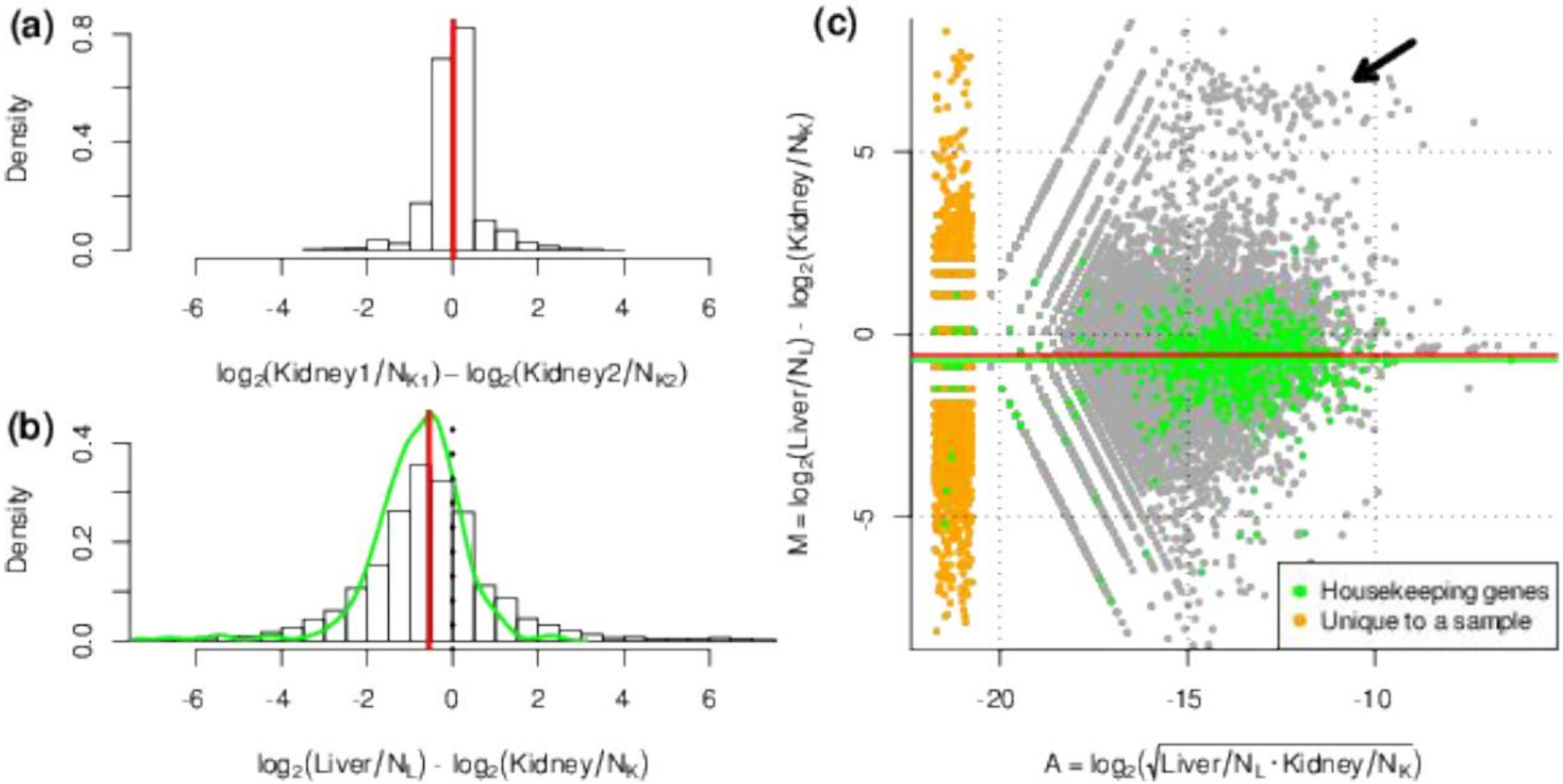
Sequence call quality

Sequence bias



# Normalization Required



# 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

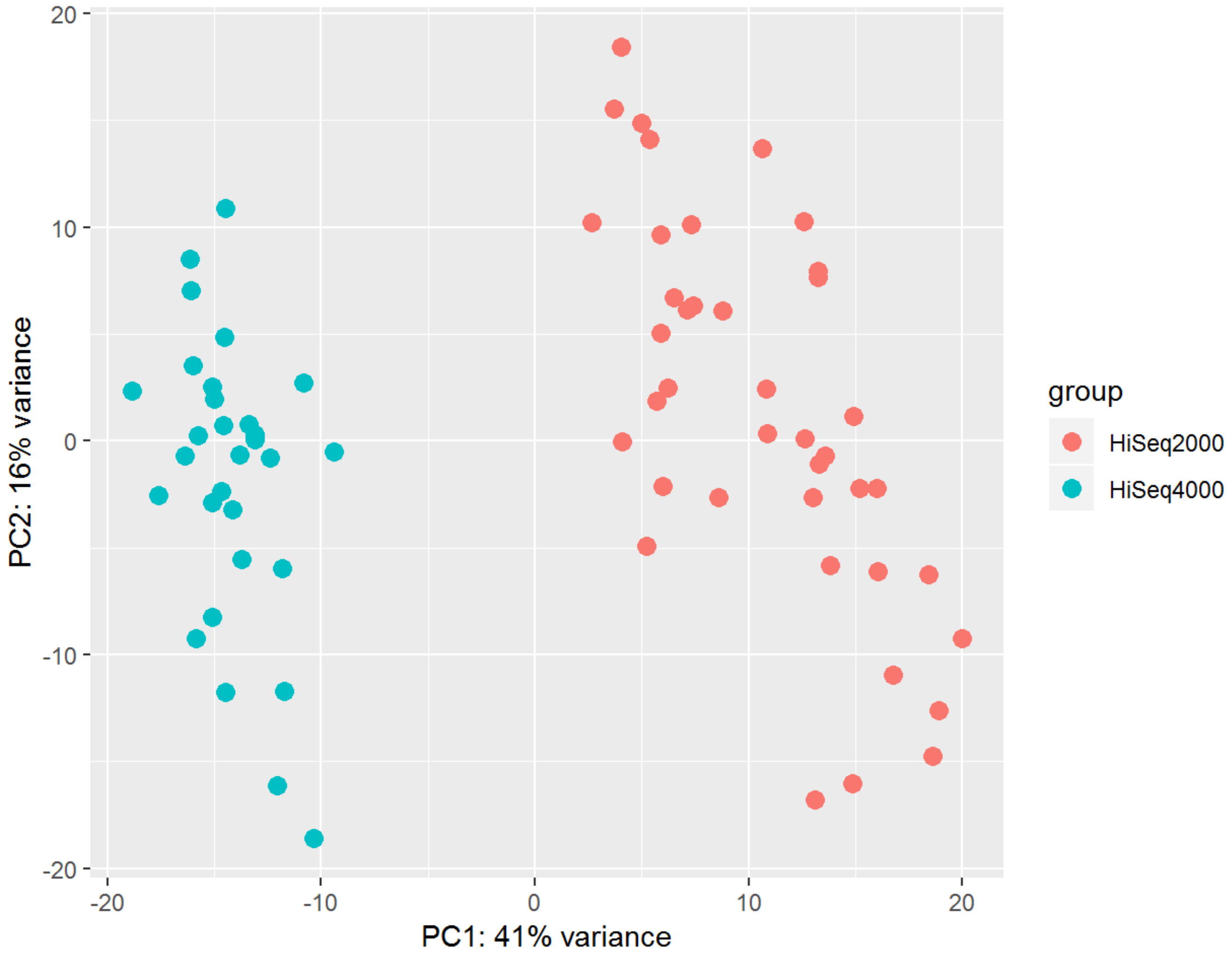
- **Log and relative log transformation;**

- 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);

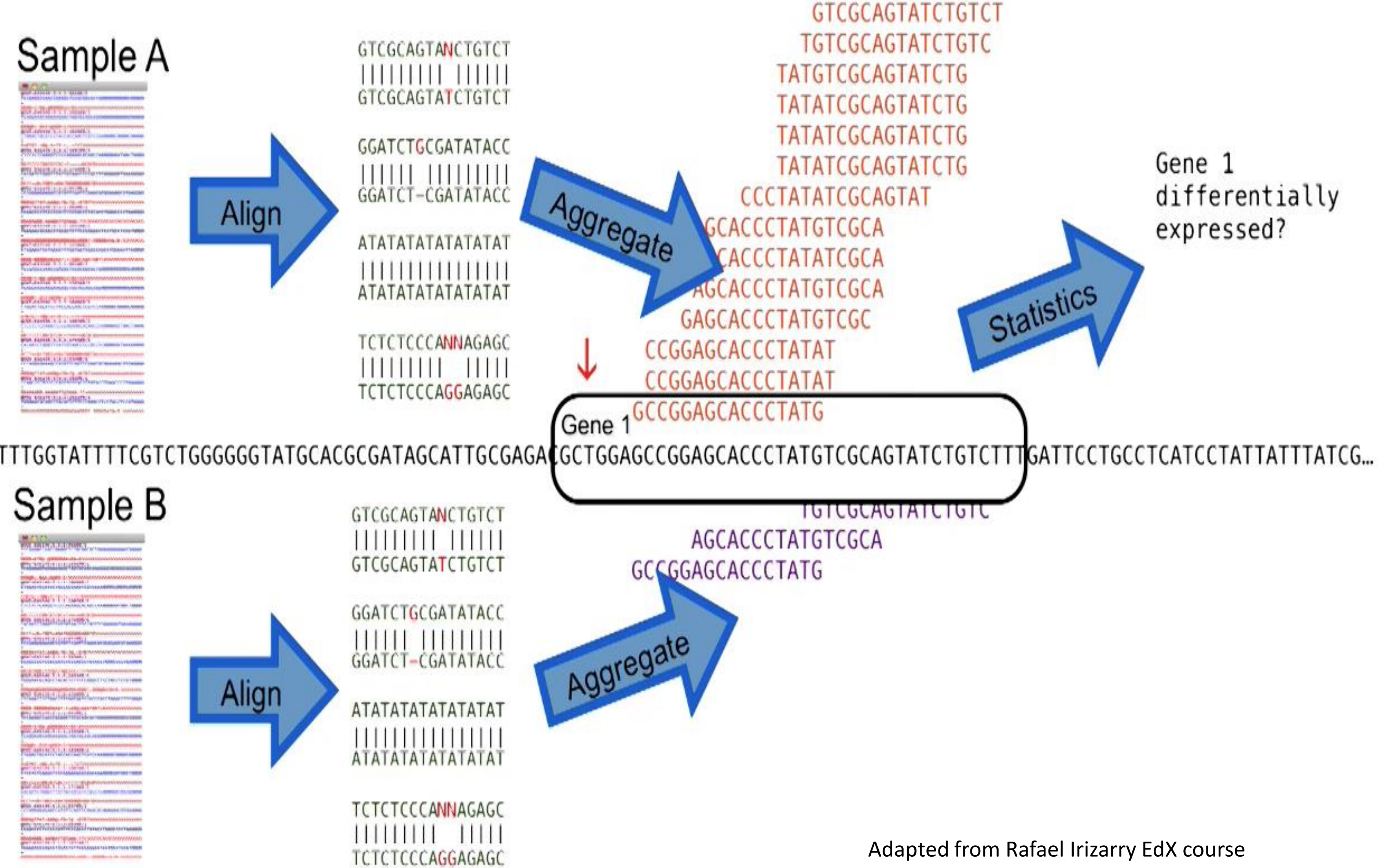
- Median ratio method (size factor);

- Quantile normalization methods;

# Batch effect

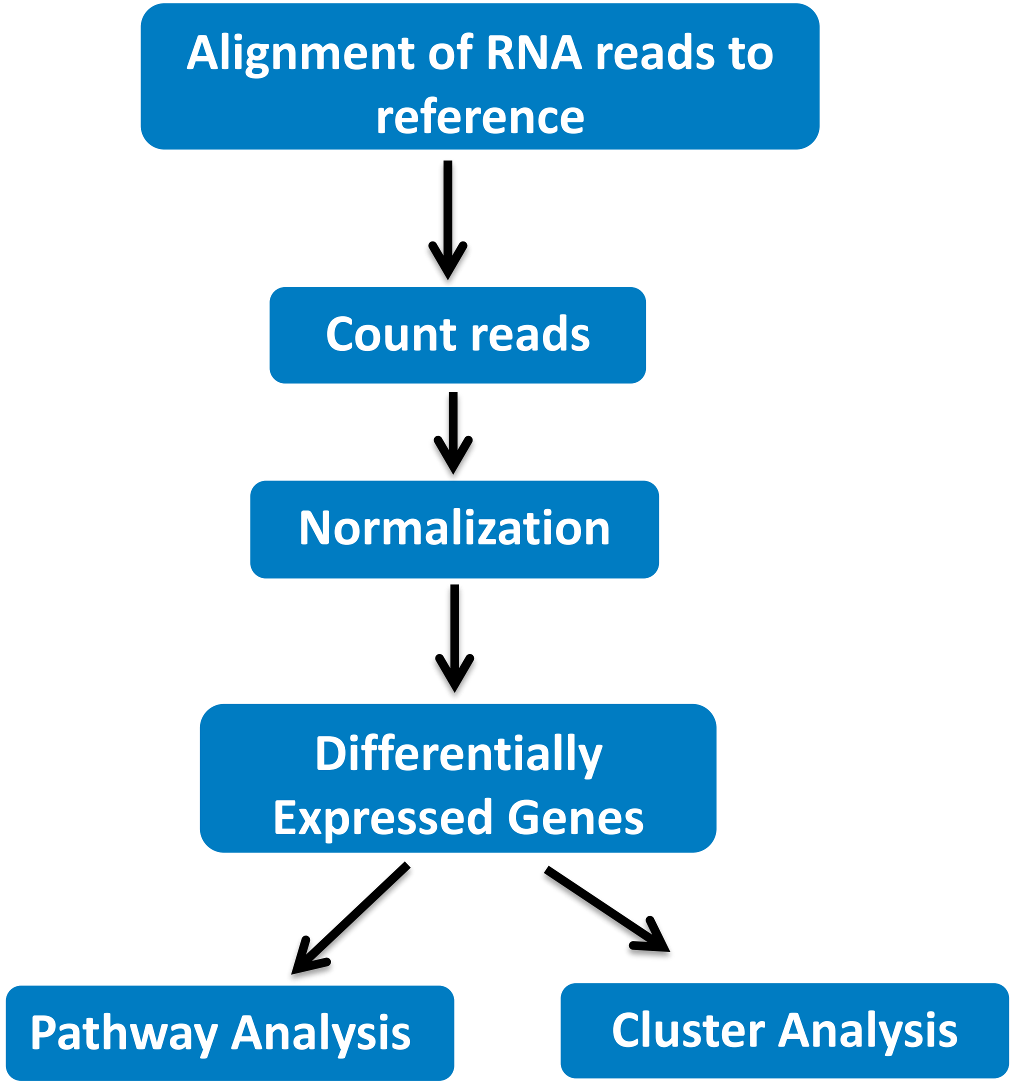


# RNA-seq data analysis overview



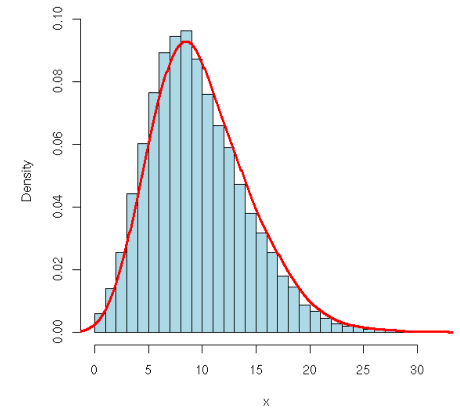

Adapted from Rafael Irizarry EdX course

# Workflow of RNA-Seq Gene Expression Data



# 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

# 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 control for batch effects)
- Batch effects
- ANOVA-like tests

# Pathways database

REACTOME
A CURATED PATHWAY DATABASE

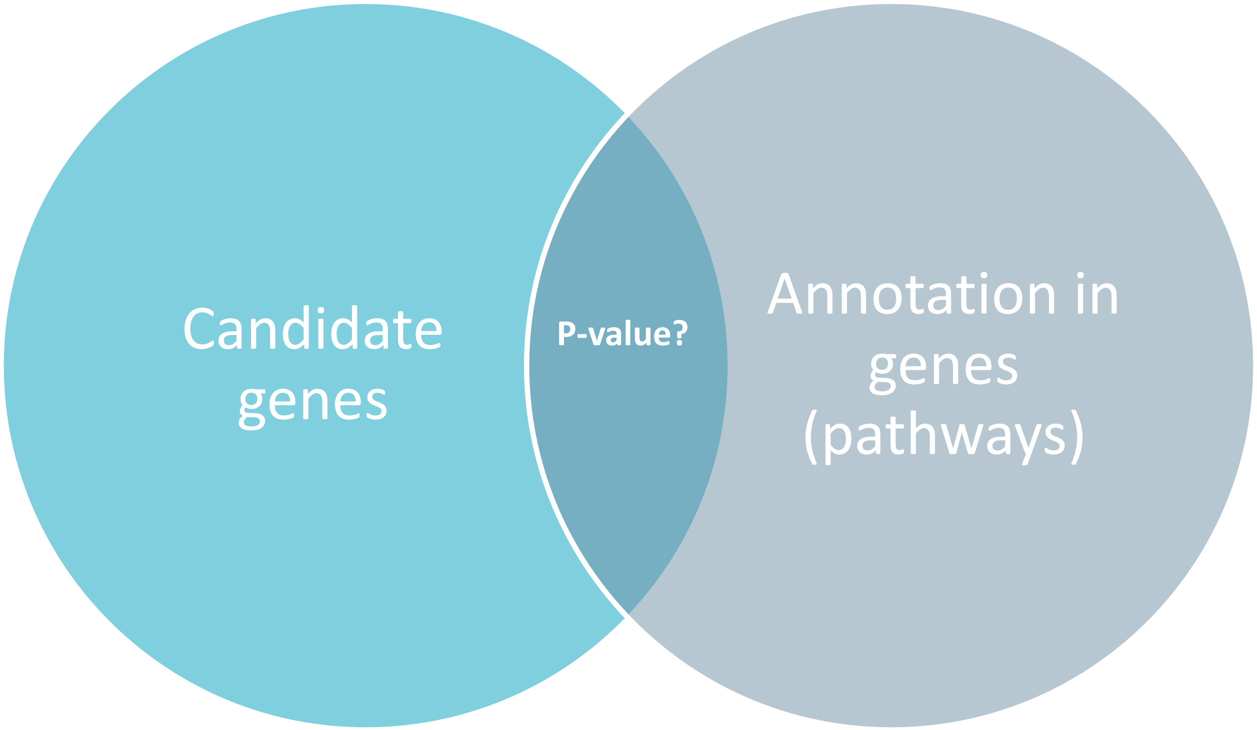
KEGG
Kyoto Encyclopedia of Genes and Genomes

PC Pathway Commons

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:

DAVID BIOINFORMATICS DATABASE

Enrichr

PANTHER Classification System

g:Profiler

| | |
|---|---|
| g:GOSt | Gene Group Functional Profiling |
| g:Cocoa | Compact Compare of Annotations |
| g:Convert | Gene ID Converter |
| g:Sorter | Expression Similarity Search |
| g:Orth | Orthology search |
| g:SNPense | Convert rsID |

Welcome! | Contact | FAQ | R / APIs | Beta | Archive

GOrilla

*Gene Ontology enRIchment anaLysis and visuaLizAtion tool*

GSEA

Gene Set Enrichment Analysis

Categorical Statistics

Most popular system (mostly historic)

Has been behind the latest annotation

- Was recently updated though

Lots of support for different IDs and Species

Configurable gene sets

Simple output presentation

# Enrichr

- 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

# Questions & Practicals

QUESTIONS?