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

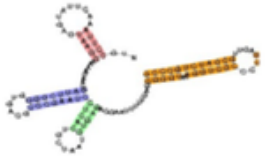
r.coutinho_de_almeida@lumc.nl

Outline

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

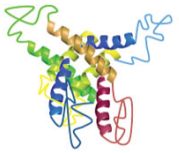
The Central Dogma of Molecular Biology

snoRNAs



snRNAs

tRNAs



DNA

Replication

~ 80%

RNA

Transcription

~2%

Protein

Translation

Alternative spliced mRNAs

lncRNAs

microRNAs

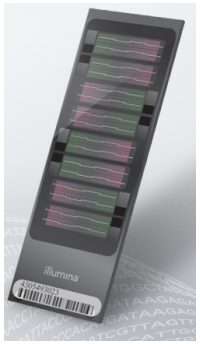
Transcriptomics

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

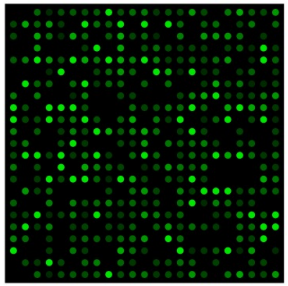
What can the transcriptome tell us?

- 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

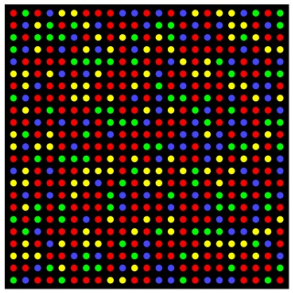


Microarray chip



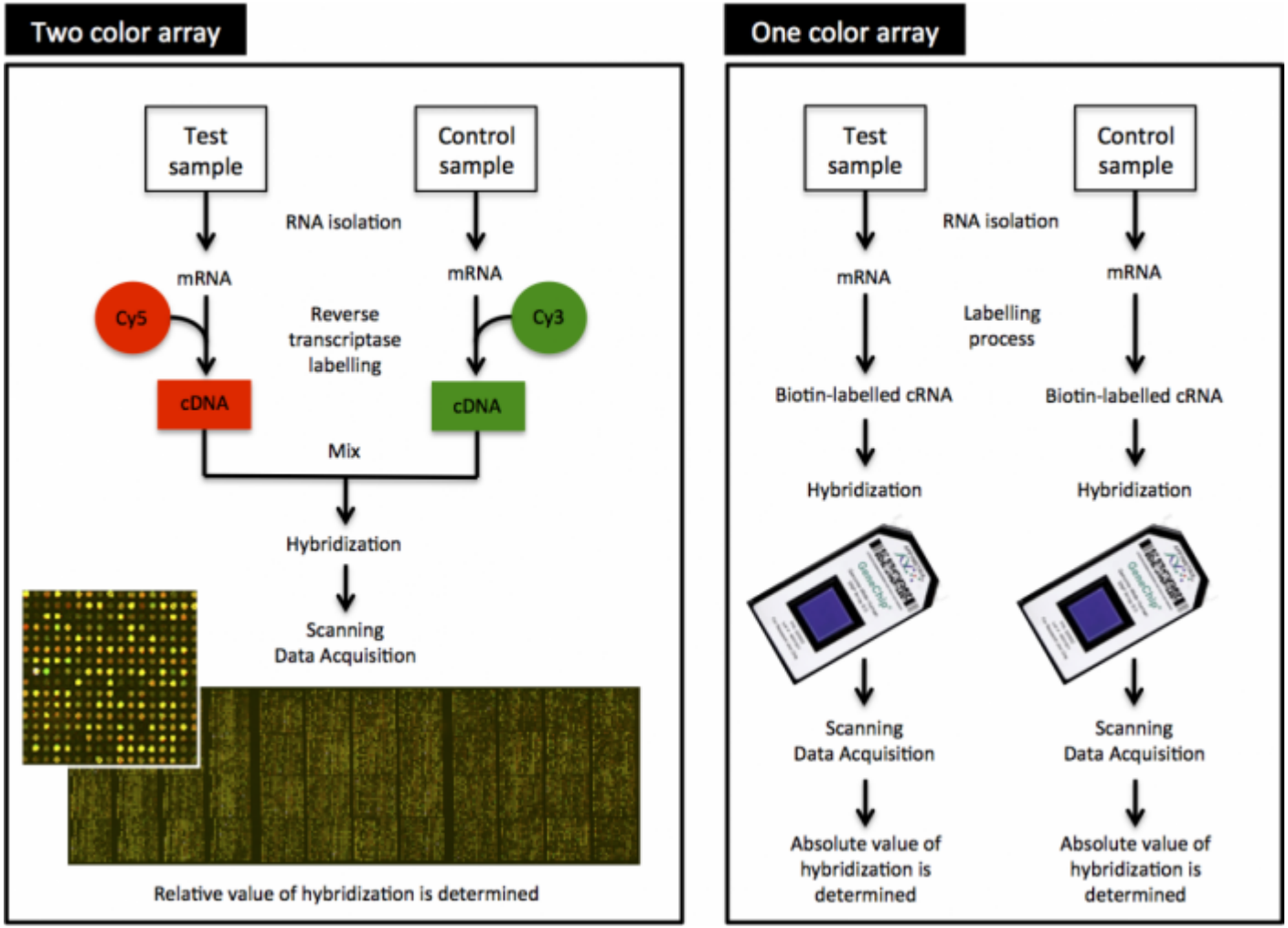
Microarray

Sequencing flow cell



RNA-seq

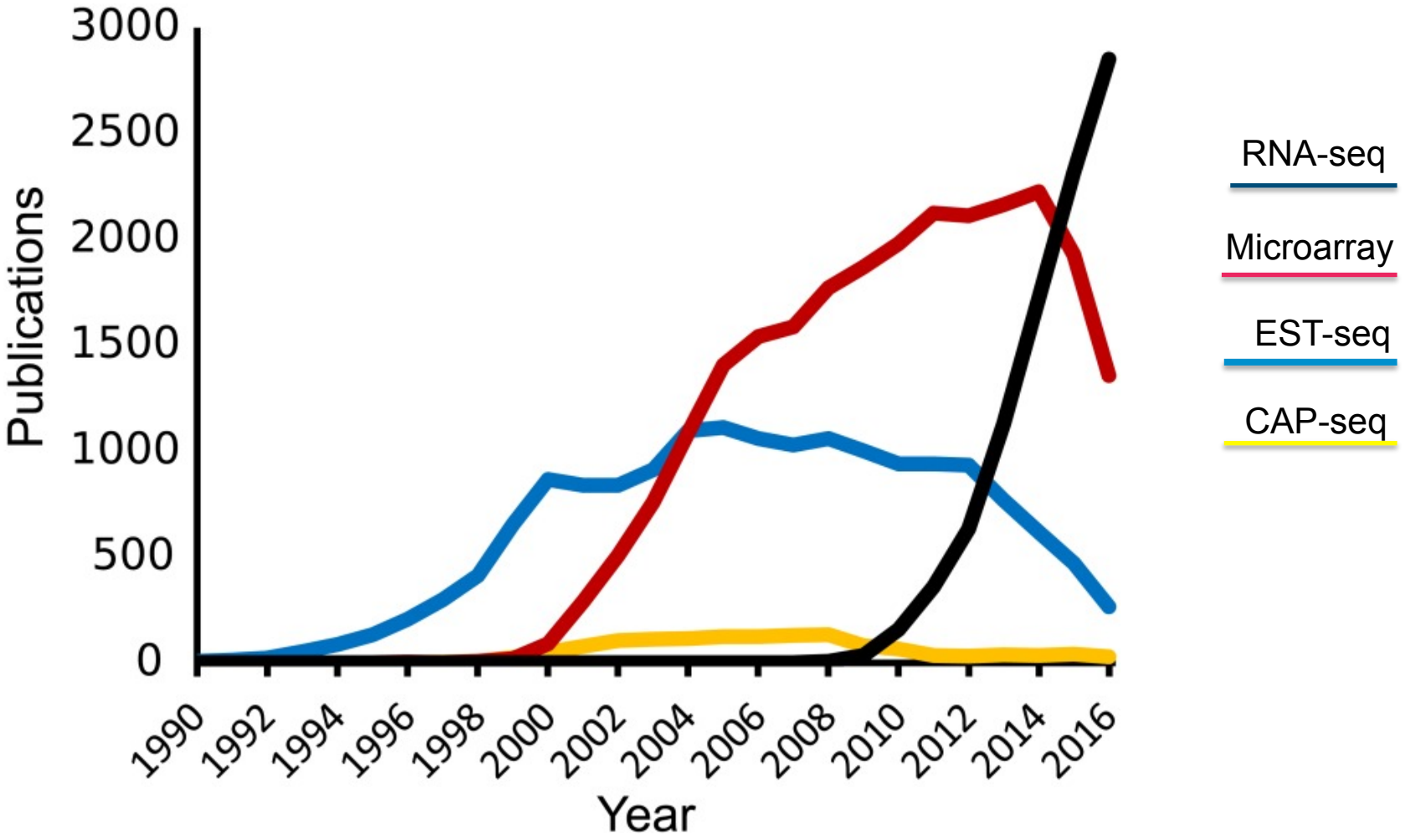
Microarray


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/
- SRA: https://www.ncbi.nlm.nih.gov/sra

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);

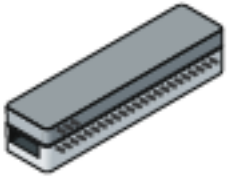
Applications of RNA-seq

- 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;

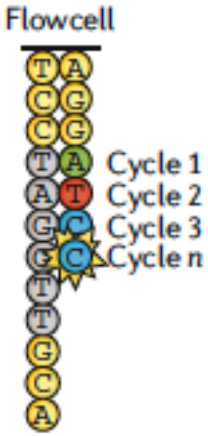
Main sequencing technologies for RNA-seq

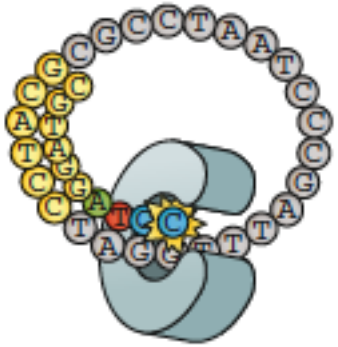
Illumina

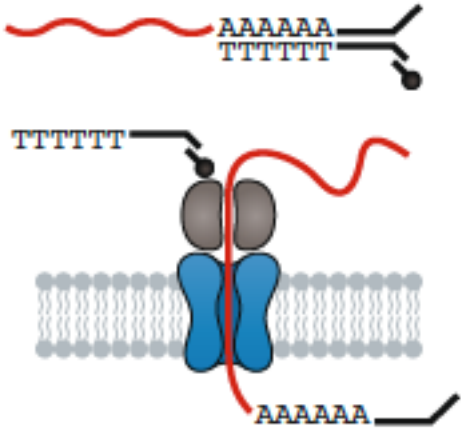
Pacific Biosciences

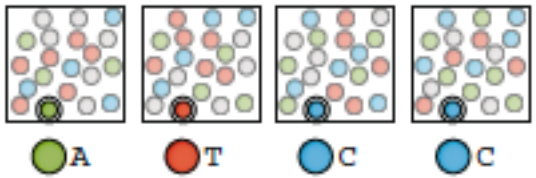


Oxford Nanopore

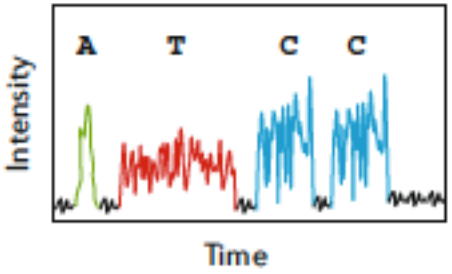




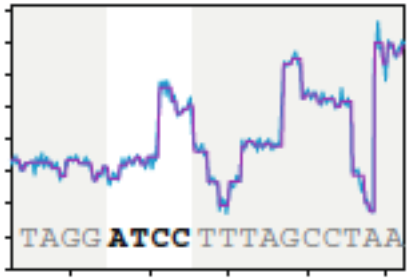




Short-read



Long-read

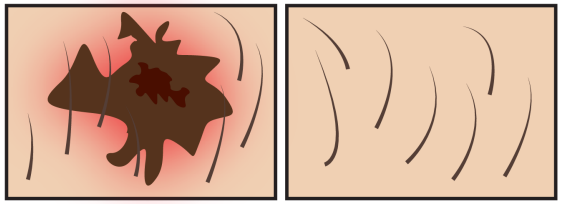


Direct

Stark et al., Nat Rev Gen 2019

Typical RNA-seq experiments (Short-read)

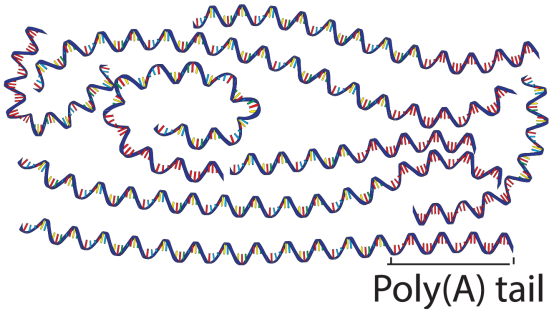
Samples of interest



Condition 1 (e.g. tumor)

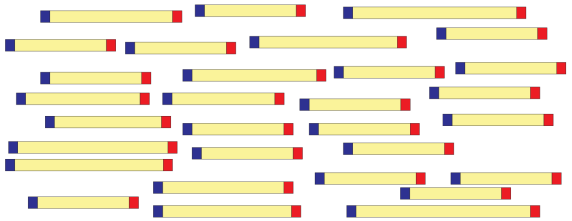
Condition 2 (e.g. normal)

Isolate RNAs

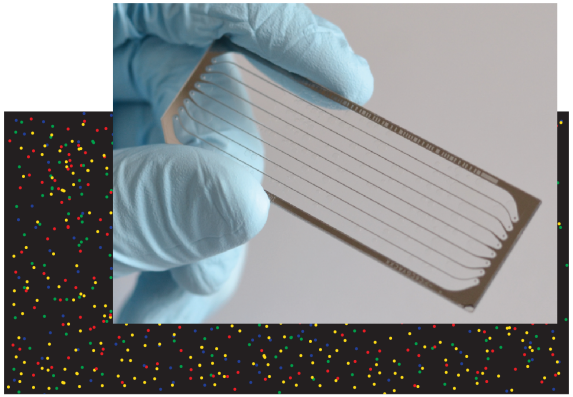


Poly(A) tail

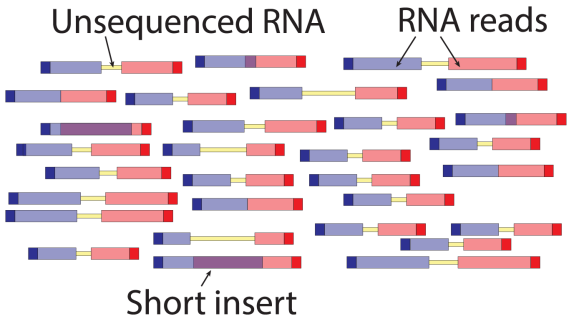
Generate cDNA, fragment, size select, add linkers



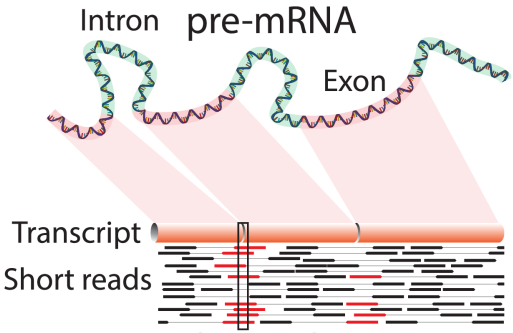
Sequence ends



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



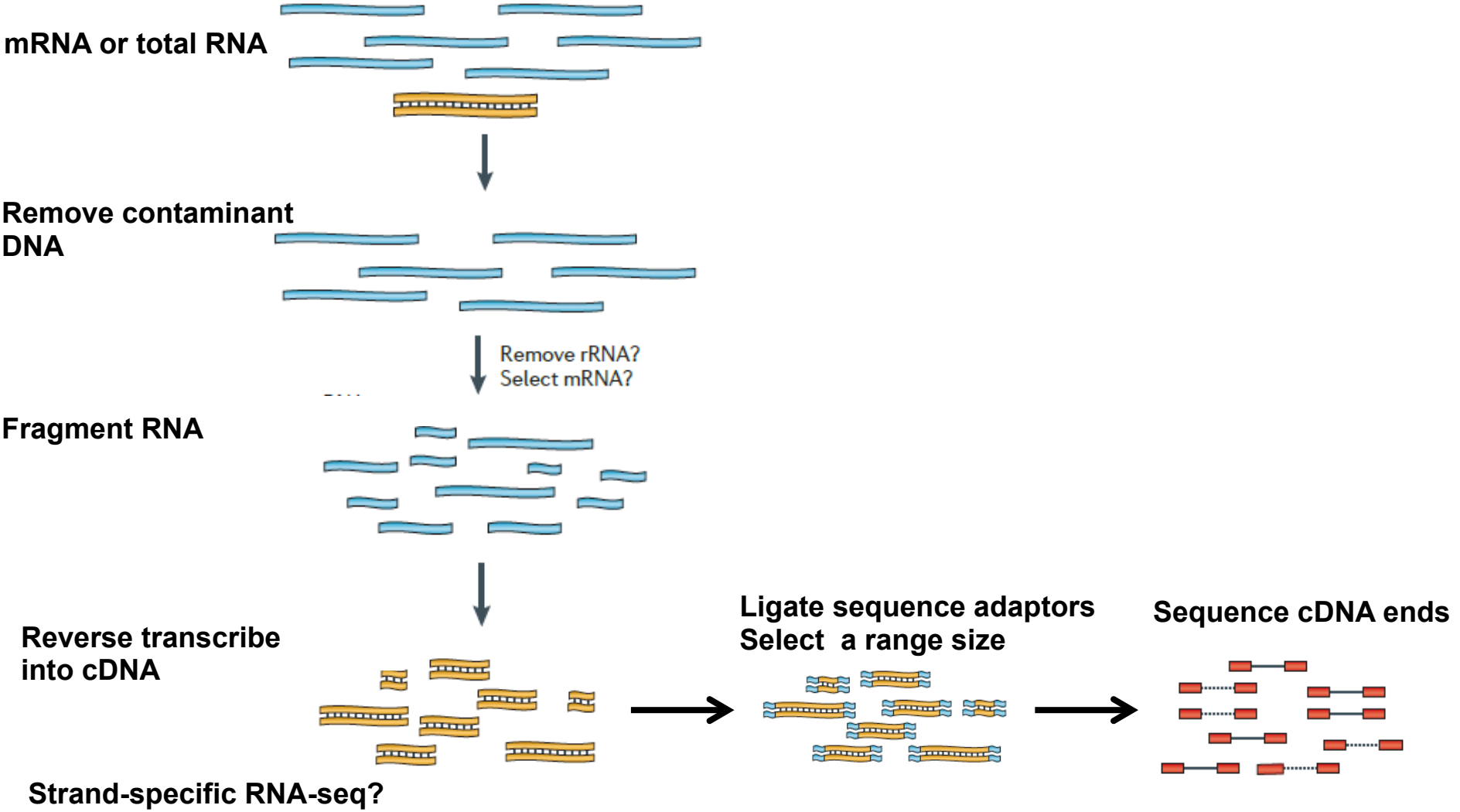
Map to genome, transcriptome, and predicted exon junctions



Downstream analysis

Source: Wikipedia

RNA-seq data generation

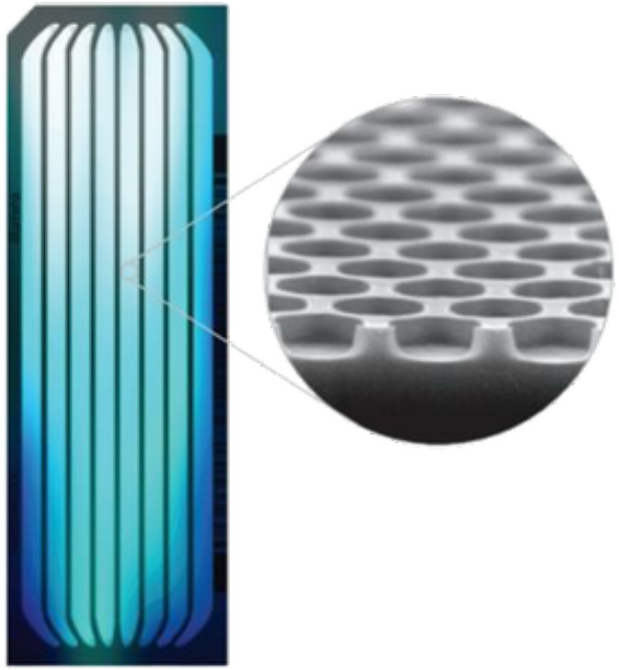


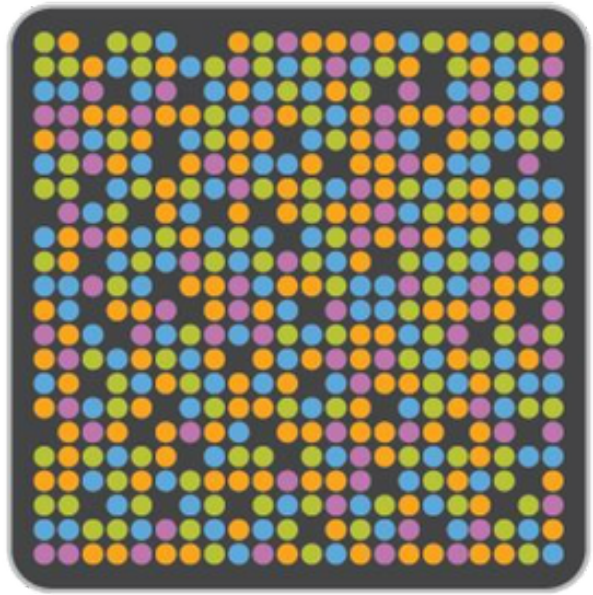
Adapted from Martin and Wang., Nat.Rev.Gen. 2011

Illumina flow cell

A
G
C
T

G	G	G	T
C	C	C	T
A	A	A	C
G	G	G	A
A	A	A	G
C	C	C	A
A	A	A	G





http://yourgene.pixnet.net/album

FASTQ file

The first line start with @ and is the ID for the sequence

```
@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=BDIIIIIIIIIFHHHHIIIHEFFHGGDD8A>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?
```

The second line contains the base call for sequenced fragment

After the + are the quality scores for each base in the sequence fragment

RNA-seq analysis

- Quality Control;
- Alignment;
- Count the number of reads per gene;
- Normalization;
- Sample and gene QC;
- Differential expression;
- Pathway analysis;

Quality Control (QC)

- Discard reads that have low quality base calls;
- Discard read that are artifacts of the chemistry;

Typical read is a cDNA fragment

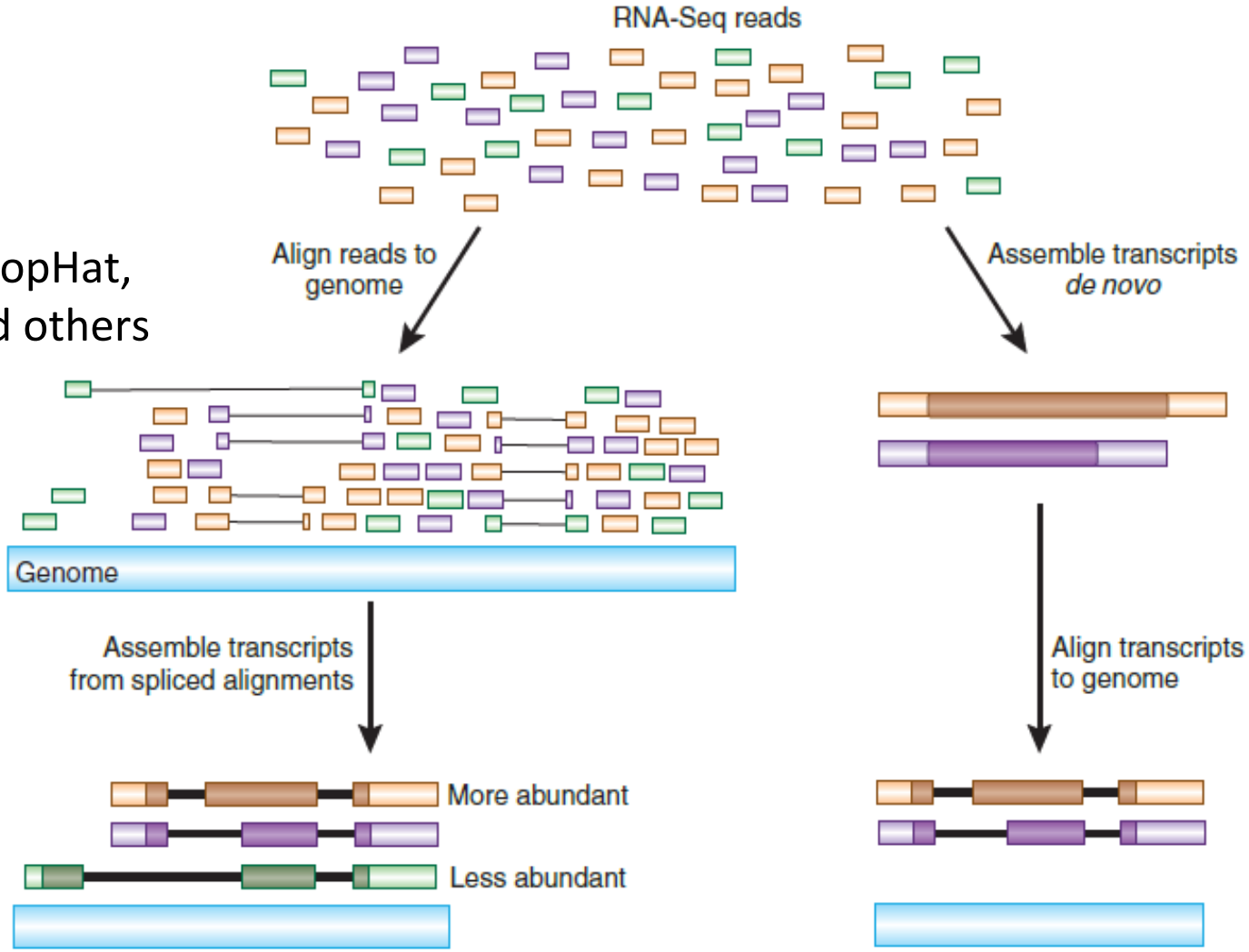
Adapter sequence

Sometimes the adapters bind to each other and the "read" is an adapter sequence

This should be discard

RNA-seq align and assemble

GSNAP, TopHat, STAR and others



Trinity, Kallisto and others

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

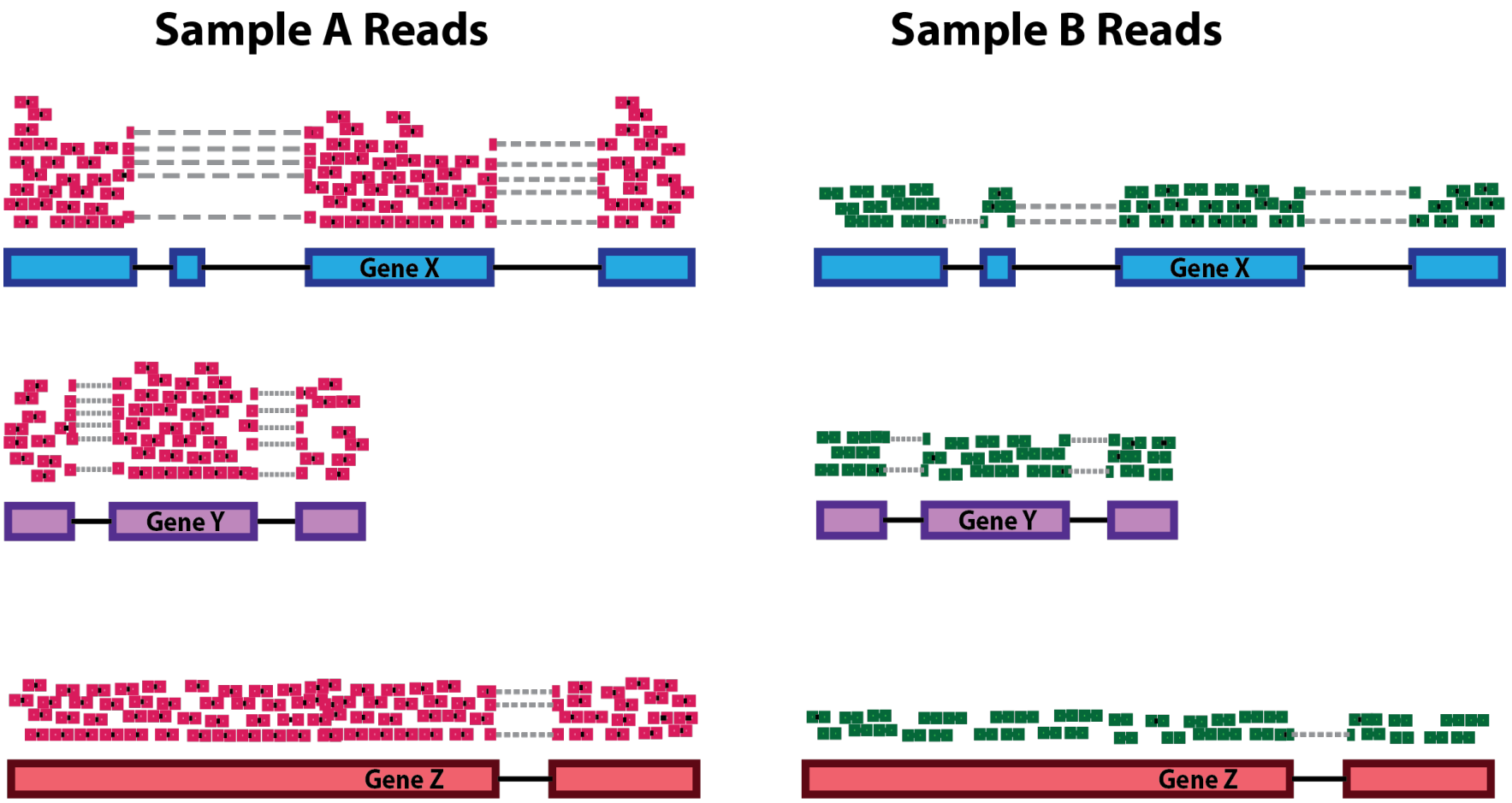
Read counts matrix

Gene	Sample1	Sample2	Sample3…
A1BG	30	5	13…
A1BG-AS1	24	10	18…
A1CF	0	0	0…
A2M	5	9	7…
A2M-AS1	3563	5771	4123…
A2ML1	13	8	7…
...	...	...	...

Normalization required

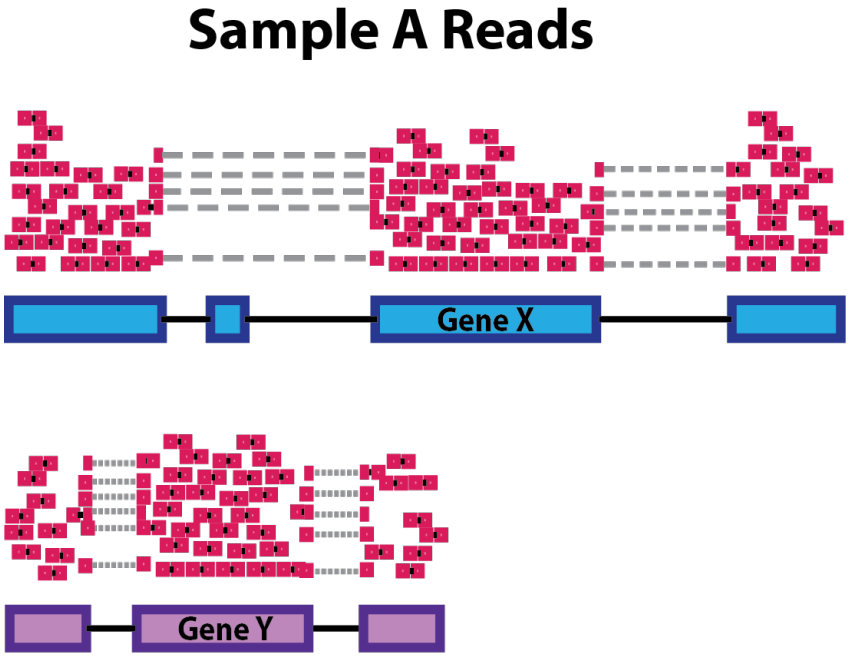
Gene	Sample #1 635 reads	Sample #2 1,270 reads
A1BG	30	60
A1BG-AS1	24	48
A1CF	0	0
A2M	563	1126
A2M-AS1	5	10
A2ML1	13	26

Normalization: Sequencing depth



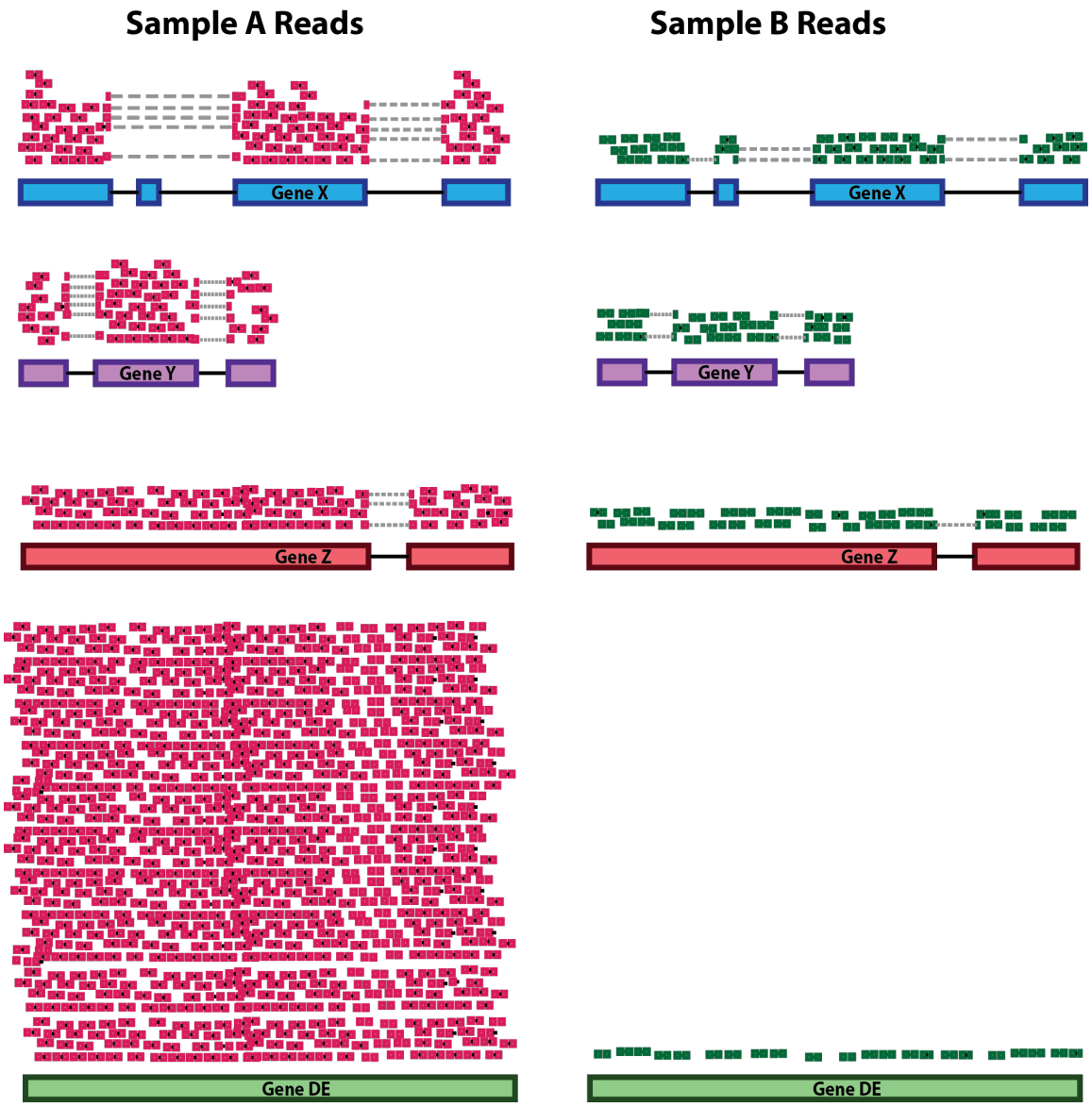
Harvard Chan Bioinformatics Core (HBC).

Normalization: Gene length



Harvard Chan Bioinformatics Core (HBC).

Normalization: RNA composition



Harvard Chan Bioinformatics Core (HBC).

Normalization methods

Not recommended for statistical testing:

- RPKM (reads per kb per million mapped reads) -
- FPKM (fragment per kb per million mapped reads);
- CPM (counts per million reads);

For statistical testing:

- EdgeR's TMM (trimmed mean of M values);
- **DESeq2 Median ratio method (size factor)**;

DESeq2 Normalization

Step 1: Creates a pseudo-reference sample (row-wise geometric mean)

gene	sampleA	sampleB	pseudo-reference sample
EF2A	1489	906	sqrt(1489 * 906) = **1161.5**
ABCD1	22	13	sqrt(22 * 13) = **17.7**
...	...	...	...

DESeq2 Normalization

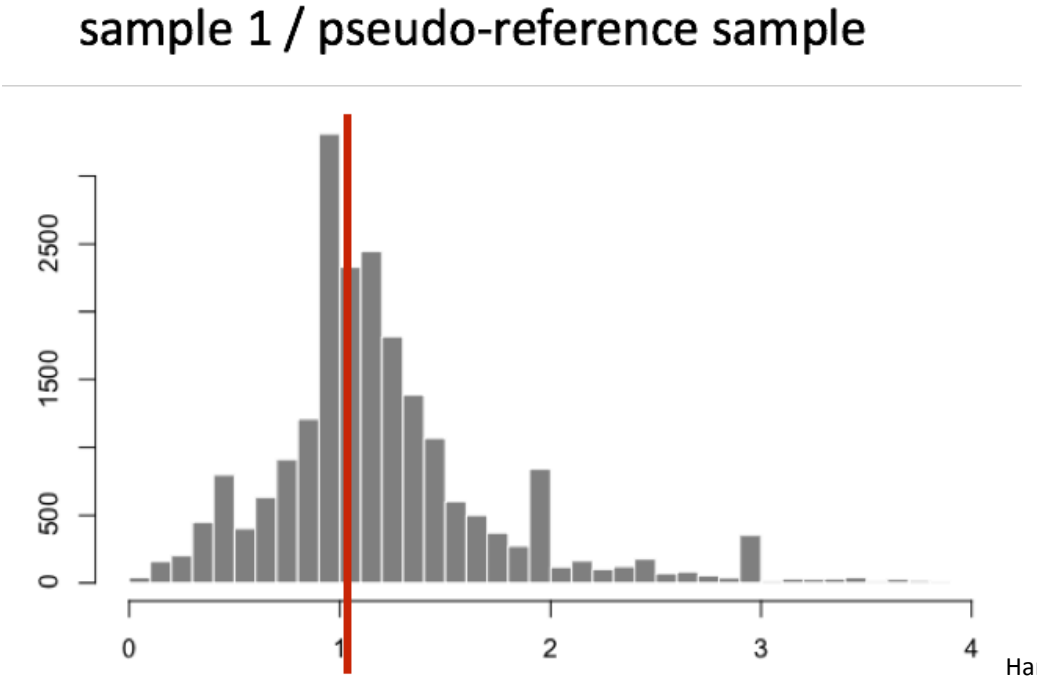
Step 2: Calculates ratio of each sample to the reference

gene	sampleA	sampleB	pseudo-reference sample	ratio of sampleA/ref	ratio of sampleB/ref
EF2A	1489	906	1161.5	1489/1161.5 = **1.28**	906/1161.5 = **0.78**
ABCD1	22	13	16.9	22/16.9 = **1.30**	13/16.9 = **0.77**
MEFV	793	410	570.2	793/570.2 = **1.39**	410/570.2 = **0.72**
BAG1	76	42	56.5	76/56.5 = **1.35**	42/56.5 = **0.74**
MOV10	521	1196	883.7	521/883.7 = **0.590**	1196/883.7 = **1.35**
...	...	...	...		

DESeq2 Normalization

Step 3: calculate the normalization factor for each sample (size factor)

```
normalization_factor_sampleA <- median(c(1.28, 1.3, 1.39, 1.35, 0.59))
normalization_factor_sampleB <- median(c(0.78, 0.77, 0.72, 0.74, 1.35))
```



Harvard Chan Bioinformatics Core (HBC).

DESeq2 Normalization

Step 4: calculate the normalized count values using the normalization factor

For example, if the median ratio for SampleA was 1.3 and the median ratio for SampleB was 0.77, you could calculate normalized counts as follows:
SampleA median ratio = 1.3
SampleB median ratio = 0.77

gene	sampleA	sampleB
EF2A	1489 / 1.3 = **1145.39**	906 / 0.77 = **1176.62**
ABCD1	22 / 1.3 = **16.92**	13 / 0.77 = **16.88**
...	...	...

DESeq2 Normalization

Raw counts

gene	sampleA	sampleB
EF2A	1489	906
ABCD1	22	13
...	...	...

Normalized counts

gene	sampleA	sampleB
EF2A	1489 / 1.3 = **1145.39**	906 / 0.77 = **1176.62**
ABCD1	22 / 1.3 = **16.92**	13 / 0.77 = **16.88**
...	...	...

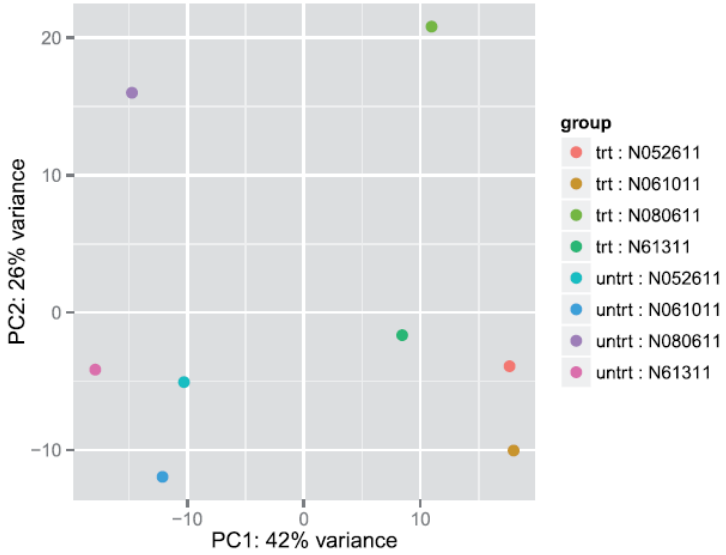
Sample level QC

- Which samples are similar to each other, which are different?
- Does this fit to the expectation from the experiment's design?
- What are the major sources of variation in the dataset?

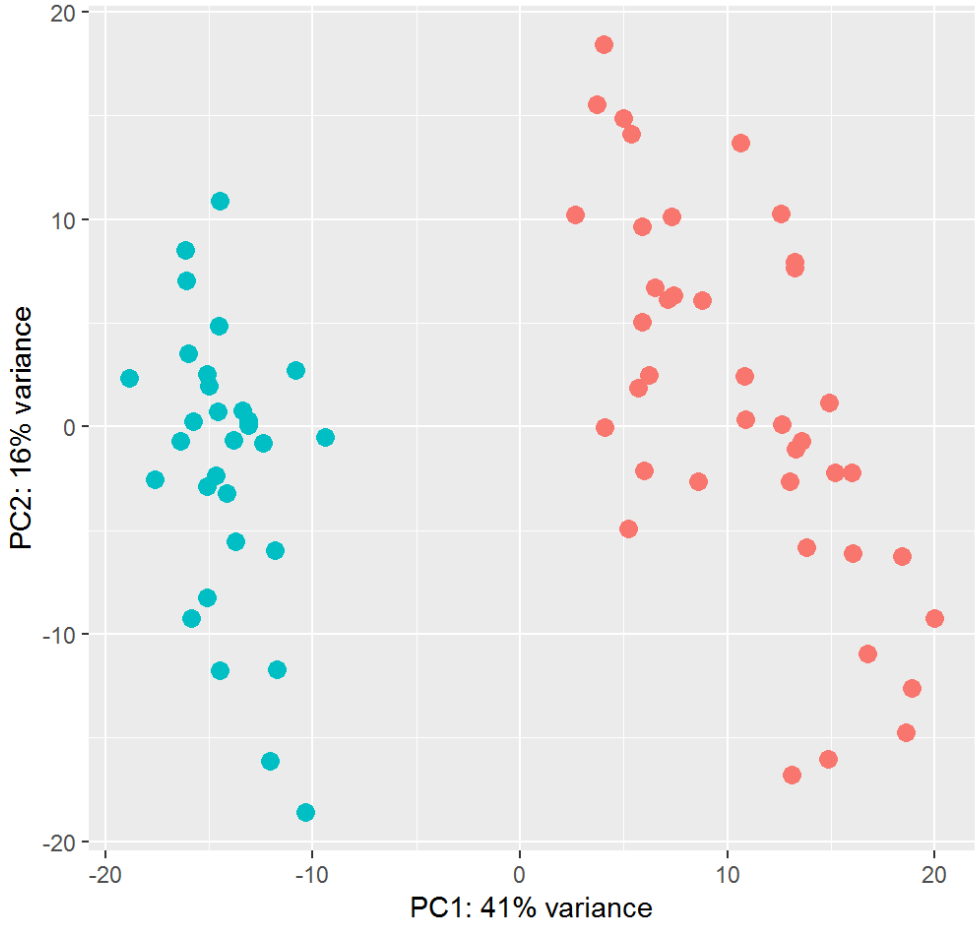
Sample level QC

Principal Component Analysis (PCA): A technique used to emphasize variation and bring out strong patterns in a dataset (dimensionality reduction

- Project a line through the data points in n dimensional space (n = genes)
- Measure how much variance there is from that line (the distance from each point to the line).
- PC1 explains highest variance, PC2 next highest etc.



Batch effect



To remove batch effect:

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

Sofwares:

Combat, limma, sva, and others

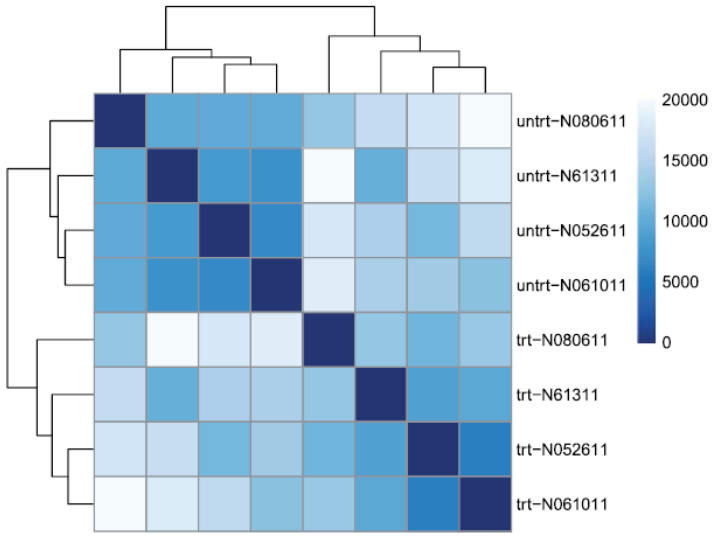
In DESeq2 uses:

- Variance Stabilizing Transformation
- Regularized log Transformation

Sample level QC

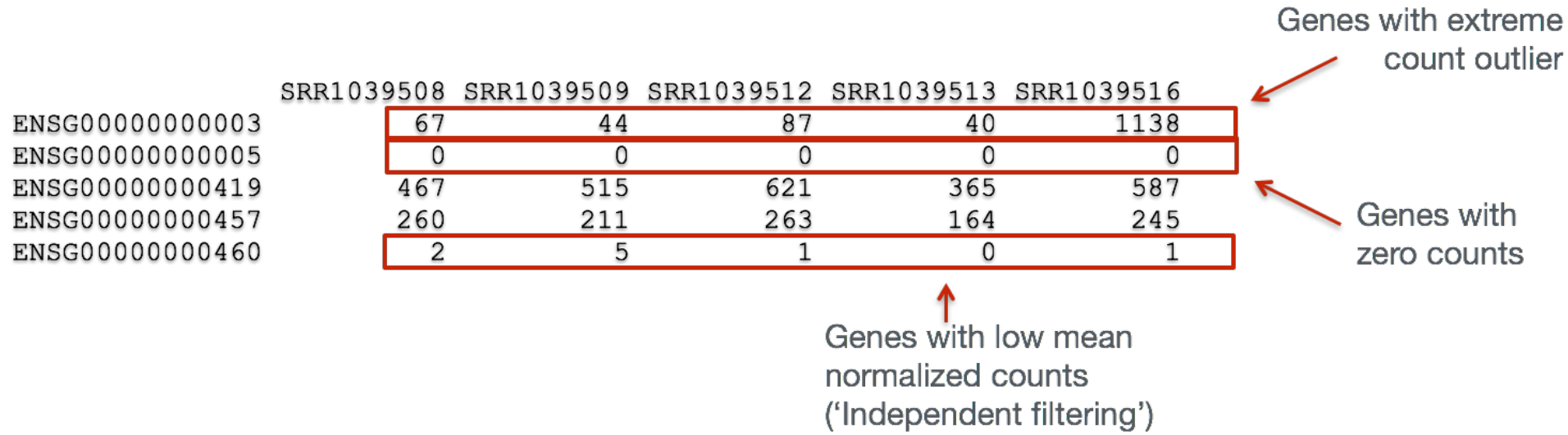
Hierarchical clustering

- Identify strong patterns in a dataset and potential outliers
- Correlation or distances for all pairwise combinations of samples.
- Generally high correlations with each other (values higher than 0.80)
- ‘Blocks’ indicate substructure in the data



Gene-level QC

- Genes with zero counts in all samples
- Genes with an extreme count outlier
- Genes with a low mean normalized counts

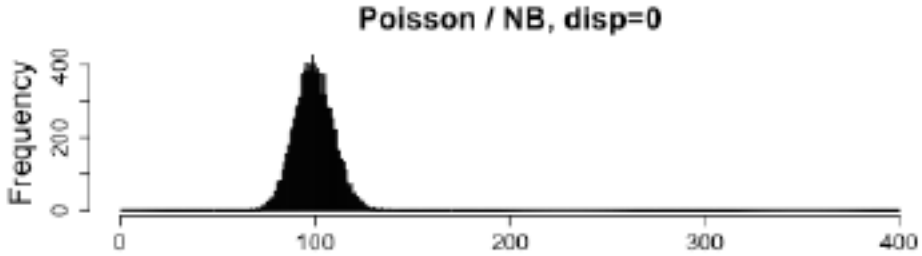


DESeq2 will perform this filtering by default; however other DE tools, such as EdgeR will not.

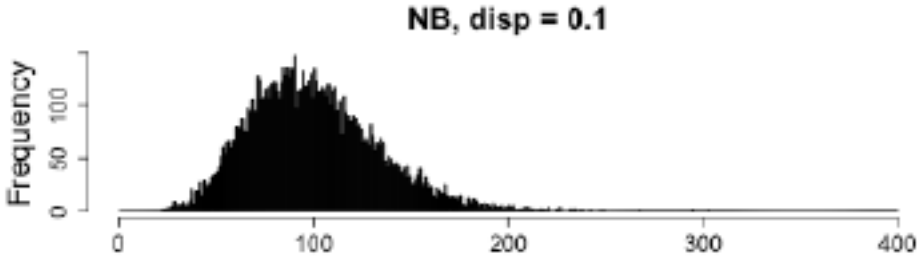
Statistical Testing in DEG Analysis

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

Poisson will underestimate the variability and the effect of the observed differences

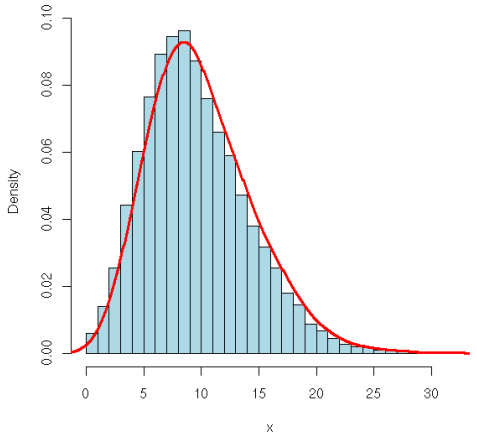


Instead we use the Negative Binomial.



Statistical Testing in DEG Analysis

- **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



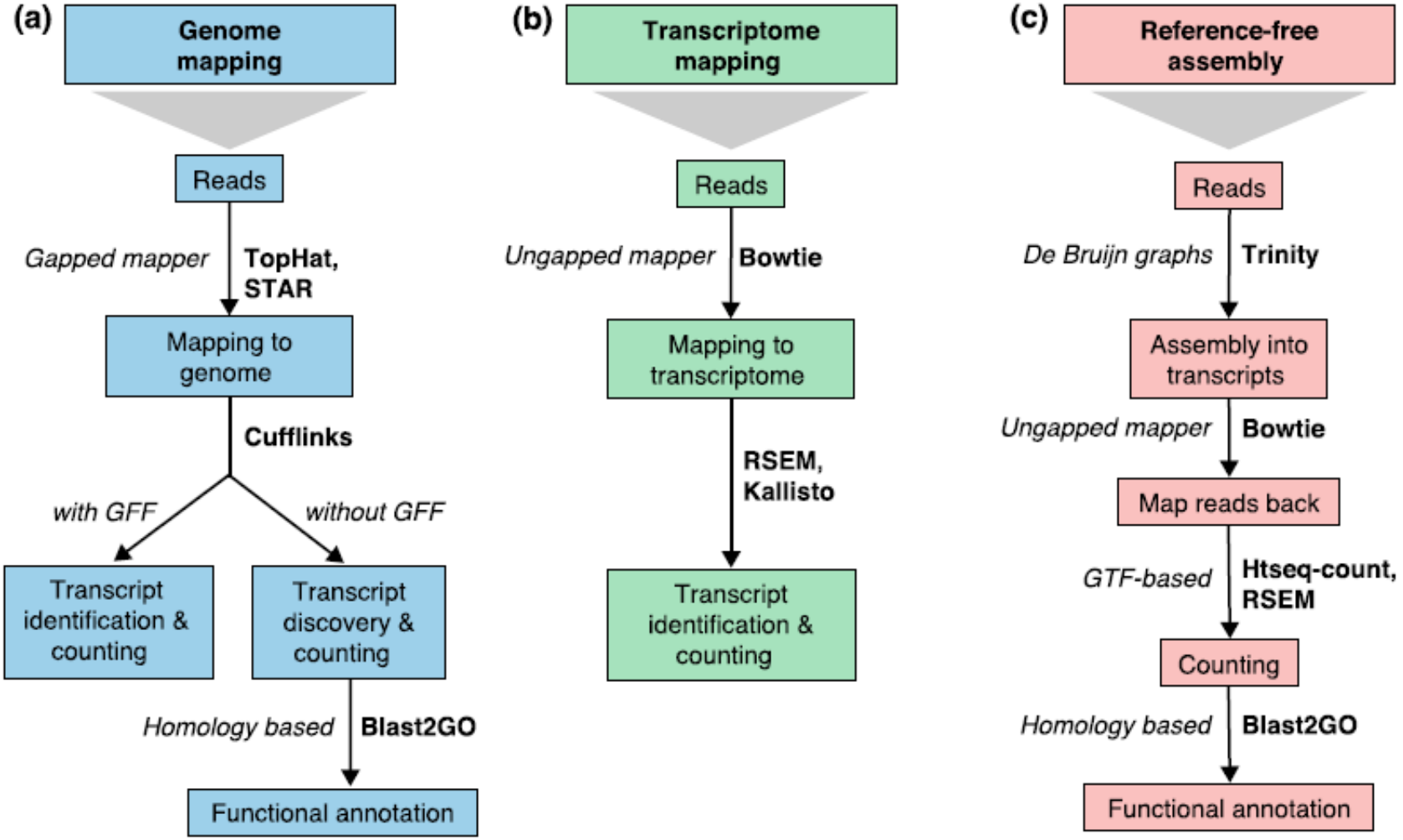
Steps in DEG Analysis

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)

RNA-seq options

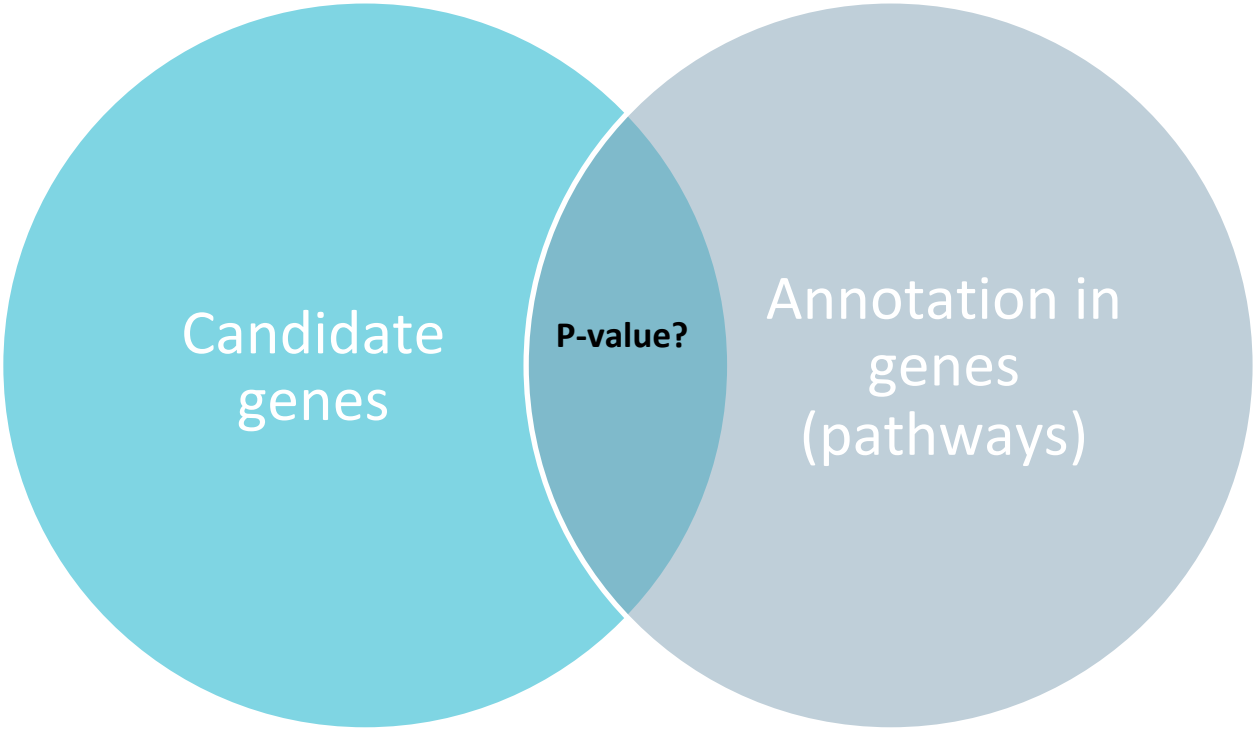


Conesa et al. Genome Biology (2016)

Pathways database

Pathways analysis

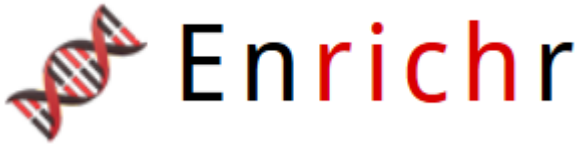
- Are there more annotations in a gene list than expected?

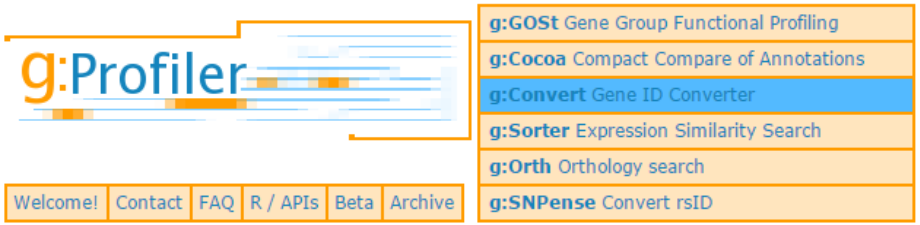


Tools for functional gene list analysis

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