

# Module 8: RNA-sequencing

Presented by:

**Ximena Ibarra-Soria**

GSK

[ximena.x.ibarra-soria@gsk.com](mailto:ximena.x.ibarra-soria@gsk.com)

@xIbarraSoria 

Based on materials by:

Victoria Offord

**Next Generation Sequencing Bioinformatics Course**  
18-22 January 2021 - Santiago - Chile



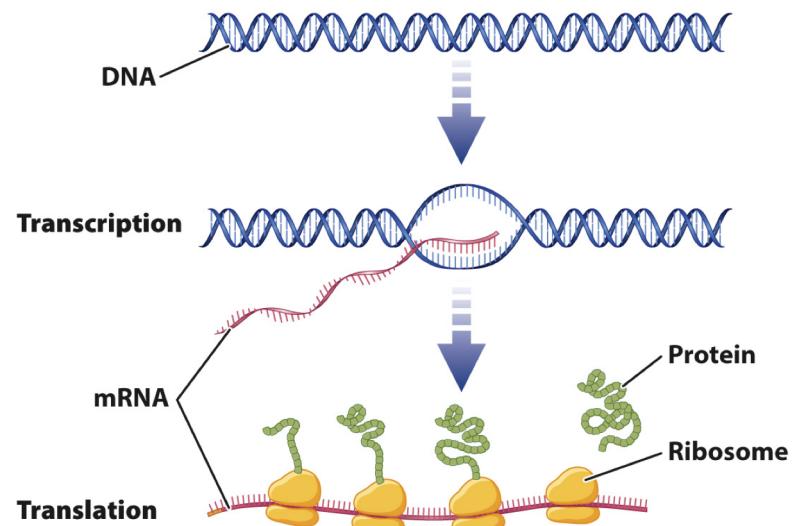
FACULTAD DE  
CIENCIAS BIOLÓGICAS  
PONTIFICIA UNIVERSIDAD  
CATÓLICA DE CHILE

WELLCOME GENOME CAMPUS

CONNECTING  
SCIENCE  
ADVANCED  
COURSES +  
SCIENTIFIC  
CONFERENCES

# Overview

- RNA-seq background.
- Experimental design.
- Data alignment.
- Quantification.
- Normalisation.
- Differential expression.
- Interpretation of results.

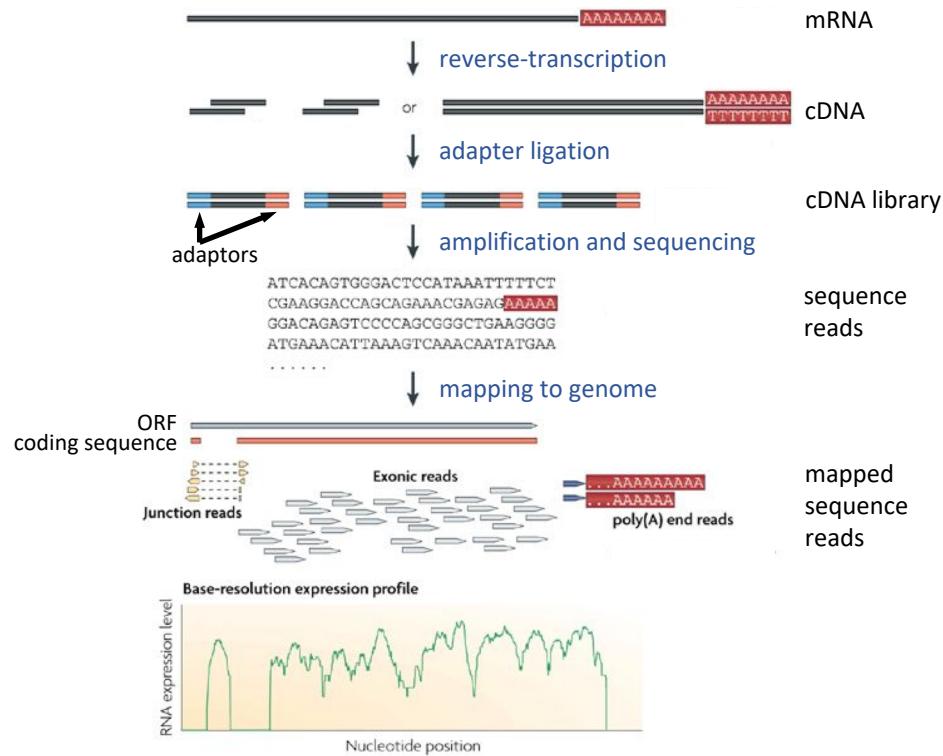
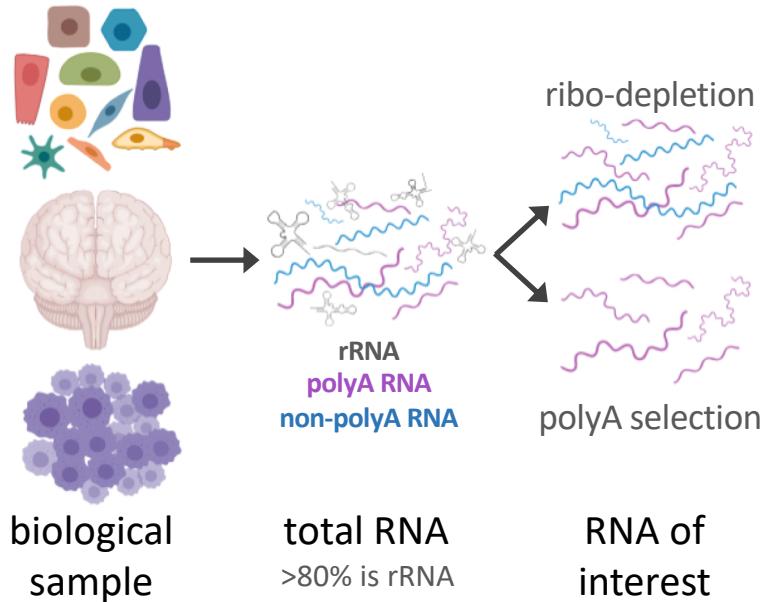


[http://www.macmillanhighered.com/BrainHoney/Resource/6716/digital\\_first\\_content/trunk/test/morris2e/asset/img\\_ch3/morris2e\\_ch03\\_fig\\_03\\_03.html](http://www.macmillanhighered.com/BrainHoney/Resource/6716/digital_first_content/trunk/test/morris2e/asset/img_ch3/morris2e_ch03_fig_03_03.html)

# RNA-sequencing

- Uses high-throughput sequencing to profile the **transcriptome** of a biological sample at a given time.
  - Transcriptome = set of RNA molecules present in a cell.
- Allows to
  - Catalogue all species of transcripts (messenger, small, non-coding).
  - Annotate the **structure** of genes (start, end, splice isoforms, UTRs).
  - Compare the **types and quantities** of the RNA molecules across time and conditions.

# RNA-sequencing



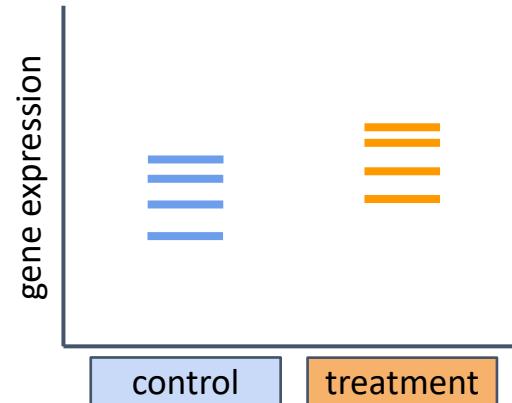
Modified from Wang, Gerstein and Snyder, *Nat Rev Genetics* 10 (2009)  
doi.org/10.1038/nrg2484

# Experimental design

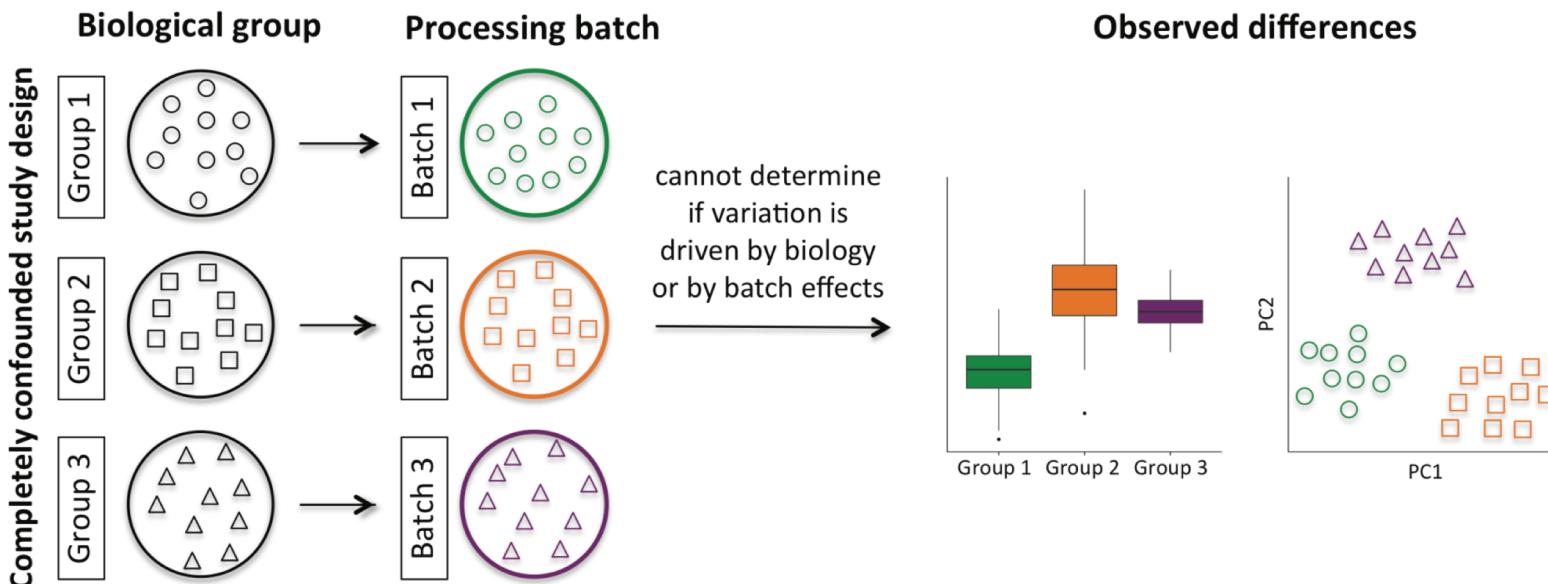
- Start by identifying **what is the question to answer** and what type of information is required.
  - **Single vs paired-end** (isoform analysis).
  - **Stranded vs unstranded** (antisense and overlapping transcripts).
  - **Sequencing depth** (detection of low abundance transcripts).
  - **Number of replicates:**
    - **Biological** = independent, biologically distinct samples.
    - **Technical** = repeated measurement of the same sample.
      - For differential expression analysis, increasing the number of replicates is better than increasing depth.

# Experimental design

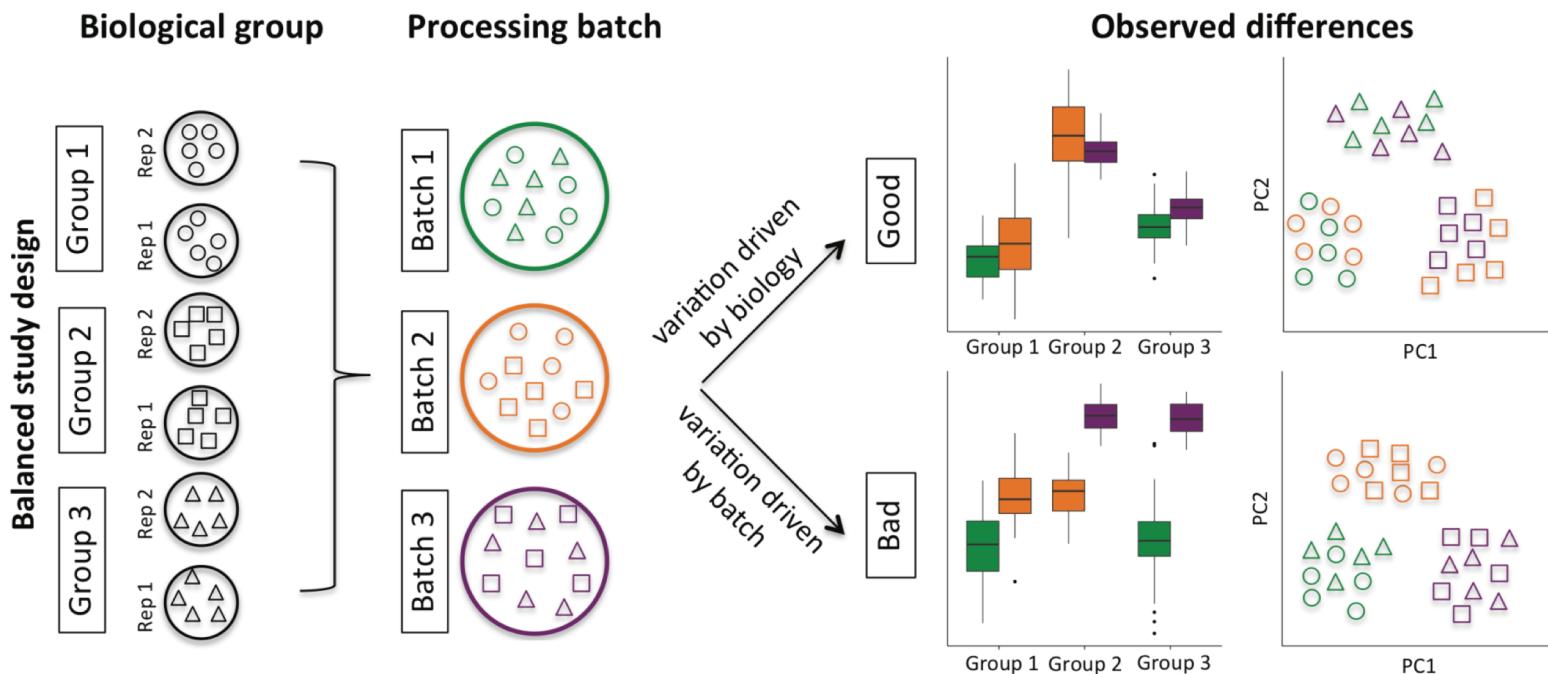
- Start by identifying **what is the question to answer.**
- Consider what are possible **sources of variation.**
  - Biological: sex, age, genetic background, ethnicity...
  - Technical: sample processing date, reagent's batch, time of sample collection...
- To estimate variation we need **biological replicates.**
- **Power** calculations.
  - Number of replicates needed to observe an effect.



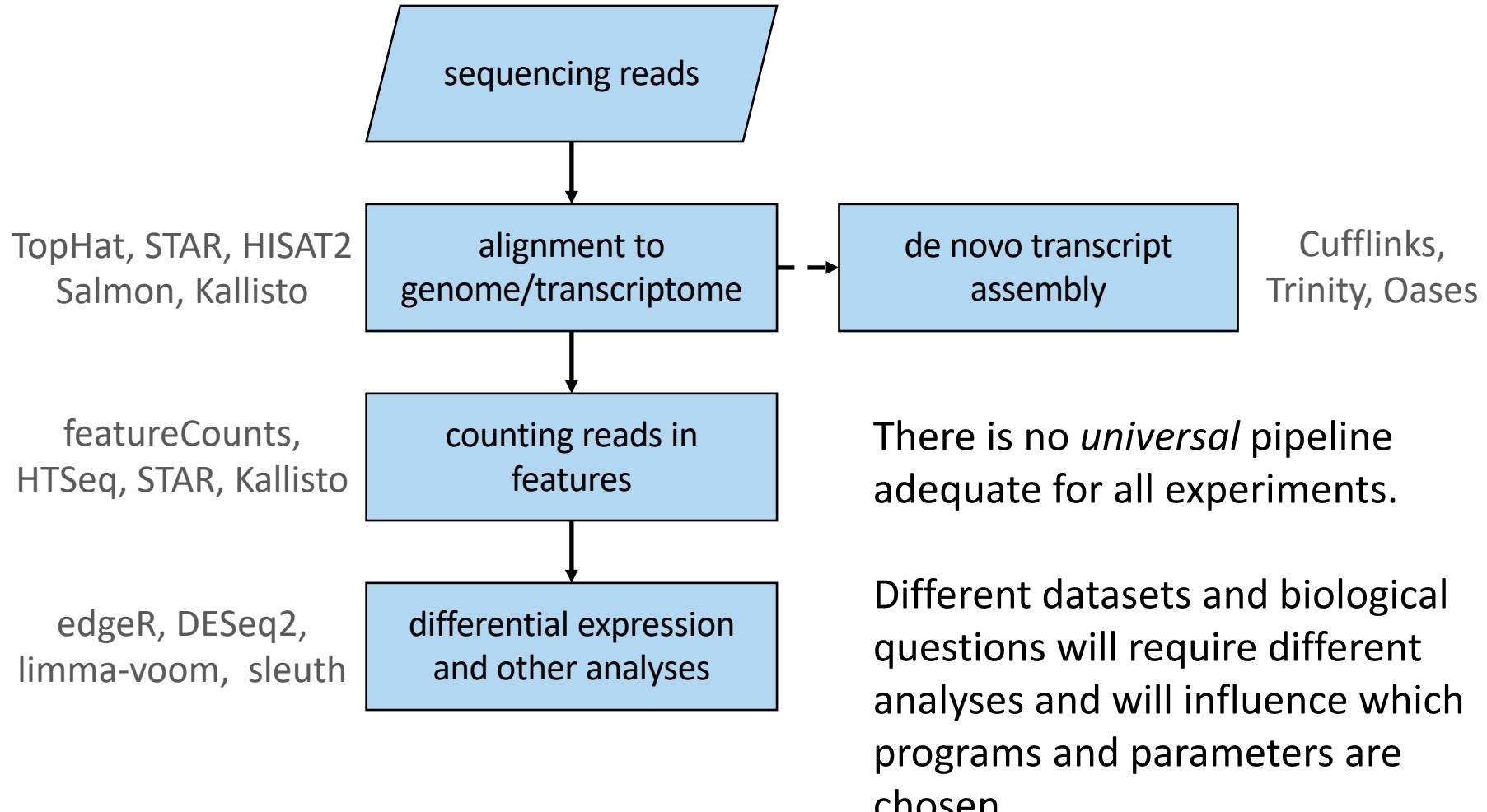
# Experimental design



# Experimental design

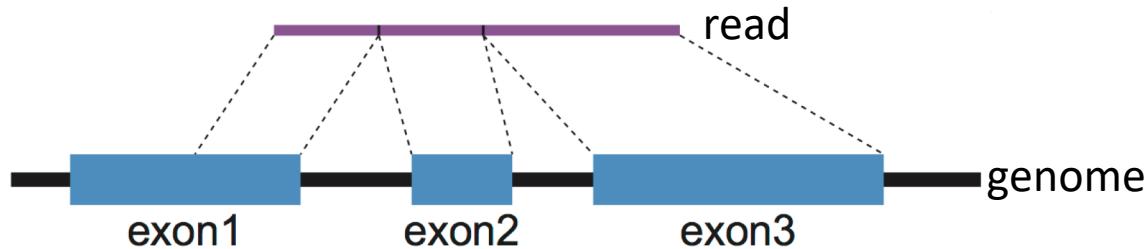


# RNA-seq data analysis pipeline



# RNA-seq data alignment

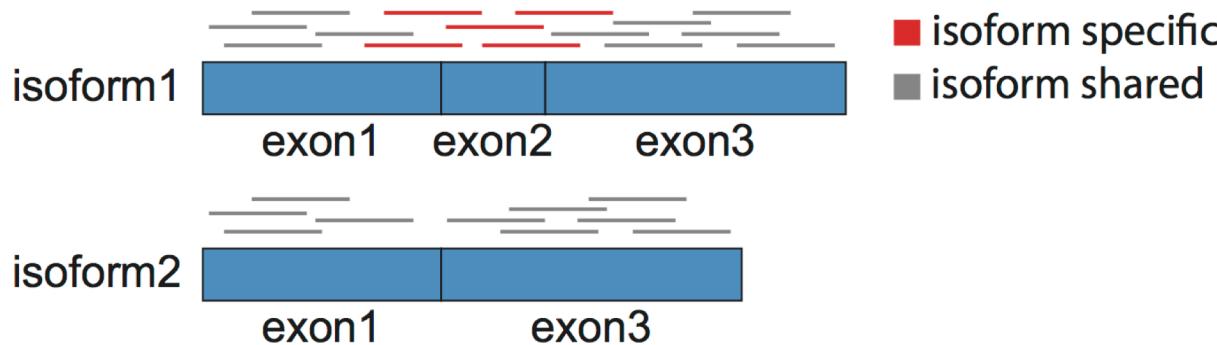
- RNA-seq reads come from spliced mRNAs.
- Therefore, their alignment in the genome is interrupted by introns.



- Two solutions:
  - Map reads to the transcriptome instead of the genome.
  - Allow gapped alignments.

# Map to the transcriptome

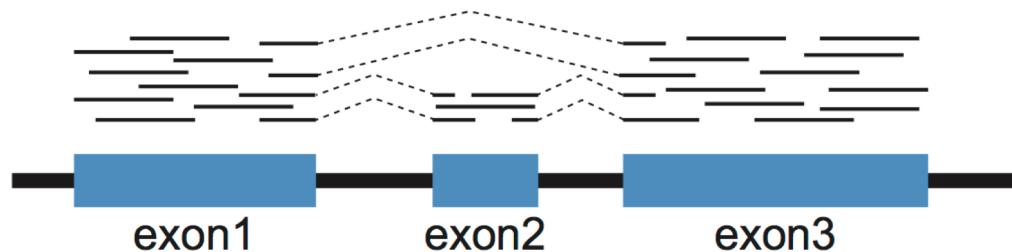
- If the RNA-seq reads are mapped to the transcriptome, reads in exons that are shared across transcript isoforms will map multiple times.



- Only possible for organisms with a well-annotated transcriptome.
  - Any novel genes or isoforms will be lost.

# Map allowing large gaps

- Instead, we can map to the genome but allowing the alignments to have large gaps.
  - Intron size ranges from  $10^2$  to  $\sim 10^5$  in eukaryotes.



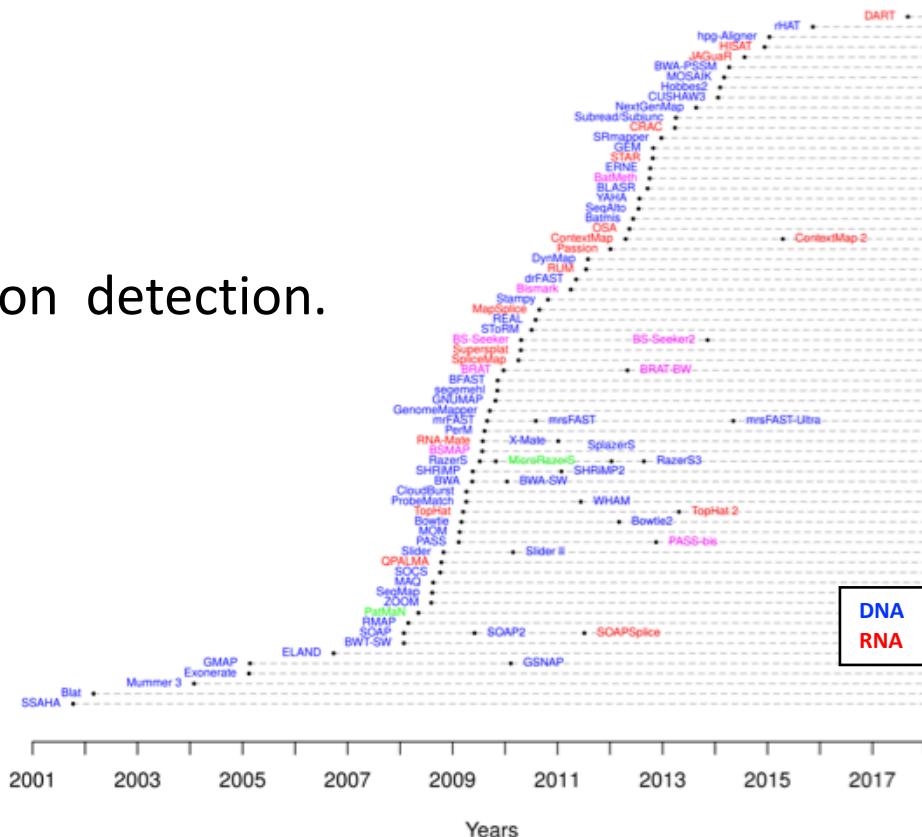
- Many different mappers.
  - TopHat, STAR, HISAT2, GSNAp, subread, MapSplice.

# RNA-seq data aligners

- There are many different programs to align RNA-seq data.
- There is no best aligner.
  - Memory usage.
  - Speed.
  - Accuracy of splice junction detection.

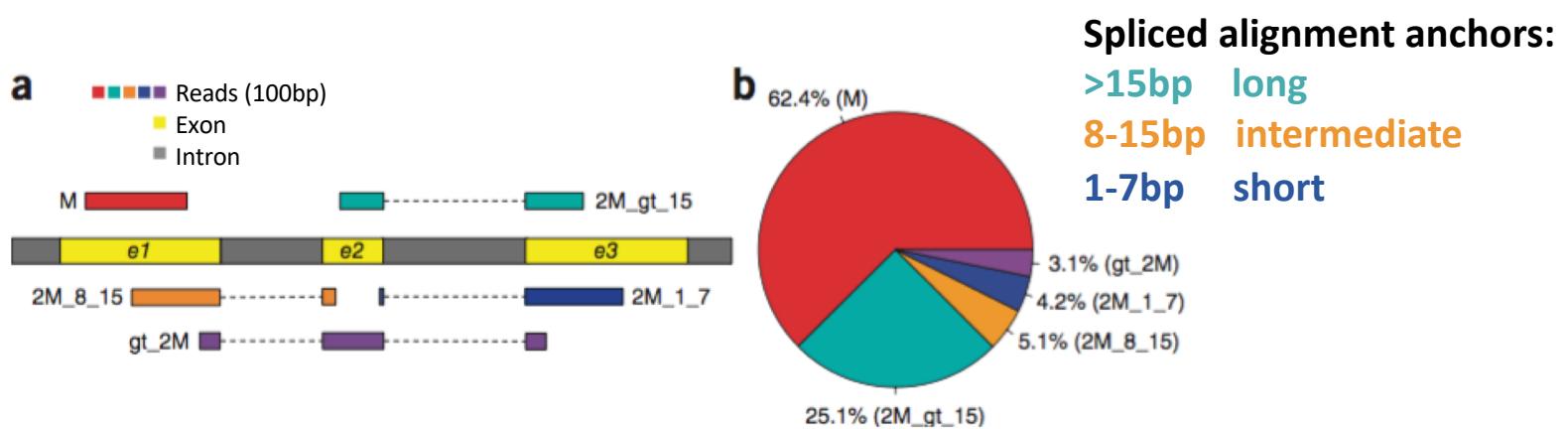
Simulation-based comprehensive  
benchmarking of RNA-seq aligners.

Baruzzo et al., *Nat Methods* 14 (2017)  
[doi.org/10.1038/nmeth.4106](https://doi.org/10.1038/nmeth.4106)

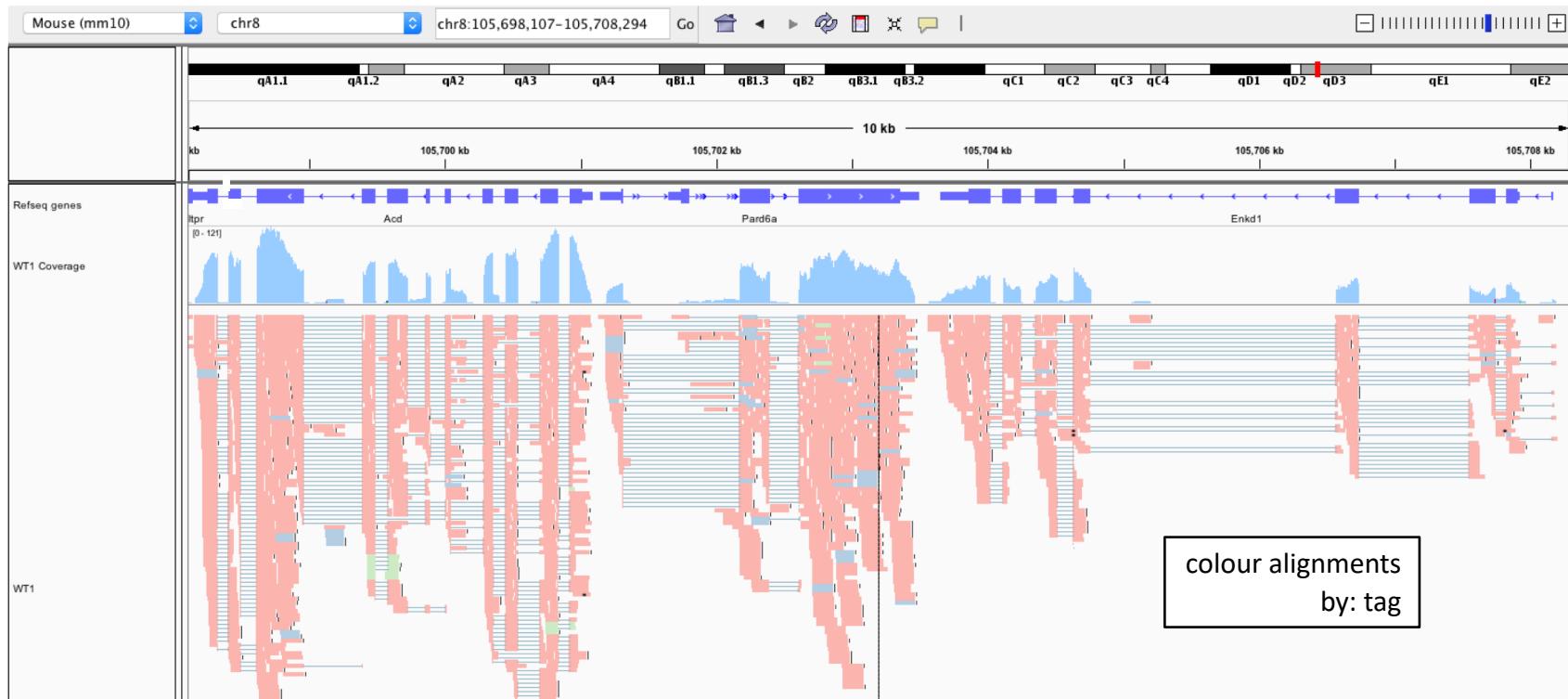


# HISAT2

- Fast and requires low memory.
  - Combines the Burrows-Wheeler transform and the FM index.
  - Two indices: **one global** FM index of the whole genome.  
**many small** overlapping FM indices of 56kb-long regions.

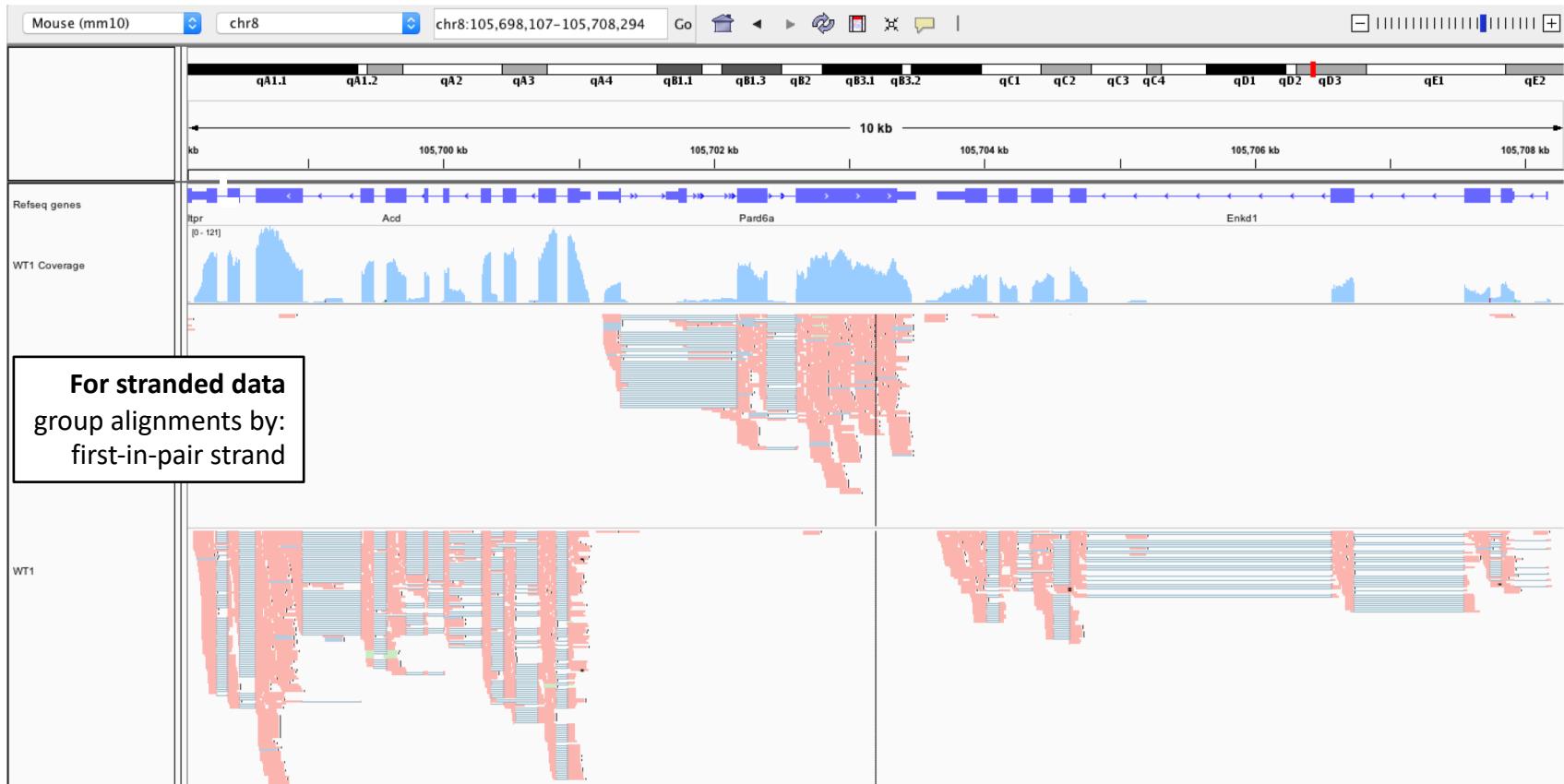


# Visualisation of aligned RNA-seq reads

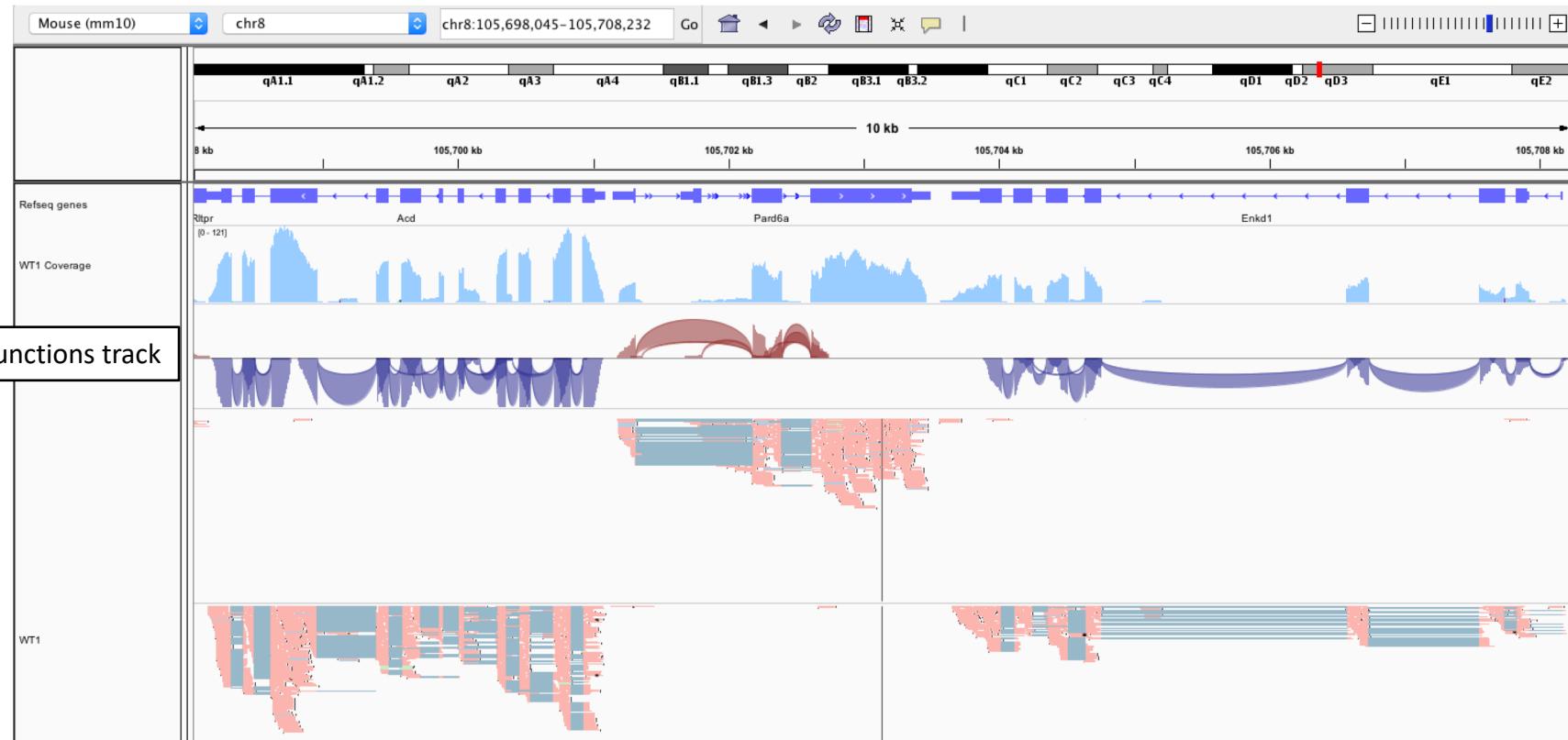


In this case reads are coloured by the NH tag, indicating the number of alignments:  
**unique   two valid alignments   three valid alignments ...**

# Visualisation of aligned RNA-seq reads



# Visualisation of aligned RNA-seq reads

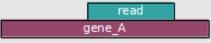
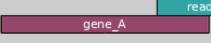
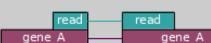
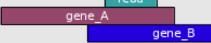


# Quantifying reads in features

- Once reads have been aligned to the genome, we want to quantify how many overlap features of interest (genes).

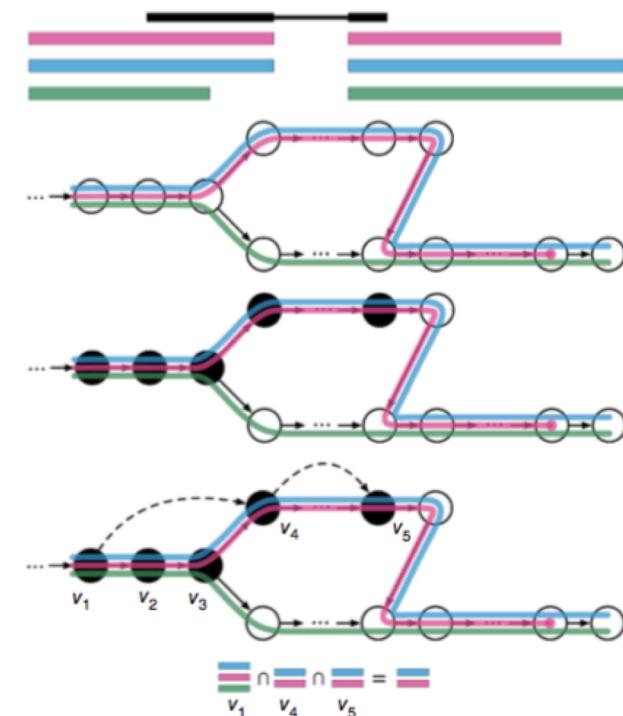
number of reads  $\propto$  transcript abundance

- Many available programs to do this (take a BAM and an annotation file).
  - HTSeq, FeatureCounts.
- Some aligners also do the quantification while mapping.
  - STAR, Kallisto.

	union	intersection _strict	intersection _nonempty
 gene_A	gene_A	gene_A	gene_A
 gene_A	gene_A	no_feature	gene_A
 gene_A gene_A	gene_A	no_feature	gene_A
 gene_A gene_A	gene_A	gene_A	gene_A
 gene_A gene_B	gene_A	gene_A	gene_A
 gene_A gene_B	ambiguous (both genes with --nonunique all)	gene_A	gene_A
 gene_A gene_B	ambiguous (both genes with --nonunique all)		
 gene_A gene_B	alignment_not_unique (both genes with --nonunique all)		

# Quantification through pseudo-alignment

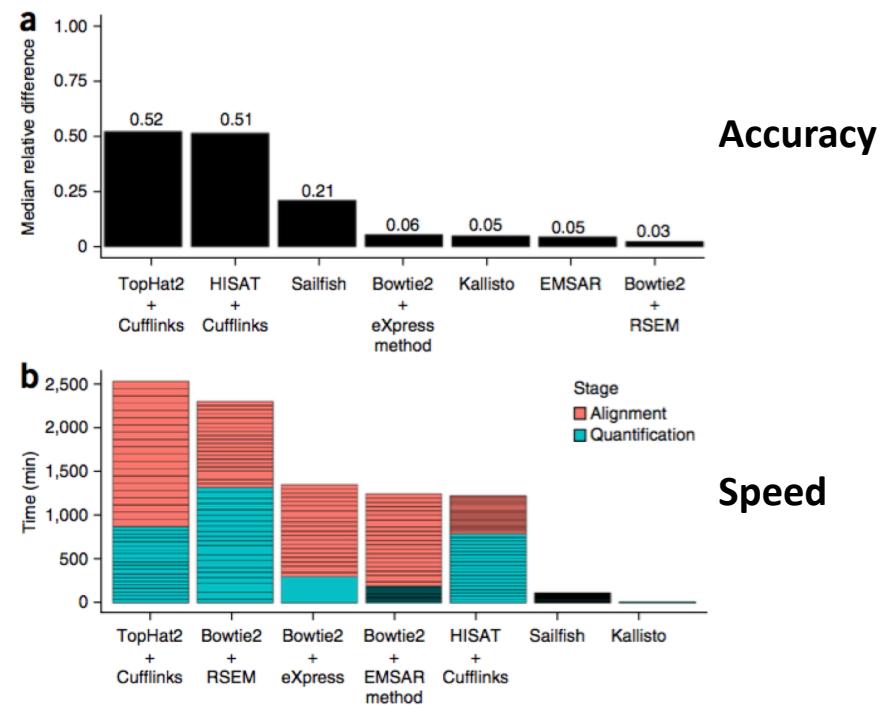
- Quantification of gene expression doesn't require knowing where a read originated from, but which transcripts could have generated it.
- Kallisto uses this principle to pseudo-align RNA-seq reads.
  - de Bruijn graph to represent the transcriptome.
  - exact matching of k-mers from reads to identify their compatibility with a set of transcripts.
- Pseudo-aligned reads are used to quantify the abundance of each compatibility class.



# Quantification through pseudo-alignment

- Quantification of gene expression doesn't require knowing where a read originated from, but which transcripts could have generated it.
- Kallisto
  - Fast and accurate.
  - Cannot discover new genes/transcripts/splice junctions.
- Salmon is a similar program.

Patro et al., *Nat Methods* 14 (2017)  
doi.org/10.1038/nmeth.4197



# Normalisation

- To compare data from different samples, counts need to be normalised to **remove systematic technical effects**.
- **Sequencing depth bias:** the most obvious difference between samples is how deep they are sequenced.
  - Larger libraries have larger counts. These need to be scaled to be comparable.

	sample 1	sample 2	sample 3
gene 1	6	12	9
gene 2	10	20	15
gene 3	2	4	3
...	...	...	...
library size	18	36	27
size factor	1	2	1.5

counts  
size factor →

	sample 1	sample 2	sample 3
gene 1	6	6	6
gene 2	10	10	10
gene 3	2	2	2
...	...	...	...
library size	18	18	18

# Normalisation

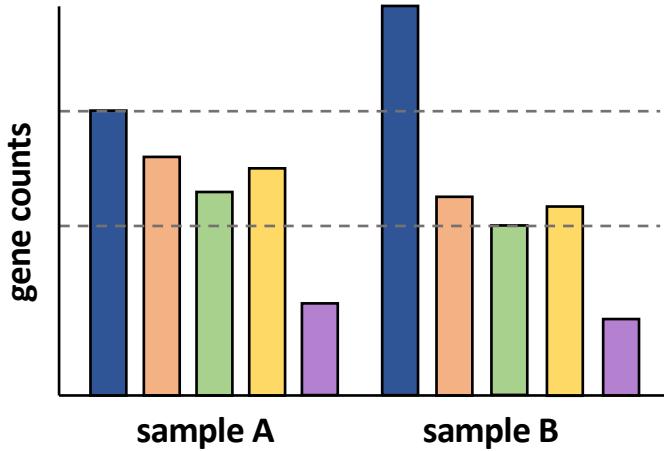
- To compare data from different samples, counts need to be normalised to **remove systematic technical effects**.
- **Sequencing depth bias**: the most obvious difference between samples is how deep they are sequenced.
  - Larger libraries have larger counts. These need to be scaled to be comparable.
- **Gene length bias**: longer genes will also produce more reads.
  - Irrelevant when comparing across samples.
  - Important when comparing across genes, but these comparisons are difficult to interpret.

# Measures of normalised counts

- **CPM** – counts per million.
  - Sequencing depth normalisation.
- **RPKM** – reads per kilobase per million mapped.
  - Sequencing depth and length normalisation.
- **FPKM** – fragments per kilobase per million mapped.
  - Equivalent to RPKM but for paired end data, where both reads of the same fragment are only counted once.
- **TPM** – transcripts per million
  - First normalise for length and then scale by library size.
  - Proportion of each transcript in the sample. All samples have the same normalised library size.

# Normalisation

- **Composition biases:** these arise when there is substantial differential expression of a set of transcripts.



Quantification is relative.

**Increase in the expression of the blue gene leads to a proportional decrease in all other genes.**

Changes in highly expressed genes can have a large impact on total library size.

- FPKM normalisation becomes misleading.

# Normalisation

- To account for sequencing depth **and** composition biases, size factors are calculated to normalise systematic differences between samples.
  - The assumption is that the majority of the transcriptome is not differentially expressed.
  - Therefore, any systematic differences must be technical.
- Two popular methods:
  - Median-of-ratios (DESeq2).
  - Trimmed mean of M values (edgeR).

# Normalisation – DESeq2

	counts				size factors			
	sample 1	sample 2	sample 3	pseudo-ref		sample 1	sample 2	sample 3
gene 1	45	88	66	63.94	gene 1	0.70	1.38	1.03
gene 2	1268	5072	3804	2903	gene 2	0.44	1.75	1.31
gene 3	2	4	3	2.88	gene 3	0.69	1.39	1.04
...	...	...	...	...	...	...	...	...
gene $n-2$	6	11	8	8.08	gene $n-2$	0.74	1.36	0.99
gene $n-1$	740	1470	1101	1061.97	gene $n-1$	0.70	1.38	1.04
gene $n$	39	26	19.5	27.04	gene $n$	1.44	0.96	0.72
library size	1	2	1.5		median	0.70	1.38	1.04

nonDE gene  
DE gene

1. Construct a **pseudo-reference** by taking the geometric mean of each gene across samples.



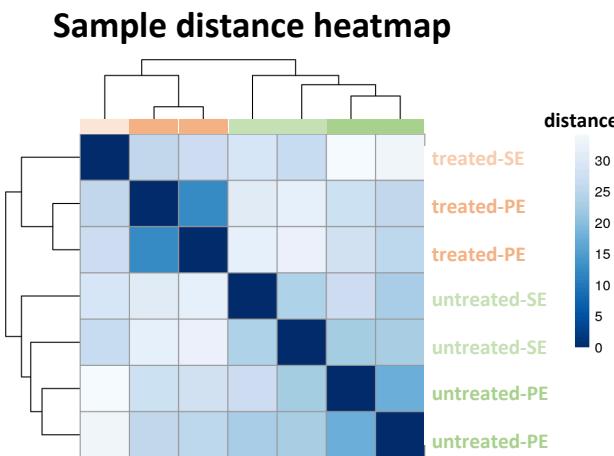
2. Compute the **fold-change** of each sample over the pseudo-reference.

3. Take the **median** of the fold-changes as the size factors to scale the counts.

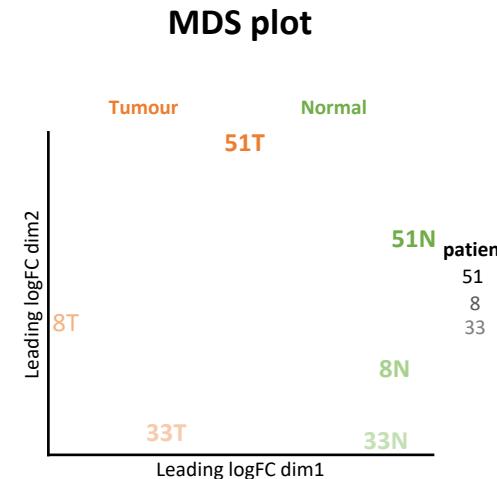
- The median protects against DE genes.

# Batch effects

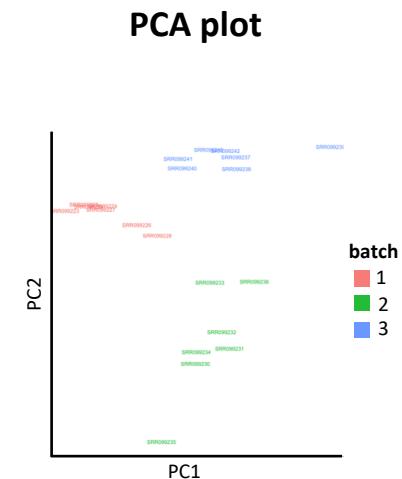
- Once the data has been normalised, it is useful to have an initial quick look at the overall transcriptomes.
  - Make sure samples from different conditions behave as expected.
  - Check if there are still systematic technical effects.



Modified from:  
<http://bioconductor.org/packages/devel/bioc/vignettes/DESeq2/inst/doc/DESeq2.html>



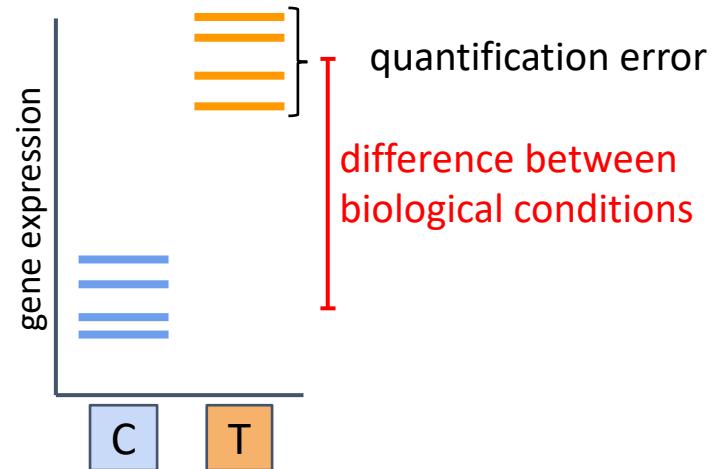
<http://bioconductor.org/packages/release/bioc/vignettes/edgeR/inst/doc/edgeRUsersGuide.pdf>



[https://pachterlab.github.io/sleuth\\_walkthroughs/bottomly/analysis.html](https://pachterlab.github.io/sleuth_walkthroughs/bottomly/analysis.html)

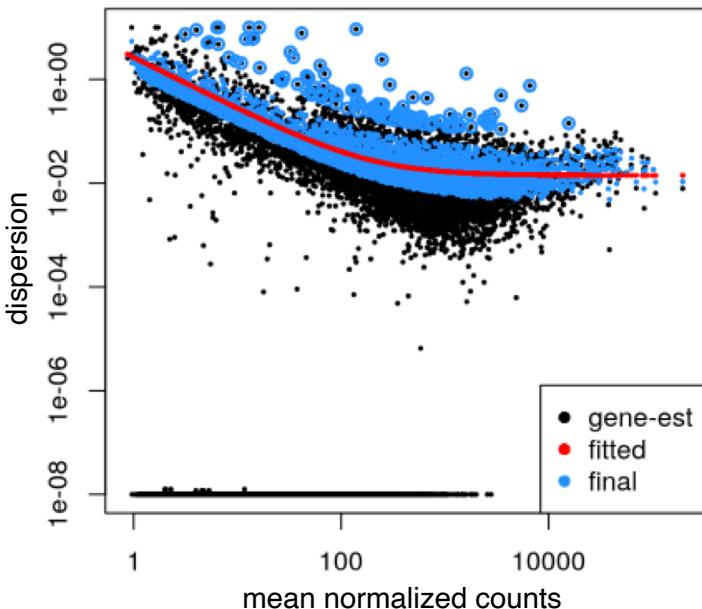
# Differential expression

- One of the main applications of RNA-seq is to compare samples under different conditions.
  - Healthy vs diseased.
  - Control vs exposure to treatment/drug/condition.
  - Wild-type vs gene knockout.
  - Changes across development/ageing.
- Differential expression analysis assesses whether the difference in expression levels between groups is larger than quantification error.



# Differential expression

- The low number of replicates makes estimating the variation accurately very difficult.
  - This is overcome by pooling information across genes, to make estimates more robust.



There is a strong relationship between the mean expression level and the dispersion.

Genes expressed at lower levels are more variable.

This relationship is not preserved after normalisation.

**Always use raw counts for DE testing.**

# Differential expression

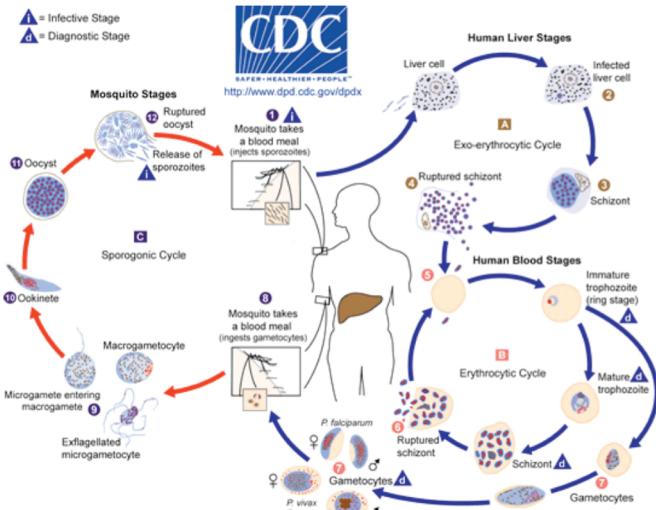
- There are many programs to perform differential expression analysis.
  - DESeq2
  - edgeR
  - limma-voom
  - sleuth (can use output from kallisto and perform DE analysis at the *transcript* level)
- It is possible to account for complex experimental designs and control for batch effects and confounding factors.

<https://www.bioconductor.org/packages/release/bioc/html/DESeq2.html>  
<http://bioconductor.org/packages/release/bioc/html/edgeR.html>  
<http://bioconductor.org/packages/release/bioc/html/limma.html>

# Interpretation of results

- What to do when you have a list of differentially expressed genes:
  1. If you already have a hypothesis, test it.
  2. Perform gene set enrichment analysis (GSEA, TopGO, InnateDB, Ingenuity Pathway Analysis, etc.)
  3. Work through the list, google, read papers.
  4. Cross-reference with other features/datasets: Pfam domains, chromosomal location, proteome, correlation with ChIP-seq / mutation / GWAS hits...
- Make new hypotheses.
- Go back to the lab.

# The exercise



Spence et al., *Nature* 498 (2013)  
doi.org/10.1038/nature12231

**IS THE TRANSCRIPTOME OF  
MOSQUITO TRANSMITTED PARASITES  
DIFFERENT FROM ONE WHICH HAS  
NOT PASSED THROUGH A MOSQUITO?**

- *Plasmodium chabaudi*: a rodent malaria parasite.
  - Exhibits many characteristic associated with the pathogenesis of human infection.
- Serial blood passage (SBP).
  - Direct injection from mouse to mouse.
  - Severe disease.
- Infection with parasite via mosquitos (MT).
  - Lower parasitaemia (presence of parasites in the blood).
  - Mild, chronic disease.