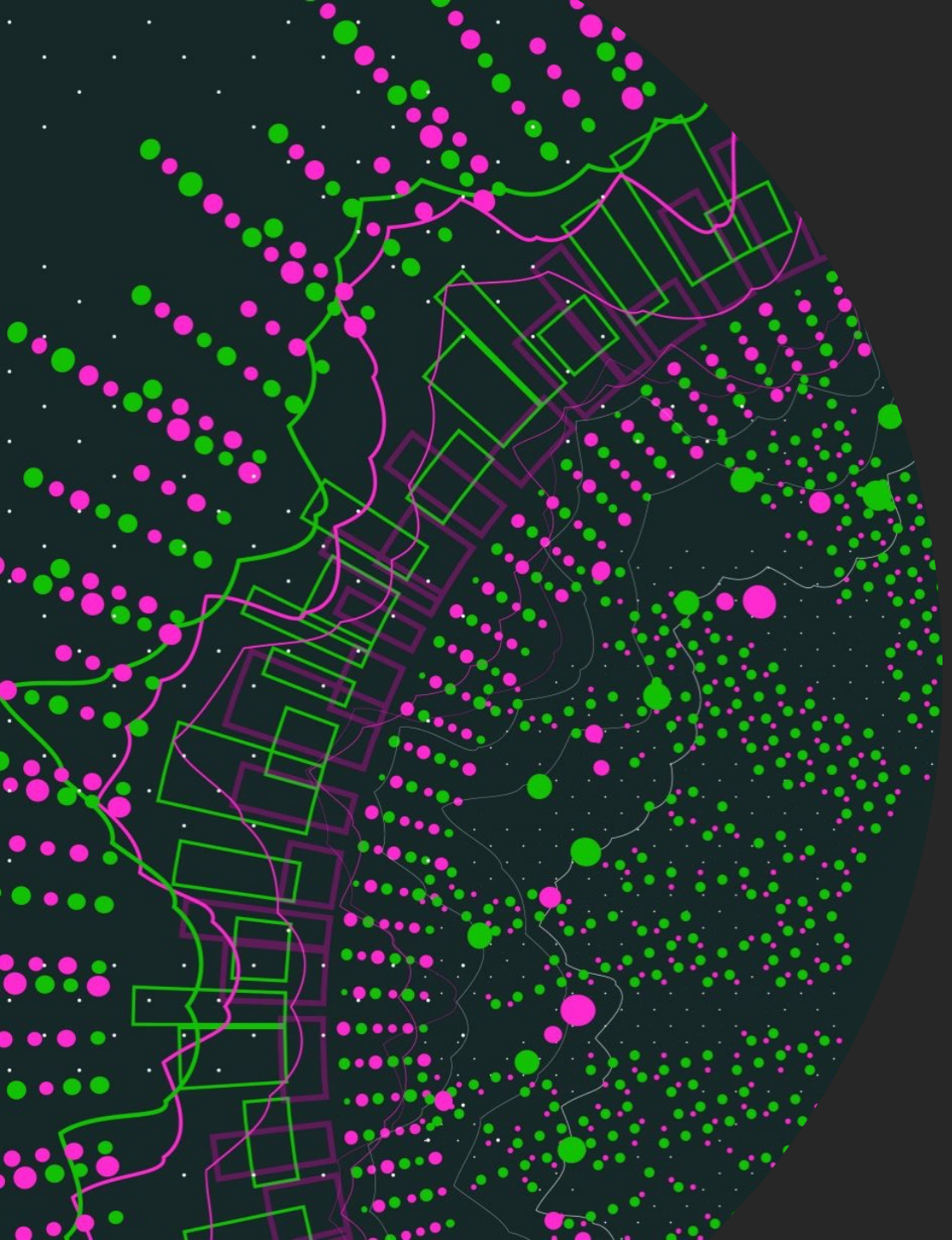


# *Intro to RNA-Seq and differential expression analyses*

Victoria Offord

NGS Bioinformatics

# Overview

- Background
- Experimental design
- The workflow…
  - *Mapping to the genome (HISAT2 and IGV)*
  - *Pseudoalignment and counting reads (Kallisto)*
  - *Read count normalisation*
  - *Differential expression and QC (Sleuth)*
- What to do with your gene list?
- Today's practical exercise

# *Background*

***RNA-Seq analyses are a diverse topic!***

*Today we will only have time to cover the most common type:* ***differential expression***

---

*Disclaimer*

# What is a transcriptome?

*"The complete set of transcripts in a cell*

*and their quantity*

*for a specific developmental stage or condition"*

Wang *et al*. (2009)
Nature Reviews Genetics
(PubMed: 19015660)

# Central dogma

**DNA**


transcription

**Pre-mRNA**

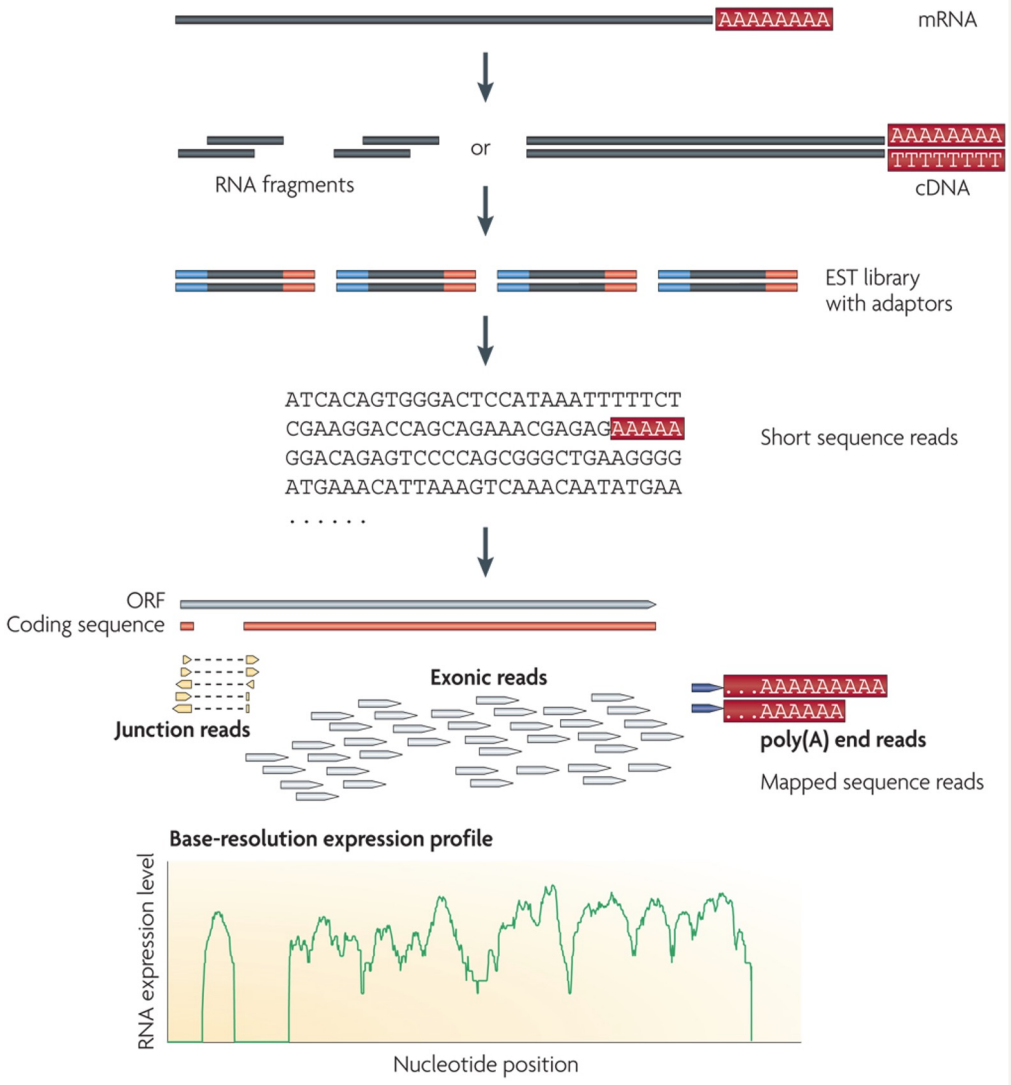

splicing
capping
tailing

**mRNA**


translation

**Protein**

# RNA sequencing




Generic overview:

1. Convert RNA into a library of cDNA fragments
2. Add sequencing adapters to cDNA fragments
3. Use high-throughput sequencing to get short reads representing the cDNA fragments
4. Assemble and/or align to reference genome or transcriptome
5. Perform downstream analyses

Wang *et al.* (2009)
Nature Reviews Genetics
(PubMed: 19015660)

# *Experimental design*

# Experimental design

- Successful RNA-Seq studies start with a good study design
- Considerations for generating data to answer your biological question include:
  - *library type*
  - *sequencing depth*
  - *number of replicates*
  - *avoiding biases*

# Experimental design - library preparation

- Total RNA = mRNA + rRNA + tRNA + regulatory RNAs...
- Ribosomal RNA can represent > 90% total RNA
- Can enrich for the 1-2% mRNA or deplete rRNA
  - *enrichment typically needs good RIN and high RNA proportion*
  - *some samples (e.g. tissue biopsies) may not be suitable*
  - *bacterial mRNA not polyadenylated -> ribosomal depletion*
- Be aware of protocol being used (e.g. some will remove small RNAs)

# Experimental design – library type

Single vs paired end reads

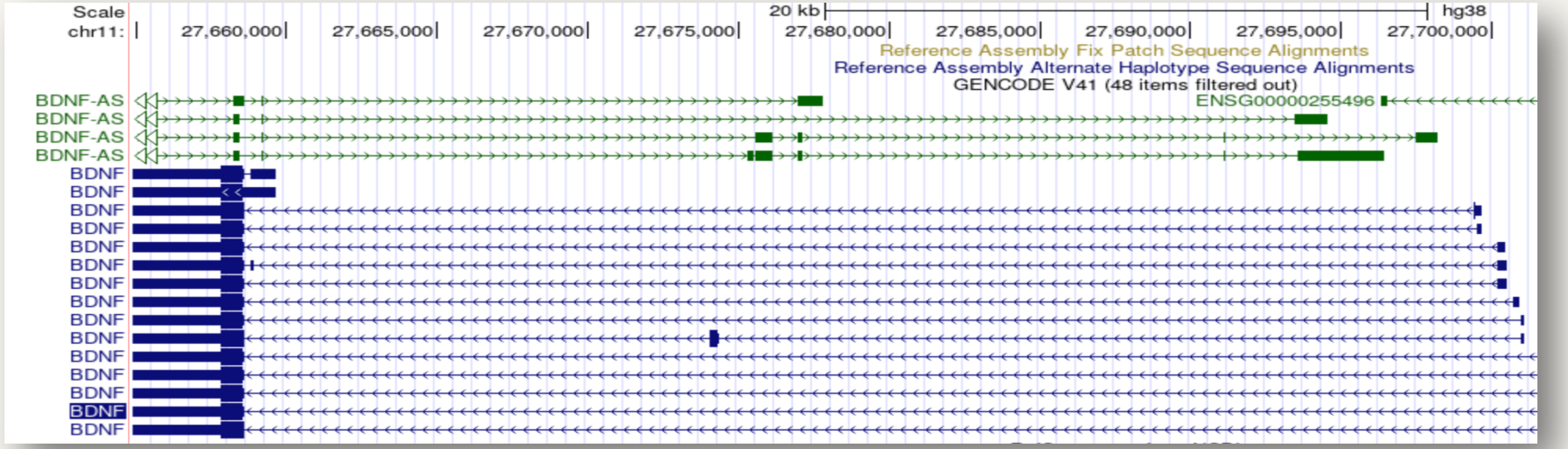
- Depends on your biological question and budget
- Paired end gives better resolution for de novo transcript discovery or isoform expression analysis because < 55% reads will span 2 or more exons

Stranded vs unstranded library

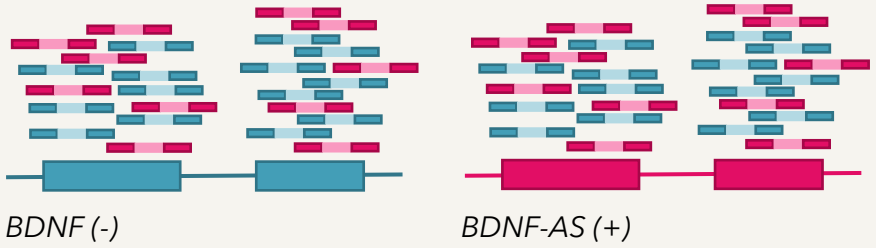
- Strand-specific better for detangling antisense or overlapping transcripts
- Always check your sequencing data:
  - *Some software may require you to explicitly state the strandedness of your library*
  - *Some software may detect strandedness automatically*
  - *Tools such as RSeQC can be used to predict strandedness*
  - *Incorrect assignment can result in >10% false negatives and over 6% false positives in DE results*

# Experimental design – strandedness example

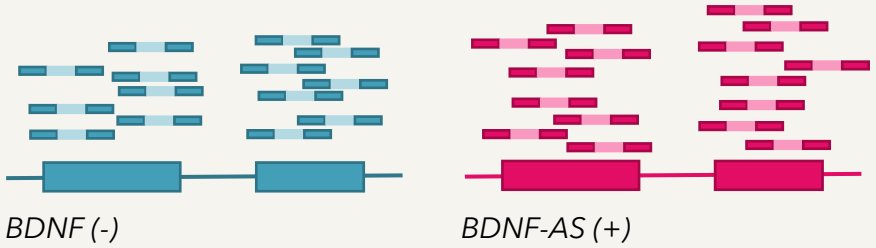
Example: *BDNF* (negative strand) and *BDNF-AS* (positive strand)




Unstranded library




Stranded library



# Experimental design – replicates

Biological replicates

- biologically distinct samples
- same type of organism treated or grown in the same condition
- understand biological variation (e.g. variation between individuals)
- relevant biological replicates are required

Technical replicates

- repeated measurements of the same sample
- understand the variation in equipment or protocols
- technical replicates are not generally required, but try to arrange samples on plates to minimise potential problems

# Experimental design – sequencing depth vs replicates

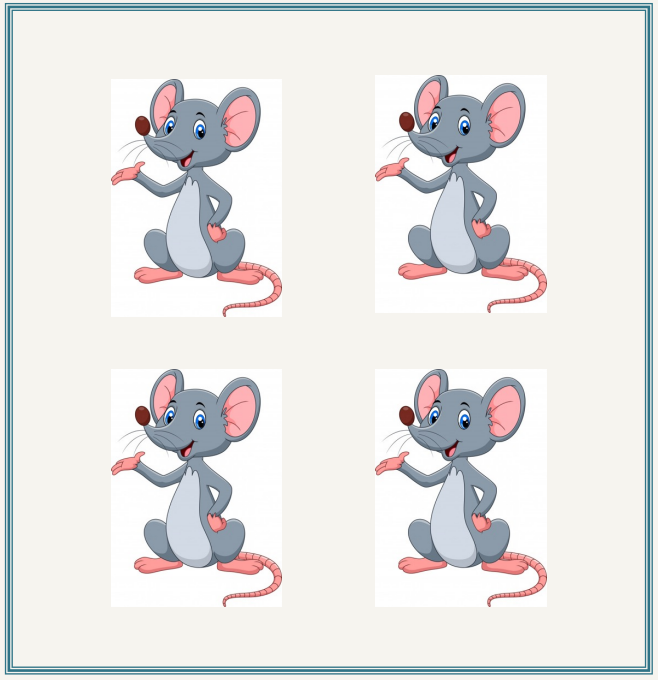
- Increasing sequencing depth can increase the ability to detect low expression transcripts
- Returns diminish beyond a certain sequencing depth
- Increasing biological replicates increases the accuracy of logFC and absolute expression levels (particularly in low expression transcripts)
- Reduces the coefficient of variation

*Is this a good experimental design?*

Wild type (WT)

VS

Knock out (KO)



 = FEMALE

 = MALE

# Try to avoid confounding factors….

Wild type (WT)

VS

Knock out (KO)

 = FEMALE

 = MALE

# *Workflows*

***No single workflow is suitable for all types of RNA-Seq experiments….***

---

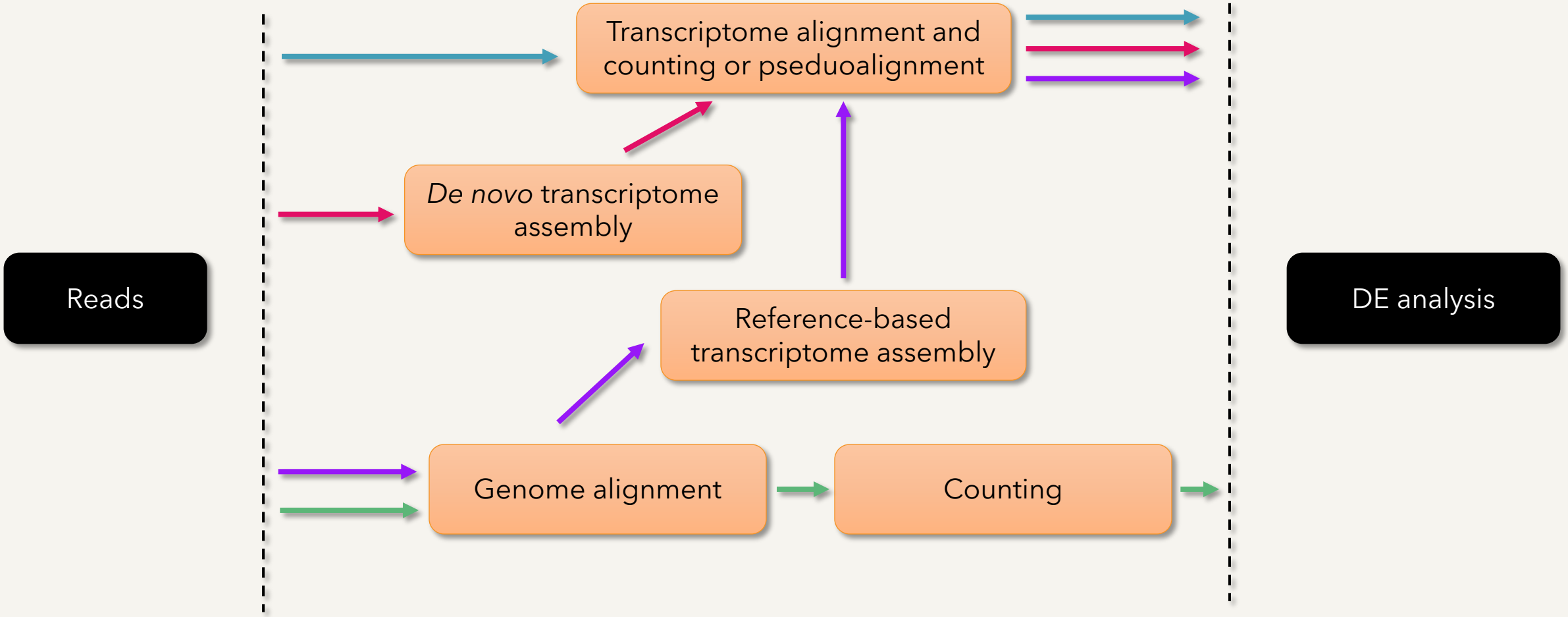
*Disclaimer (yup…another one!)*

# *DE analysis*

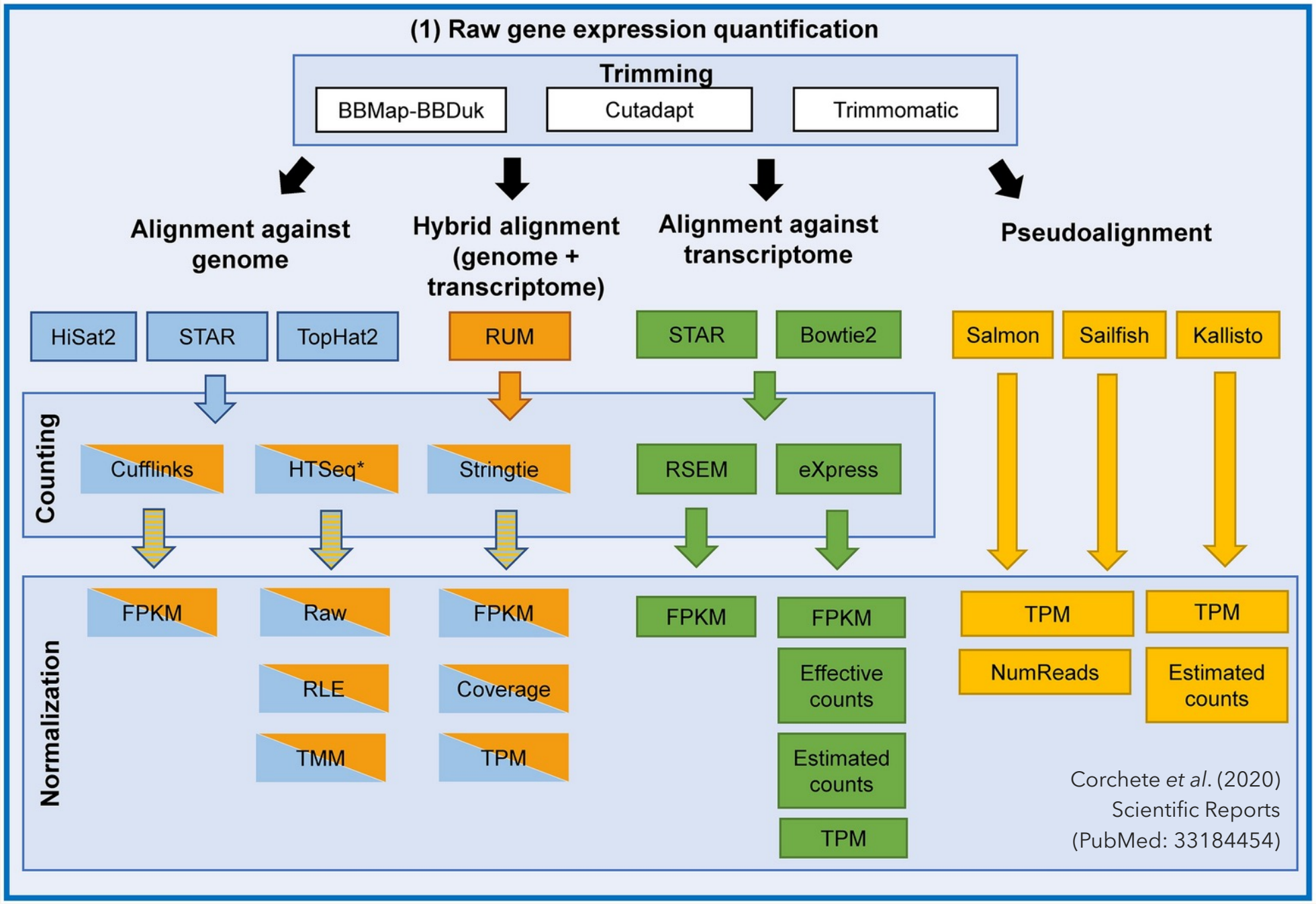
# What do we need to know?

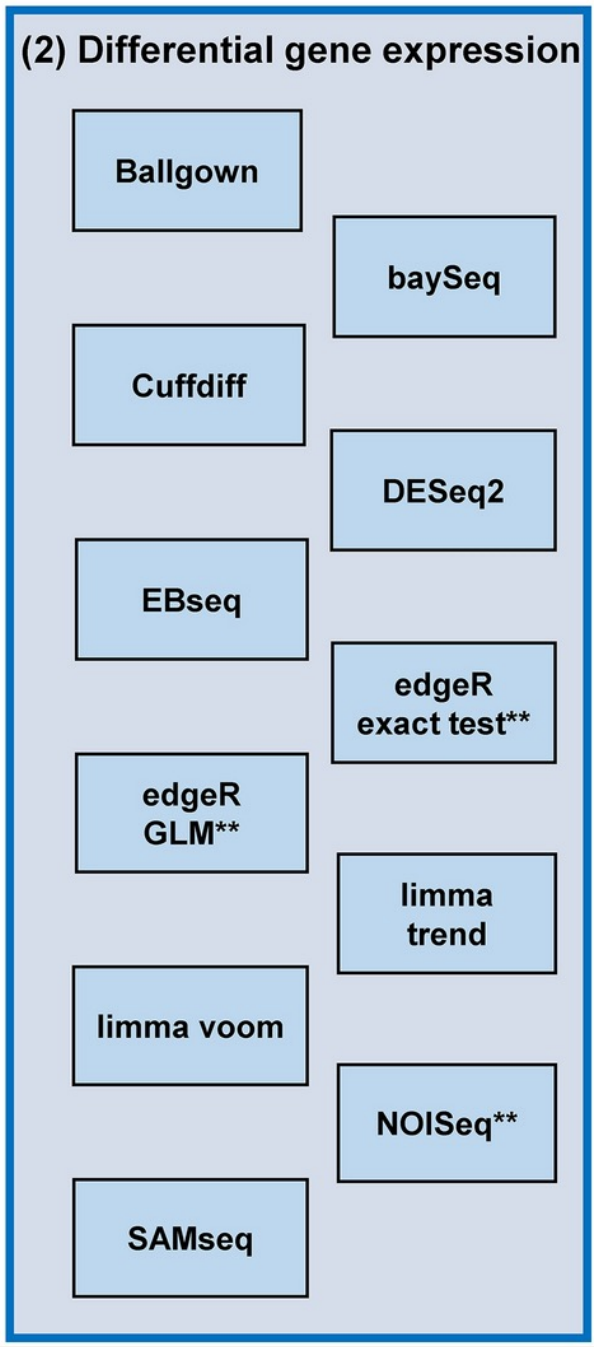
1. Which genes/transcripts do our reads belong to? *mapping / assembly*
2. How many reads belong to a specific gene/transcript? *quantification*
3. Do different sample groups express genes/transcripts differently? *DGE analysis*

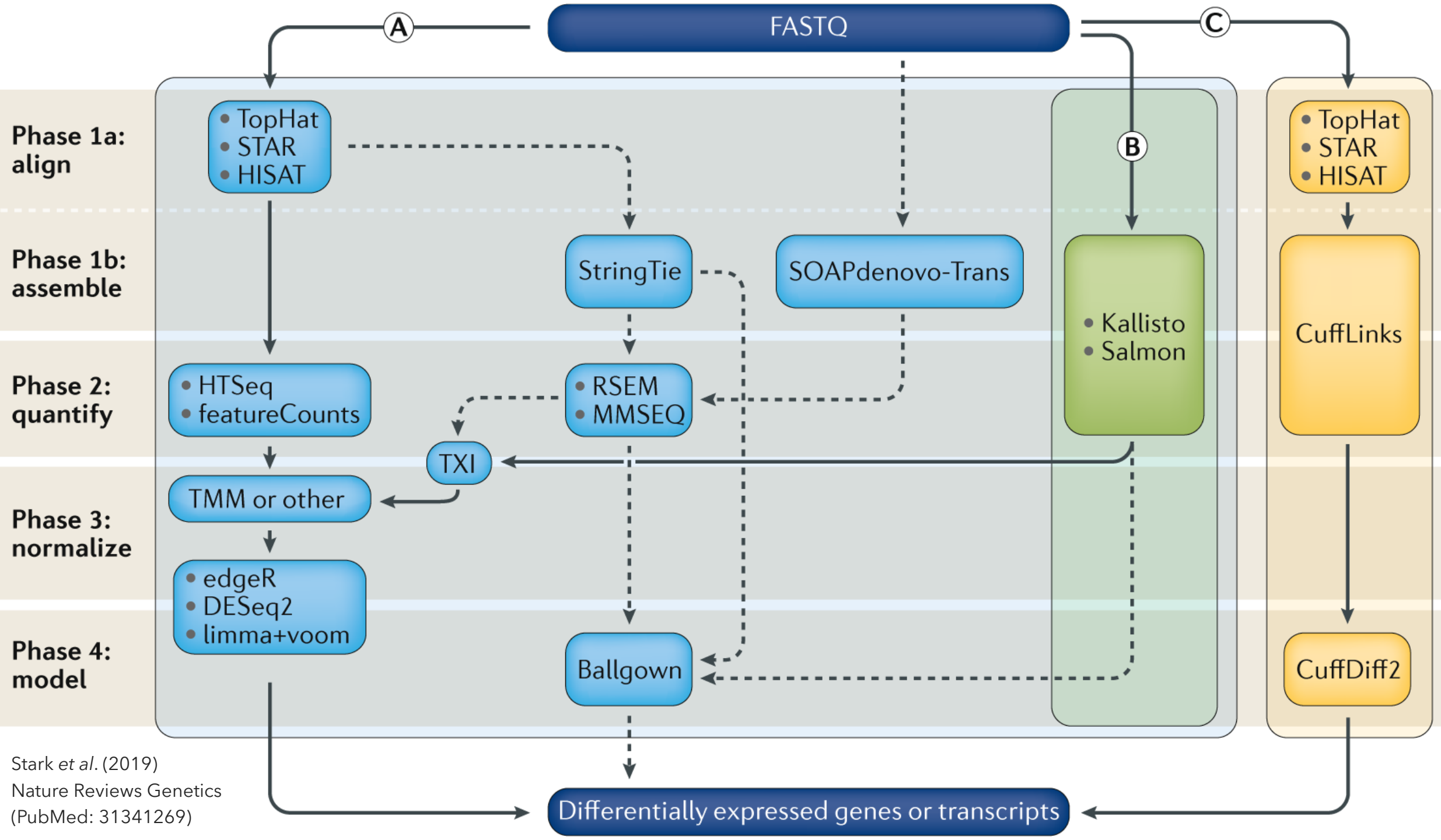
# Example DE analysis workflows



# RNA-seq data analysis workflow

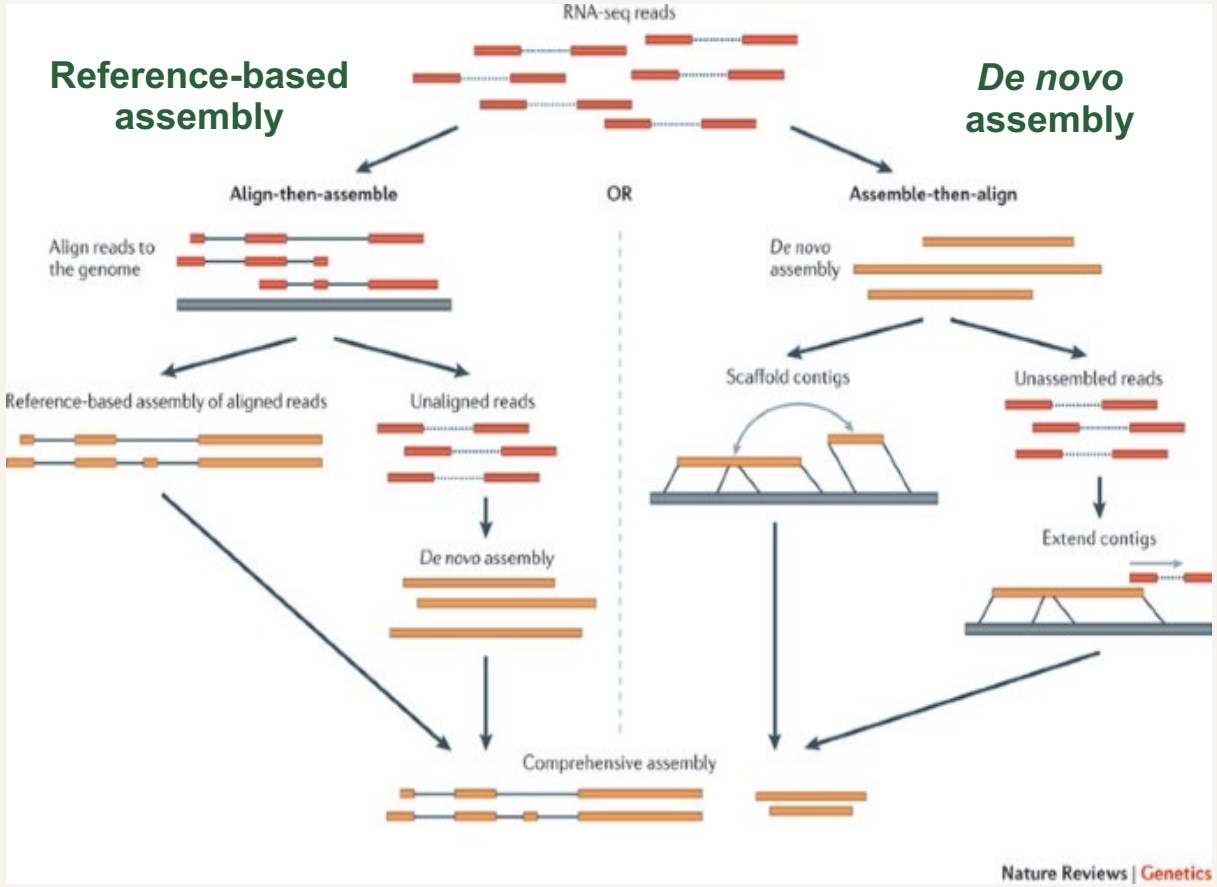





Stark *et al*. (2019)
Nature Reviews Genetics
(PubMed: 31341269)

# Wait!! I'm not studying a model organism…




<u>Reference-based assembly</u>

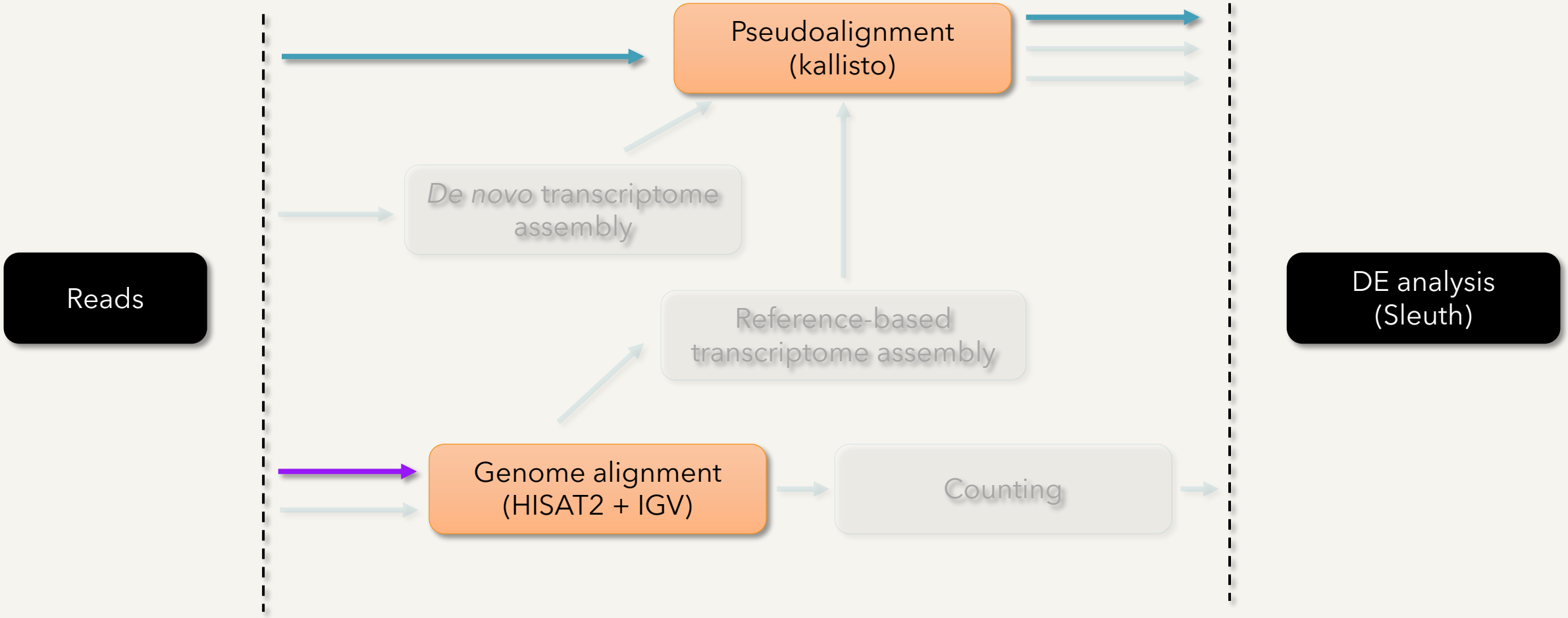
- Genome is available
- Transcriptome not available / poor quality
- Example tools: Cufflinks and Scripture

<u>*De novo (reference-free) assembly*</u>

- Genome is not available / poor quality
- Transcriptome is not available / poor quality
- Note: requires more data than reference-based
- Example tools: Oases, TransABySS and Trinity

Martin *et al*. (2011)
Nature Reviews Genetics
(PubMed: 21897427)

# Which workflow are we looking at today



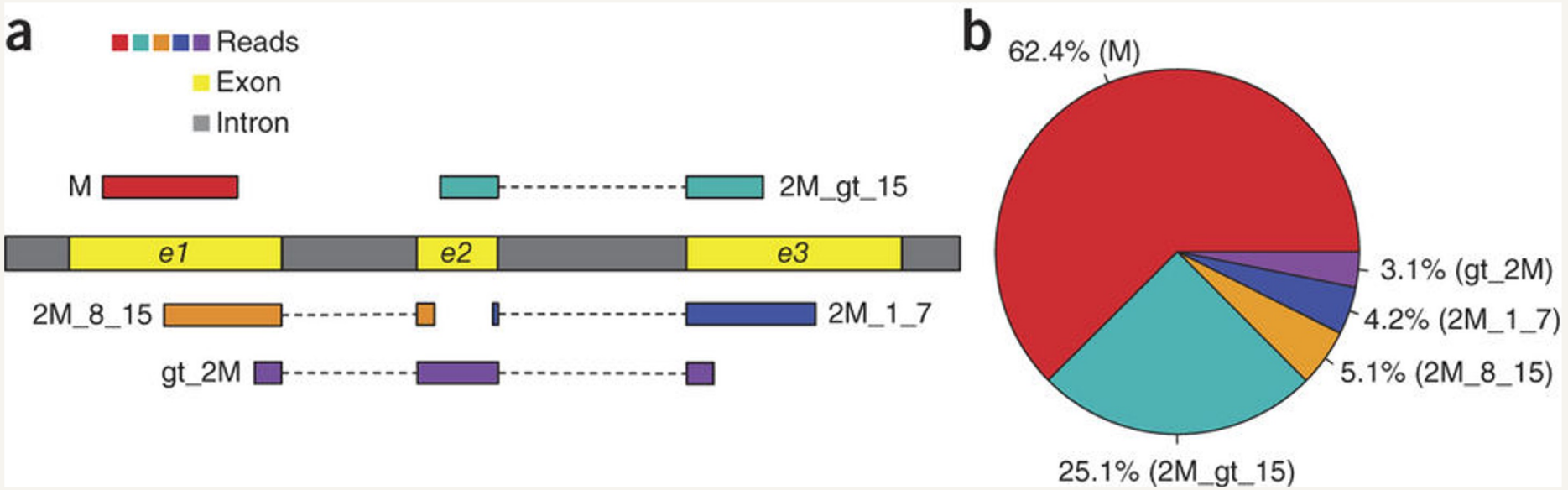
# *Genome alignment & visualisation*

# Mapping RNA-seq reads to the genome (HISAT2)

- Mapping to the genome is great for determining whether your RNA-seq data is of high quality and exploring the structure of genes of interest
- Eukaryotic genes have introns, which are not present in mature mRNA so special mapping algorithms are required (**splice-aware aligners**)
- **HISAT2** is only one such algorithm, but is *accurate*, *fast* and *easy to use*

# Splice aware alignment (example)

- Most reads are simple to place, mapping to a single exon (M => ~62%)
- Remainder (~38%) are mapped across to two (~35%) or more (~3%) exons and require the aligner to be “splice-aware”
- Splice-aware aligners use different approaches to map reads across multiple exons



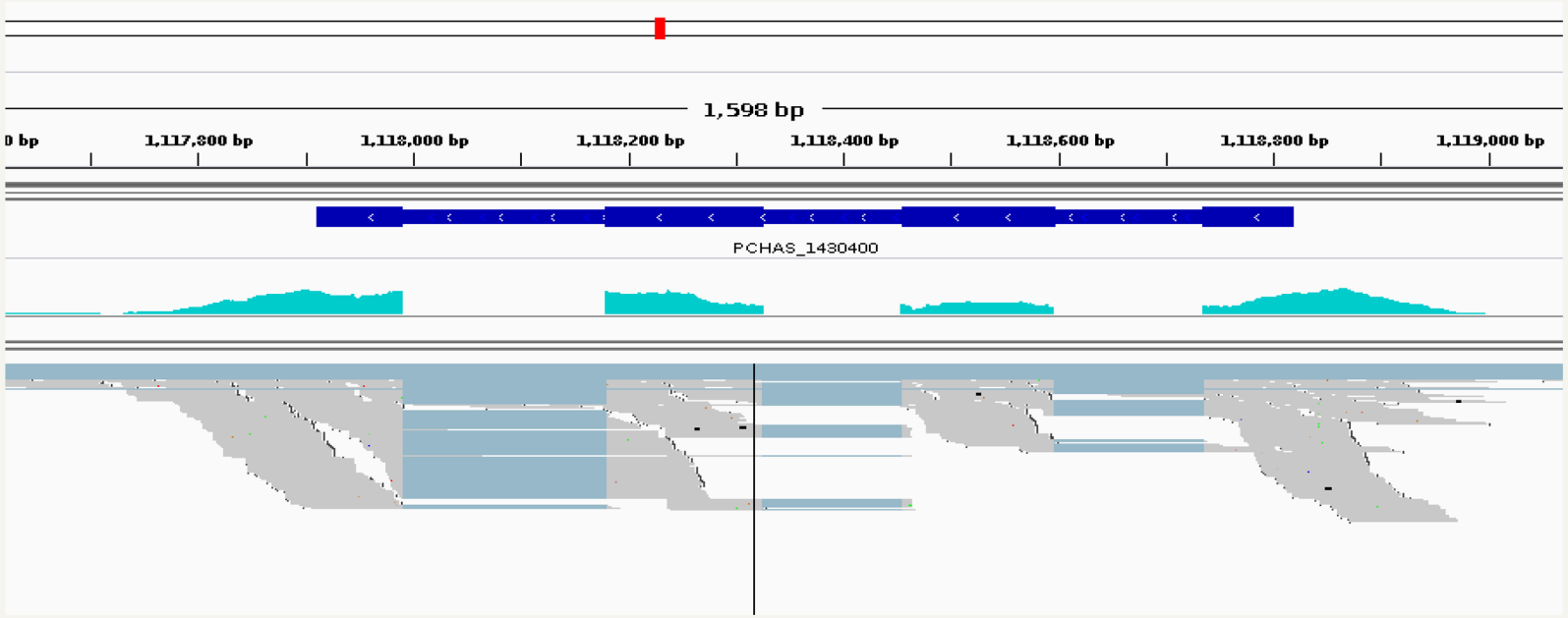

Kim et al. (2015)
Nature Methods
(PubMed: 25751142)

# Splice aware alignment (HISAT2)

- Builds indexes from compressed genomes
- Two types of index:
  - Single global index
  - Tens of thousands of small local indexes
- HISAT2 alignment strategy:
  - Longer part of the read which maps contiguously (anchor) is mapped to a global index
  - This identifies the relevant local index which can be used to assign the remaining (smaller) part of the read to a single index (rather than searching the whole genome)
  - Uses information from other reads to help place smallest (1-7bp) read fragments

# Visualising the genome alignment (IGV)

- Genome alignment may be a dependency for downstream processing
- But…it is also useful for checking your data looks "OK" (quality control)
- Visualising alignments is best done with a genome browser (e.g. IGV)




*Not all species will have a reference genome available (so this may not always be possible)*

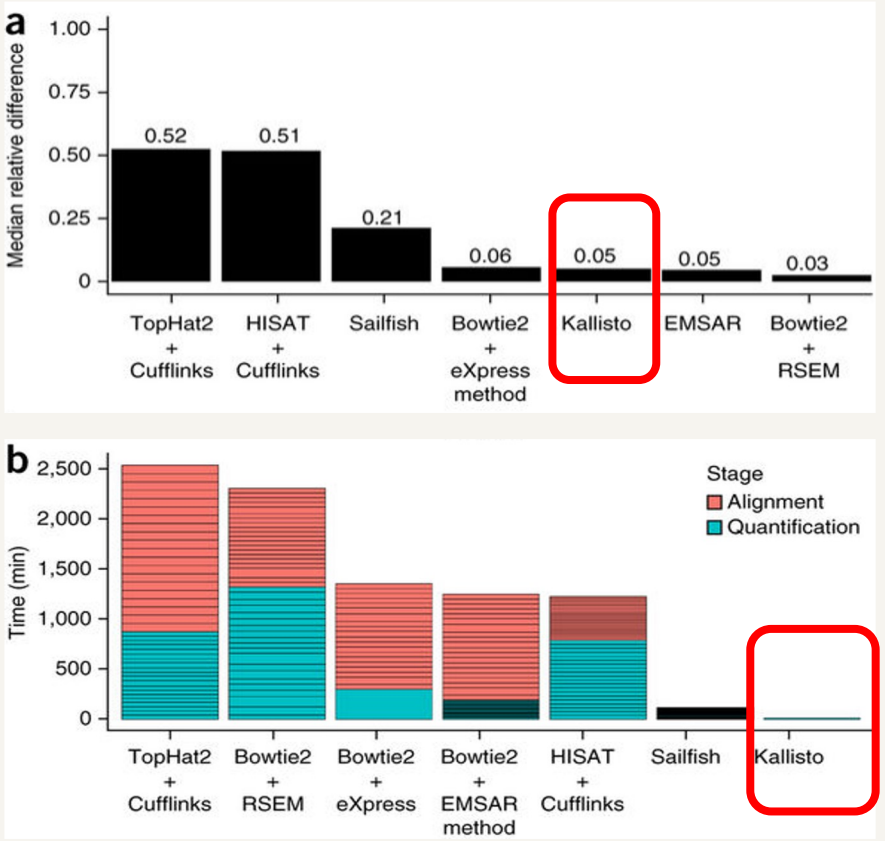
# *Transcript abundance*

# Transcriptome alignment

- Multiple splice forms per gene introduce ambiguity into the mapping
- Mapping to the spliced transcripts allows us to generate transcript-specific read counts
- Faster because there is less target sequence
- Recent improvements in algorithms (pseudoalignment) make this even faster
- Pseudoalignment doesn’t care where in each transcript reads map to, just which of the transcripts they are assigned to
- With pseudoalignment, counting comes for free

# Pseudoalignment (kallisto)

- kallisto constructs a target de Bruijn graph (or path through the transcriptome)
- Breaks down both the reads and transcripts into k-mers for indexing
- kallisto is quick because:
  - *it ignores kmers which don't change between transcripts*
  - *it only looks for exact matches*
  - *it discards reads which have errors or can't be mapped*
- For ~60% of the reads, the first and last kmers are needed for assignment
- Only 2% of the reads require all kmers to assign them to a transcript
- Bootstrapping (resampling of the data) to estimate technical variance which is subtracted downstream to improve understanding of biological variance
- Calculates transcript abundance in the form of TPMs




Bray et al. (2016)
Nature Biotechnology
(PubMed: 27043002)

# *Normalisation*

# Why is normalisation important?

- Raw counts will have both intra- and inter-sample biases
- Greater depth means more reads mapping to each gene (sequencing depth bias)
- Longer genes have more reads mapping to them (gene length bias)
- Common practice is to normalise for both sequencing depth and gene length:
  - *RPKM - reads per kilobase per million*
  - *FPKM - fragments per kilobase per million*
  - *TPM - transcripts per million*
- The order of the normalization processes is important:
  - RPKM and FPKM normalise for sequencing depth and then gene length
  - TPM normalises for gene length and then sequencing depth

# RPKM (reads per kilobase per million)

BEFORE

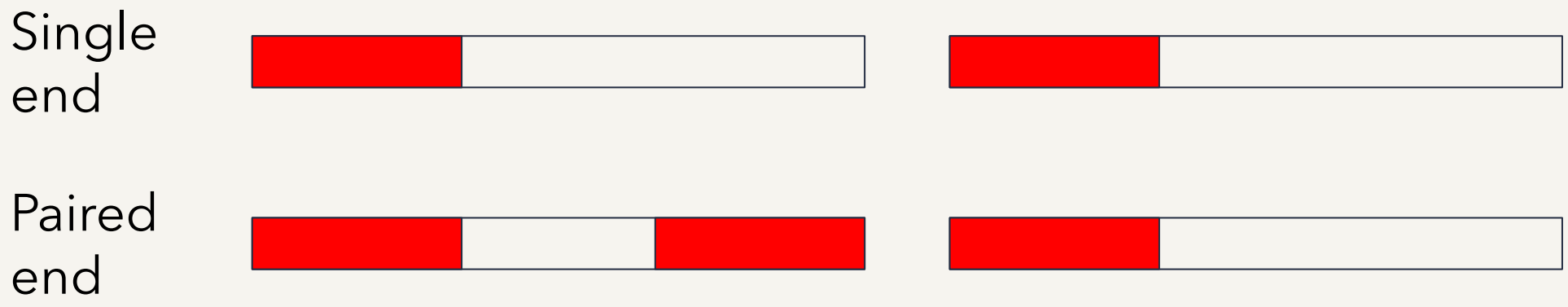
| Gene | Replicate 1 Counts | Replicate 2 Counts | Replicate 3 Counts |
|---|---|---|---|
| A (2,000 bases) | 10 | 12 | 30 |
| B (4,000 bases) | 20 | 25 | 60 |
| C (1,000 bases) | 5 | 8 | 15 |
| D (10,000 bases) | 0 | 0 | 1 |

AFTER

| Gene (bases) | Replicate 1 RPKM | Replicate 2 RPKM | Replicate 3 RPKM |
|---|---|---|---|
| A (2,000 bases) | 1.43 | 1.33 | 1.42 |
| B (4,000 bases) | 1.43 | 1.39 | 1.42 |
| C (1,000 bases) | 1.43 | 1.78 | 1.42 |
| D (10,000 bases) | 0 | 0 | 0.009 |

# FPKM (fragments per kilobase per million)

- Essentially RPKM for paired reads
- Takes into account that two reads can map to one fragment (and so it doesn't count this fragment twice)



# RPKM vs TPM (transcripts per million)

**RPKM**

| Gene | R1 | R2 | R3 |
|---|---|---|---|
| **A** | 1.43 | 1.33 | 1.42 |
| **B** | 1.43 | 1.39 | 1.42 |
| **C** | 1.43 | 1.78 | 1.42 |
| **D** | 0 | 0 | 0.009 |
| **Total** | **4.29** | **4.5** | **4.25** |

**TPM**

| Gene | R1 | R2 | R3 |
|---|---|---|---|
| **A** | 3.33 | 2.96 | 3.326 |
| **B** | 3.33 | 3.09 | 3.326 |
| **C** | 3.33 | 3.95 | 3.326 |
| **D** | 0 | 0 | 0.02 |
| **Total** | **10** | **10** | **10** |

Easier to see the proportion of each gene within a sample as sum of TPMs same across samples

# *DE analysis*

# Differential expression (DE) analysis

- We don’t normally have enough replicates to do traditional tests of significance for RNA-seq data
- Most methods look for outliers in the relationship between average abundance and fold change
- Assume most genes are not differentially expressed

# Differential gene expression analysis

**TABLE 1.** RNA-seq differential gene expression tools and statistical tests

| Name | Assumed distribution | Normalization | Description | Version | Citations[d] | Reference |
|---|---|---|---|---|---|---|
| *t*-test | Normal | DEseq[a] | Two-sample *t*-test for equal variances | – | – | – |
| log *t*-test | Log-normal | DEseq[a] | Log-ratio *t*-test | – | – | – |
| Mann-Whitney | None | DEseq[a] | Mann-Whitney test | – | – | Mann and Whitney (1947) |
| Permutation | None | DEseq[a] | Permutation test | – | – | Efron and Tibshirani (1993a) |
| *Bootstrap* | Normal | DEseq[a] | Bootstrap test | – | – | Efron and Tibshirani (1993a) |
| *baySeq*[c] | Negative binomial | Internal | Empirical Bayesian estimate of posterior likelihood | 2.2.0 | 159 | Hardcastle and Kelly (2010) |
| *Cuffdiff* | Negative binomial | Internal | Unknown | 2.1.1 | 918 | Trapnell et al. (2012) |
| *DEGseq*[c] | Binomial | None | Random sampling model using Fisher's exact test and the likelihood ratio test | 1.22.0 | 325 | Wang et al. (2010) |
| *DESeq*[c] | Negative binomial | DEseq[a] | Shrinkage variance | 1.20.0 | 1889 | Anders and Huber (2010) |
| *DESeq2*[c] | Negative binomial | DEseq[a] | Shrinkage variance with variance based and Cook's distance pre-filtering | 1.8.2 | 197 | Love et al. (2014) |
| *EBSeq*[c] | Negative binomial | DEseq[a] (median) | Empirical Bayesian estimate of posterior likelihood | 1.8.0 | 80 | Leng et al. (2013) |
| *edgeR*[c] | Negative binomial | TMM[b] | Empirical Bayes estimation and either an exact test analogous to Fisher's exact test but adapted to over-dispersed data or a generalized linear model | 3.10.5 | 1483 | Robinson et al. (2010) |
| *Limma*[c] | Log-normal | TMM[b] | Generalized linear model | 3.24.15 | 97 | Law et al. (2014) |
| *NOISeq*[c] | None | RPKM | Nonparametric test based on signal-to-noise ratio | 2.14.0 | 177 | Tarazona et al. (2011) |
| *PoissonSeq*[c] | Poisson log-linear model | Internal | Score statistic | 1.1.2 | 37 | Li et al. (2012) |
| *SAMSeq*[c] | None | Internal | Mann-Whitney test with Poisson resampling | 2.0 | 54 | Li and Tibshirani (2013) |

[a]See Anders and Huber (2010).
[b]See Robinson and Oshlack (2010).
[c]R (v3.2.2) and bioconductor (v3.1).
[d]As reported by PubMed Central articles that reference the listed reference (December 21, 2015).

| DGE Tool | Citation Count | Percentage | Publish Year |
|---|---|---|---|
| edgeR [34] | 7175 | 32.3% | 2010 |
| Cuffdiff/Cuffdiff2 [35, 36] | 6103 | 27.5% | 2012/2013 |
| DESeq2 [37] | 4355 | 19.6% | 2014 |
| limma [38] | 2451 | 11.0% | 2015 |
| DEGseq [39] | 1244 | 5.6% | 2009 |
| baySeq [40] | 567 | 2.6% | 2010 |
| SAMseq [41] | 279 | 1.3% | 2013 |
| sleuth [42] | 45 | 0.2% | 2017 |
| NOIseq [43] | 39 | 0.2% | 2012 |

McDermaid et al. (2019)
Briefings in Bioinformatics
(PubMed: 30099484)

Schurch et al. (2016)
RNA
(PubMed: 27022035)

# DE analysis (Sleuth)

- Sleuth is a fast, lightweight Shiny R package (i.e. it has a user interface)
- Sleuth takes as input bootstrapping and estimates of transcript abundances from pseudoaligners (e.g. kallisto, salmon, sailfish)
- Bootstrapping can be used as a proxy technical replicates to account for variability in estimates of transcript abundances due to:
  - random processes underlying the RNA-Seq experiment
  - statistical processes used in the read assignment

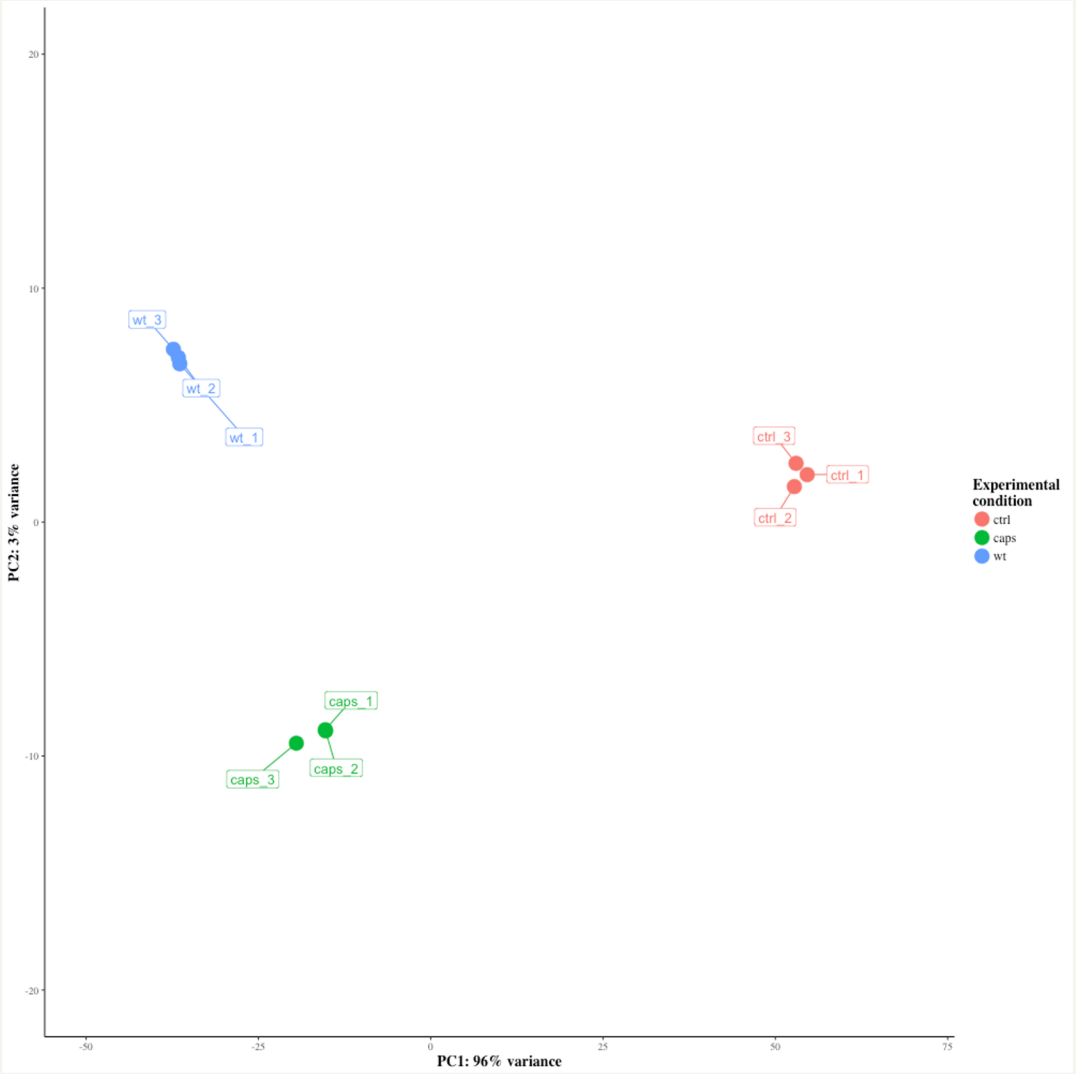
# How does Sleuth identify DE transcripts?

- Sleuth assumes that most genes are not differentially expressed
- Sleuth tests two models
  - a “reduced” model which assumes that the abundances are equal between conditions
  - a “full” model which assumes that the abundances differ between conditions
- Sleuth identifies DE transcripts as those which fit better to the “full” model:
  - beta value (b) – broadly analogous to log2 fold change
  - p-value (pval)
  - q-value (qval)
- If we had 10,000 genes, using a p-value of 0.05 would mean we expect 500 genes to be significant just by chance
- Thus, we use the q-value to assess significance (i.e. the p-value adjusted for multiple hypothesis testing)

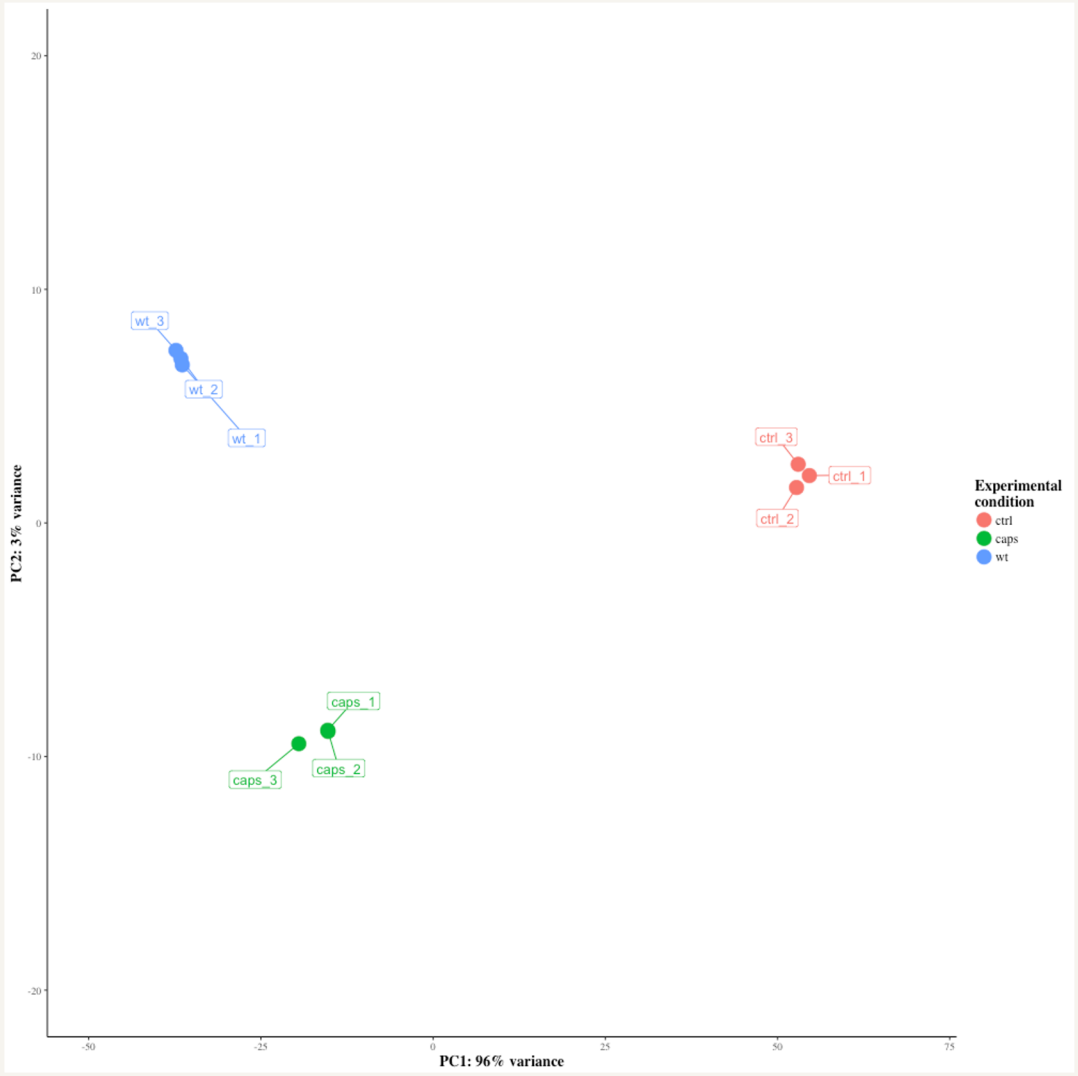
# *Quality control*

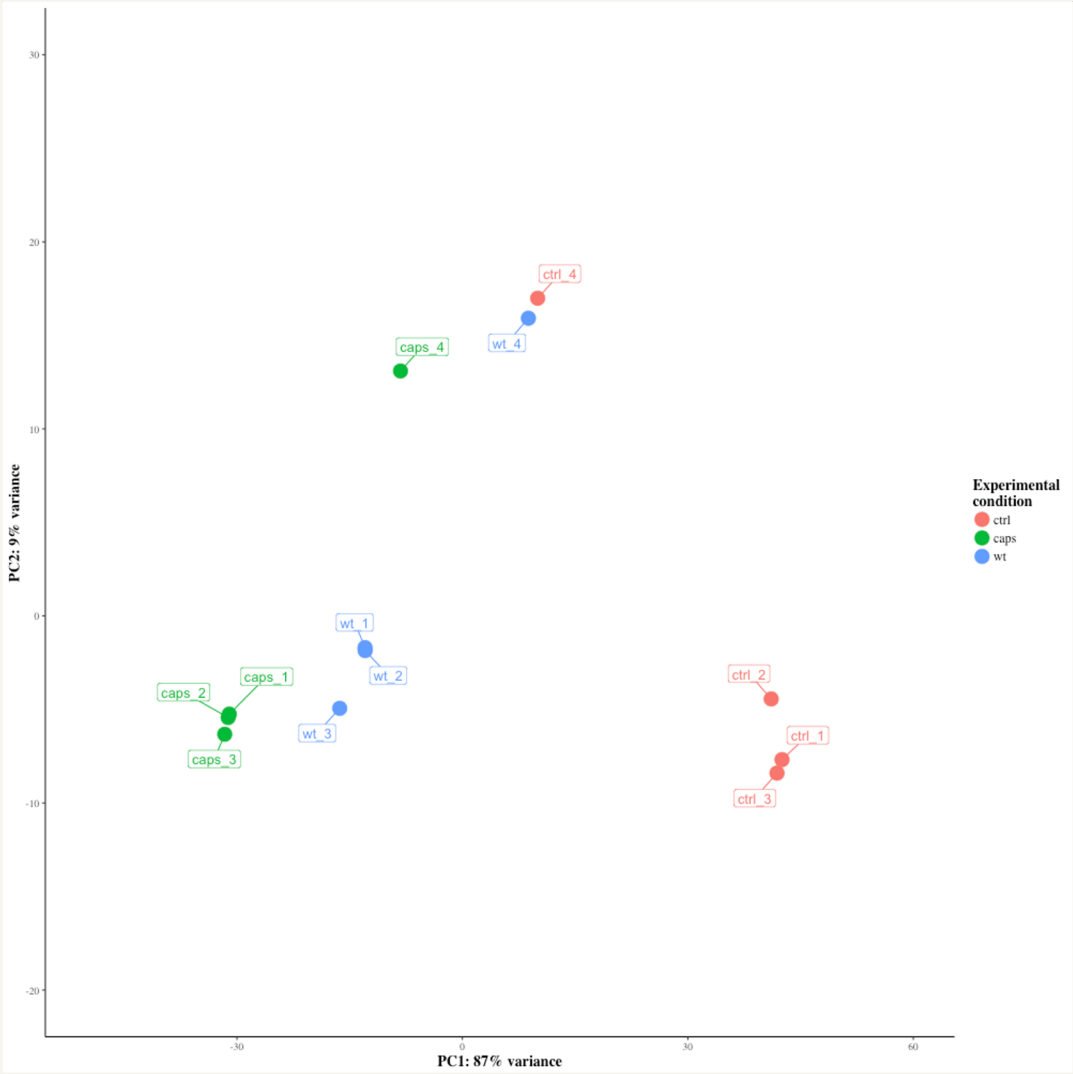
# Principal component analysis (PCA)

- Use to look at variation and strong patterns within data
- Identifies uncorrelated variables or principal components (PC)
- Tries to explain the maximum amount of variance with the smallest number of principal components



# Why QC our data?



# *What next?*

# What to do next with your gene list?

When you have a list of differentially expressed genes, things start to get difficult.

What to do:

1. Have a hypothesis already? Test it.
2. GO term/pathway/gene-set enrichment analysis (GSEA, TopGO, InnateDB, Ingenuity Pathway Analysis etc.)
3. Work through list, Google, read papers
4. Overlay datasets on essentiality, populations, mutations, Pfam domains, chromosomal location, expression, proteome…

Then make a hypothesis about what genes are interesting and why. Can you test/explore this further bioinformatically? Design the next wet lab experiment

# GO term and gene set enrichment analyses

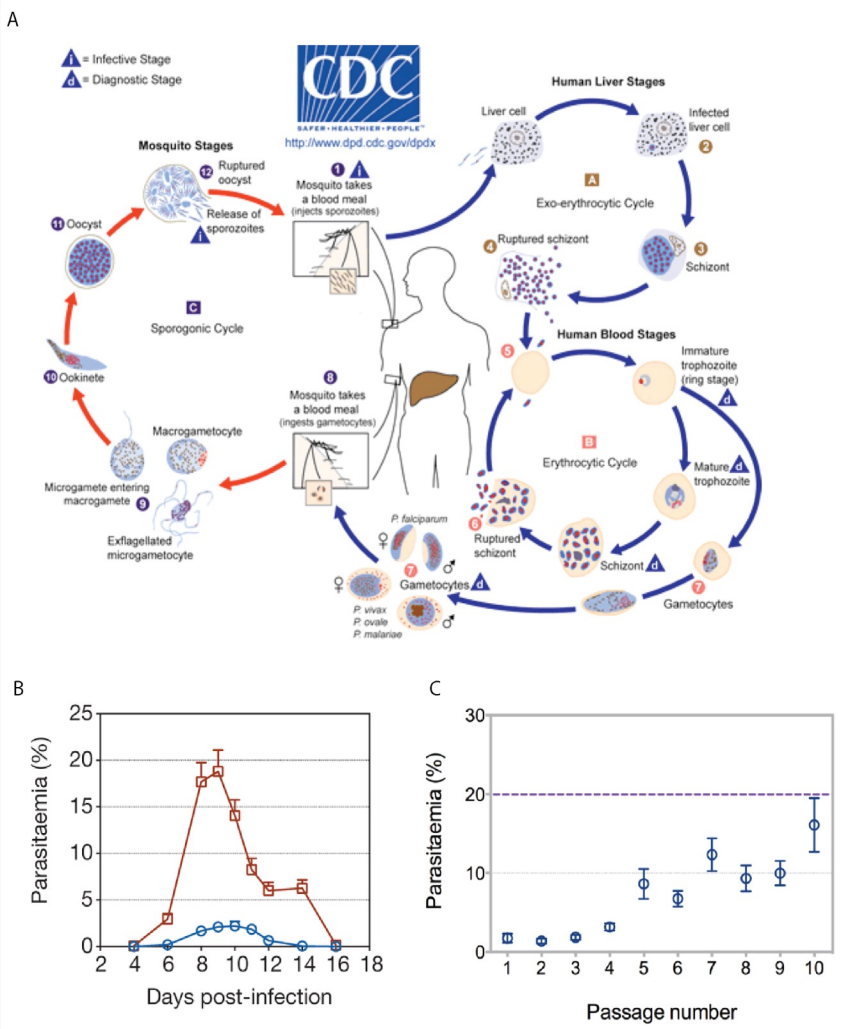
- Large lists of DEGs are difficult to interpret by eye
- Biological patterns may only emerge once we group these genes by their functional annotations or pathways
- GO terms are a curated, hierarchical vocabulary of functional annotations
  - cellular component – where the gene is localised in a cell
  - molecular function – function of the gene product
  - biological process – series of events the gene product is involved in

# Enrichment approaches

- over-representation analysis (e.g. DAVID)
  - statistical test (e.g. Fisher's exact) of overlap between a gene set and genes of interest
  - requires a threshold to identify genes of interest which can strongly affect results
- ranked enrichment analysis (e.g. GSEA, fgsea)
  - uses the entire gene list regardless of significance (i.e. no threshold required)
  - genes are ordered by a metric (e.g. fold change, q-value)
  - do does gene set fall at the top or bottom of the list more frequently than randomly picked gene lists of the same size
- conditional enrichment analysis (e.g. topGO)
  - GO terms are organised hierarchically into nested gene sets
  - a gene set may look enriched just because one of its child gene set is enriched
  - if a child gene set is enriched, then the genes it contains are excluded from the parental gene set before testing
  - works from gene sets with no children all the way to the root (top of the hierarchy)

# *Today's practical*

# Practical exercise




- Rodent malaria parasite *Plasmodium chabaudi* exhibits many characteristics associated with the pathogenesis of human infection
- We would like you to identify differentially expressed genes between two conditions:
  - serial blood passage (**SBP**)
    - *direct injection from mouse to mouse*
    - *results in severe disease*
  - infection with parasites via mosquitoes (**MT**)
    - *develop lower parasitaemia (presence of parasites in the blood)*
    - *mild, chronic disease*
- To determine whether the transcriptome of mosquito transmitted parasite different from one which has not passed through a mosquito