H3ABioNet
Pan African Bioinformatics Network for H3Africa

WELLCOME GENOME CAMPUS
CONNECTING SCIENCE
ADVANCED COURSES+ SCIENTIFIC CONFERENCES

Next Generation Sequencing Bioinformatics Course 2021

Transcriptome Sequencing (RNA-Seq)

Module 7: Human | Module 8: Pathogen

H3ABioNet
Pan African Bioinformatics Network for H3Africa

WELLCOME GENOME CAMPUS
CONNECTING SCIENCE
ADVANCED COURSES+ SCIENTIFIC CONFERENCES

NGS Bioinformatics Course Africa 2021
Jon Ambler
Nyasha Chambwe
Phelelani Mpangase

Lecture Outline

I RNA-Seq Overview

- RNA-Seq Background
- Experimental Design Considerations

II RNA-Seq Analysis Workflow

- Mapping
- Expression Quantification
- Read Count Normalisation
- Differential Expression Analysis

III Downstream Analysis

- Functional Enrichment Analysis

Learning Outcomes

By the end of this module you should be able to:

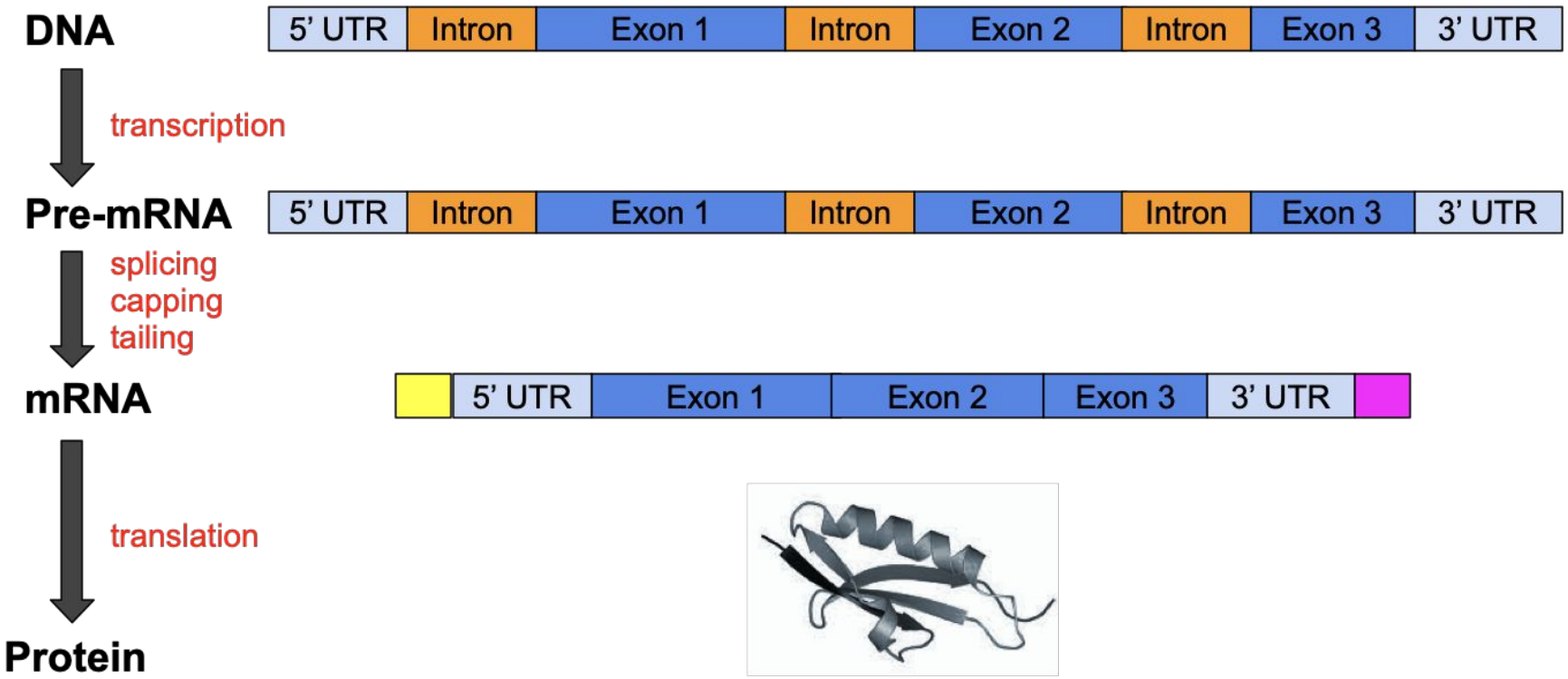
- Understand the key factors to consider when designing an RNA-seq experiment
- Describe and execute a bioinformatics analysis workflow for bulk RNA-seq from sequenced reads to differentially expressed genes

What is the transcriptome?

"The complete set of transcripts in a cell and their quantity for a specific developmental stage or condition"

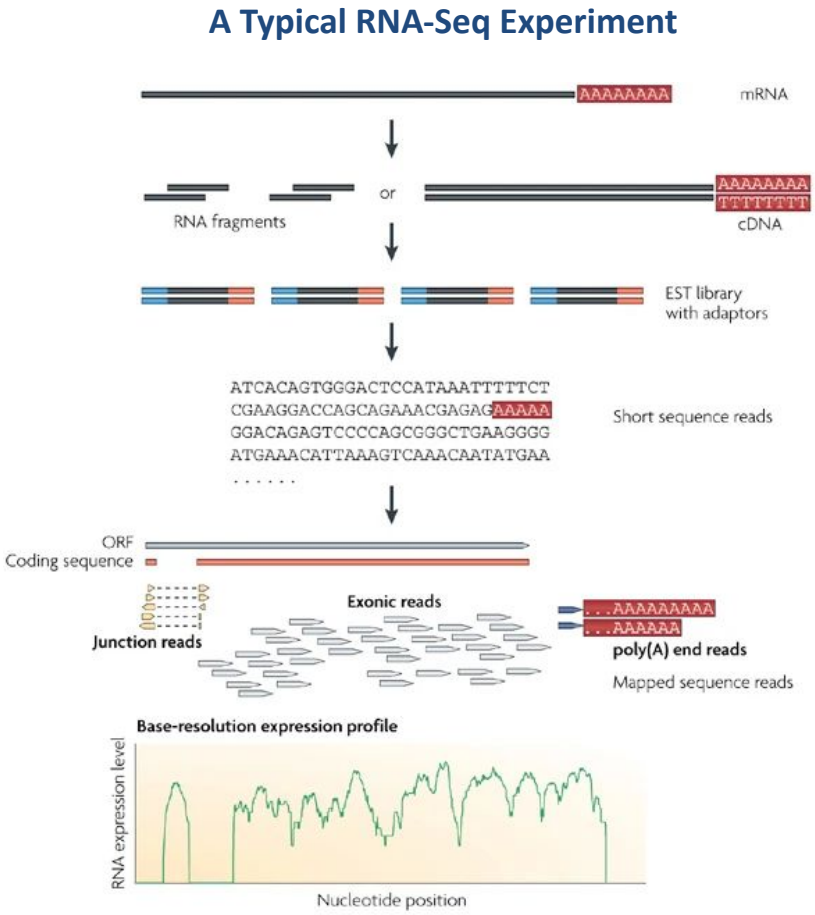
Wang *et al.* (2009) Nature Reviews Genetics (PubMed: PMID:19015660)

The Central Dogma of Biology



RNA-sequencing

- **RNA-seq:** comprehensive profiling of the transcriptome using sequencing technologies
- Measures transcript (isoform) abundance in a biological sample



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

Experimental Design

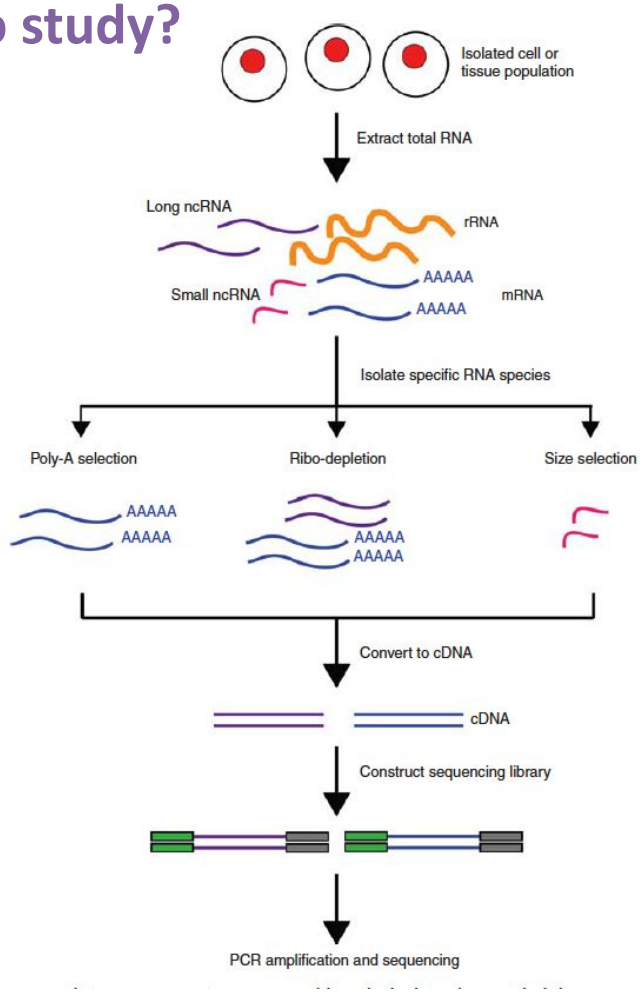
Successful RNA-Seq studies start with a good study design driven by research question

Image source: https://bit.ly/2SH9Ag9

Experimental Design: Library Preparation

What RNAs do you want to study?

- **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)
- PCR amplification can reduce coverage of transcripts or regions with a high GC content (can use amplification-free protocol)



Kukurba KR, Montgomery SB. (2015) Cold Spring Harb Protoc. (PubMed:PMID: 25870306)

Experimental Design: Library Type

What type of sequencing library is best for your experimental question?

Stranded vs unstranded

strand-specific protocols better for detangling antisense or overlapping transcripts

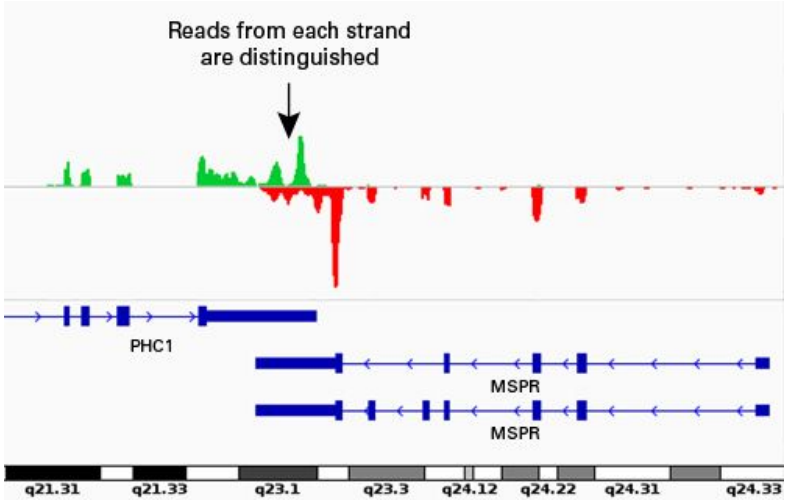


Image source: https://bit.ly/3w4xbGi

Single or paired end

paired end better for de novo transcript discovery or isoform expression analysis

< 55% reads will span 2 or more exons

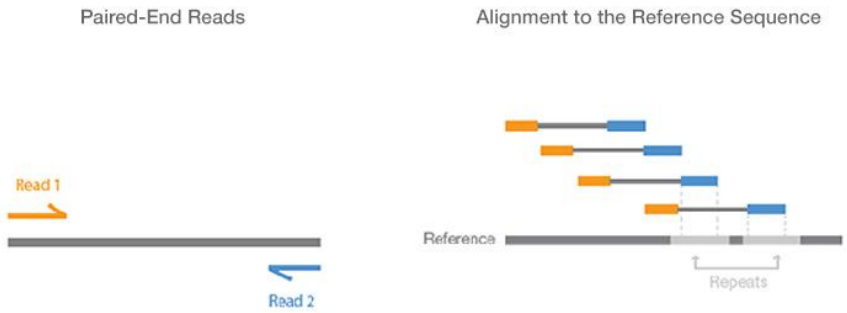


Image source: https://bit.ly/3hjNPNT

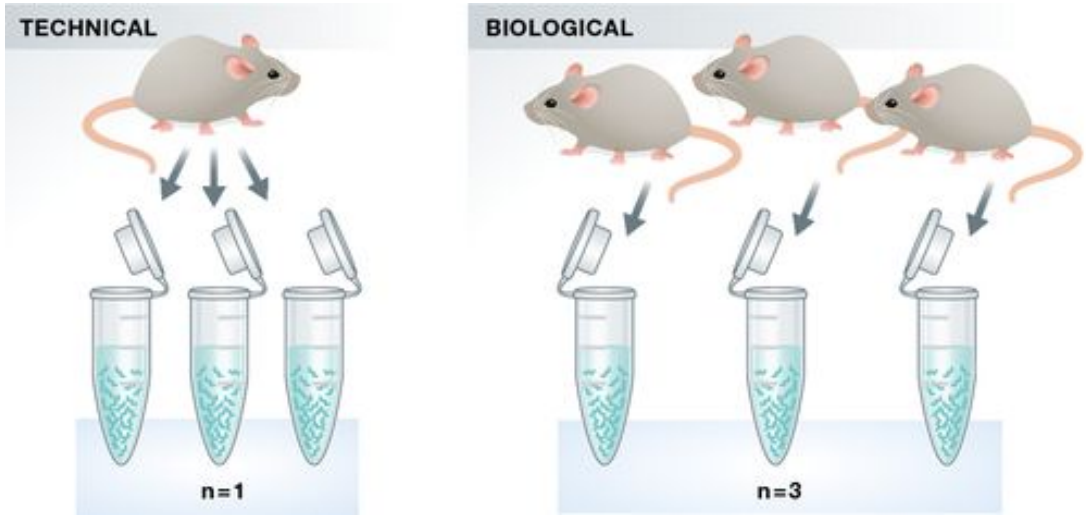
Experimental Design: Replication

How many samples do I need to sequence?

Two Group Comparison

Our ability to detect meaningful differences between groups is dependent our ability to determine variability between measurements → **replicates**

Technical vs. Biological 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

- 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**

Image source: https://bit.ly/3bqb3Oq

Experimental Design: Replication

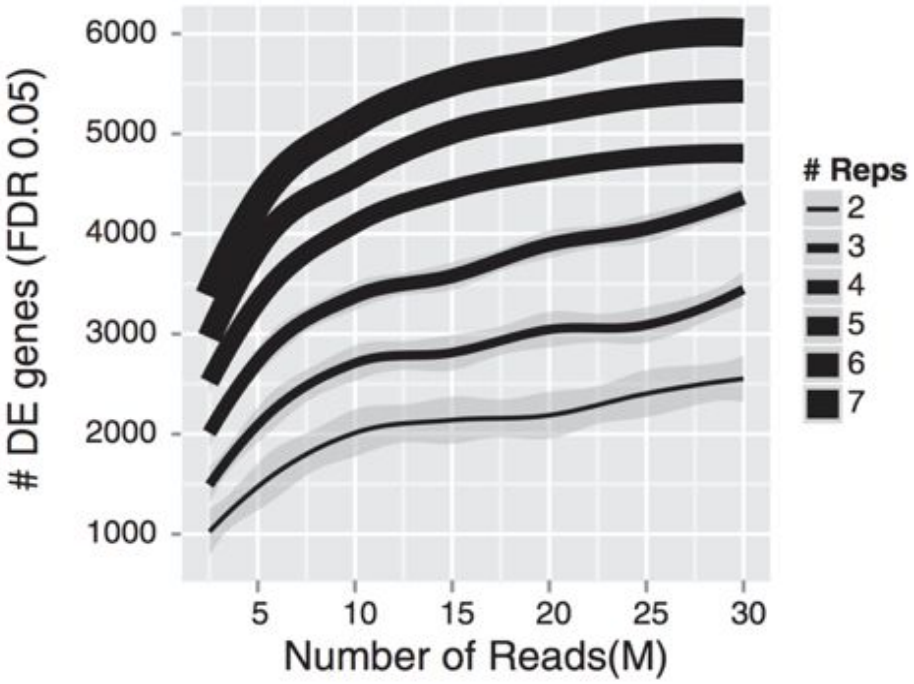
How many replicates do I really need?

$$$$$$
Sample Availability

Biological replicates
Sequencing depth

Experimental Design: Sequencing Depth/Replicate

How many replicates do I really need?



Liu Y et. al. (2014) (PubMed:PMID: 24319002)

- Increasing sequencing depth can increase the ability to detect low expression transcripts (i.e. increases ability to detect DE genes)
- 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

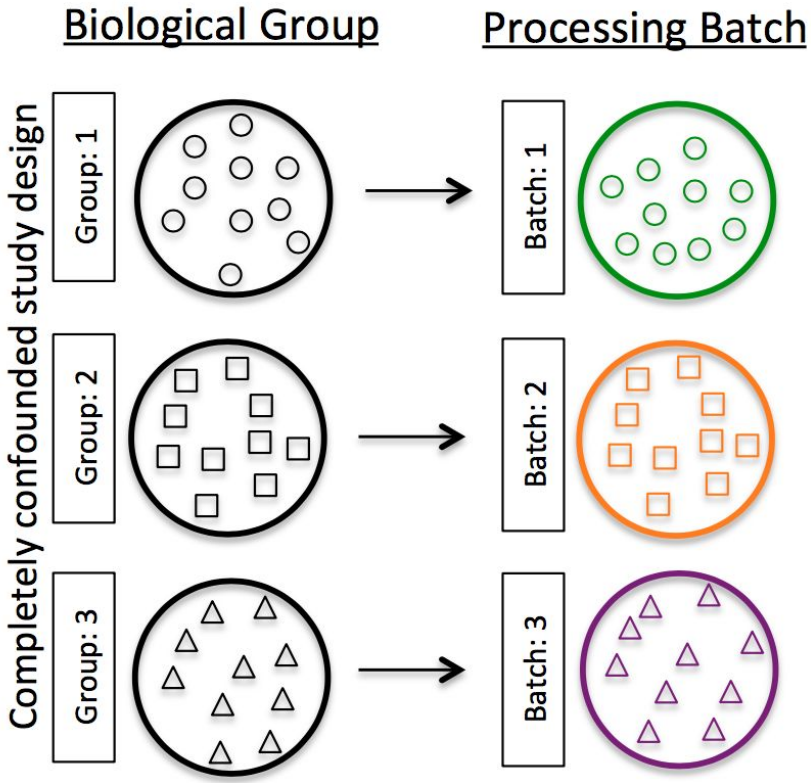
General rule of thumb:

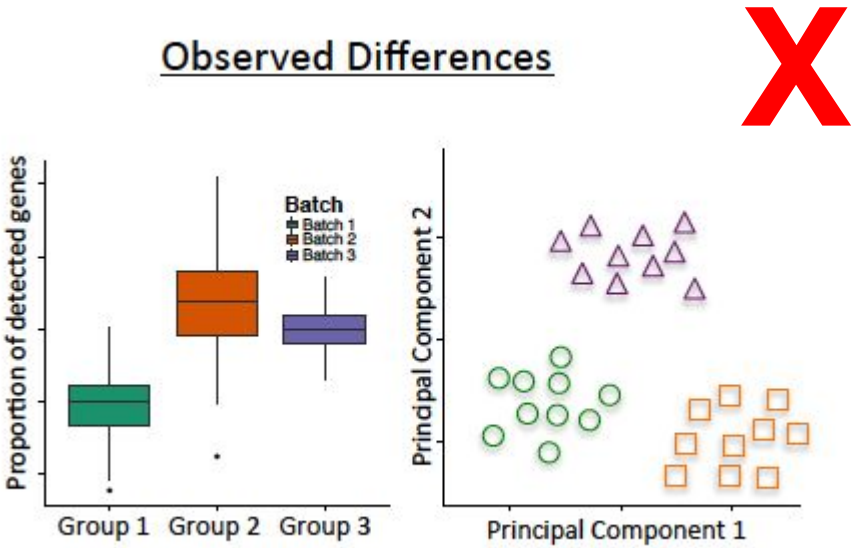
- Differential Gene Expression: 10-30M reads (SE 50-75bp)
- Alternative Splicing: 50-100M reads (PE, 2x75bp)

Consider each experiment carefully.

Experimental Design: Avoid Confounded Study Designs

Confounded Design



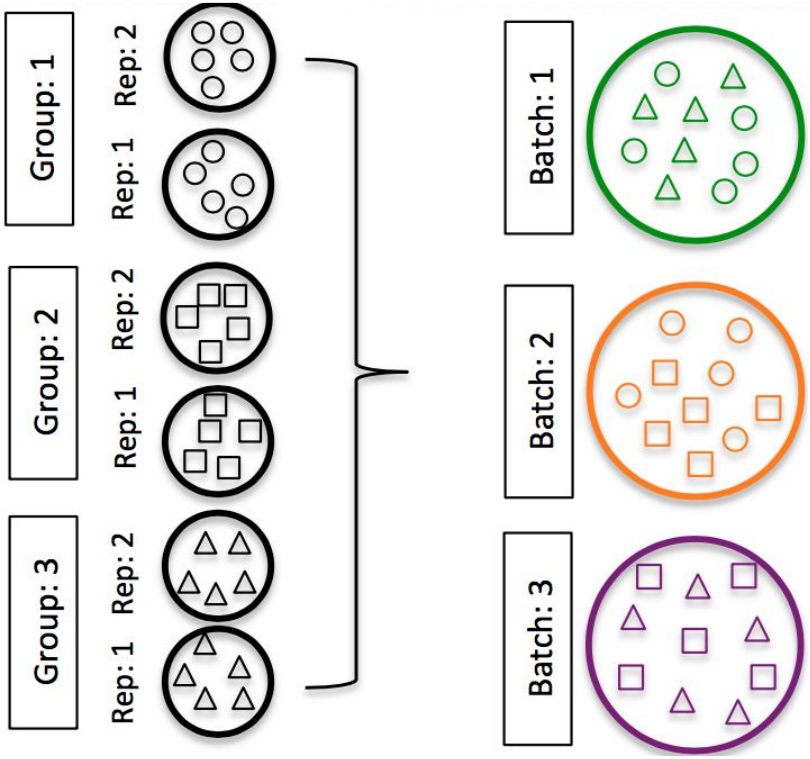


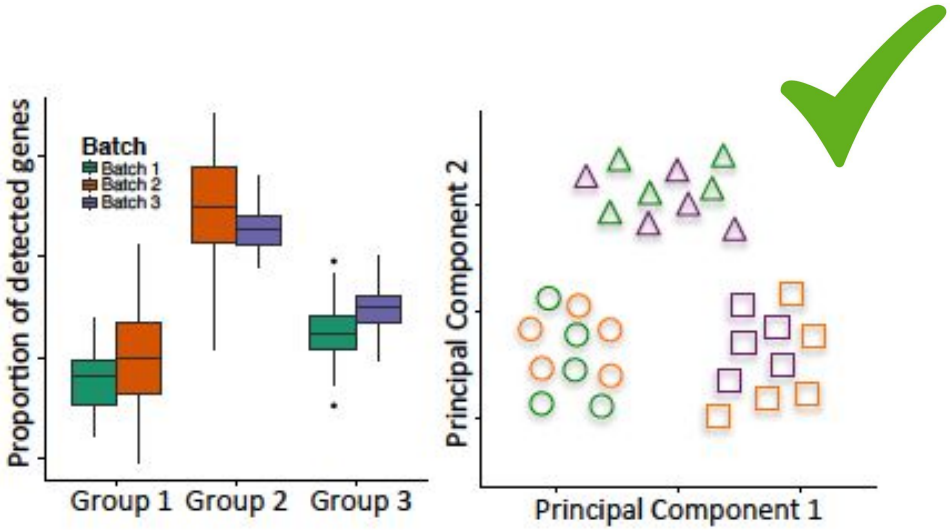
X

Hicks SC, Townes FW, Teng M, Irizarry RA. (2018) Biostatistics. (PubMed:PMID: 29121214)

Experimental Design: Avoid Confounded Study Designs

Balanced Design





Hicks SC, Townes FW, Teng M, Irizarry RA. (2018) Biostatistics. (PubMed:PMID: 29121214)

What Research Questions Can We Answer Using RNA-Seq Datasets ?

1. Which genes/transcripts do our reads belong to?

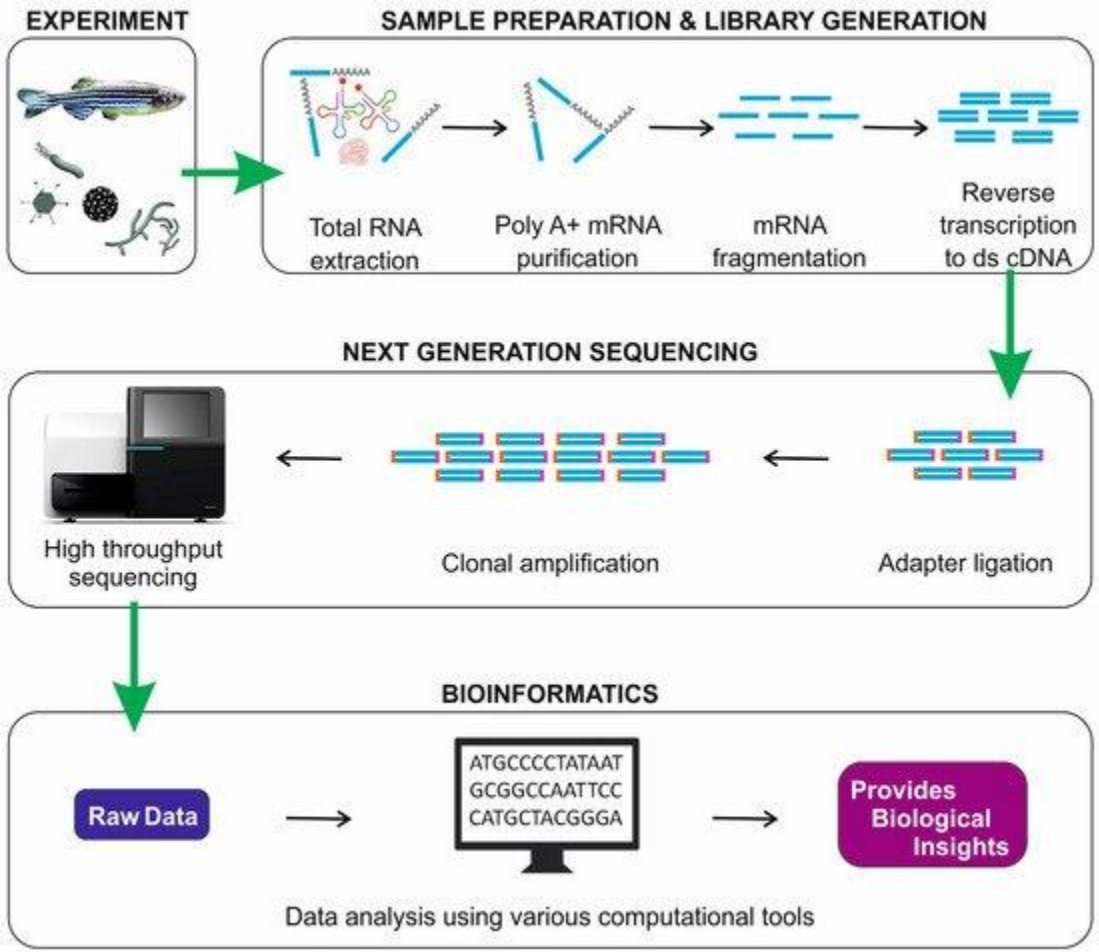
   mapping / assembly

2. How many reads align to a specific gene/transcript?

   quantification

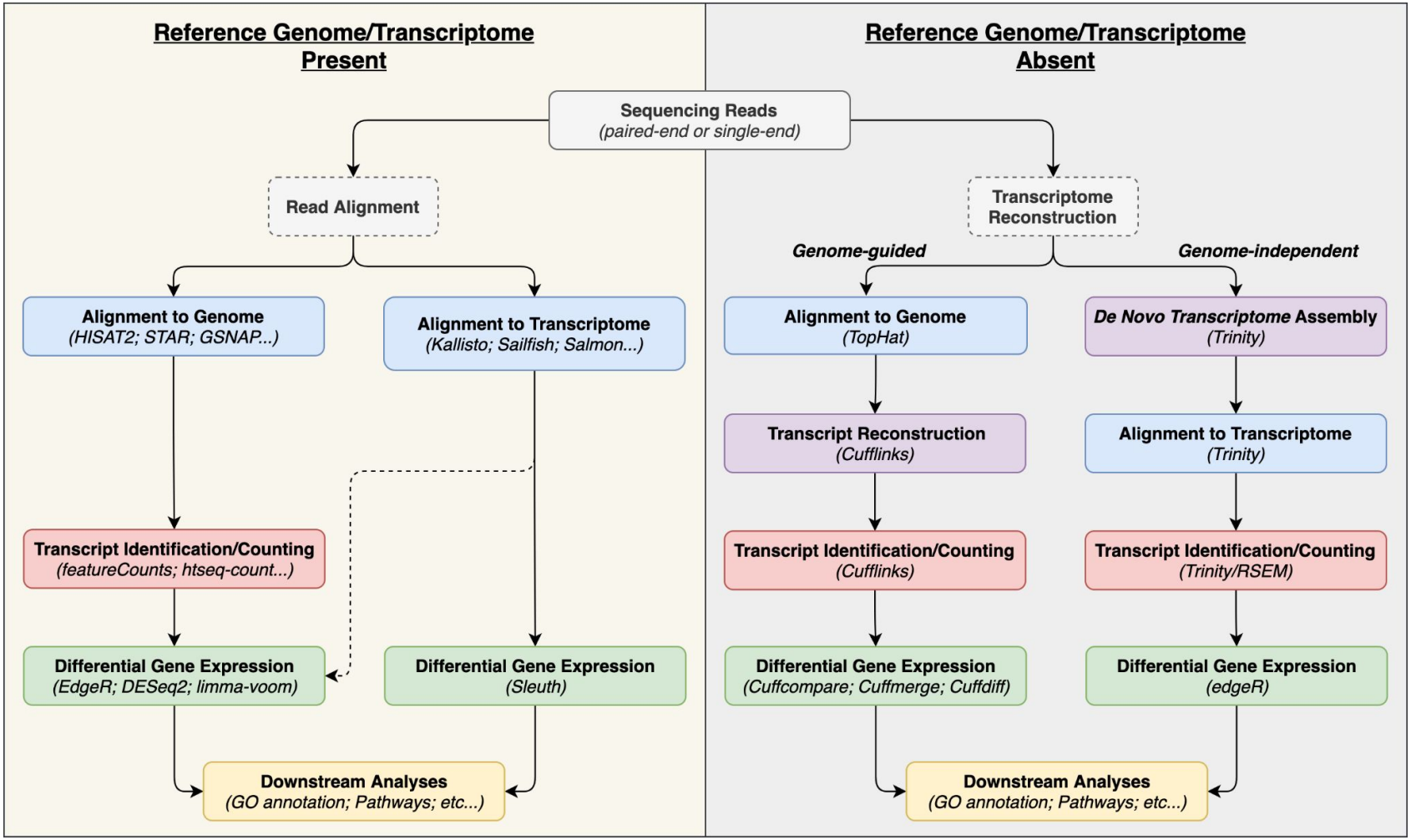
3. Do different sample groups express genes/transcripts differently? **differential gene expression (DGE) analysis**

RNA-sequencing



Sudhagar, Arun et al. (2018) Int. J. Mol Sci. (Pubmed: PMID: 29342931)

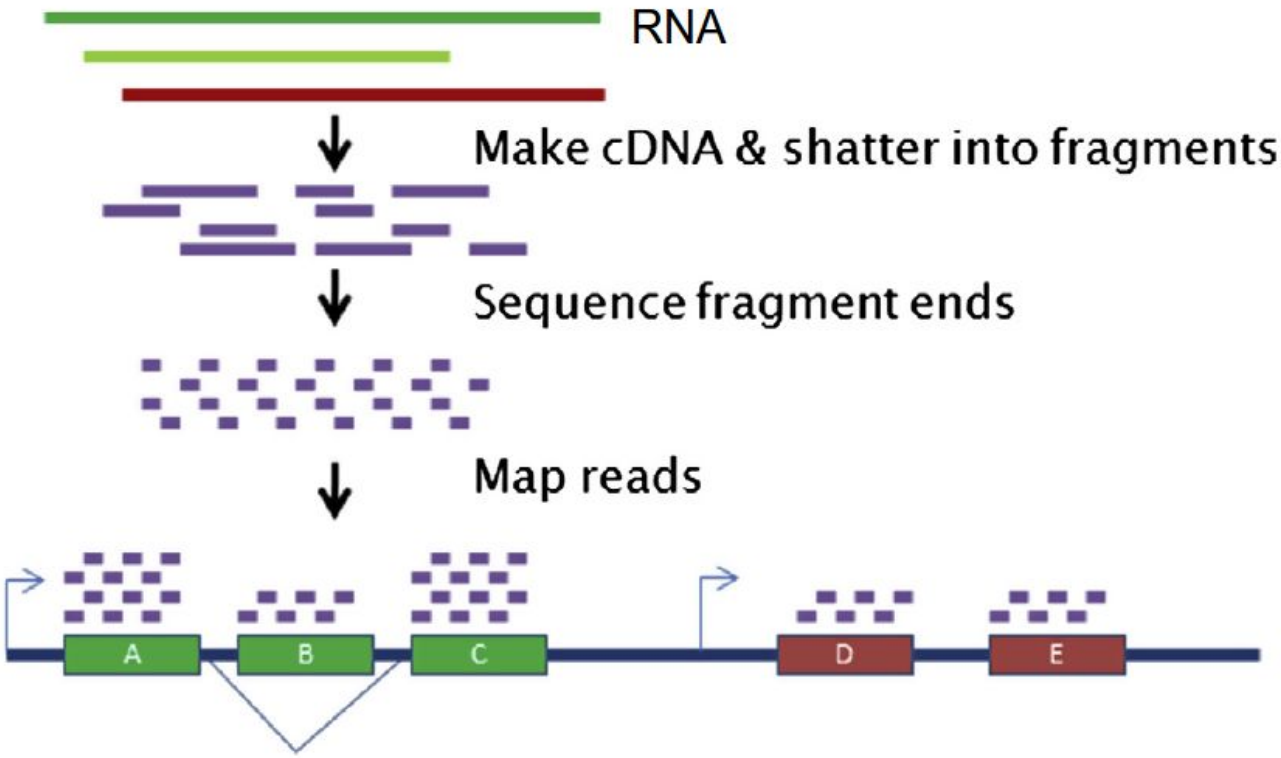
Overview of RNA-Seq Analysis Pipelines



Mapping RNA-seq reads

- Purpose of mapping:
  - Find the location in the genome where reads originated from.
  - Determine whether RNA-seq data is of high quality.
  - Explore the structure of genes of interest.

Mapping RNA-seq reads



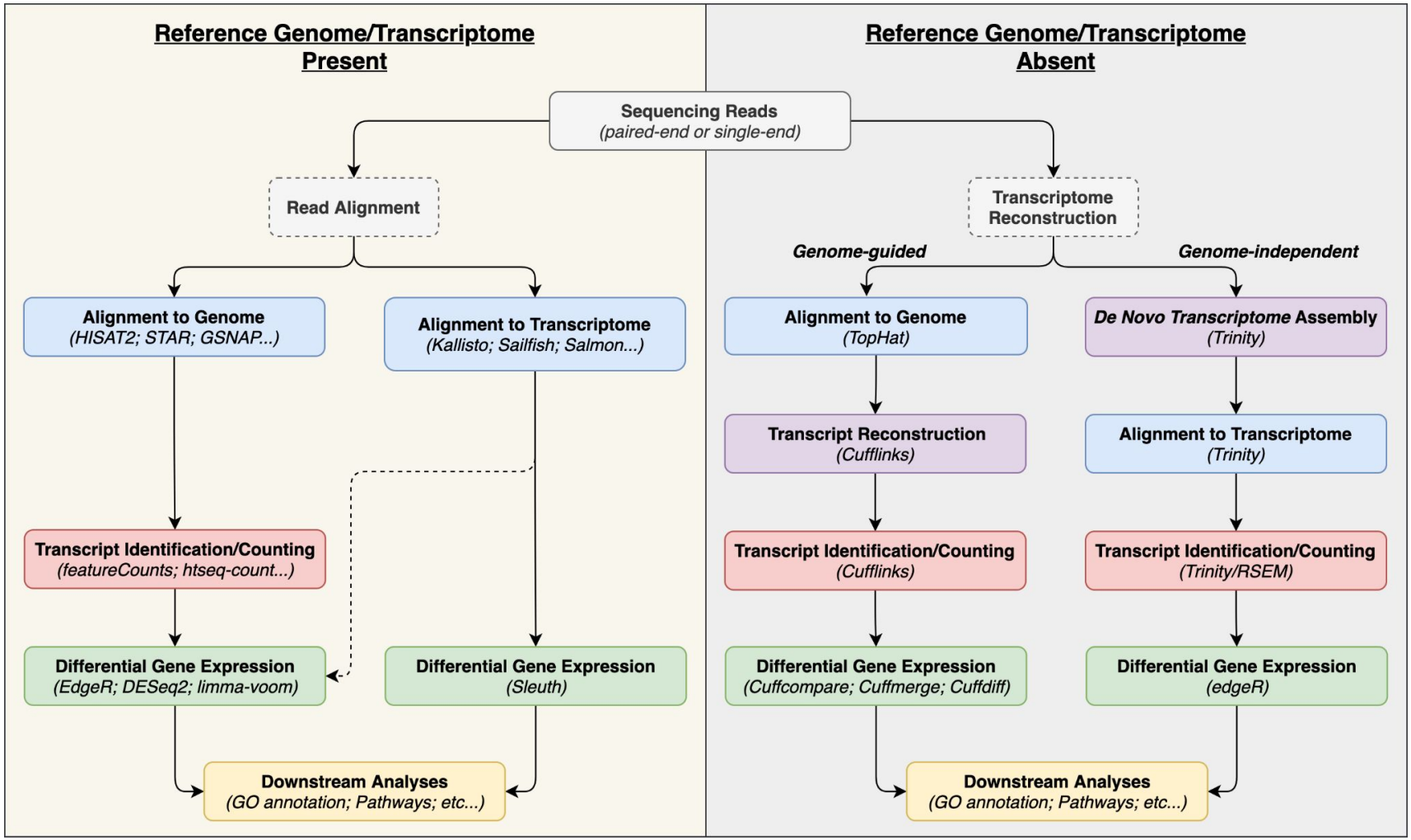
https://www.healthcarebusinesstoday.com/here-are-all-the-steps-involved-in-the-rna-seq-data-analysis/

Mapping RNA-seq reads

- Purpose of mapping:
  - Find the location in the genome where reads originated from.
  - Determine whether RNA-seq data is of high quality.
  - Explore the structure of genes of interest.

- Mapping options:
  - Reference genome
  - Reference transcriptome
  - Reconstruct transcriptome (from your data):
    - Genome guided
    - *De novo*

Mapping RNA-seq reads



Mapping RNA-seq reads

- Purpose of mapping:
  - Find the location in the genome where reads originated from.
  - Determine whether RNA-seq data is of high quality.
  - Explore the structure of genes of interest.

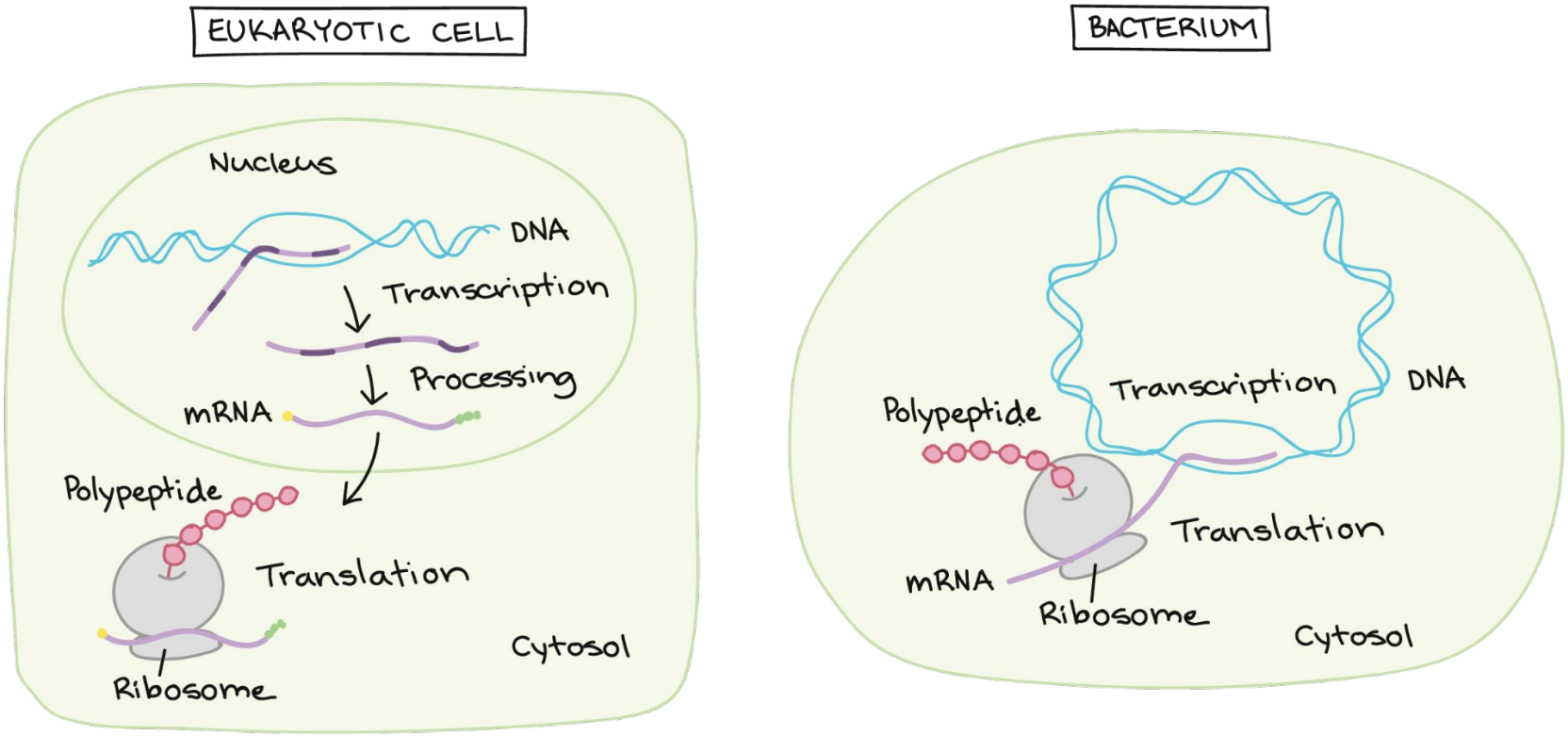
- Mapping options:
  - Reference genome
  - Reference transcriptome
  - Reconstruct transcriptome (from your data):
    - Genome guided
    - *De novo*

- Challenges:
  - Genomic variation, sequencing errors & non-unique sequences.
  - Eukaryotic genes have introns (not present in mature mRNA).

- Special "splice-aware" mapping algorithms are required.

Mapping RNA-seq reads

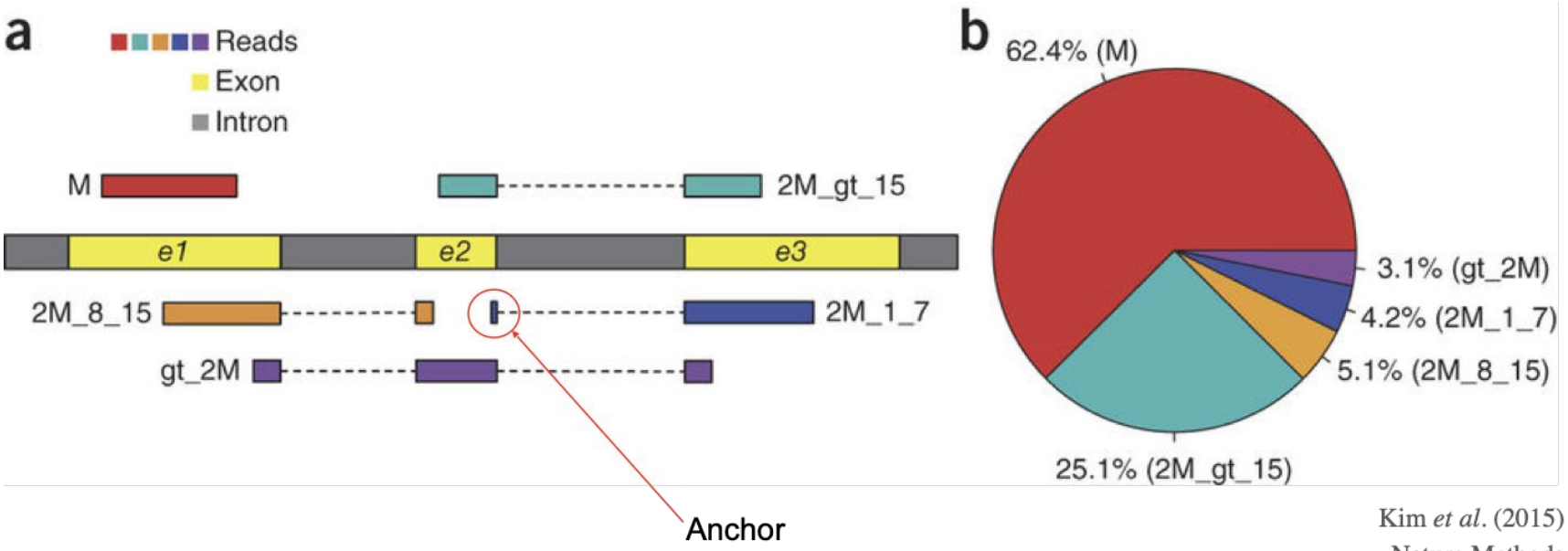
mRNA pre-processing (eukaryotic vs. prokaryotic)



https://www.khanacademy.org/science/ap-biology/gene-expression-and-regulation/transcription-and-rna-processing/a/eukaryotic-pre-mrna-processing

Mapping RNA-seq reads

Splice-aware Aligners

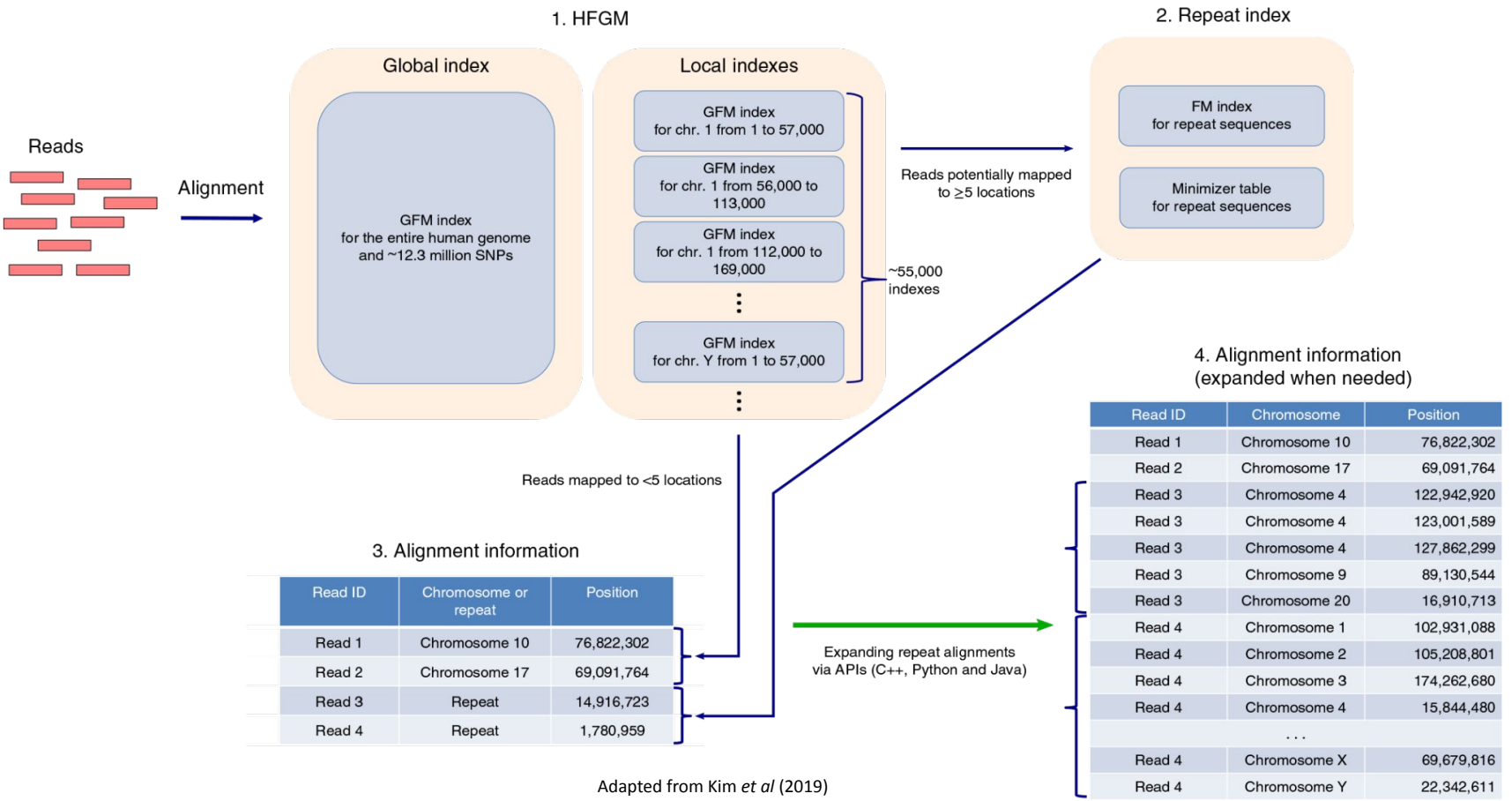


Kim *et al*. (2015)
Nature Methods
(PubMed: 25751142)

- **STAR** (Spliced Transcripts Alignment to a Reference) - Dobin *et al,* 2013
- **TopHat2** - Kim *et al*, 2013
- **HISAT2** (Hierarchical Indexing for Spliced Alignment of Transcripts) - Kim *et al*, 2019

Mapping RNA-seq reads

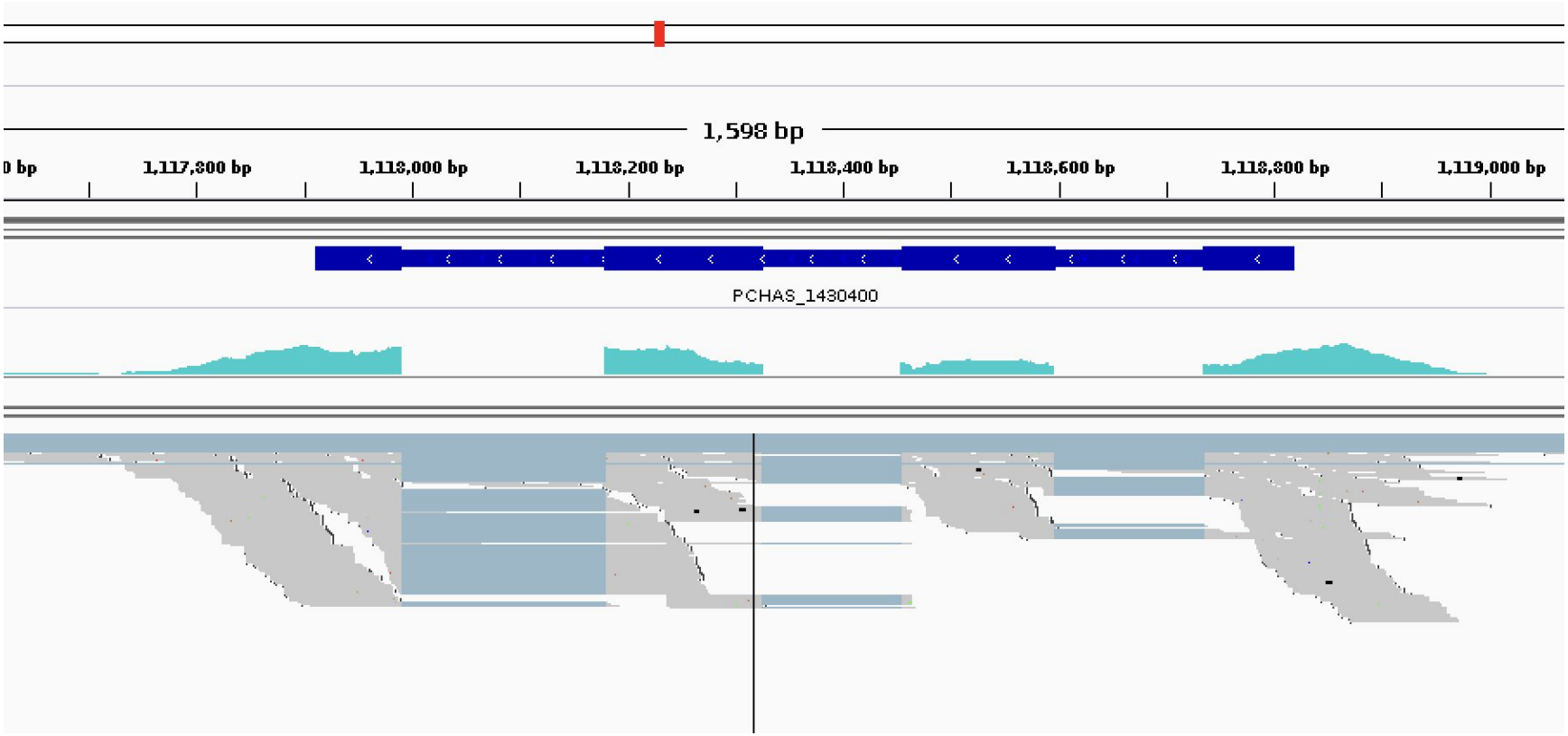
HISAT2 Indexing and Alignment



Adapted from Kim *et al* (2019)

Mapping RNA-seq reads

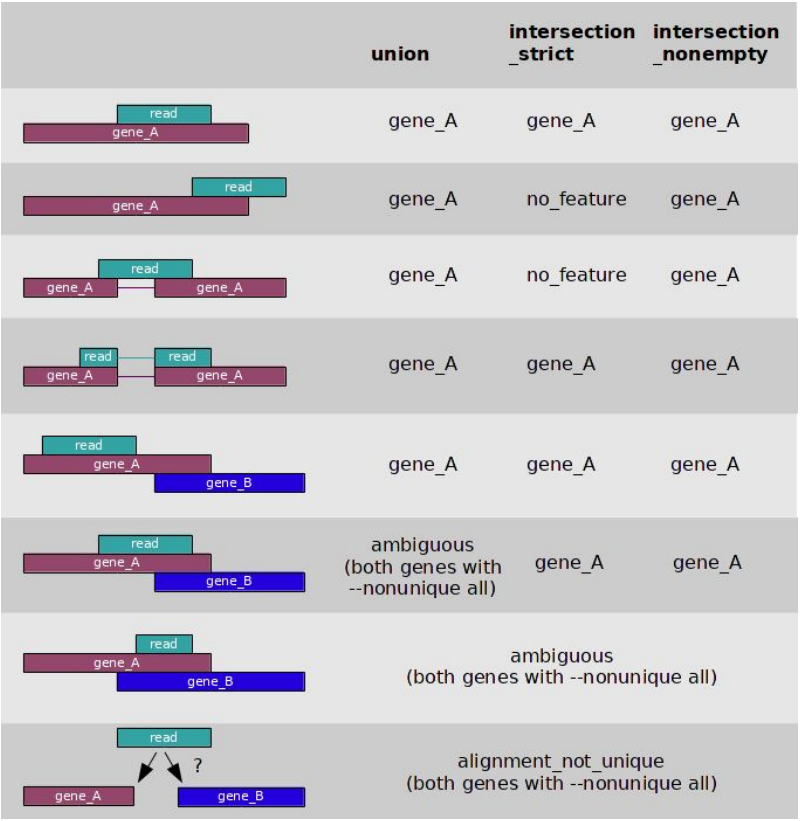
Visualisation - Integrative Genomics Viewer (IGV)



Mapping RNA-seq reads

Read Counting/Quantification

- Alignment (BAM) file as input
- How many reads align to map feature/gene.
- How to deal with reads that align to or overlap with more than one feature/gene.
- htseq-count (Anders *et al,* 2015):
  - *union*
  - *intersect-strict*
  - *intersect non-empty*



https://htseq.readthedocs.io/en/release_0.11.1/count.html#

Mapping RNA-seq reads

Mapping to the transcriptome (pseudoalignment)

- Multiple splice forms per gene introduce ambiguity.
- Mapping to the spliced transcript sequences:
  - Allows ambiguity to be taken into account.
  - Allows transcript-specific read counts.
  - Faster - less target sequence.
- Recent improvements in algorithms make this even faster:
  - Algorithms don’t care where in each transcript reads map to.
  - Only consider which of the transcripts they map to.
- Counting comes for free.

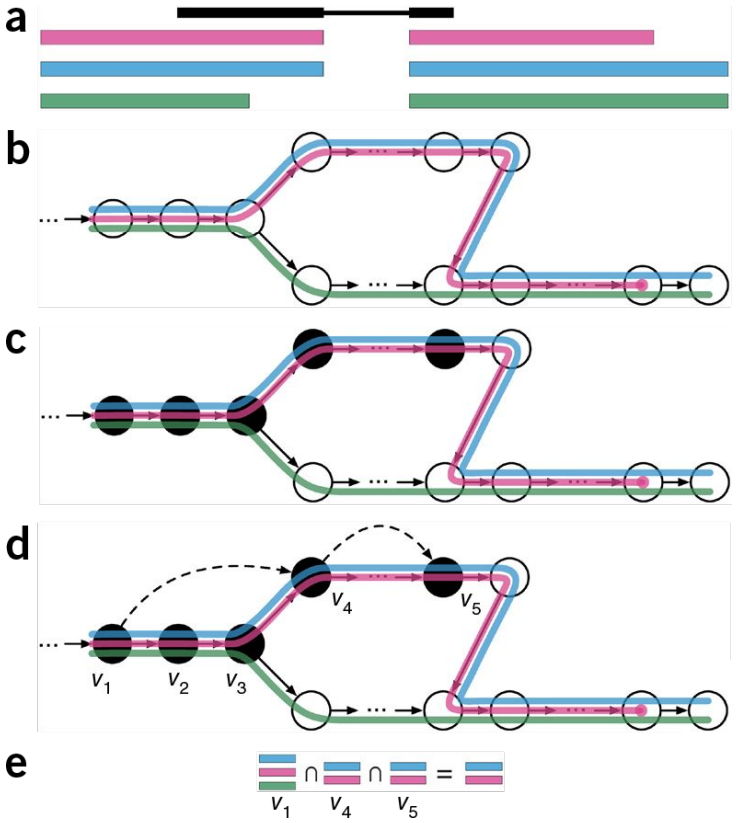
Mapping RNA-seq reads

Mapping to the transcriptome (pseudoalignment)

- Kallisto has two steps:
  1. Building an index from the spliced transcript sequences.
  2. Quantify reads against the index.
- However, Kallisto cannot be used to identify novel transcripts.

Mapping RNA-seq reads

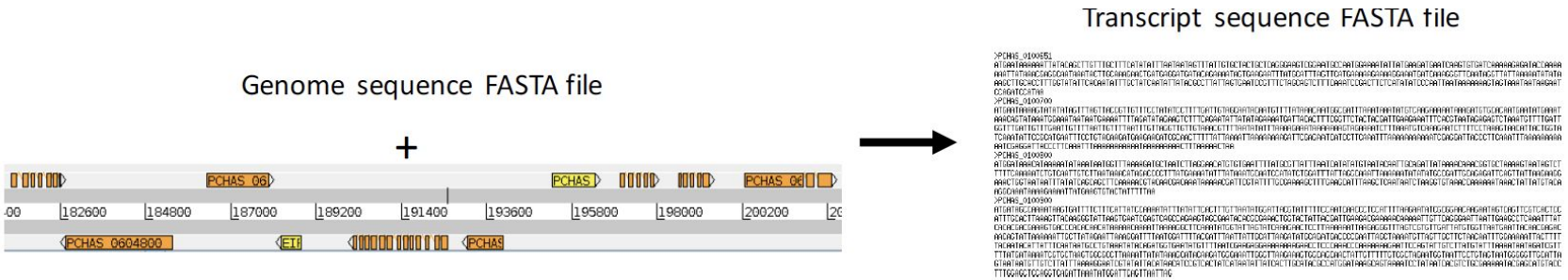
Mapping to the transcriptome (pseudoalignment)

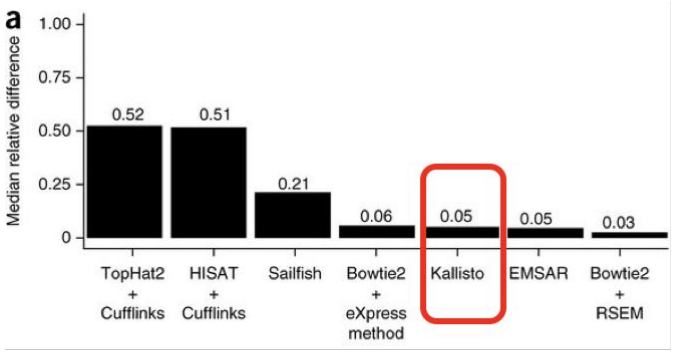


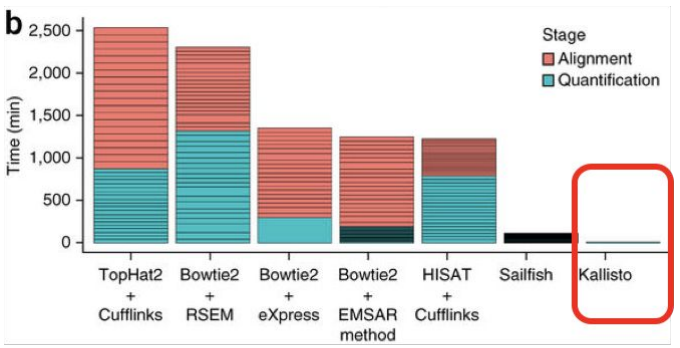
Adapted from Bray *et al* (2016)

Mapping RNA-seq reads

Mapping to the transcriptome (pseudoalignment)

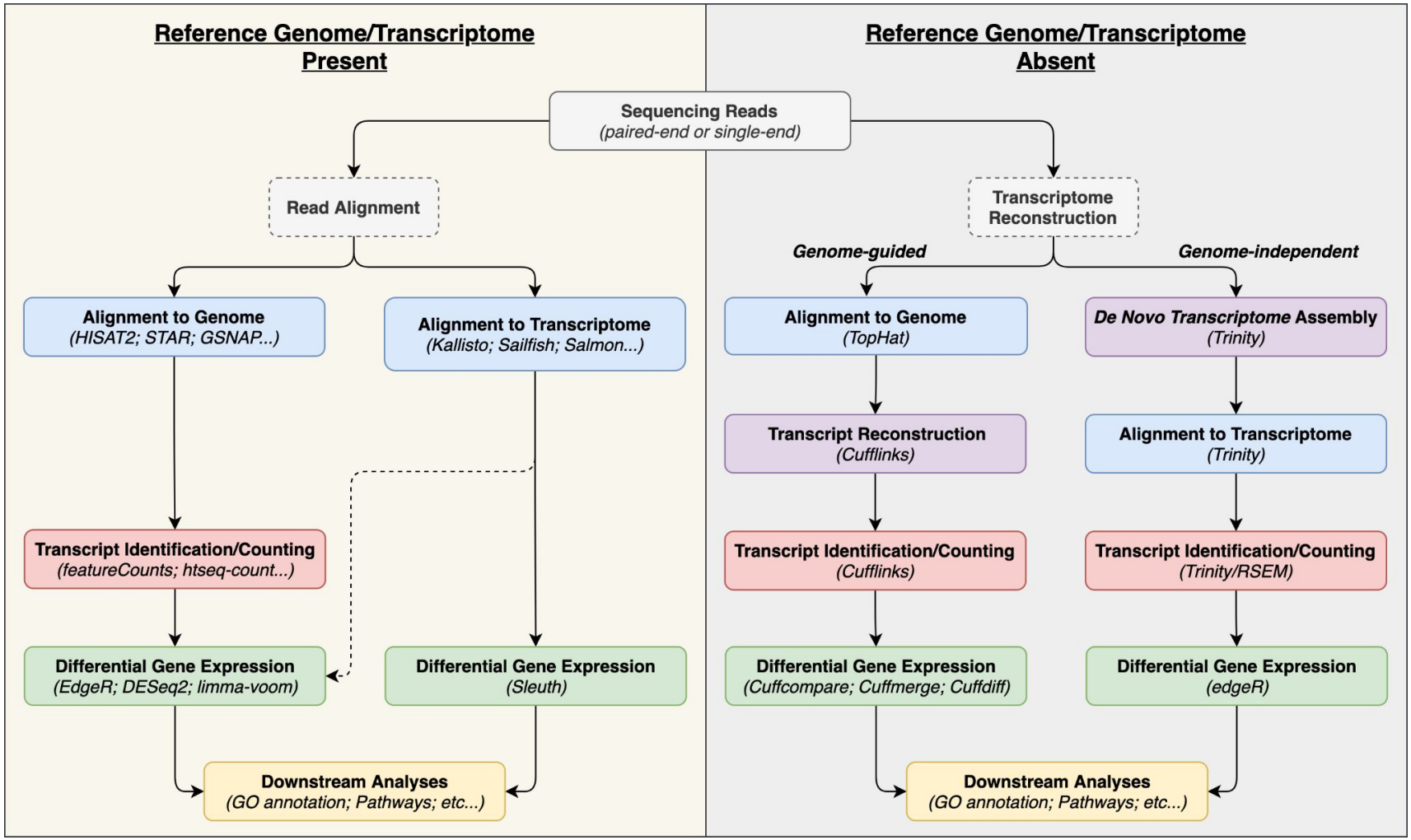






Bray *et al*. (2016)
Nature Biotechnology
(PubMed: 27043002)

Mapping RNA-seq reads

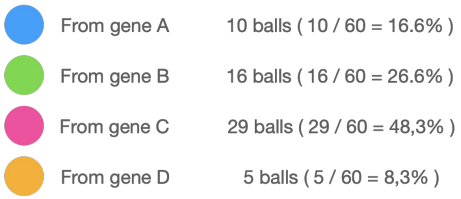


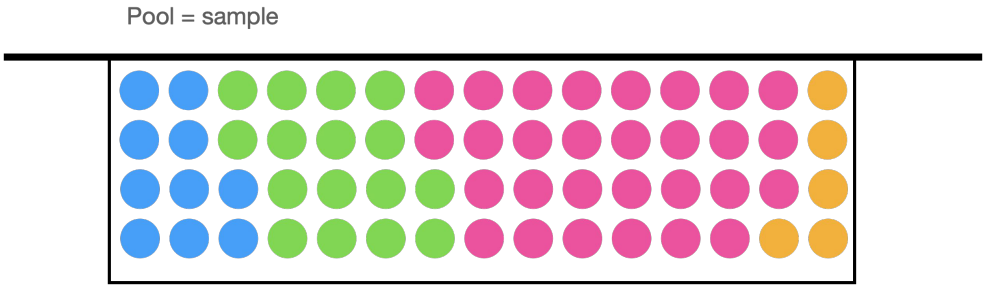
Normalisation

- Most tools have built in normalisation methods
- It is important to understand why data needs to be normalised
- Understand where the bias comes from
  - Runs with more depth will have more reads mapping to each gene
    - (**sequencing depth bias**)
  - Longer genes will have more reads mapping to them
    - (**gene length bias**)

Normalisation - Seq depth bias

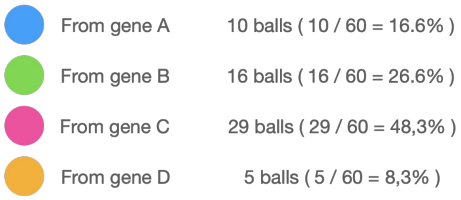
- A sample contains a large amount of RNA
- Each gene contributes a portion of the total amount of reads

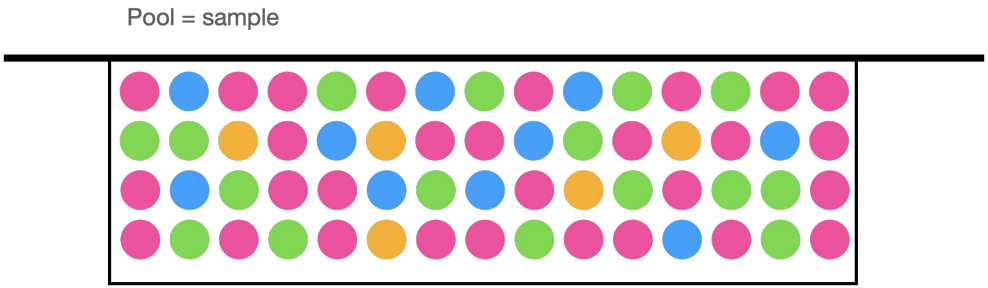




Normalisation - Seq depth bias

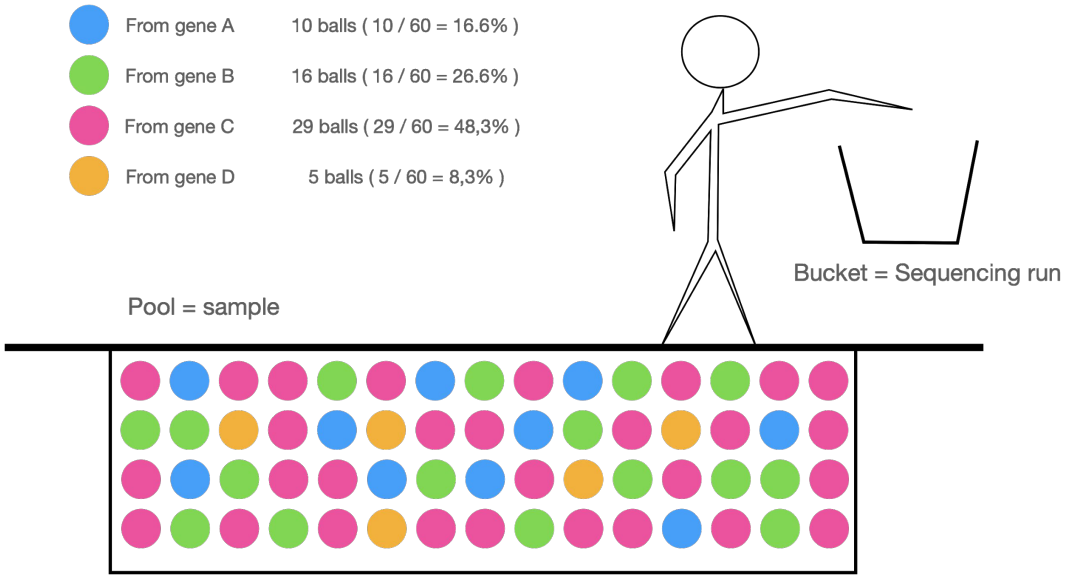
- A sample contains a large amount of RNA
- Each gene contributes a portion of the total amount of reads
- We assume the reads are randomly distributed in the sample





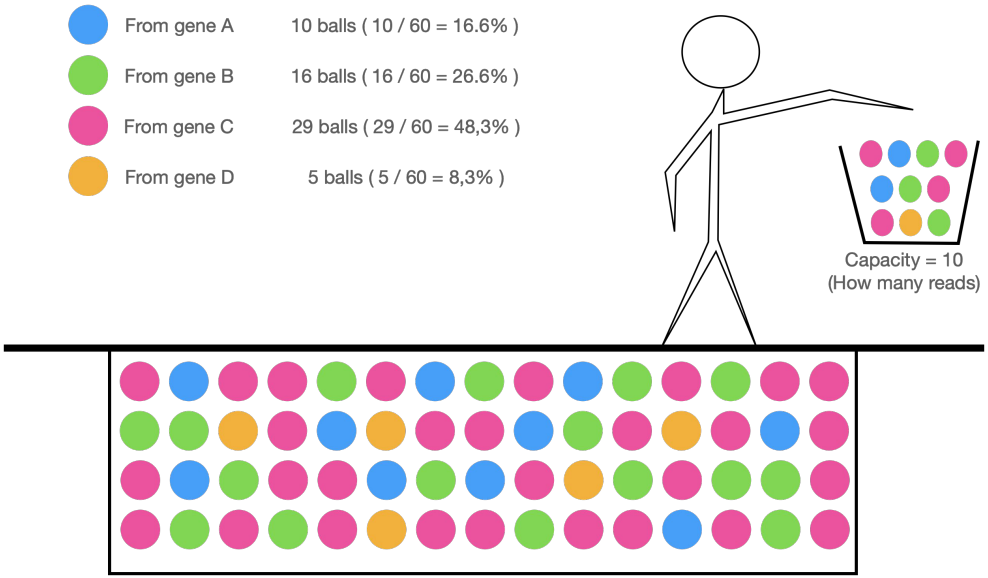
Normalisation - Seq depth bias

- A sample contains a large amount of RNA
- Each gene contributes a portion of the total amount of reads
- We assume the reads are randomly distributed in the sample



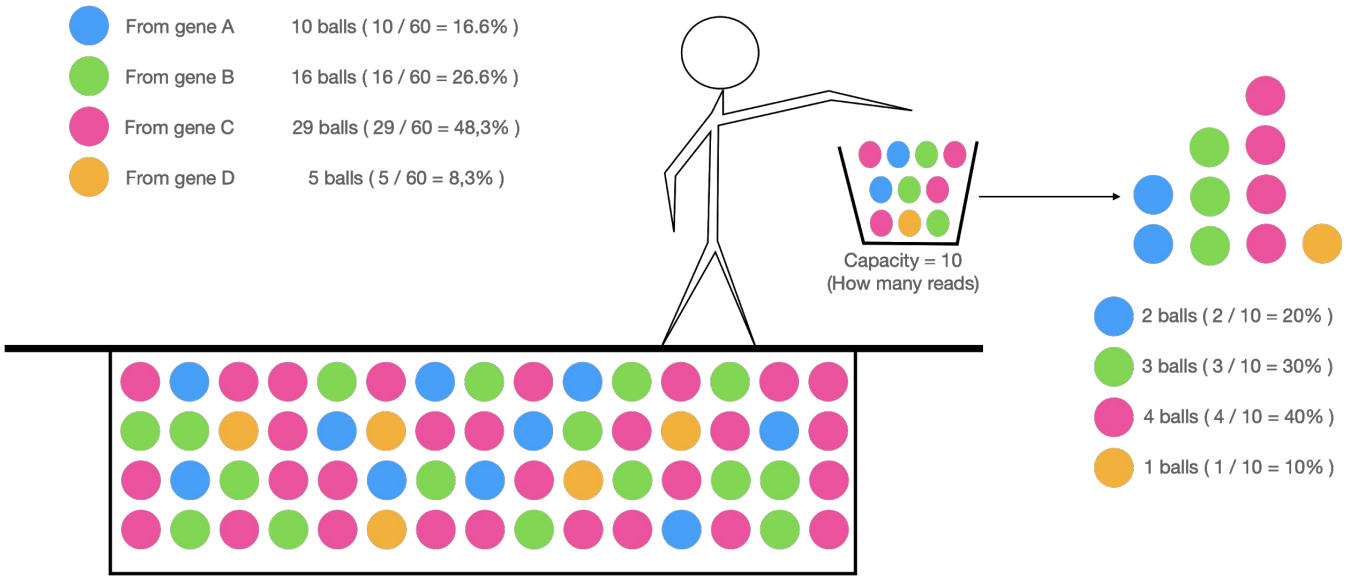
Normalisation - Seq depth bias

- A sample contains a large amount of RNA
- Each gene contributes a portion of the total amount of reads
- We assume the reads are randomly distributed in the sample



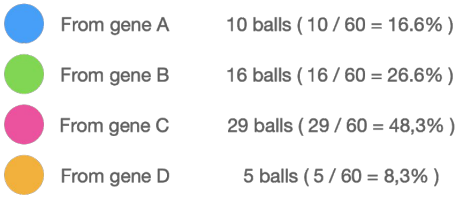
Normalisation - Seq depth bias

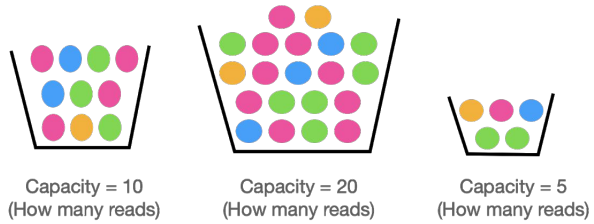
- A sample contains a large amount of RNA
- Each gene contributes a portion of the total amount of reads
- We assume the reads are randomly distributed in the sample
- The sample is an approximation of the true proportions

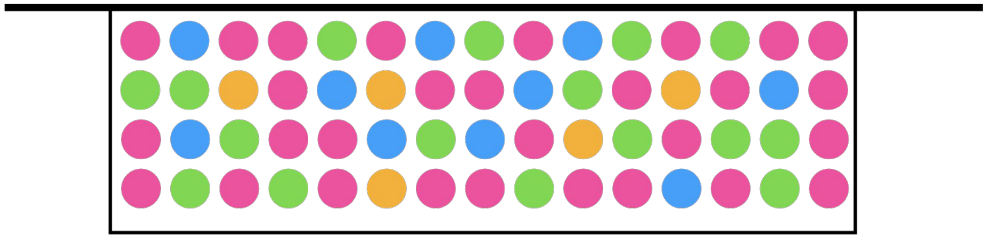


Normalisation - Seq depth bias

- Genes from higher depth samples appear to have higher expression
- We need to compensate for different sequencing depths
- Account for more uncertainty in samples with lower depths

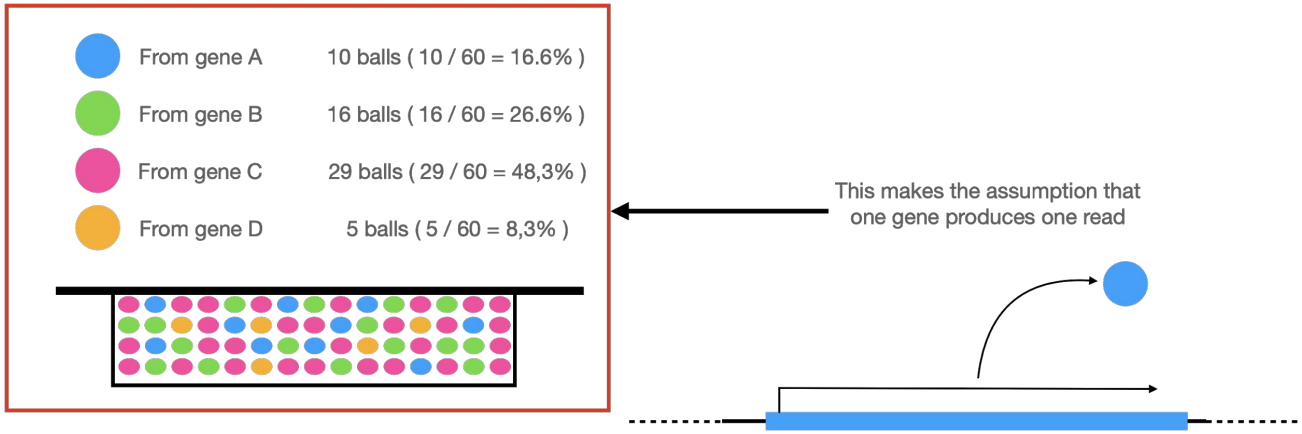






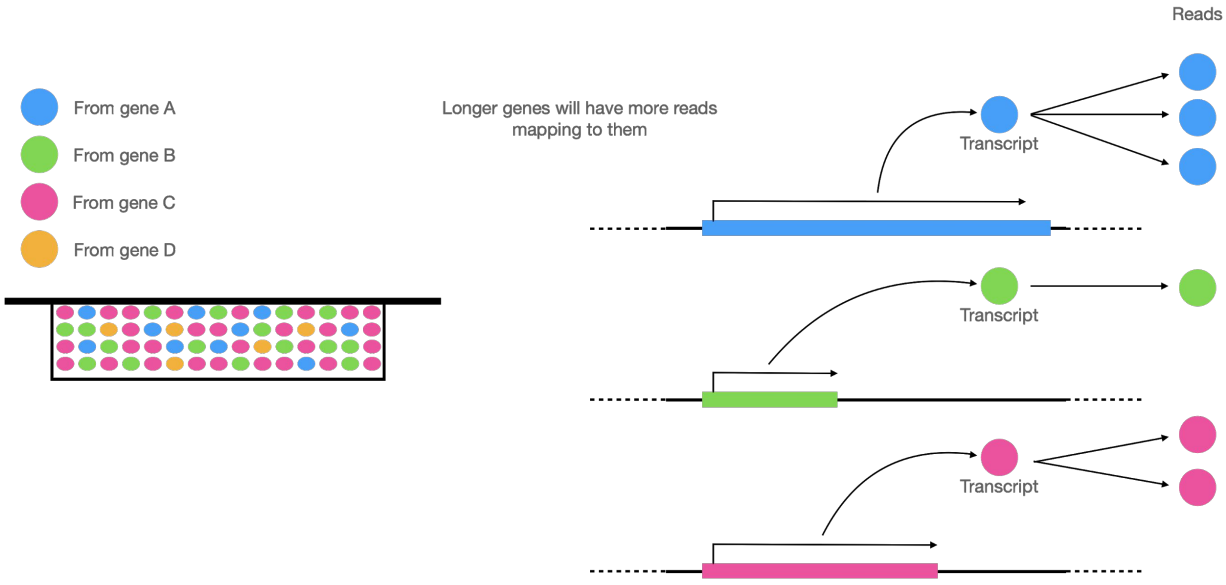
Normalisation - gene length

Within samples: Different gene lengths



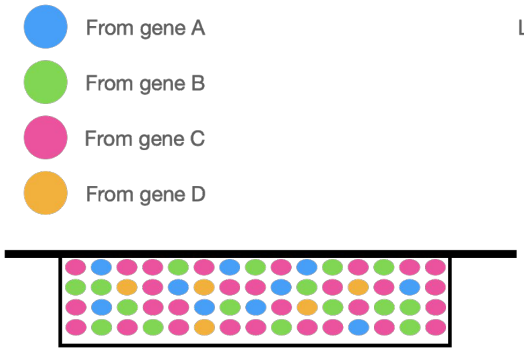
Normalisation - gene length

Within samples: Different gene lengths

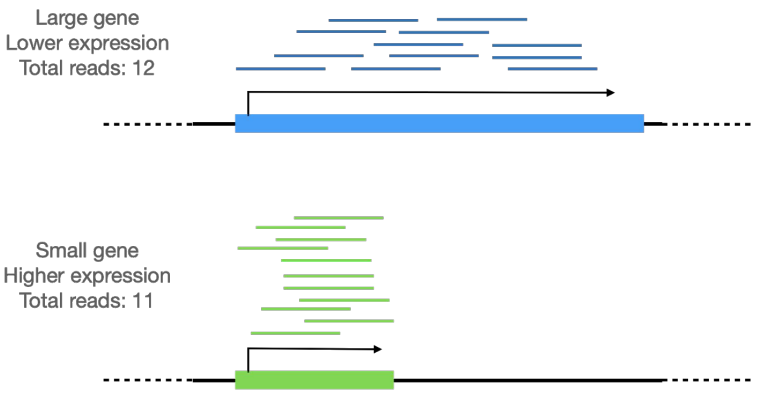


Normalisation - gene length

Within samples: Different gene lengths

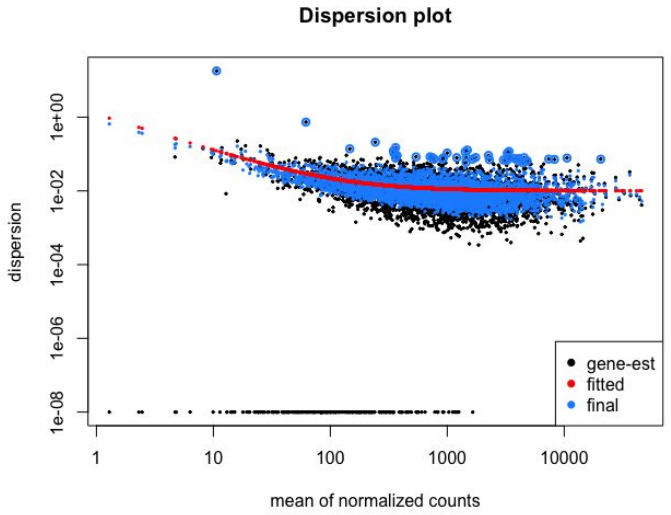


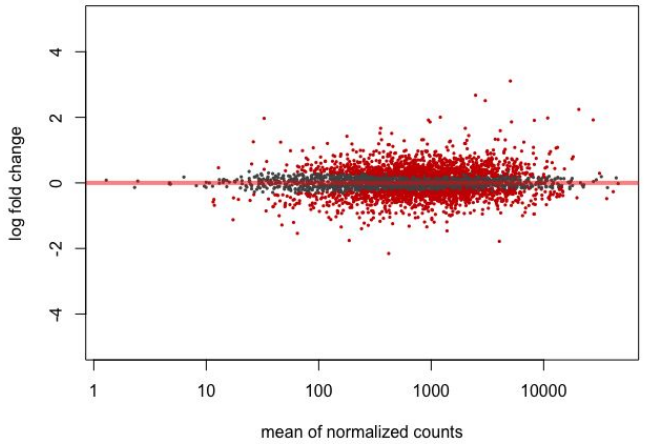
Longer genes will have more reads mapping to them



Normalisation methods

- Old Normalisation methods:
  - **RPKM** - reads per kilobase per million
  - **FPKM** - fragments per kilobase per million
  - **TPM** - transcripts per million
- Some of these methods have problems with highly expressed genes
- It's better to use more sophisticated normalisation procedures (**DESeq2 rlog**, **Sleuth**)
- Model based normalisation
- Visualise data to see if the normalisation was successful





Normalisation methods - RPKM

BEFORE

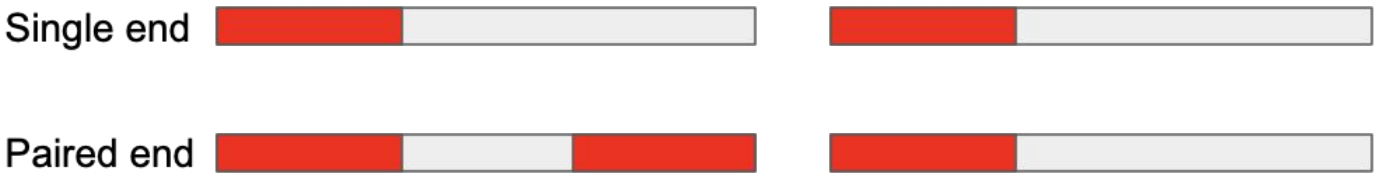
Gene	Replicate 1 Counts	Replicate 2 Counts	Replicate 3 Counts
A (2,000 bases)	10	12	30
B (4,000 bases)	20	25	60
C (1,000 bases)	5	8	15
D (10,000 bases)	0	0	1

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

Normalisation methods - FPKM

- FPKM (fragments per kilobase million)
- RPKM for paired reads
- takes into account that two reads can map to one fragment (and so it doesn't count this fragment twice)



Normalisation methods - RPKM vs TPM

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

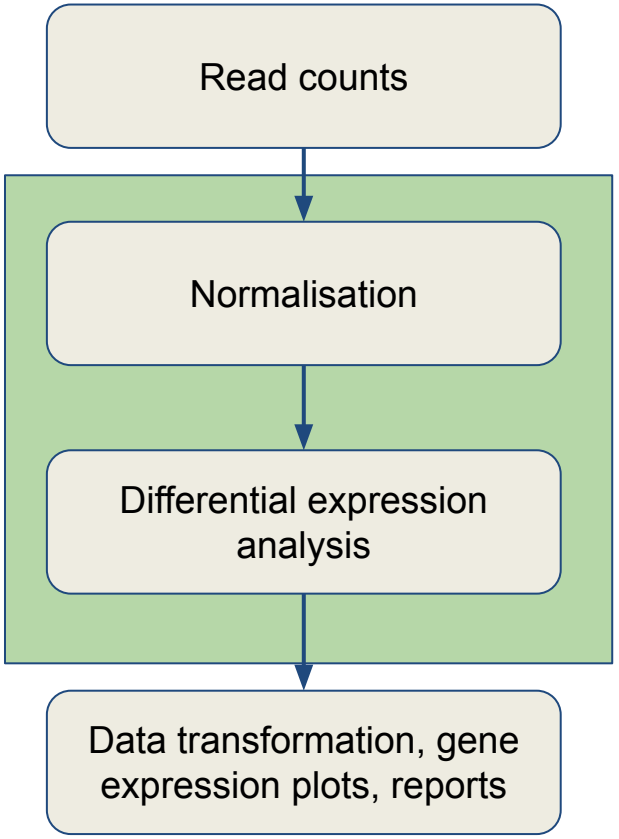
Gene	R1	R2	R3
A	3.33	2.96	3.326
B	3.33	3.09	3.326
C	3.33	3.95	3.326
D	0	0	0.02
Total	**10**	**10**	**10**

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

Adapted from StatQuest (http://statquest.org)

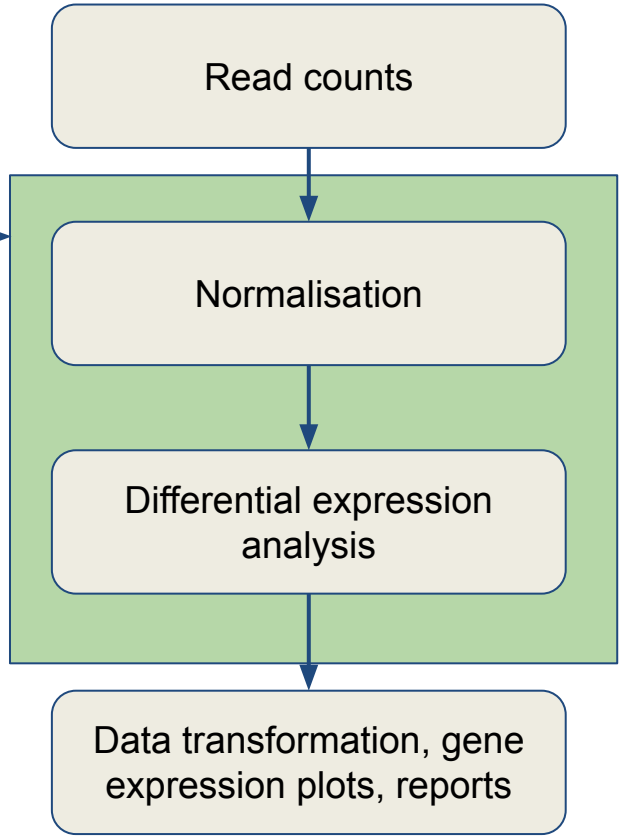
Differential expression analysis

- 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



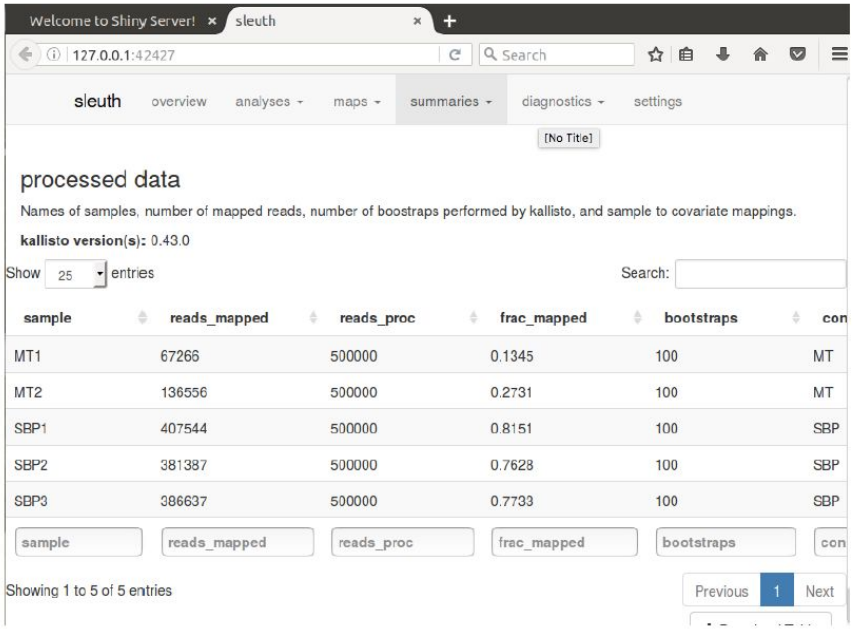
Differential expression analysis

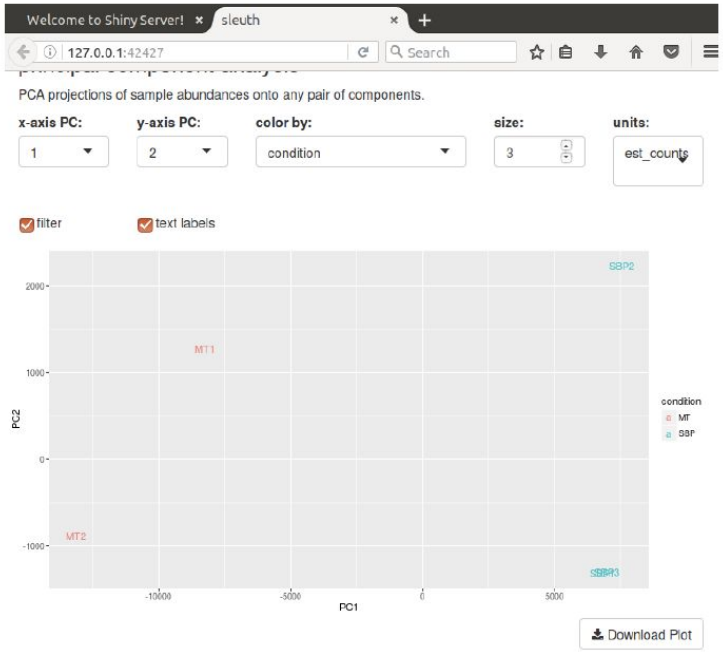
- Normalisation and differential expression analysis handled by Deseq2, edgeR, Sleuth
- Deseq2
  - Applies a Generalised linear model
  - Modeled using a negative binomial distribution



Differential expression analysis

QC with Sleuth

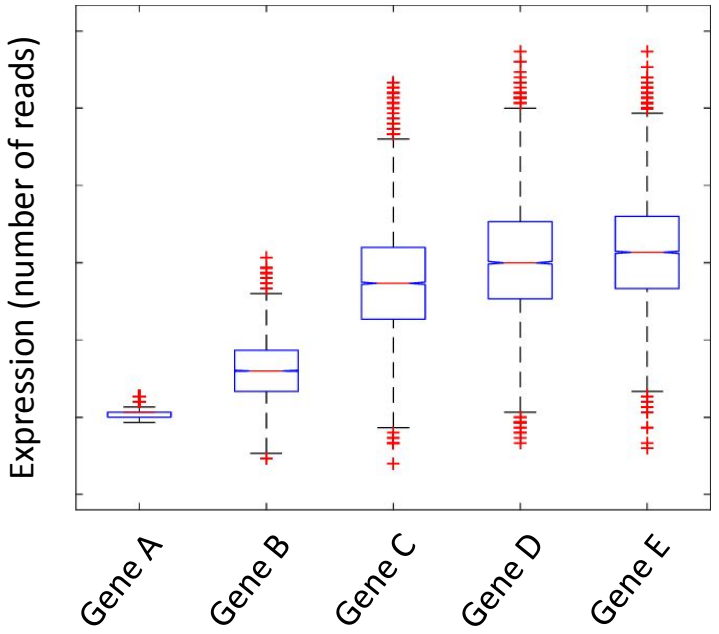




Data transformation before PCA

- PCA generally requires data with the same range of variance at different ranges of the mean values
- Genes with higher levels of expression have higher absolute differences between samples
- This means that they will influence the PCA more than other genes
  - Gene D will have more influence than Gene A
- Data transformations with:
  - Log2
  - Regularised log (rlog)
  - Variance stabilising transform (vst)

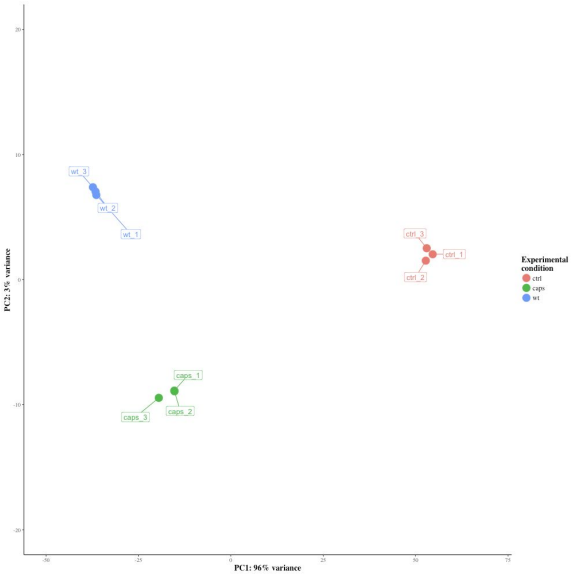
Try compensate for this effect.



Differential expression analysis

Principal component analysis (PCA)

- What is it?
  - A dimensionality reduction method
  - Identifies uncorrelated variables or principal components (PC)
  - Tries to explain the maximum amount of variance with the smallest number of principal components
- Why do it?
  - Samples / replicated should cluster by condition / phenotype
  - Look for batch effects
  - Exclude outliers



Differential expression analysis

Principal component analysis (PCA)

- What is meant by dimensionality reduction?
- How similar are samples in terms of their expression profiles?

Differential expression analysis

Principal component analysis (PCA)

- What is meant by dimensionality reduction?
- How similar are samples in terms of their expression profiles?

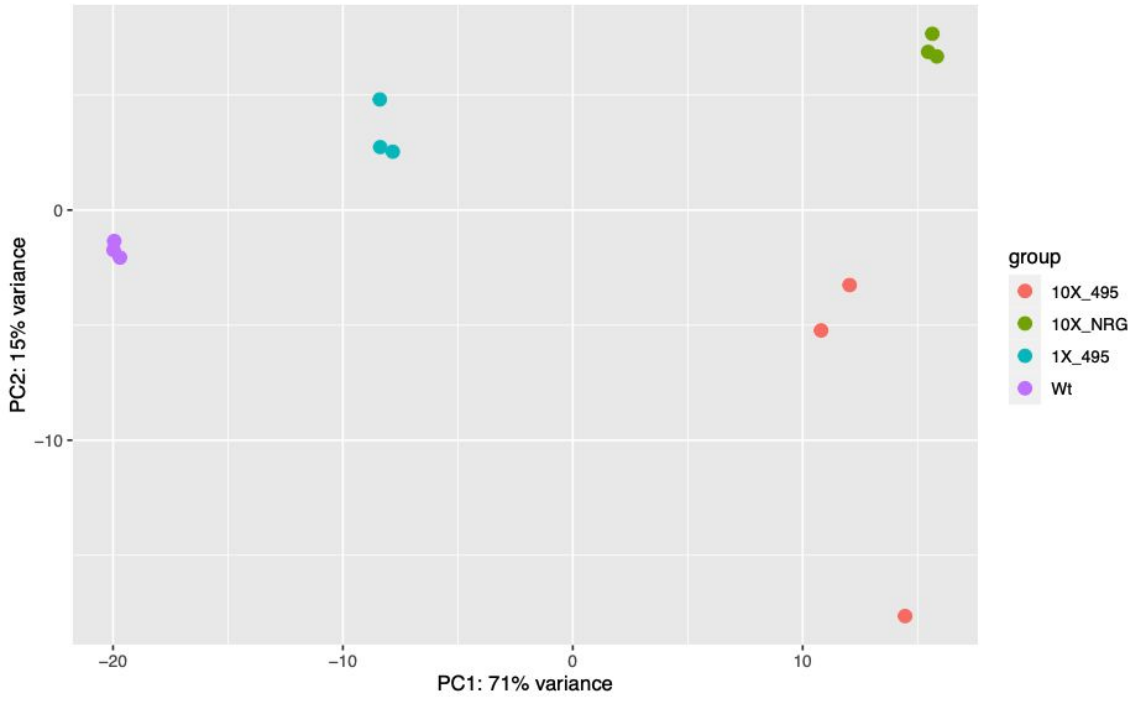
Differential expression analysis

Principal component analysis (PCA)

- What is meant by dimensionality reduction?
- How similar are samples in terms of their expression profiles?

Differential expression analysis

Principal component analysis (PCA)



Downstream Analysis

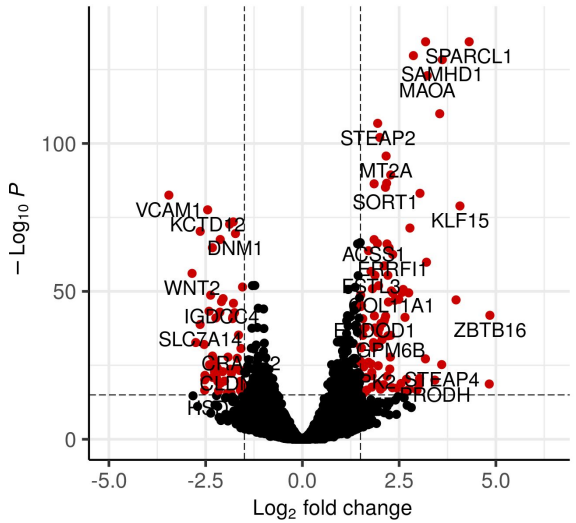
What do you do next with your list of differentially expressed genes?

Condition A
normal
untreated
control

VS

Condition B
disease
treated
perturbed

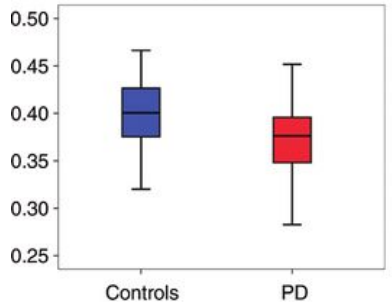
Differential Expression Analysis Result



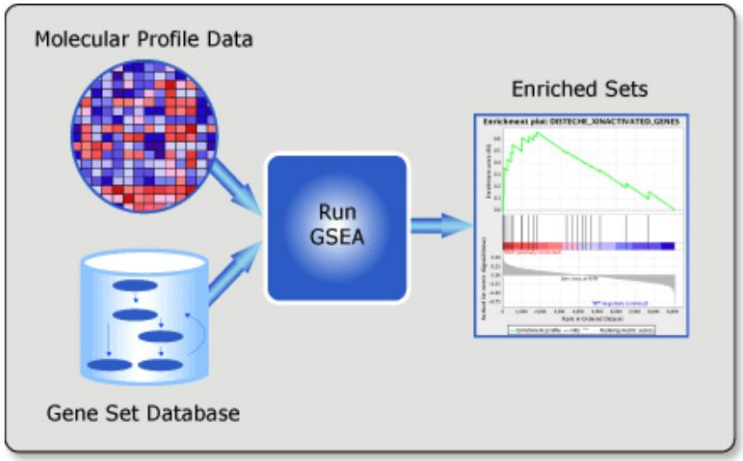
Hypothesis Testing/ Literature Curation

Have a hypothesis already? -- **Test it.**

Read papers



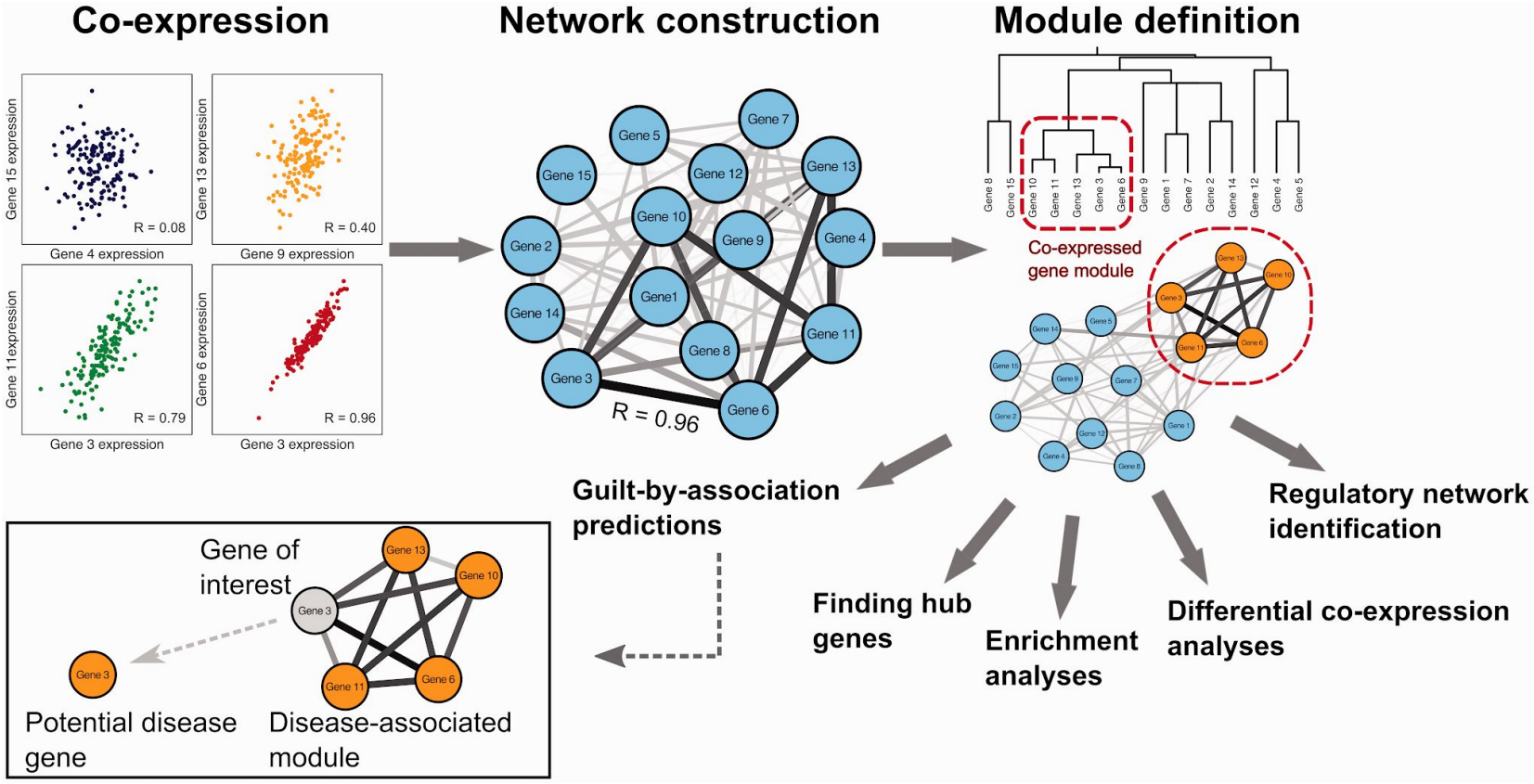
Functional Enrichment Analysis



www.gsea-msigdb.org

Downstream Analysis

What else can you do with RNA-seq data?



van Dam, S., *et. al.* (2018) Briefings in Bioinformatics(PubMed: PMID:28077403)

Module Summary

I RNA-Seq Overview

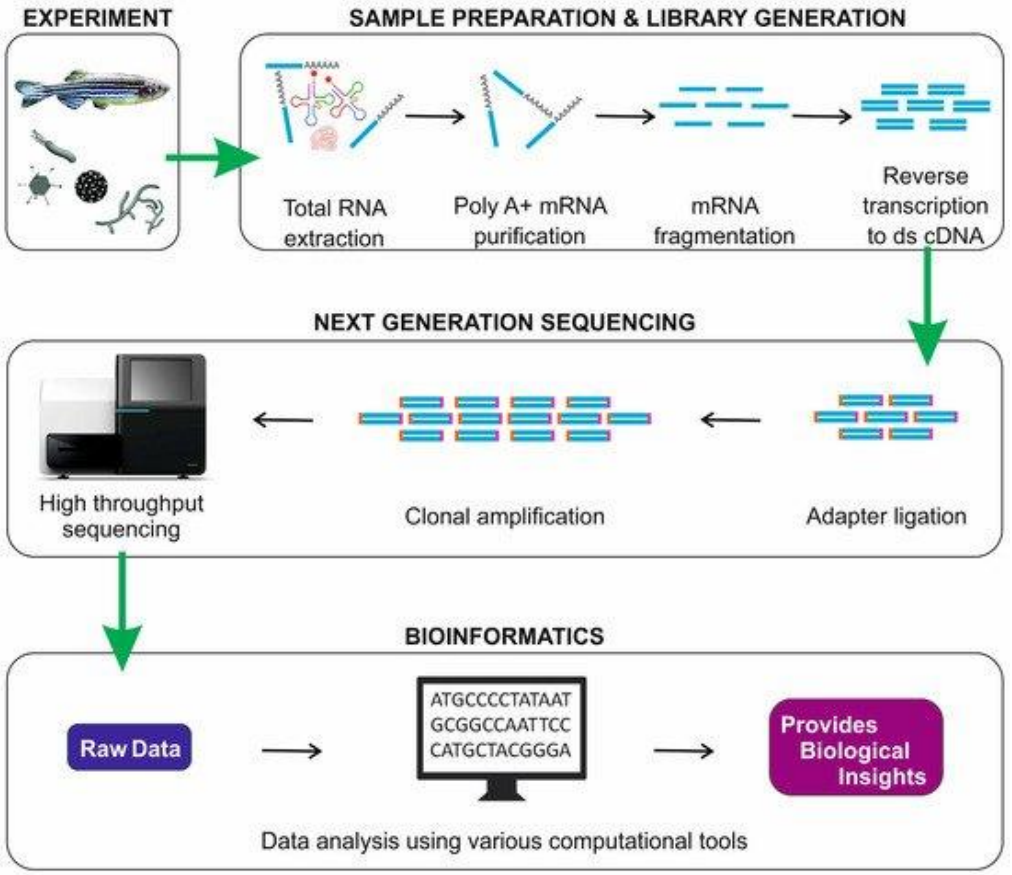
- Protocol Overview
- Experimental Design Considerations

II RNA-Seq Analysis Workflow

- Mapping
- Expression Quantification
- Read Count Normalisation
- Differential Expression Analysis

III Downstream Analysis

- Functional Enrichment Analysis



Sudhagar, Arun et al. (2018) Int. J. Mol Sci. (Pubmed: PMID: 29342931)

Acknowledgements

Slides originally developed by:

- Victoria Offord (Hinxton 2019)

Updated and expanded (2021):

- Nyasha Chambwe
- Jon Ambler
- Phelelani Mpangase