Getting started with RNA-Seq: Transforming raw reads into biological insights

6 September 2023

biocommons.org.au AustralianBioCommons @AusBiocommons

Australian BioCommons

Australian BioCommons

Actively supporting Australian life sciences research through bioinformatics and bioscience data infrastructure

biocommons.org.au AustralianBioCommonsChannel @AusBiocommons

Acknowledgement of Country

We acknowledge the Traditional Owners and their custodianship of the lands on which we meet today.

We pay our respects to their Ancestors and their descendants, who continue cultural and spiritual connections to Country.

We recognise their valuable contributions to Australian and global society.

Getting started with RNA-Seq: Transforming raw reads into biological insights

Dr Nandan Deshpande

Senior Research Bioinformatician

Sydney Informatics Hub

Australian BioCommons

RNA-Seq and me

- Experience with “omics” technologies when working at:
  - Previous position as a research associate (UNSW)
  - My current professional role at Sydney Informatics Hub (USyd)

- Application of RNA-Seq to answer various biological questions:
  - In model organisms e.g Human and Mouse

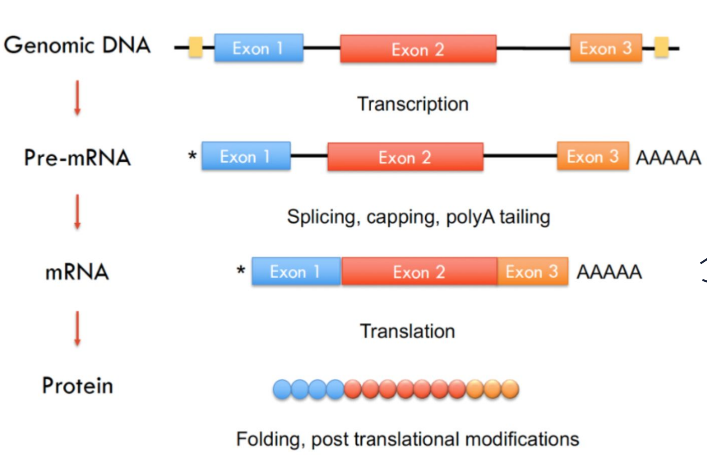
    infectious diseases, heart disease, cancer-related studies
  - In non-model organisms e.g. bacteria, fungi, fruit fly, sea urchins

- Working with different experimental designs
  - Simple “control versus treatment” designs
  - Time series designs
  - Expression profiling across independent experimental cohorts

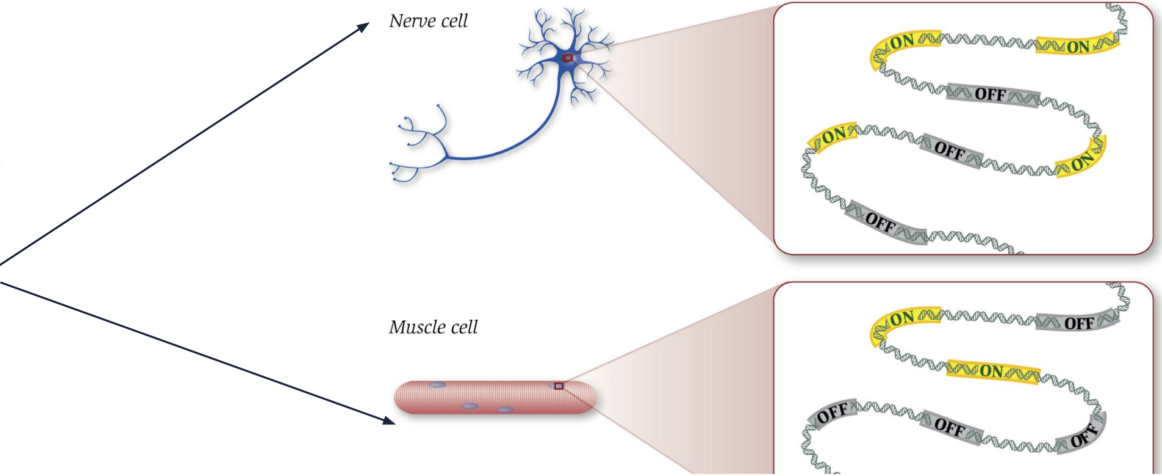
Outline

- What is: bulk RNA-Seq and when can we use this technique?
- How to: prepare a RNA-Seq sample, sequence it?
- How can we: perform data analysis?
- What are: the general computational requirements for data-analysis?
- Can we: use pre-made RNA-Seq workflows to make data analysis simpler?
- Are there: any additional practical considerations?

What is RNA?



Central dogma of molecular biology



Same gene can express differently across cell-types

RNA-seq is an experimental technique to quantify gene expression

Can sequencing RNA answer your question?

- What is your biological question? Can it be answered using RNA-Seq?
- Are there specific types of genes which you expect to be differentially expressed?

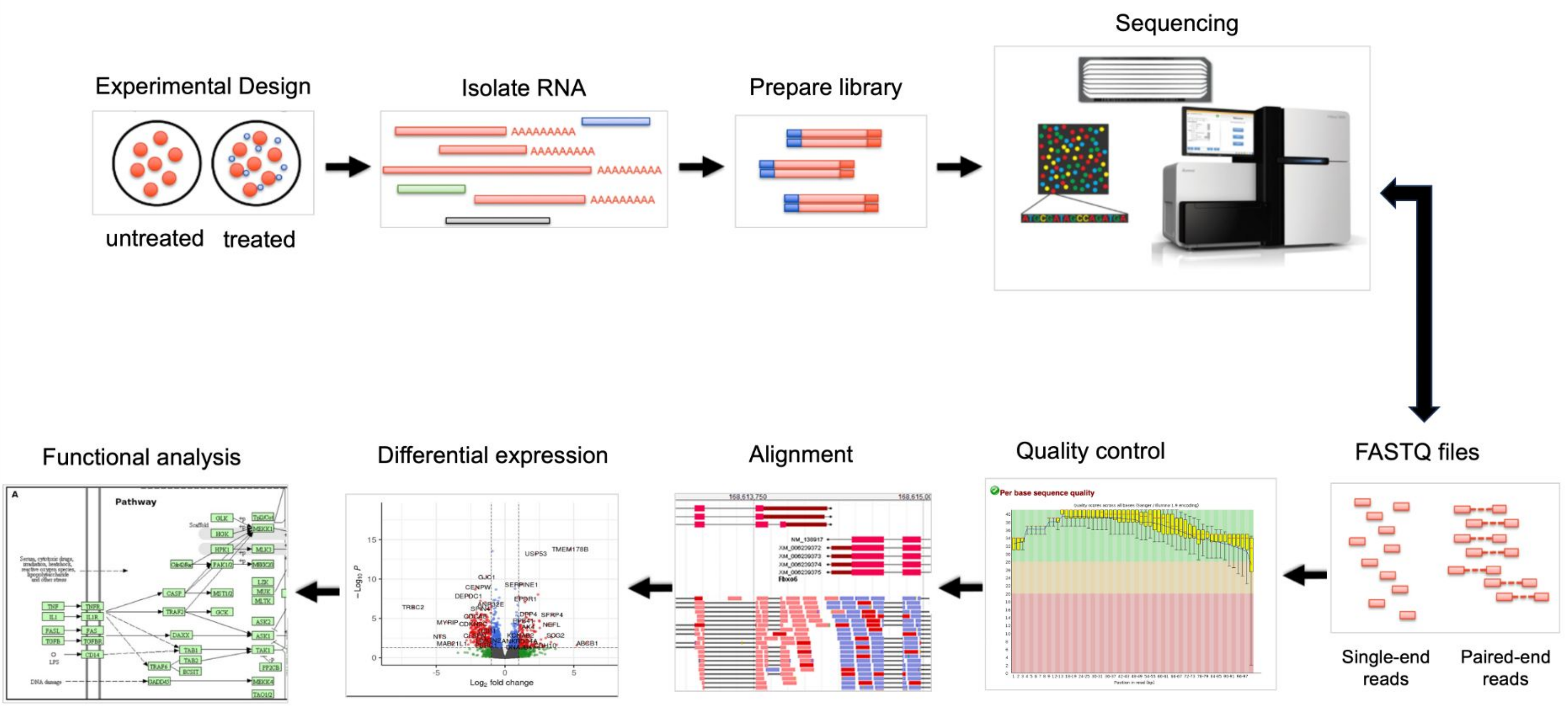
Some scenarios

1) For a particular cancer, some patients respond well to chemotherapy while the other do not. Is there an underlying gene expression signal which can answer this?
2) Patients have responded to a Covid strain differently. Are specific genes responsible for the severe/milder symptoms?

What can you do with RNA-Seq data?

- Differential Expression (DE) analysis.
- Reconstruction of novel transcripts and their annotation.
- Assembling and annotating transcriptome (non-model?) organism.
- Identify variants (SNPs) and allele specific expression.
- Identify alternative splice forms.

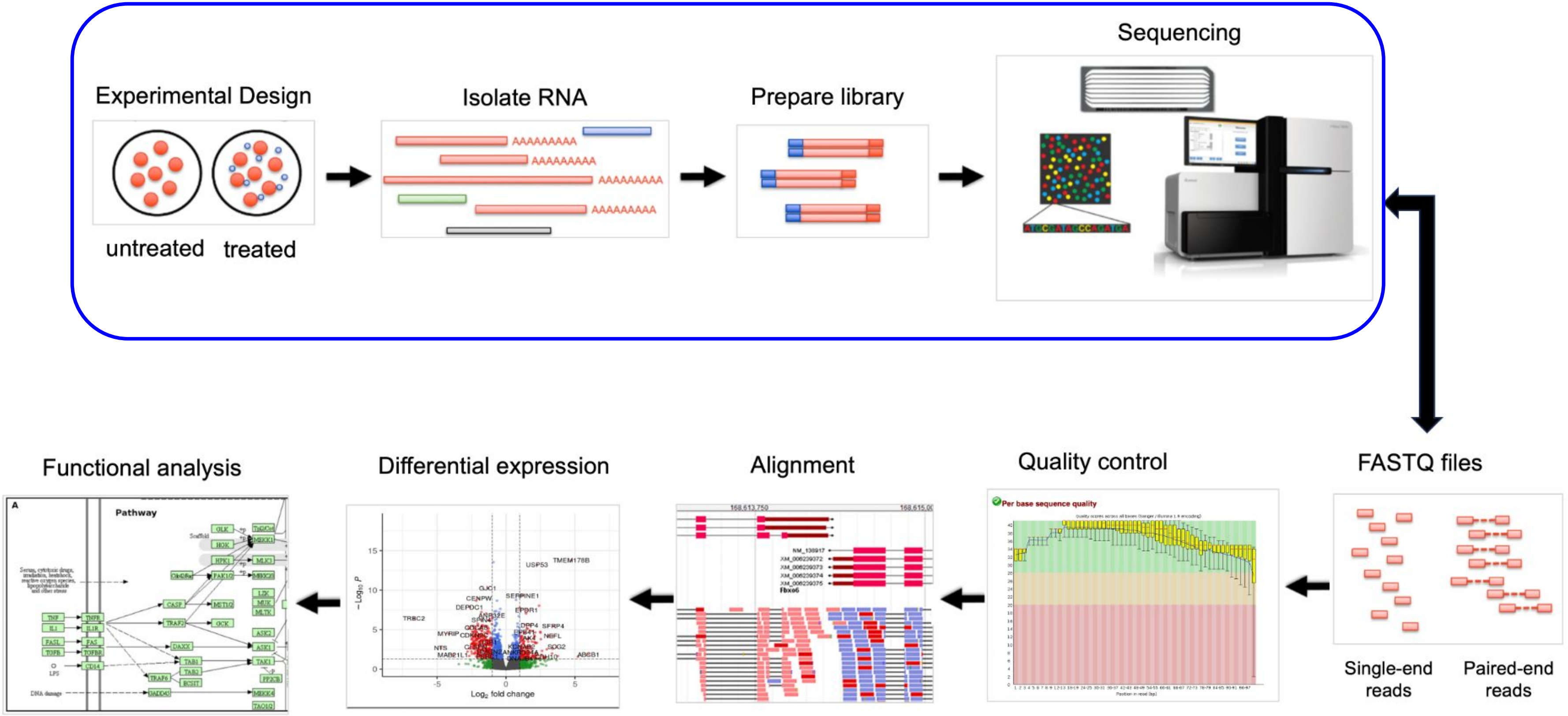
Differential expression analysis workflow



Experimental design to sequencing

Australian BioCommons

RNA-Seq: Experimental design - Sequencing



Experimental Design

untreated treated

Isolate RNA

Prepare library

Sequencing

FASTQ files

Single-end reads

Paired-end reads

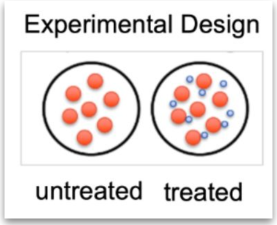
Quality control

Alignment

Differential expression

Functional analysis

Designing the experiment is important



RNA-Seq experiment produces high dimensional data

- Huge number of observations for a small number of samples.

A well designed RNA-seq experiment will:

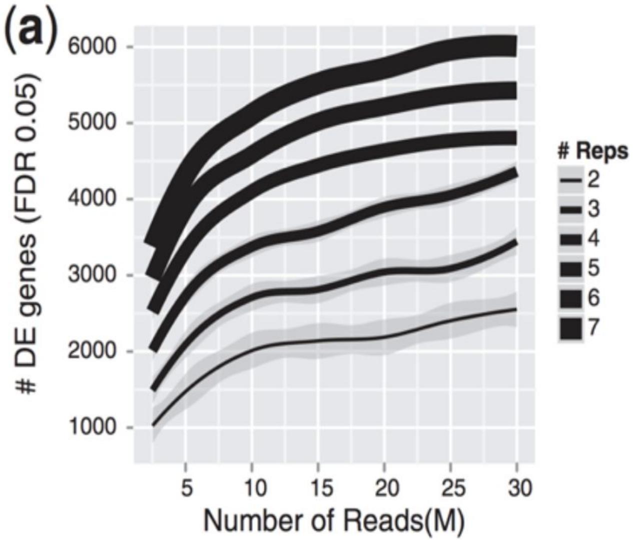
- Have enough replicates for your experiments
- Have enough sample quantity for a good sequencing depth
- Work with a bioinformatician and/or statistician in the experimental design stage

Above considerations are essential for a robust statistical analysis

Sequencing depth OR more replicates

For differential expression analysis:

- ~ 20-30 million reads/sample
- Increasing replicates helps more than increasing beyond minimum sequencing depth



Power calculators for RNA-Seq:

Estimate the sample size, based on -

- Variance across the samples
- Size and complexity of the transcriptome

Liu, Y., et al., Bioinformatics (2014) ***30****(3): 301–304*

Australian BioCommons

Control 'variation' across samples

Biological variance - Not everything in your sample is your signal of interest

Technical variance - How to control/minimise it?

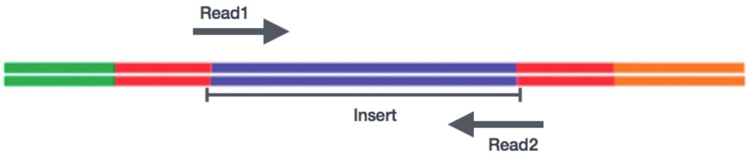
- Choose organisms from the same litter
- If sex is not a covariate, choose organisms of the same sex when possible
- Choose same sample-collection time for all samples
- Use the same laboratory technician to perform all library preps
- Randomise samples, if all samples cannot be processed at one time

Single-end or paired-end reads

Read is the basic segment of genetic information which is sequenced

Single end dataset (SE) : Only Read1

- Only one end of the RNA fragment is sequenced
- SE sequencing is more cost-effective



Paired-end dataset (PE) : Read1 + Read2

- Both ends of the RNA fragment are sequenced
- Complex transcript structures such as alternative splicing and fusion genes can demand PE data
- PE data can help in more certainty in detection of lowly expressed genes

Generally single-end sequencing is sufficient for RNA-Seq DE analysis

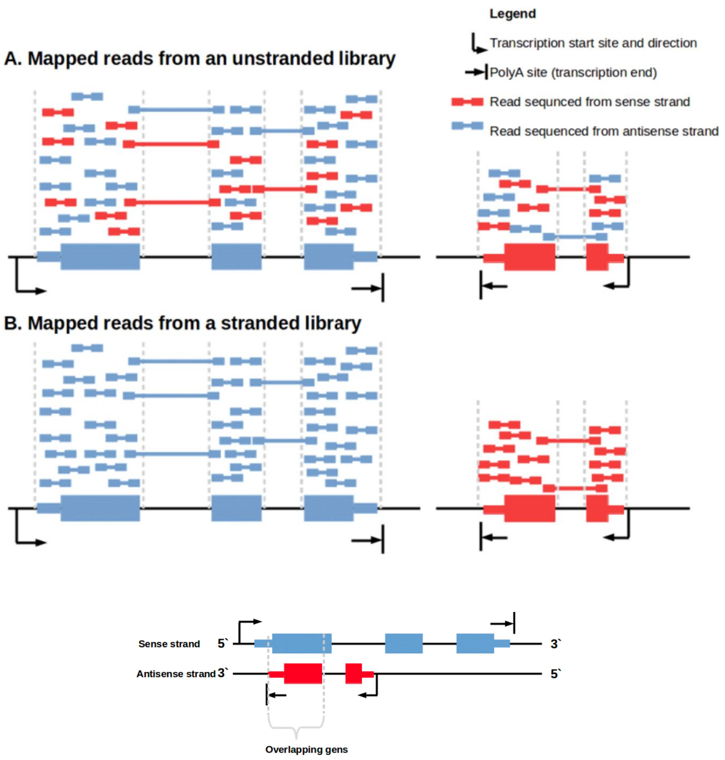
Un-stranded vs stranded protocols

Un-stranded protocol
Do not know which strand (sense/antisense) of the cDNA corresponds to the original mRNA

Stranded protocol
Involves attaching different adapters in a known orientation relative to the 5’ and 3’ ends of the RNA transcript

Advantages

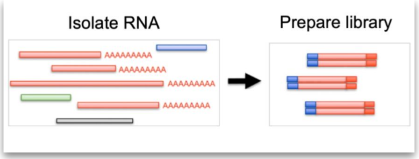
- Detecting overlapping transcripts
- Identification of non-coding RNAs
- Reduces mapping errors in general
- Assists in alternative splicing analysis



Legend
Transcription start site and direction
PolyA site (transcription end)
Read sequnced from sense strand
Read sequenced from antisense strand
A. Mapped reads from an unstranded library
B. Mapped reads from a stranded library
Sense strand
5`
3`
Antisense strand
3`
5`
Overlapping gens

Image credit: https://www.ecseq.com/support/ngs/how-do-strand-specific-sequencing-protocols-work

Methods for RNA library preparation



Parameter	PolyA Selection	Total RNA Sequencing	Ribosomal Depletion
Research Focus	Protein-coding genes, long non-coding RNAs	Non-coding RNAs, splice variants, isoforms	Comprehensive transcriptome analysis
Enrichment Focus	Mature, protein-coding transcripts	Comprehensive transcriptome (coding and non-coding)	Balanced view of coding and non-coding transcripts
Suitable for Non-Coding RNAs	Limited	Yes	Yes
rRNA Contamination	Reduced	Higher	Reduced (more effective)
Suitable for Lowly Expressed Transcripts	Moderate	Yes	Yes
Insights into RNA Processing/ Degradation	Limited	Yes	Limited
Biological Insights	Gene expression profiling	Comprehensive transcriptome analysis, RNA dynamics	Balanced transcriptome representation

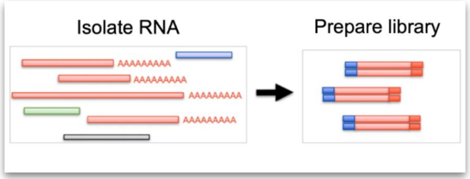
To get more out of your experiment, select the most appropriate method

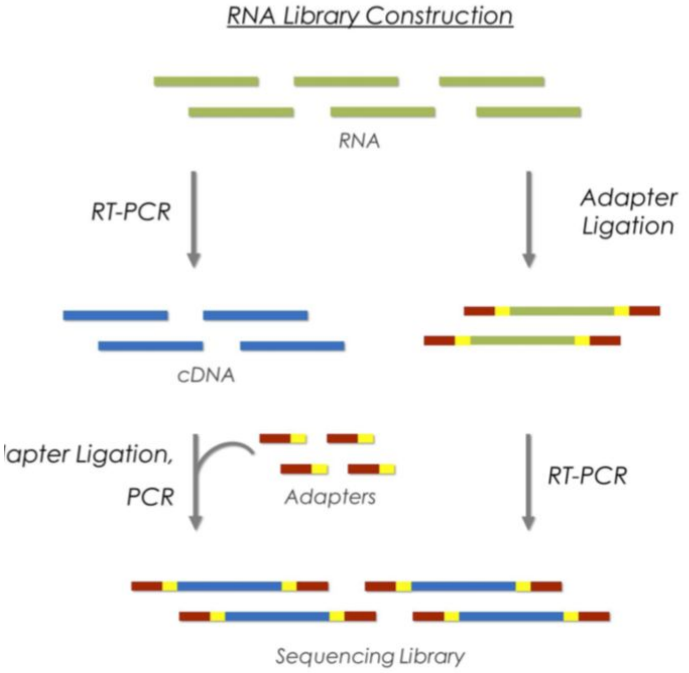
RNA sequencing: NGS library construction

The aim is to convert the RNA molecule into a NGS sequencer compatible format

There are multiple steps which include:

- RNA extraction
- Quality control
- RNA fragmentation
- cDNA synthesis
- Second strand synthesis
- Adapter ligation
- Size selection
- Library amplification
- Loading into sequencer





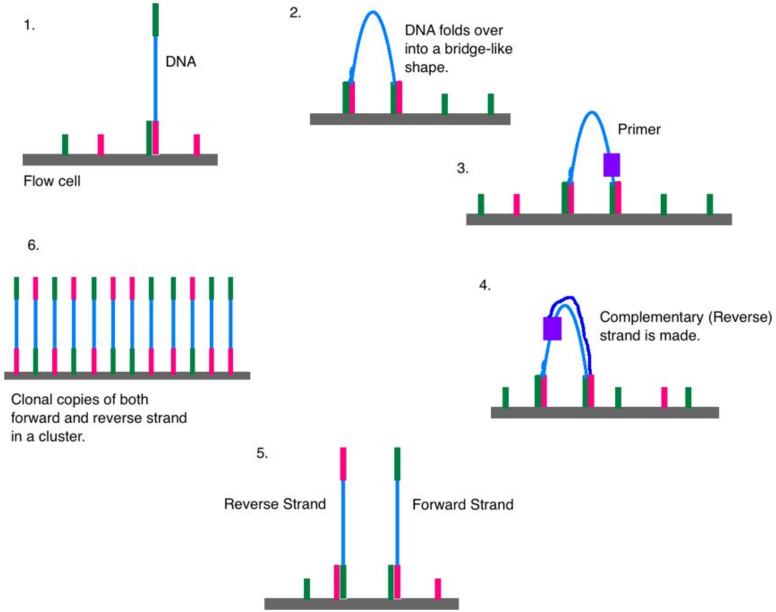
https://www.biocompare.com/Molecular-Biology/9187-Next-Generation-Sequencing/

Short-read sequencing

"Sequencing by synthesis" - get a digital copy of your RNA fragment

The main steps involved are -

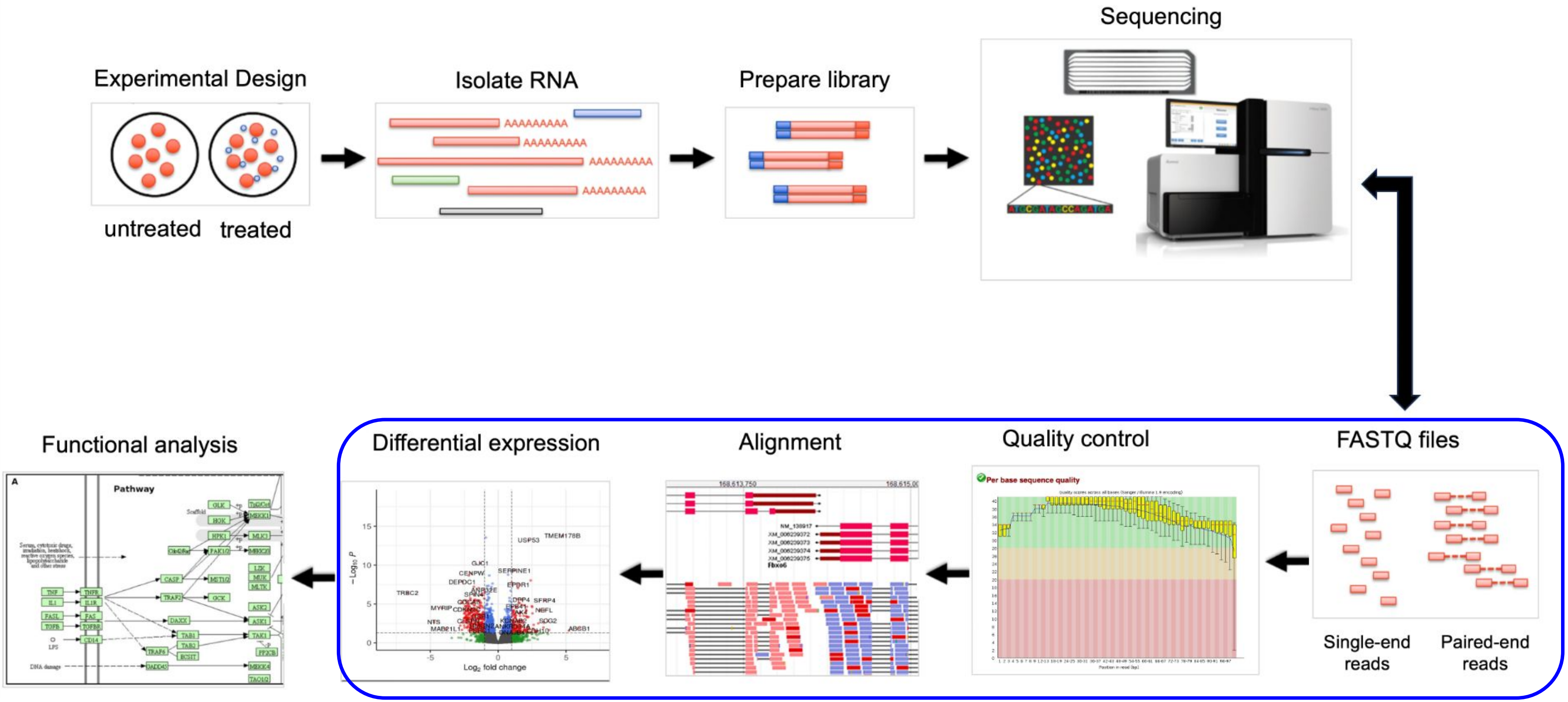
- Cluster generation
- Primer hybridisation
- Nucleotide incorporation
- Image capture
- Base calling
- *Data analysis*


1.
DNA
Flow cell
2.
DNA folds over into a bridge-like shape.
3.
Primer
4.
Complementary (Reverse) strand is made.
5.
Reverse Strand
Forward Strand
6.
Clonal copies of both forward and reverse strand in a cluster.

Credit: https://en.wikipedia.org/wiki/Illumina_dye_sequencing

Raw sequence to differential expression

Convert the raw sequence to gene lists



Experimental Design

untreated treated

Isolate RNA

Prepare library

Sequencing

FASTQ files

Single-end reads

Paired-end reads

Quality control

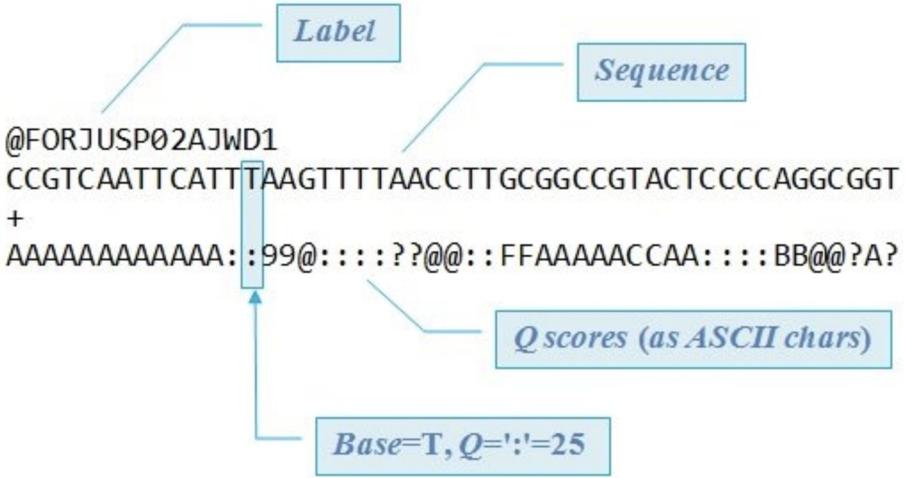
Alignment

Differential expression

Functional analysis

What is the FASTQ format ?

- A FASTQ read with 50nt in Illumina format.
- Always four lines per read:
  - **Line 1:** starts with ‘@’, followed by the label.
  - **Line 2:** contains the actual sequence.
  - **Line 3:** starts with ‘+’. In some files, the ‘+’ line contains a second copy of the label.
  - **Line 4:** contains quality (Q) scores represented as ASCII characters.



```
@FORJUSP02AJWD1
CCGTCAATTCATTTAAGTTTTAACCTTGCGGCCGTACTCCCCAGGCGGT
+
AAAAAAAAAAAA::99@::::??@@::FFAAAAACCAA::::BB@@?A?
```

- The quality score is an integer (Q) which is typically in the range 2 - 40

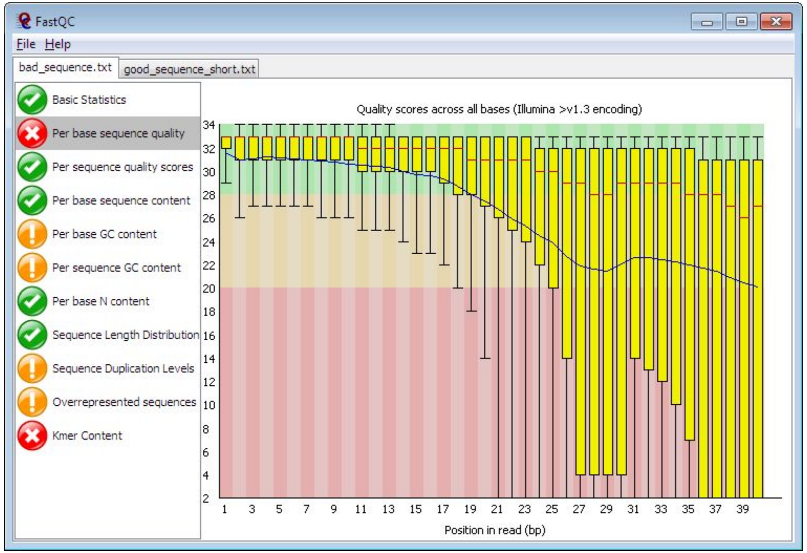
Quality assessment of raw data

Inspect the raw data

- Check the sequence format to ensure there is no data corruption
- Inspect the read-length distribution

Tools for quality assessment

- FastQC
- RSeqQC
- QualiMap


FastQC
File Help
bad_sequence.txt
good_sequence_short.txt
✓ Basic Statistics
✗ Per base sequence quality
✓ Per sequence quality scores
✓ Per base sequence content
! Per base GC content
! Per sequence GC content
✓ Per base N content
✓ Sequence Length Distribution
! Sequence Duplication Levels
! Overrepresented sequences
✗ Kmer Content
Quality scores across all bases (Illumina >v1.3 encoding)
34 32 30 28 26 24 22 20 18 16 14 12 10 8 6 4 2
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39
Position in read (bp)

Tool: FastQC

Quality assessment: other important parameters

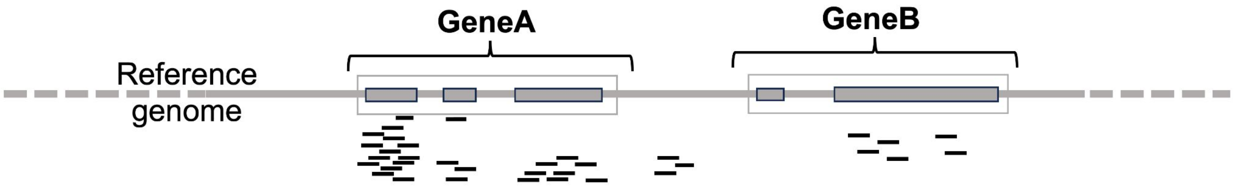
- Identify low quality reads based on average quality scores.
- Identify adapters used in library preparation which can be still attached to the reads
- Check for known contaminants such as bacterial sequences
- Check for duplicate reads resulting from PCR amplification during the library preparation step
- Check GC content distribution and identify any anomalies

Checked the quality, now trim the reads

- Remove low quality bases from ends of the read
- Discard low-quality reads with average quality scores below threshold
- Remove the adapter sequences
- Remove the contaminant sequences
- Remove the duplicate reads

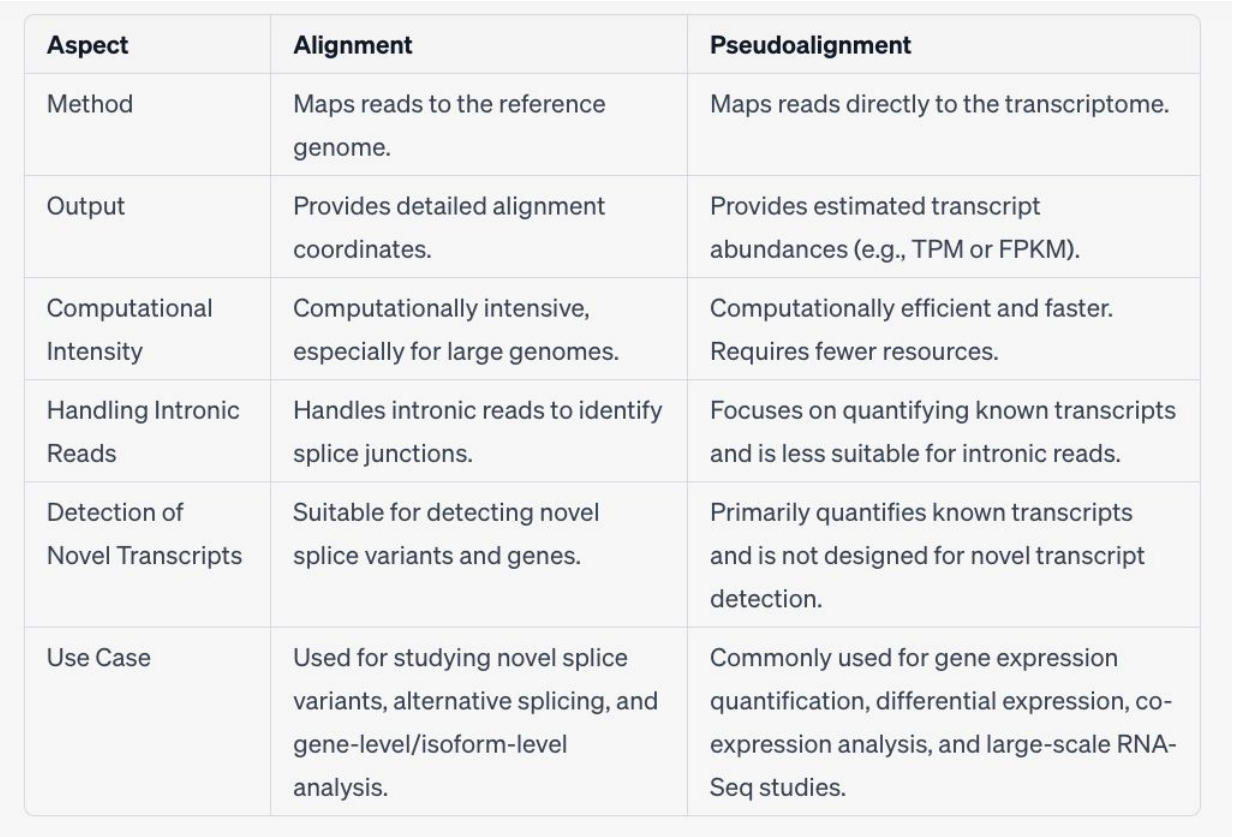
Read alignment

- The reference genome contains the most up-to-date DNA sequence of the species of interest
- Reads are mapped to the reference genome using alignment tools such as STAR and HISAT2



RNA-seq specific aligner needs to be used to accommodate splicing

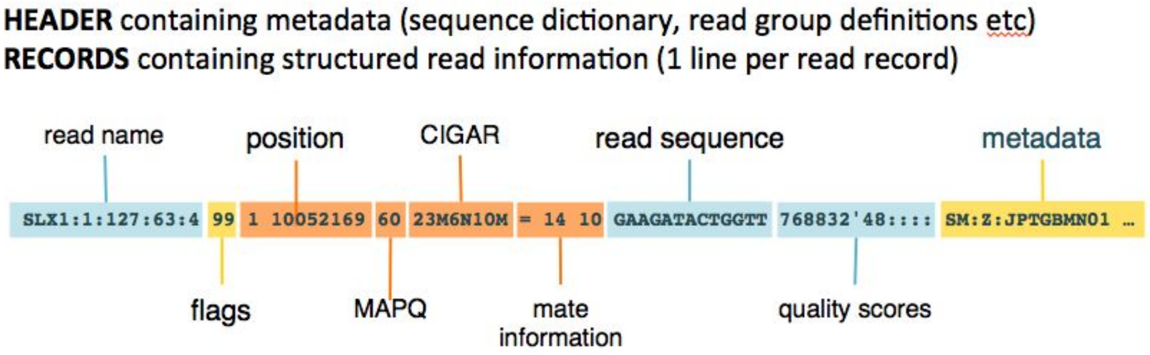
Read alignment methods



Aspect	Alignment	Pseudoalignment
Method	Maps reads to the reference genome.	Maps reads directly to the transcriptome.
Output	Provides detailed alignment coordinates.	Provides estimated transcript abundances (e.g., TPM or FPKM).
Computational Intensity	Computationally intensive, especially for large genomes.	Computationally efficient and faster. Requires fewer resources.
Handling Intronic Reads	Handles intronic reads to identify splice junctions.	Focuses on quantifying known transcripts and is less suitable for intronic reads.
Detection of Novel Transcripts	Suitable for detecting novel splice variants and genes.	Primarily quantifies known transcripts and is not designed for novel transcript detection.
Use Case	Used for studying novel splice variants, alternative splicing, and gene-level/isoform-level analysis.	Commonly used for gene expression quantification, differential expression, co-expression analysis, and large-scale RNA-Seq studies.

SAM (Sequence Alignment/Map) file format

Output of the alignment step is a BAM file - Compressed form of a SAM file



HEADER containing metadata (sequence dictionary, read group definitions etc)
RECORDS containing structured read information (1 line per read record)

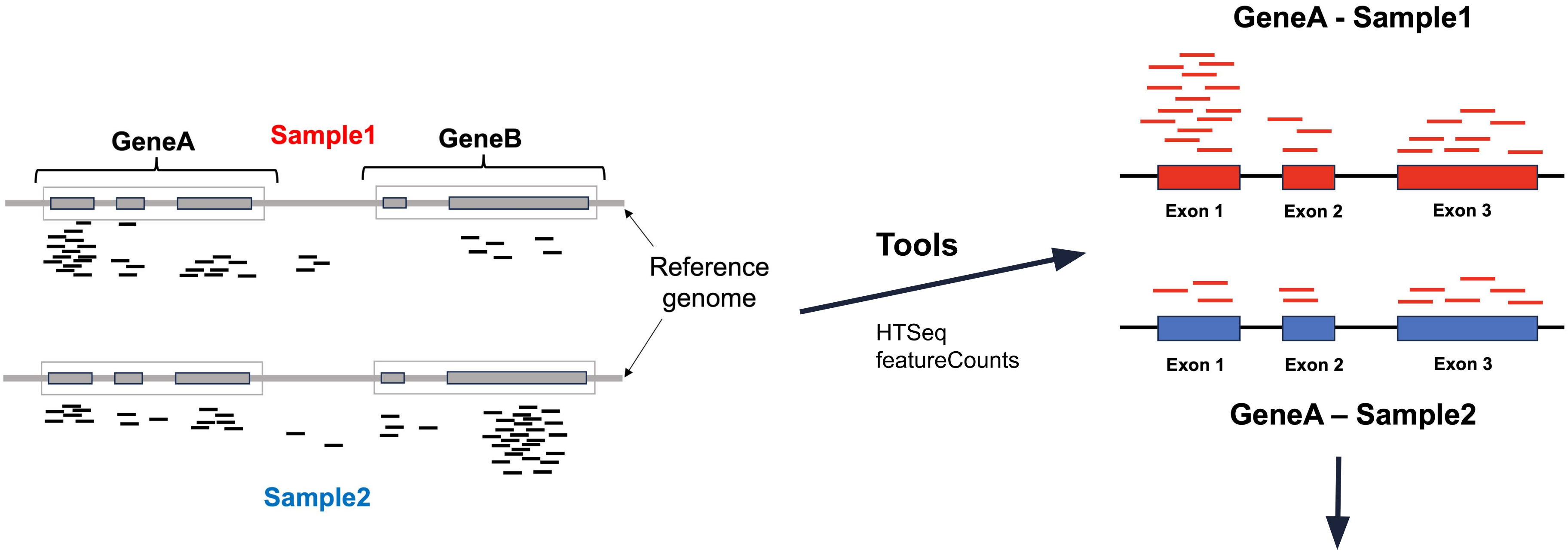
```
SLX1:1:127:63:4 99 1 10052169 60 23M6N10M = 14 10 GAAGATACTGGTT 768832'48:::: SM:Z:JPTGBMN01 …
```

The SAM format contains:

- Read name, read sequence, alignment position in the reference genome
- Mapping quality
- Additional information about the alignment

Image from GATK

Associate aligned-reads to genes



Tools

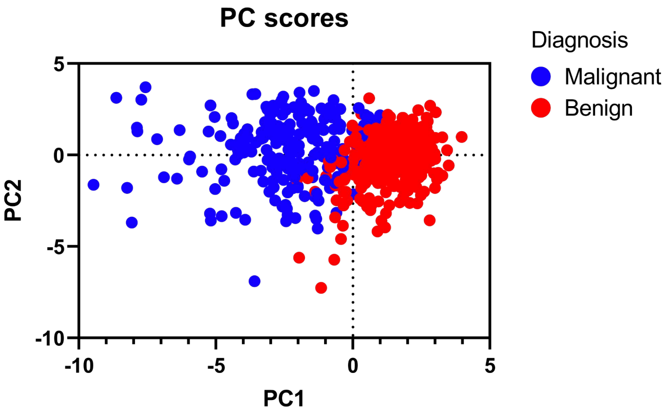
HTSeq
featureCounts

Gene-count matrix

Gene	Sample1	Sample2
GeneA	25	11
GeneB	5	43

Differentially expression analysis: Pre-processing

- **Quality control**: Filter the lowly expressed genes
- **Principal component analysis (PCA)**
  - Reduce the high-dimensionality of RNA-Seq data to 2-dimensions
  - To assess overall similarity across samples
  - To assess batch effects
  - To identify outlier samples



Batch effects

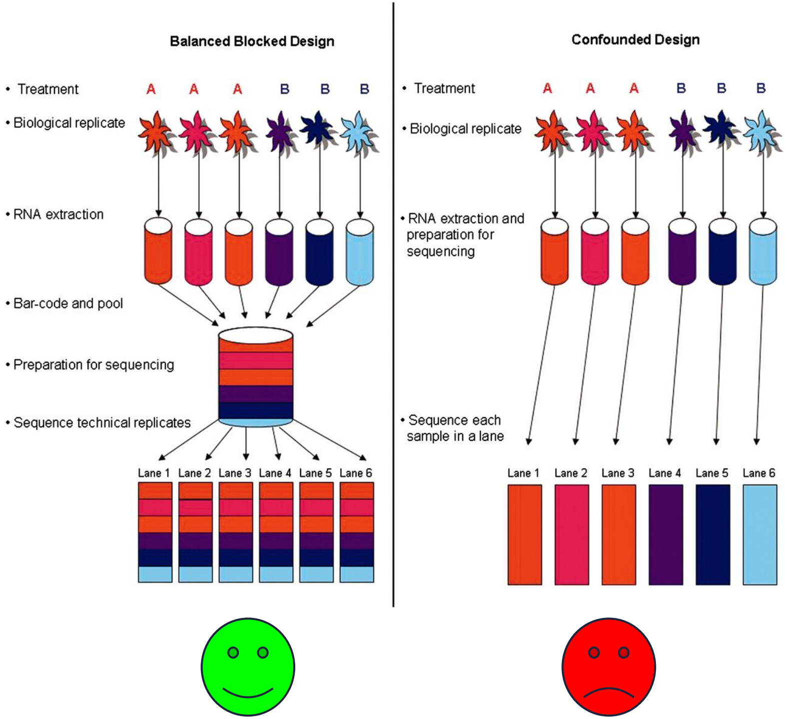
What are batch effects?

- Unbalanced experimental design
- Samples processed at different facilities

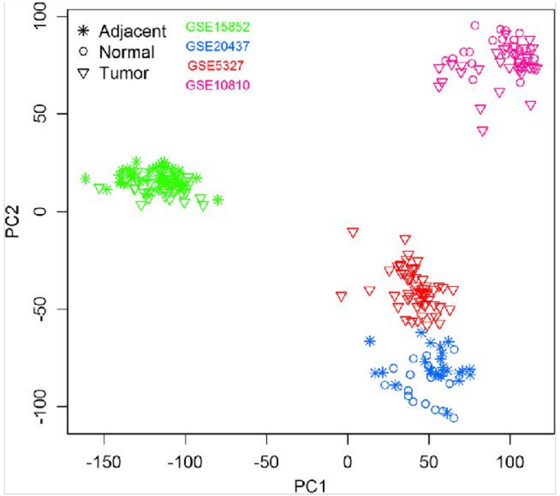
How to overcome batch effects

- **Before sequencing:** Randomise the experimental design
- **After sequencing:** Use informatics approaches

Randomisation in sequencing runs



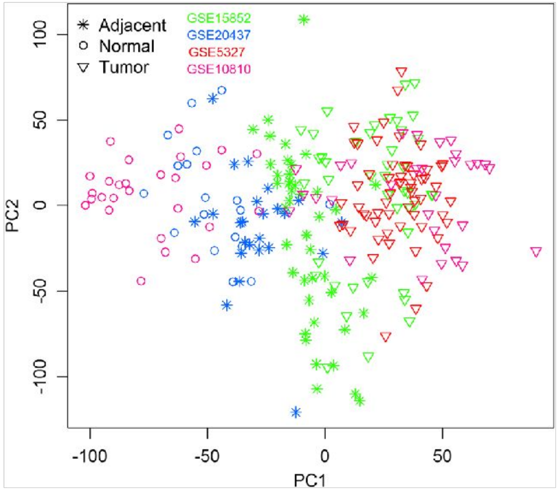
PCA plots before/after batch-effect removal



With batch-effects

Tools

- RUVSeq
- ComBat
- sva



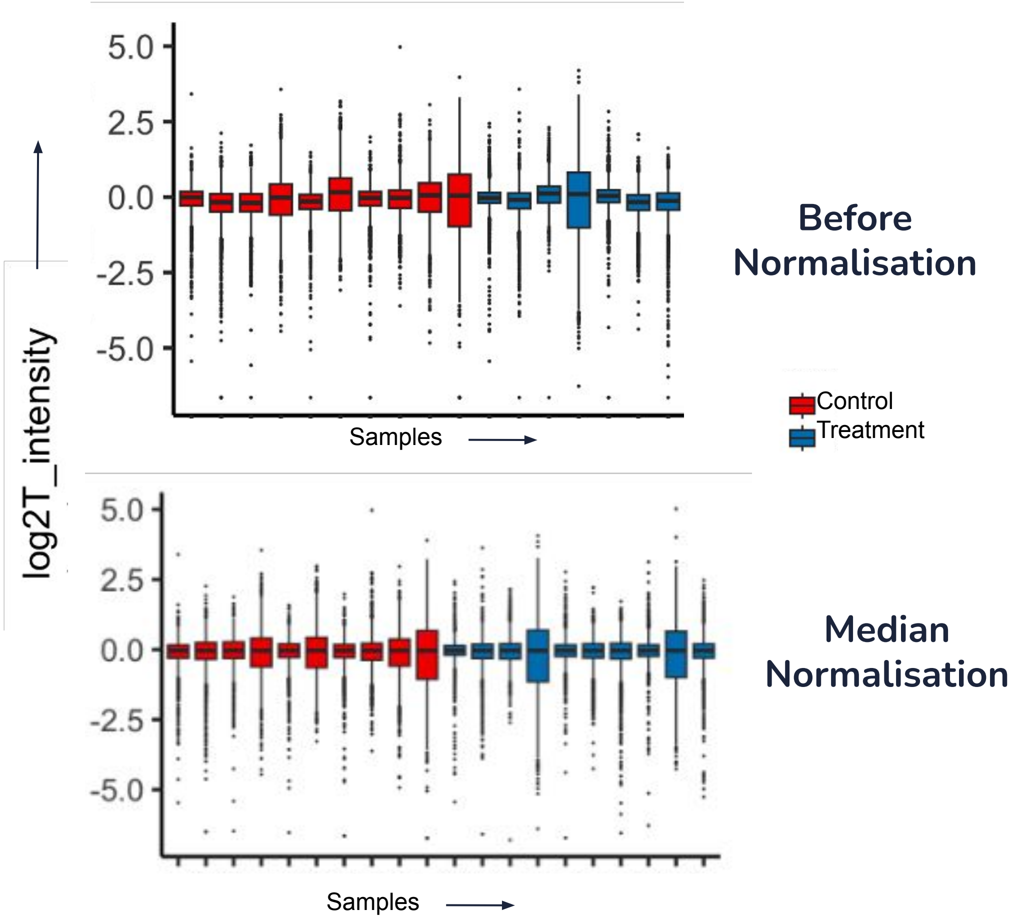
After removal of batch-effect

Image credits:
https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-023-05376-z

Normalisation

A critical step in data processing

- To correct for variation in sequencing depth
- Make the expression levels of genes comparable across samples
- Multiple methods
  - Median normalisation
  - Quantile normalisation



Identifying differentially expressed (DE) genes

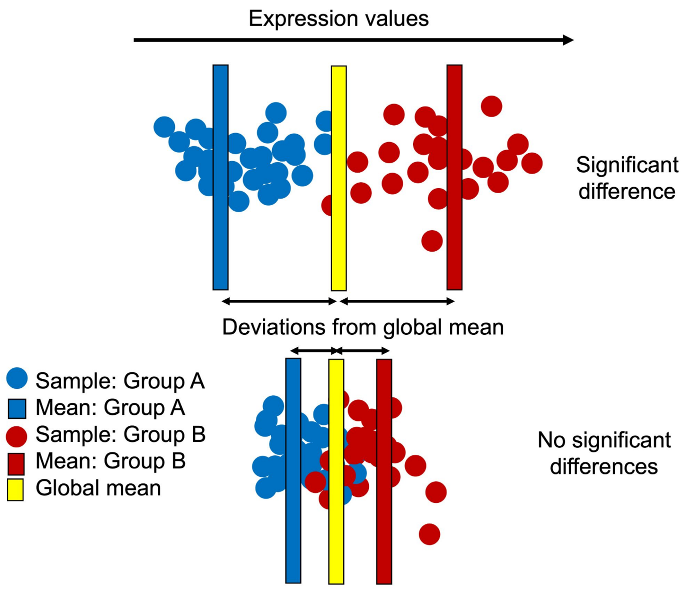
- Compare the expressions of a gene in group A versus that in group B

  Fold change: Difference is average expression of a gene across groups

- Use multiple samples (as replicates) per condition to identify statistically significant results

  P-values: Indicates the likeliness of a gene to be differentially expressed

  Adjusted p-values: Adjusting for multiple statistical tests (one per gene)



Results

Tools: DESeq2, edgeR or limma

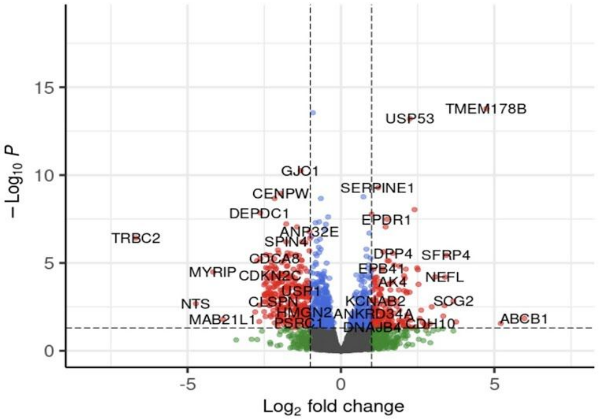
- **List of differentially expressed genes**

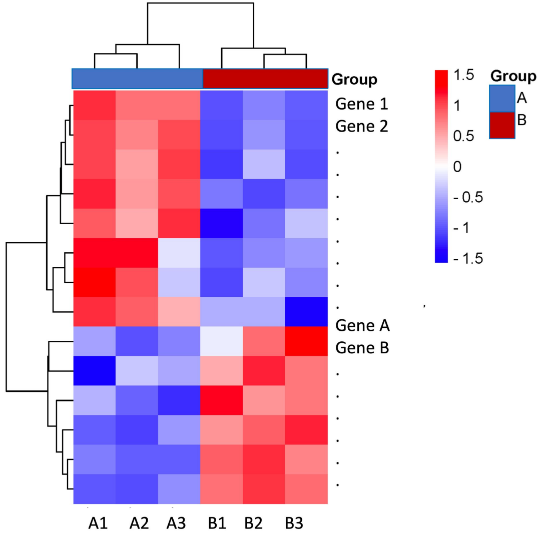
- **Volcano plot**

A scatter plot with fold-changes on the x-axis and their statistical significance on the y-axis.

- **Heatmap**

A Matrix Representation of gene expression data where rows represent genes and columns represent samples



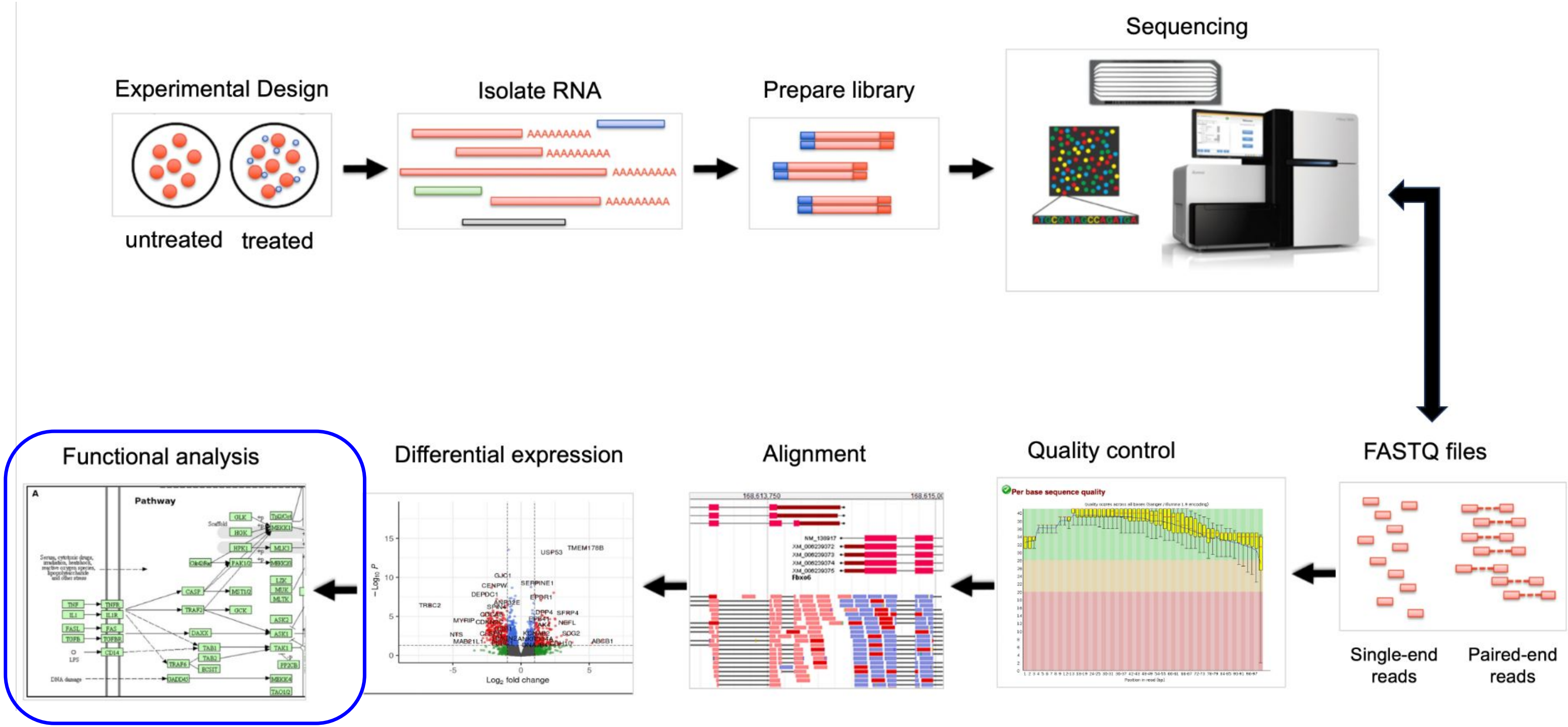


Australian BioCommons

Differentially expressed genes to function

Australian BioCommons

Associate gene lists to function

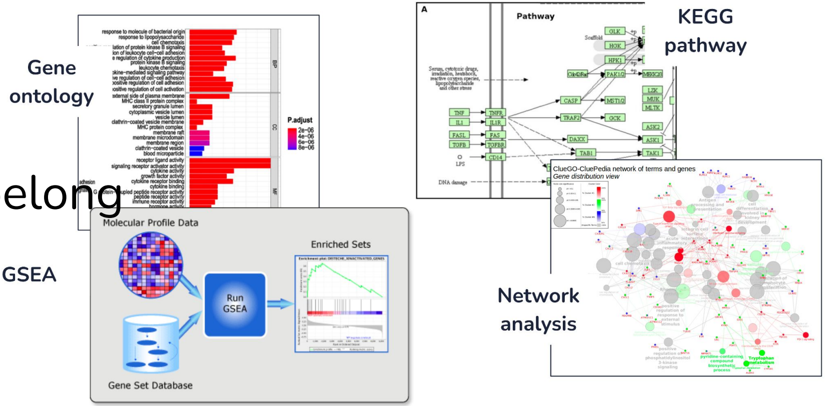


Functional analysis

Aim: Understand the biological context of the differentially expressed (DE) genes.

Multiple methods: Use pre-defined gene sets - Gene-ontologies, KEGG or Reactome pathways

Overrepresentation analysis (ORA):
Input: A list of genes of interest (e.g. DE genes)
Method: Check if a significant number of genes belong to a particular genes set



Gene set enrichment analysis (GSEA):
Input: All genes in an experiment ranked by a metric e.g. Fold change or p-values
Method: Check if genes from a particular gene set are found disproportionately at the top or bottom of the ranked list

Multiple web-based tools and R-packages are available

Hypothesis generation and biological validation

Based on enrichment analysis, we can

- Generate a hypothesis
- Discover new connections
- Get hints at upstream regulators
- Suggest potential targets

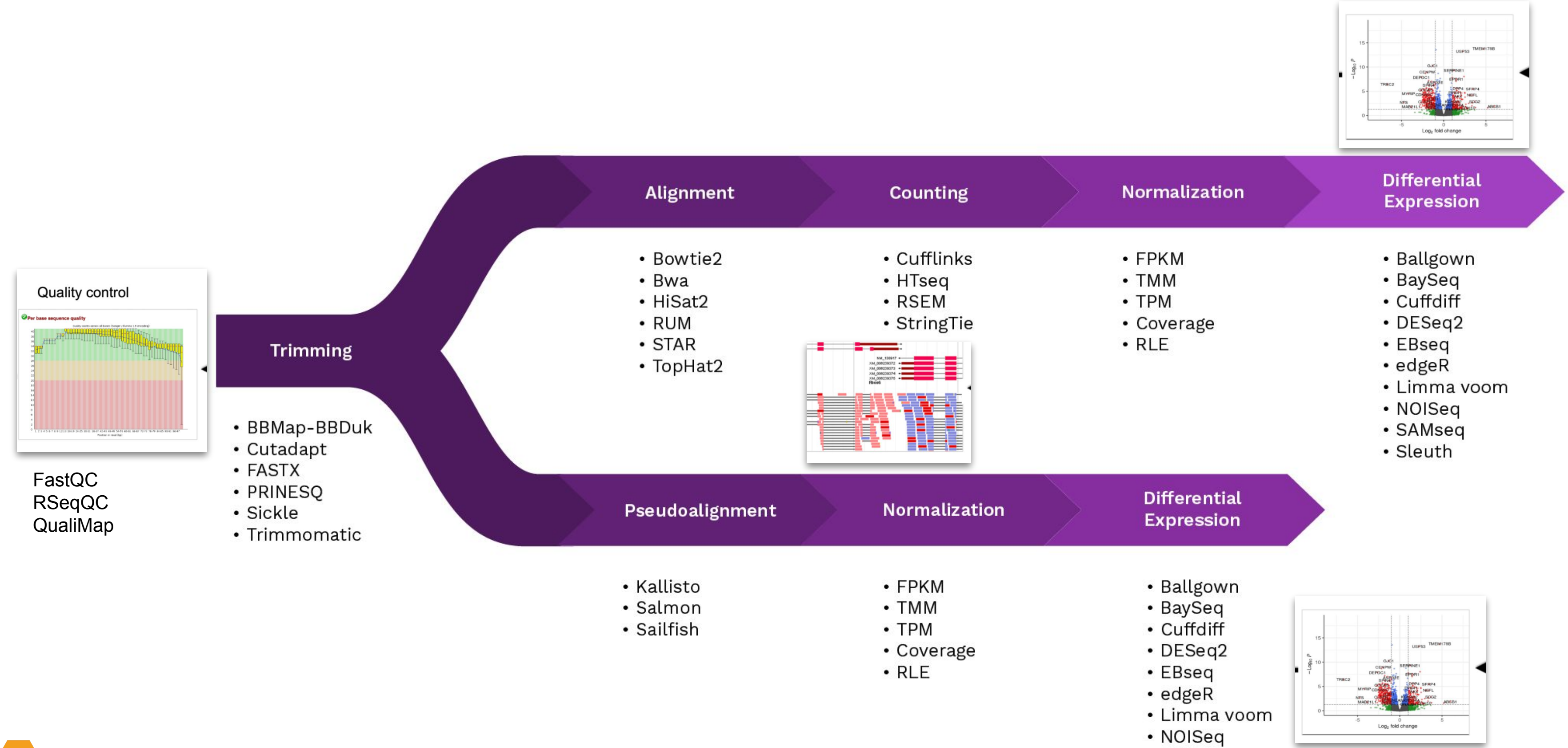
This can be followed by experimental validations

- Quantitative real-time PCR
- Western blotting
- Gene knockdown experiments
- Use clinical samples for experiments

Where to find resources

Australian BioCommons

Open-source tools for RNA-Seq



Quality control

FastQC
RSeqQC
QualiMap

Trimming

- BBMap-BBDuk
- Cutadapt
- FASTX
- PRINESQ
- Sickle
- Trimmomatic

Alignment

- Bowtie2
- Bwa
- HiSat2
- RUM
- STAR
- TopHat2

Counting

- Cufflinks
- HTseq
- RSEM
- StringTie

Normalization

- FPKM
- TMM
- TPM
- Coverage
- RLE

Differential Expression

- Ballgown
- BaySeq
- Cuffdiff
- DESeq2
- EBseq
- edgeR
- Limma voom
- NOISeq
- SAMseq
- Sleuth

Pseudoalignment

- Kallisto
- Salmon
- Sailfish

Normalization

- FPKM
- TMM
- TPM
- Coverage
- RLE

Differential Expression

- Ballgown
- BaySeq
- Cuffdiff
- DESeq2
- EBseq
- edgeR
- Limma voom
- NOISeq
- SAMseq
- Sleuth

Image credit:
https://www.elucidata.io/blog/bulk-rna-sequencing-a-comparison-of-the-most-popular-tools-and-pipelines

Understanding compute options

How you process your data depends on -

❑ **(Size of your data and size of the reference) + (Available resources)**

Local machine

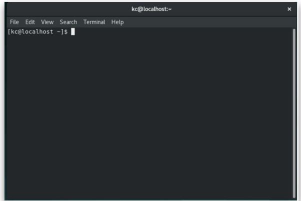
High performance computing

Cloud computing

❑ **Level of comfort with command-line**

Using open-source tools

Unix-based at command-line



```
[kc@localhost ~]$ 
```

Workflows with GUI

Galaxy Australia

Workflows at command line

nextflow

RStudio

jupyter

Using commercial tools

QIAGEN CLC Genomics Workbench

geneious by Dotmatics Bioinformatics platform

Galaxy for RNA-Seq analysis

Galaxy is a powerful and easy to use web-based platform for scientific data analysis.

- Provides a graphical web interface
- Contains customise/build pipelines using reusable modules
- No need of programming experience; can be used by bioinformaticians and novice users.
- Provides a large amount of high-quality, community-developed and maintained tools and training materials

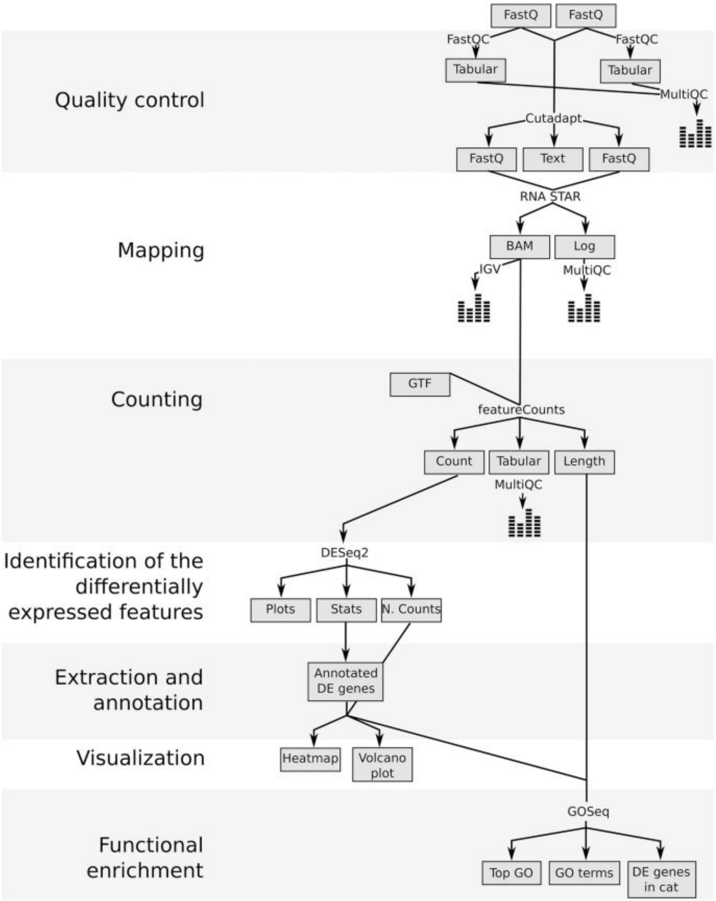
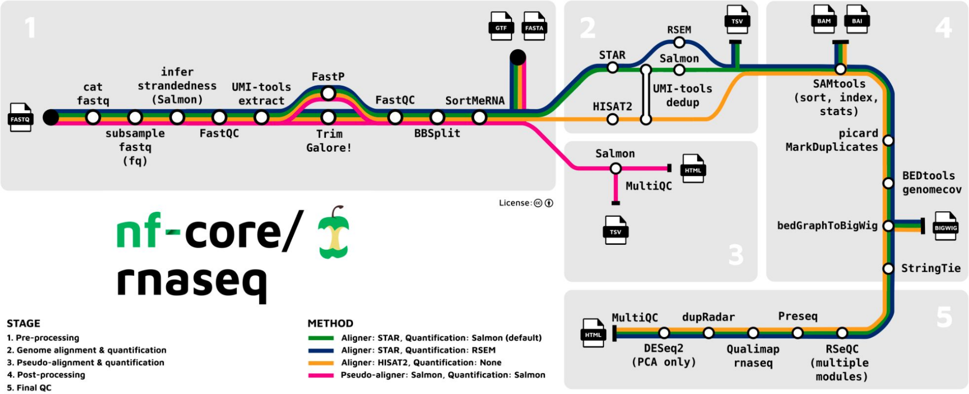


Image credit: https://link.springer.com/protocol/10.1007/978-1-0716-1307-8_20

Nextflow ‘nf-core/rnaseq’ workflow

Nextflow is a workflow management tool which enables ease in pipeline development at command-line.



STAGE
1. Pre-processing
2. Genome alignment & quantification
3. Pseudo-alignment & quantification
4. Post-processing
5. Final QC

METHOD
🟩 Aligner: STAR, Quantification: Salmon (default)
🟦 Aligner: STAR, Quantification: RSEM
🟧 Aligner: HISAT2, Quantification: None
🟥 Pseudo-aligner: Salmon, Quantification: Salmon

Nextflow

- Optimises resource usage
- Handles software installations
- Can run on multiple compute platforms
- Enable reproducibility and portability

nf-core is a community curated collection of bioinformatics pipelines in Nextflow
nf-core/rnaseq contains multiple tools to choose when deploying the pipeline.

Australian BioCommons

Where can you find resources and support?

Compute platforms

- National infrastructures
- Institutional HPCs
- Galaxy Australia
- RStudio

Training

- BioCommons events
- Galaxy training network
- Software Carpentry
- Local bioinformatics organisations

Online communities

- Biostars, seqanswers
- nf-core #rnaseq Slack channel
- GitHub

Ask a bioinformatician

- Bioinformatics core facilities
- Colleagues and collaborators
- Methods section of genomics papers

Public datasets

- Gene Expression Omnibus (GEO)
- European Nucleotide Archive (ENA)
- Sequence Read Archive (SRA)

Takeaways

- RNA-seq is an exciting experimental technique for gene expression profiling
- Designing a good RNA-Seq experiment is half battle won!
  - Choice of library prep is crucial
  - Try to keep the variance and batch-effects in your samples to the minimum
  - Attain a balance between sequencing depth and the number of replicates

- Quality assessment of sequenced data: “Crap IN Crap OUT” - trim your data
- The read-alignment step is the most resource intensive

- Identify the appropriate resources for data processing
  - Graphical user interface or command-line is a big choice
  - Which platform to work on? : Local, HPC, cloud
  - Using predefined workflows can save you a lot of time

Questions

NEXT …

WORKSHOP: RNASeq: reads to differential genes and pathways

11 & 12 October 2023

Applications close 25 September

https://www.biocommons.org.au/webinars-workshops

Keep in touch

@AusBiocommons

biocommons.org.au/subscribe

Australian BioCommons

NEXT …

WORKSHOP: WORKSHOP: Single cell RNAseq analysis in R

26 & 27 September 2023

Applications close 11 September

https://www.biocommons.org.au/webinars-workshops

Keep in touch

@AusBiocommons

biocommons.org.au/subscribe

Australian
BioCommons

Thanks for joining us!

The Australian BioCommons is enabled by NCRIS via Bioplatforms Australia funding

BIOPLATFORMS AUSTRALIA

NCRIS
National Research Infrastructure for Australia
An Australian Government Initiative

Australian BioCommons