CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Introduction to RNAseq Methods

June 2021

HTS Applications - Overview

DNA Sequencing

- Genome Assembly
- SNPs/SVs/CNVs
- DNA methylation
- DNA-protein interactions (ChIPseq)
- Chromatin Modification (ATAC-seq/ChIPseq)

RNA Sequencing

- Transcriptome Assembly
- **Differential Gene Expression**
- Fusion Genes
- Splice variants

Single-Cell

- RNA/DNA
- Low-level RNA/DNA detection
- Cell-type classification
- Dissection of heterogenous cell populations

RNAseq Workflow

Experimental Design

Library Preparation

Sequencing

Bioinformatics Analysis

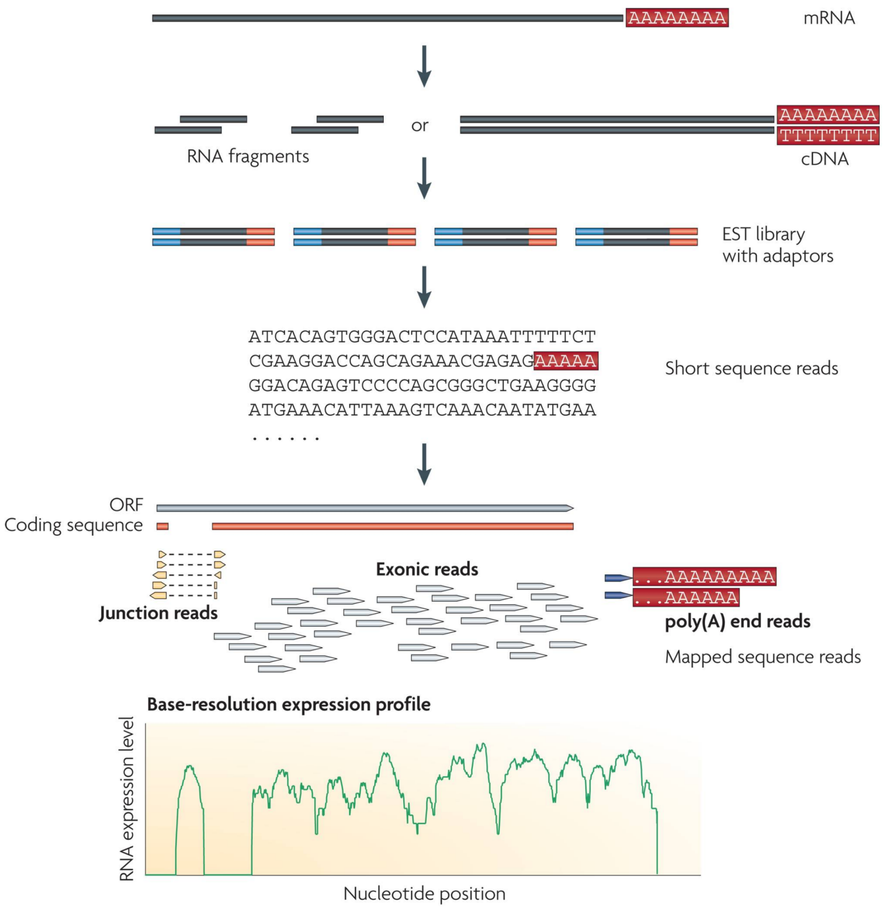


Image adapted from: Wang, Z., et al. (2009), Nature Reviews Genetics, 10, 57–63.

Designing the right experiment

A good experiment should:

- Have clear objectives
- Have sufficient power
- Be amenable to statisical analysis
- Be reproducible
- More on experimental design later

Designing the right experiment

Practical considerations for RNAseq

- Coverage: how many reads?
- Read length & structure: Long or short reads? Paired or Single end?
- Controlling for batch effects
- Library preparation method: Poly-A, Ribominus, other?

Designing the right experiment - How many reads do we need?

The coverage is defined as:

$$\frac{Read\ Length \times Number\ of\ Reads}{Length\ of\ Target\ Sequence}$$

The amount of sequencing needed for a given sample is determined by the goals of the experiment and the nature of the RNA sample.

- For a general view of differential expression: 5–25 million reads per sample
- For alternative splicing and lowly expressed genes: 30–60 million reads per sample.
- In-depth view of the transcriptome/assemble new transcripts: 100–200 million reads
- Targeted RNA expression requires fewer reads.
- miRNA-Seq or Small RNA Analysis require even fewer reads.

Designing the right experiment - Read length

Long or short reads? Paired or Single end?

The answer depends on the experiment:

- Gene expression – typically just a short read e.g. 50/75 bp; SE or PE.
- kmer-based quantification of Gene Expression (Salmon etc.) - benefits from PE.
- Transcriptome Analysis – longer paired-end reads (such as 2 x 75 bp).
- Small RNA Analysis – short single read, e.f. SE50 - will need trimming.

Designing the right experiment - Replication

Biological Replication

- Measures the biological variations between individuals
- Accounts for sampling bias

Technical Replication

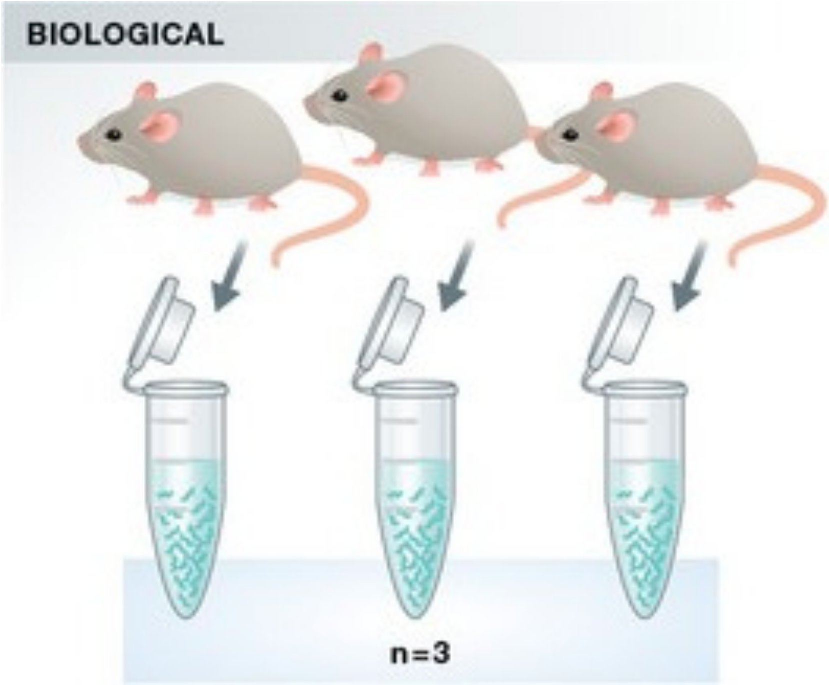
- Measures the variation in response quantification due to imprecision in the technique
- Accounts for technical noise

Designing the right experiment - Replication

Biological Replication

Each replicate is from an indepent biological individual

- *In Vivo*:
  - Patients
  - Mice
- *In Vitro*:
  - Different cell lines
  - Different passages

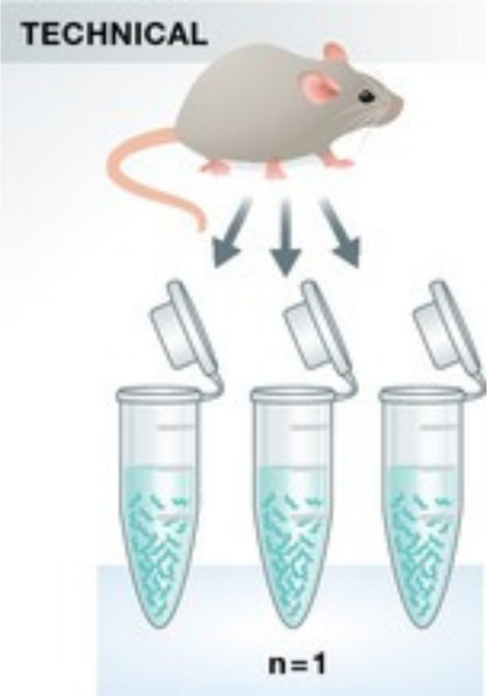


Designing the right experiment - Replication

Technical Replication

Replicates are from the same individual but processed separately

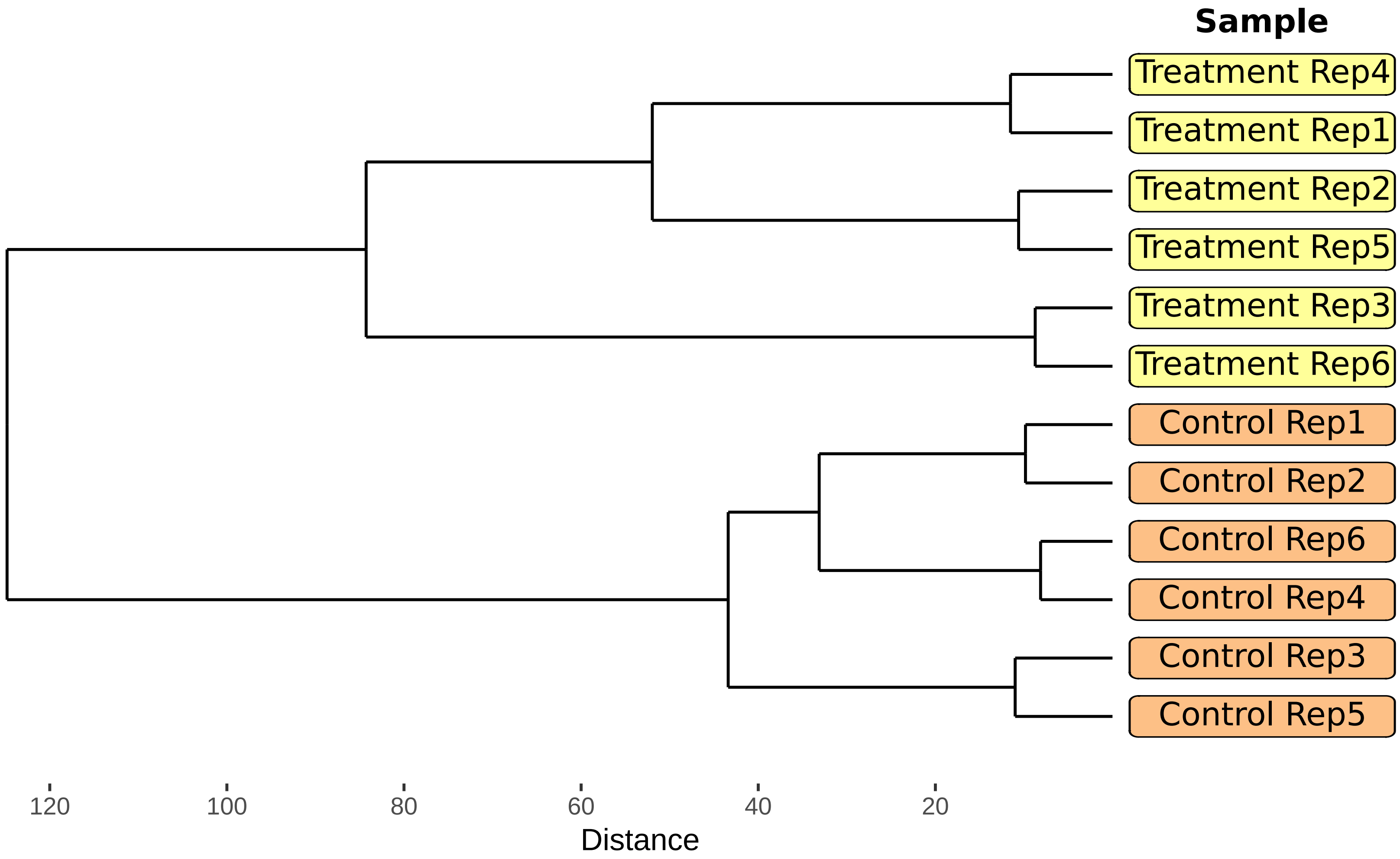
- Experimental protocol
- Measurement platform



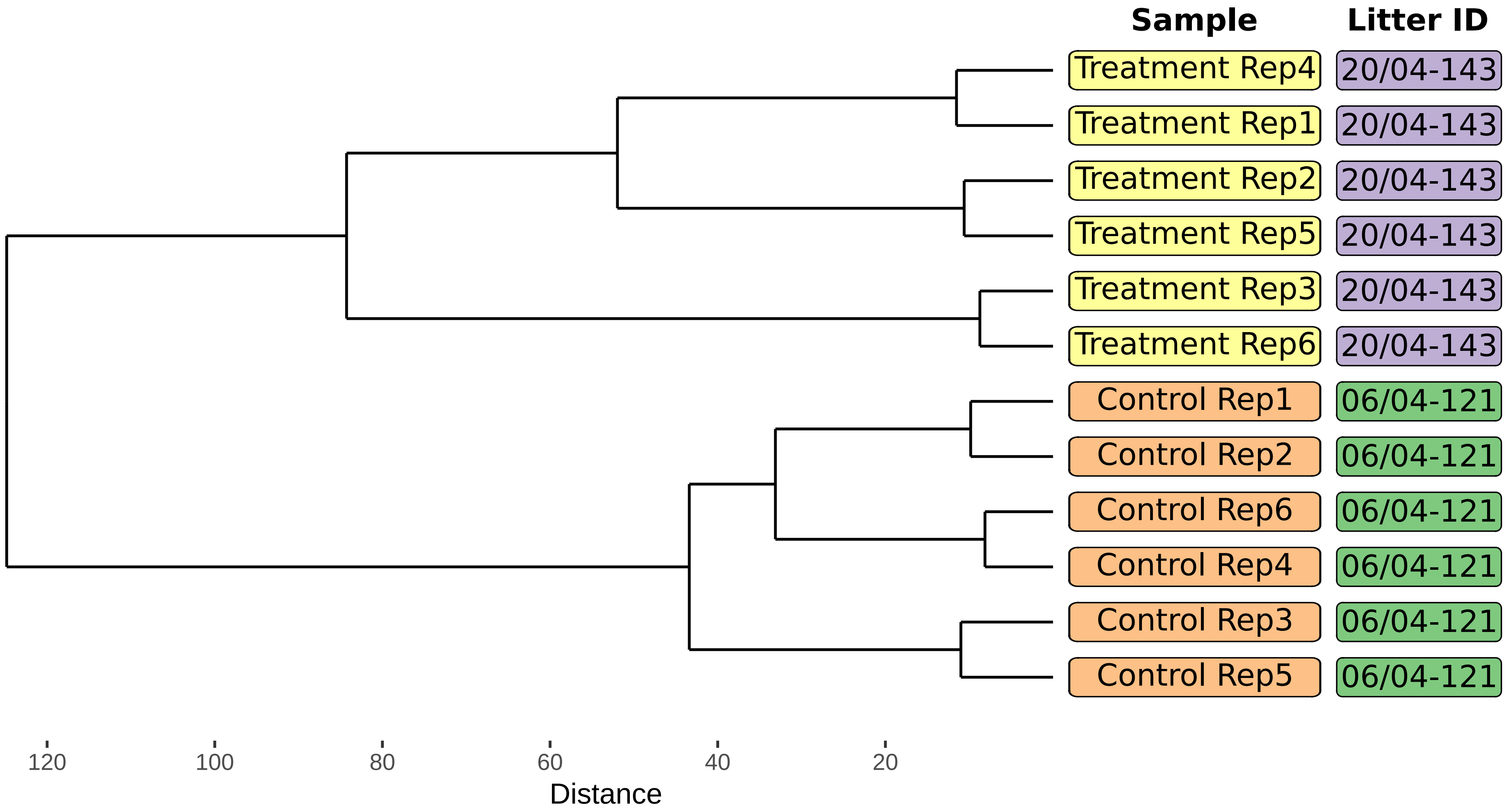
Designing the right experiment - Batch effects

- Batch effects are sub-groups of measurements that have qualitatively different behavior across conditions and are unrelated to the biological or scientific variables in a study.
- Batch effects are problematic if they are confounded with the experimental variable.

Designing the right experiment - Batch effects



Designing the right experiment - Batch effects

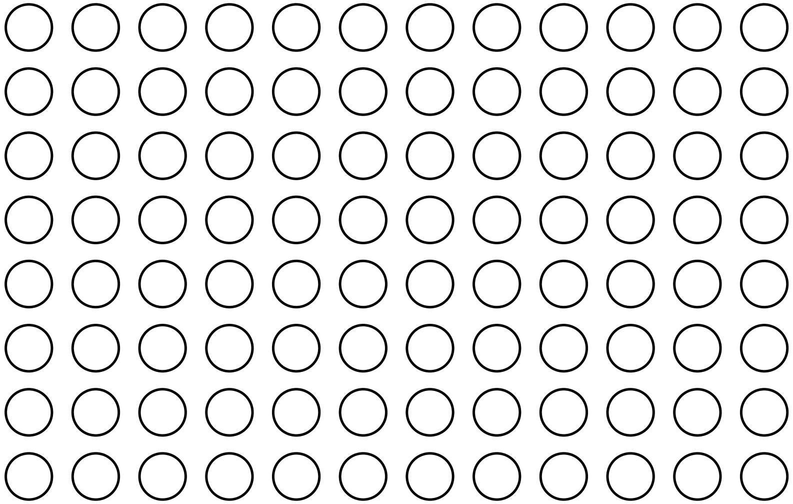


Designing the right experiment - Batch effects

- Batch effects are sub-groups of measurements that have qualitatively different behavior across conditions and are unrelated to the biological or scientific variables in a study.
- Batch effects are problematic if they are confounded with the experimental variable.
- Batch effects that are randomly distributed across experimental variables can be controlled for.

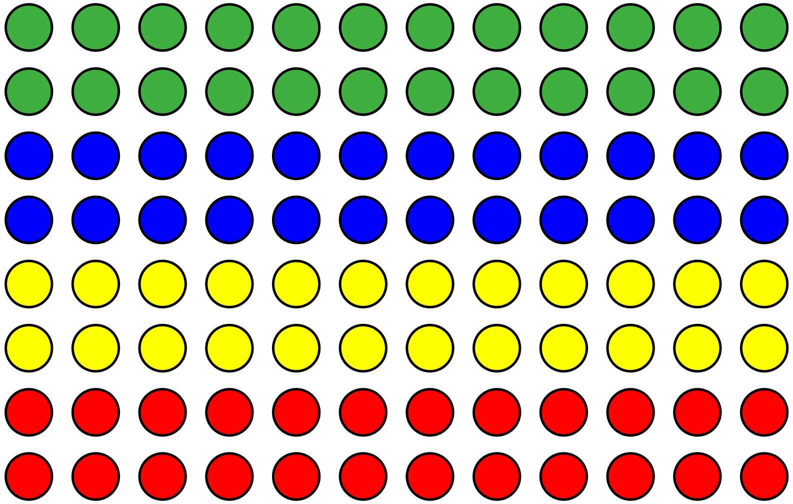
Designing the right experiment - Batch effects

- Batch effects are sub-groups of measurements that have qualitatively different behavior across conditions and are unrelated to the biological or scientific variables in a study.
- Batch effects are problematic if they are confounded with the experimental variable.
- Batch effects that are randomly distributed across experimental variables can be controlled for.
- Randomise all technical steps in data generation in order to avoid batch effects.



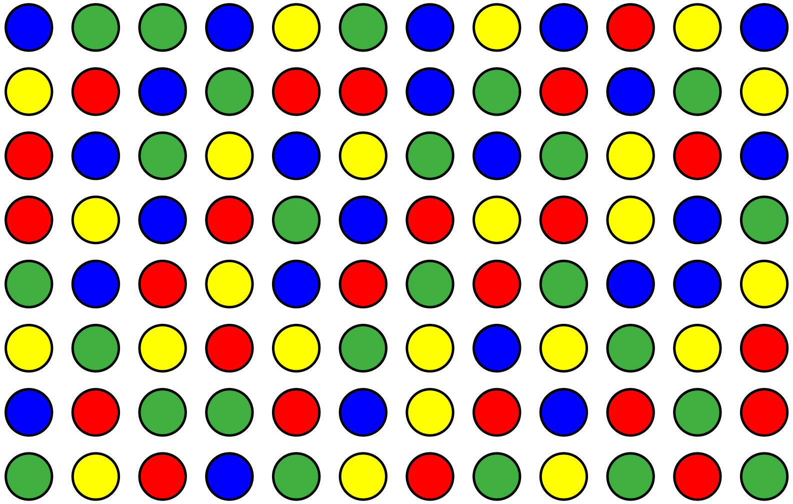
Designing the right experiment - Batch effects

- Batch effects are sub-groups of measurements that have qualitatively different behavior across conditions and are unrelated to the biological or scientific variables in a study.
- Batch effects are problematic if they are confounded with the experimental variable.
- Batch effects that are randomly distributed across experimental variables can be controlled for.
- Randomise all technical steps in data generation in order to avoid batch effects.



Designing the right experiment - Batch effects

- Batch effects are sub-groups of measurements that have qualitatively different behavior across conditions and are unrelated to the biological or scientific variables in a study.
- Batch effects are problematic if they are confounded with the experimental variable.
- Batch effects that are randomly distributed across experimental variables can be controlled for.
- Randomise all technical steps in data generation in order to avoid batch effects.



Designing the right experiment - Batch effects

- Batch effects are sub-groups of measurements that have qualitatively different behavior across conditions and are unrelated to the biological or scientific variables in a study.
- Batch effects are problematic if they are confounded with the experimental variable.
- Batch effects that are randomly distributed across experimental variables can be controlled for.
- Randomise all technical steps in data generation in order to avoid batch effects
- **Record everything:** Age, sex, litter, cell passage ..

RNAseq Workflow

Experimental Design

Library Preparation

Sequencing

Bioinformatics Analysis

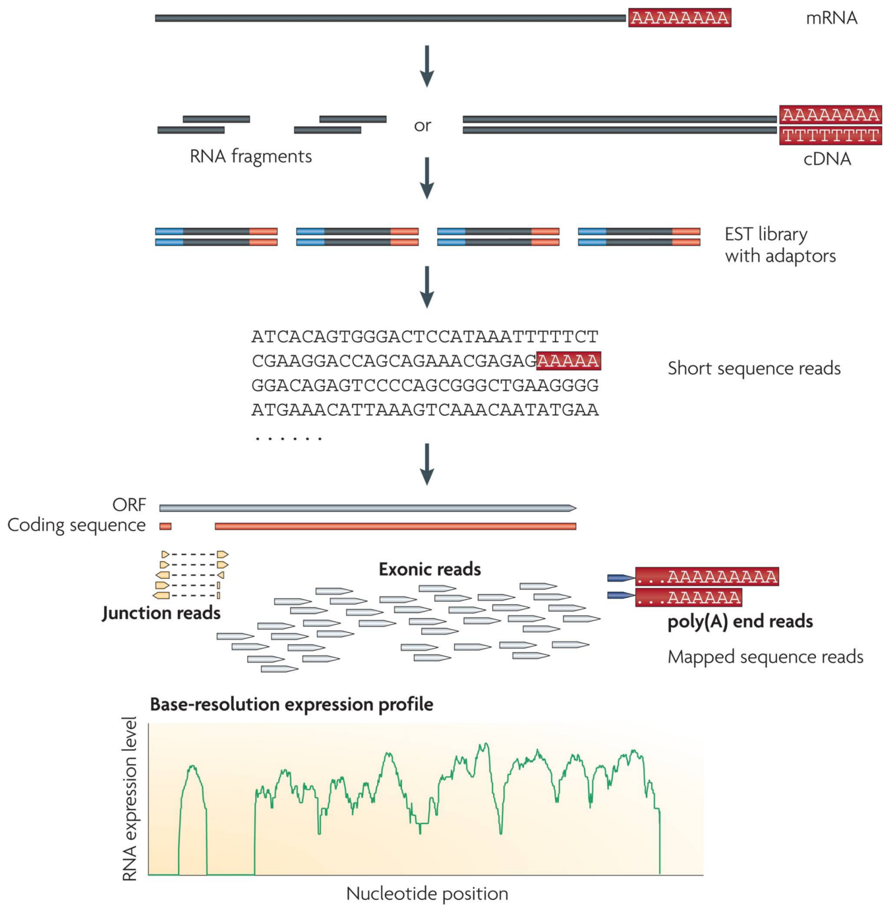
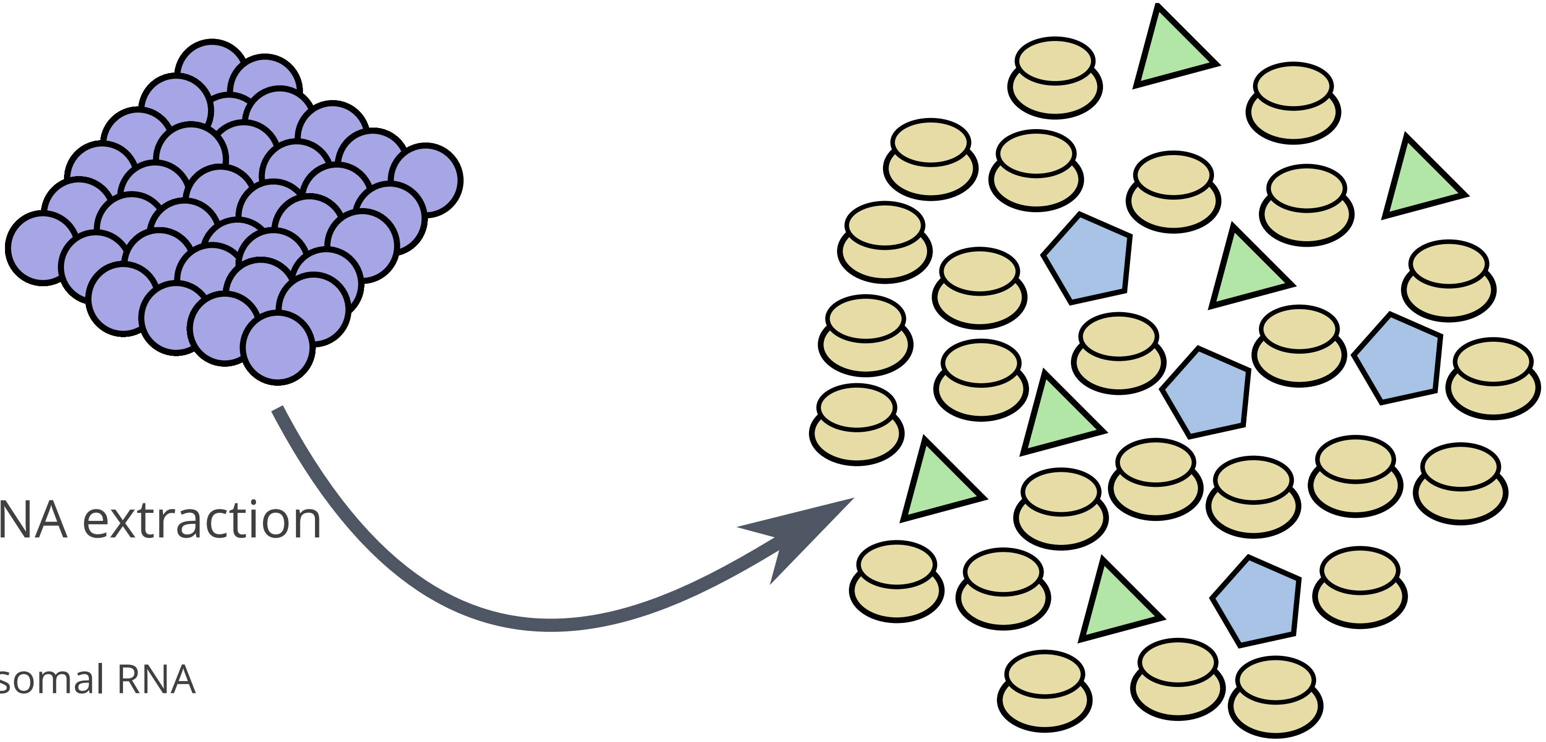


Image adapted from: Wang, Z., et al. (2009), Nature Reviews Genetics, 10, 57–63.

Library preparation



Total RNA extraction

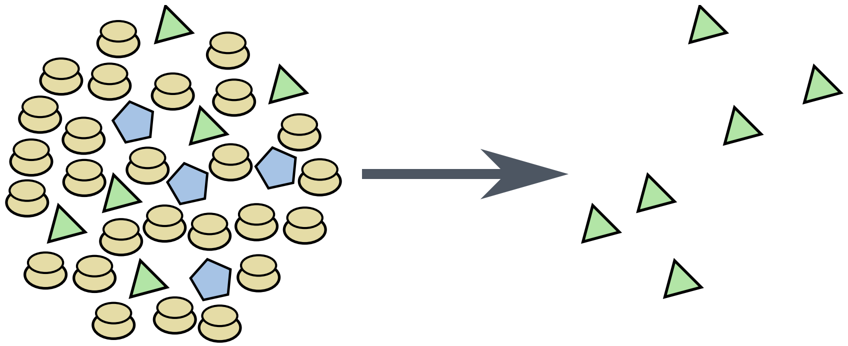
⬭ - Ribosomal RNA

▲ - Poly-A transcripts

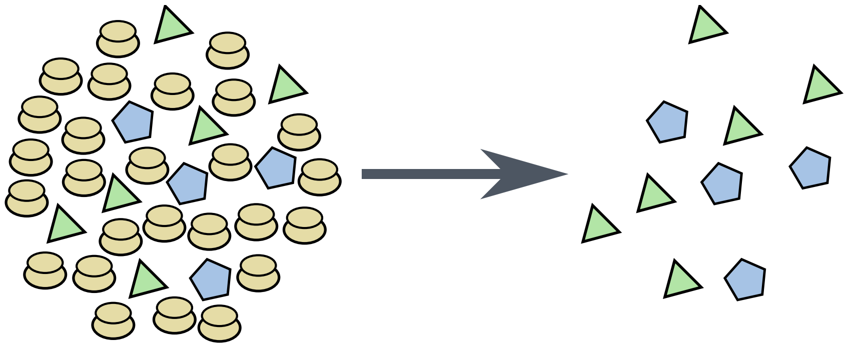
⬟ - Other RNAs e.g. tRNA, miRNA etc.

Library preparation

Poly-A Selection



Ribominus selection



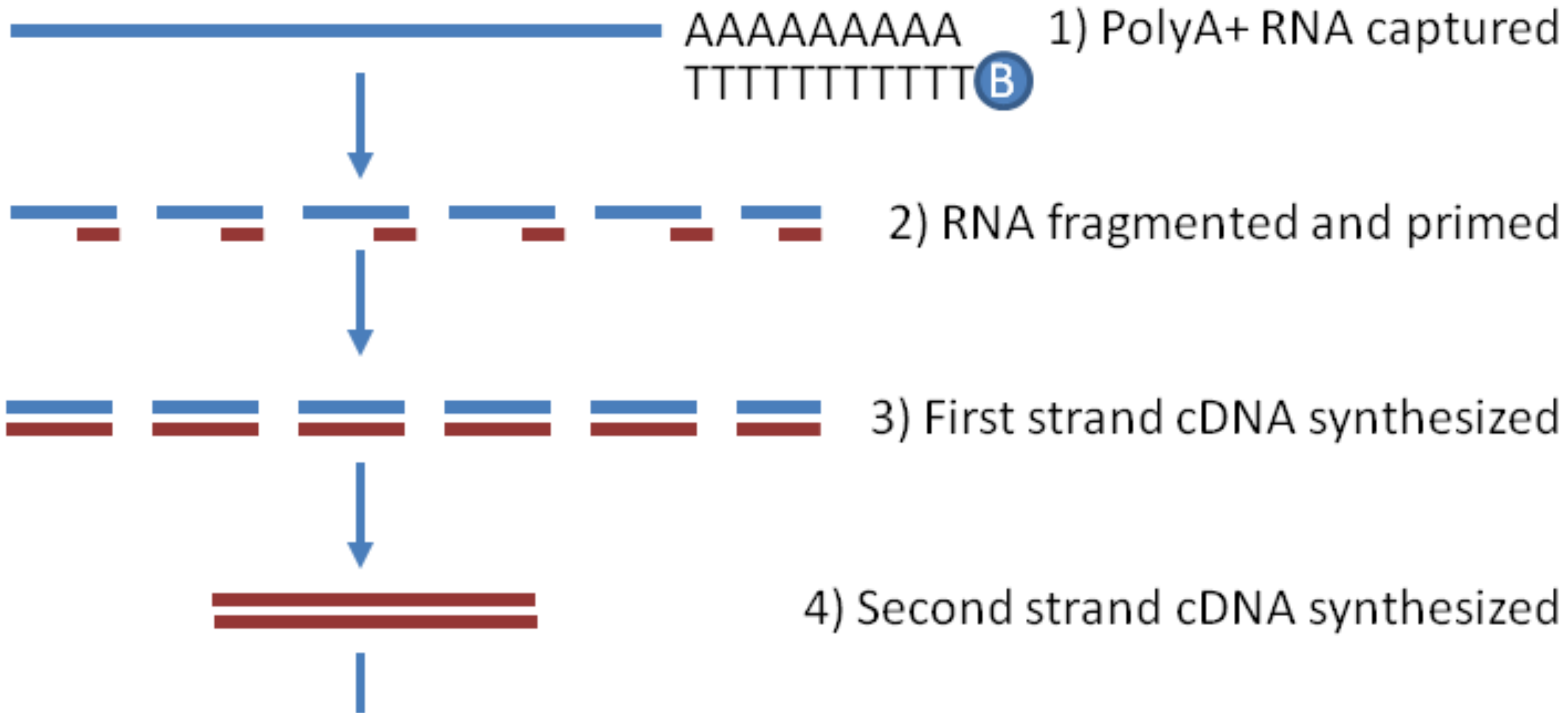
Poly-A transcripts e.g.:

- mRNAs
- immature miRNAs
- snoRNA

Poly-A transcripts + Other mRNAs e.g.:

- tRNAs
- mature miRNAs
- piRNAs

Library preparation



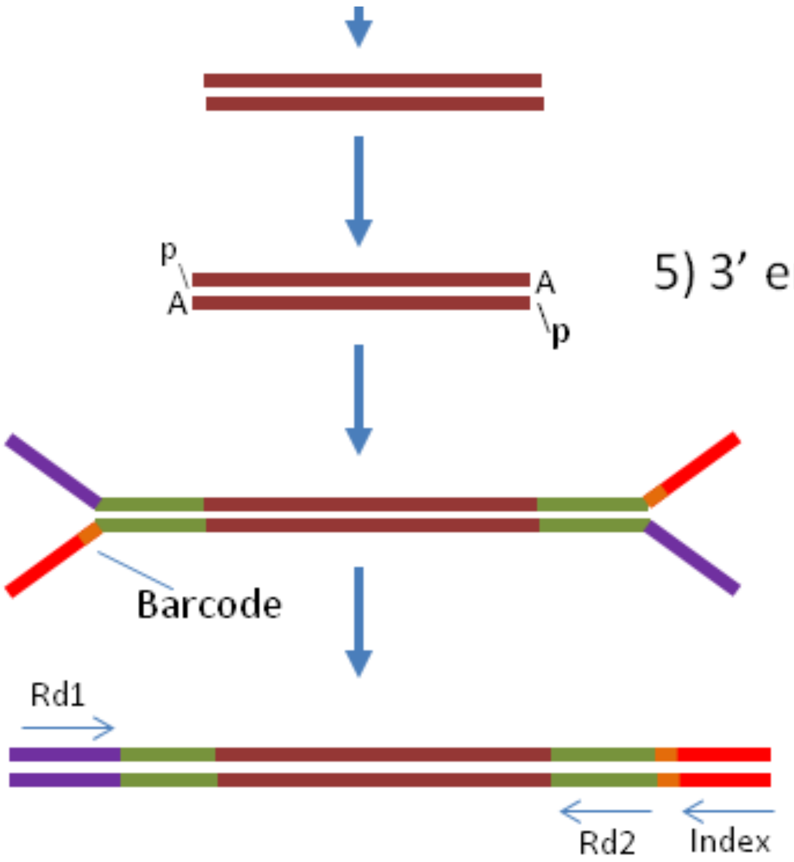
1) PolyA+ RNA captured

2) RNA fragmented and primed

3) First strand cDNA synthesized

4) Second strand cDNA synthesized

Library preparation



4) Second strand cDNA synthesized

5) 3’ ends adenylated and 5’ ends repaired

6) DNA sequencing adapters ligated

7) Ligated fragments PCR amplified

RNAseq Workflow

Experimental Design

Library Preparation

Sequencing

Bioinformatics Analysis

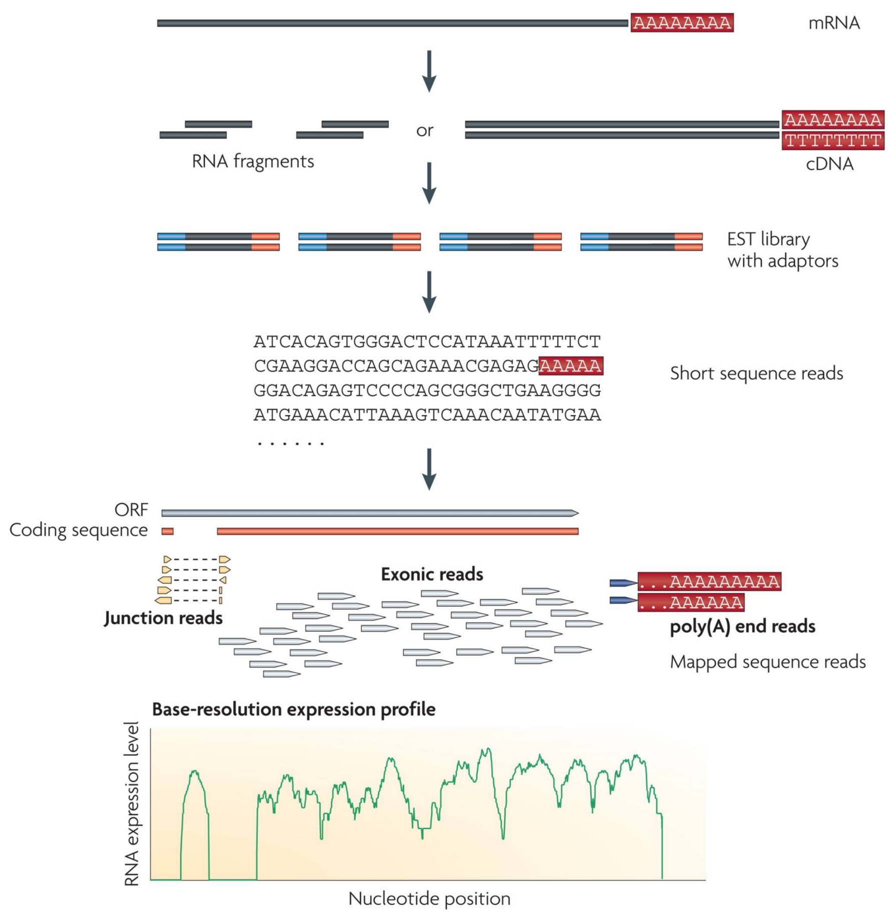
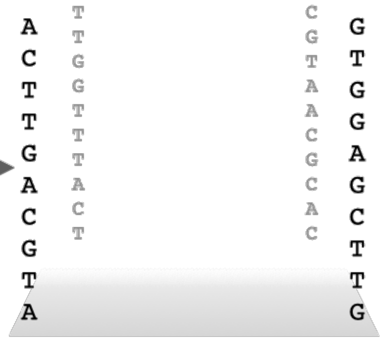


Image adapted from: Wang, Z., et al. (2009), Nature Reviews Genetics, 10, 57–63.

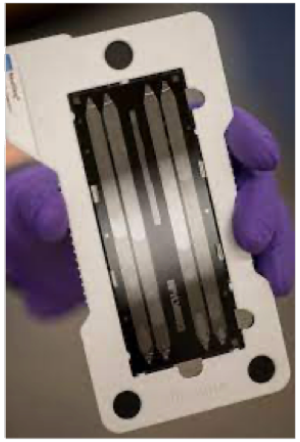
Sequencing by synthesis



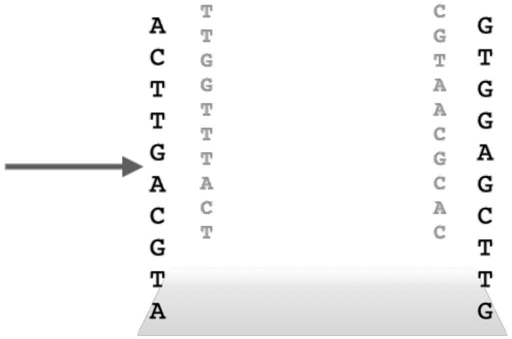
This is a fragment of cDNA we want to sequence →

Slides adapted from StatQuest

Sequencing by synthesis



This is a fragment of cDNA we want to sequence



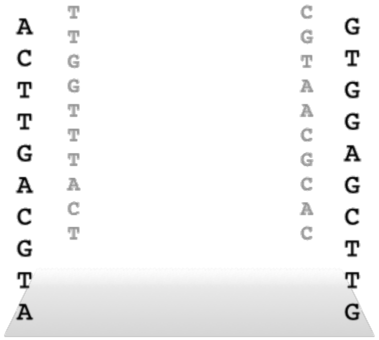
There are 4 fragments in this toy example

NextSeq550 flowcells have 400,000,000 fragments

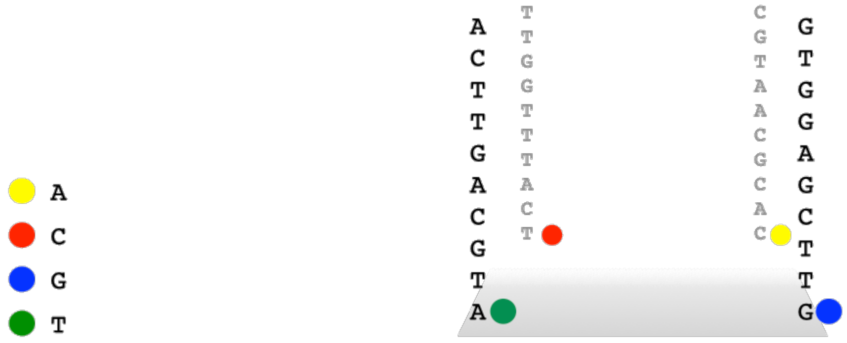
Sequencing by synthesis

Fluorescent labeled bases are incorporated to the fragments by DNA polymerase

🟡 A
🔴 C
🔵 G
🟢 T

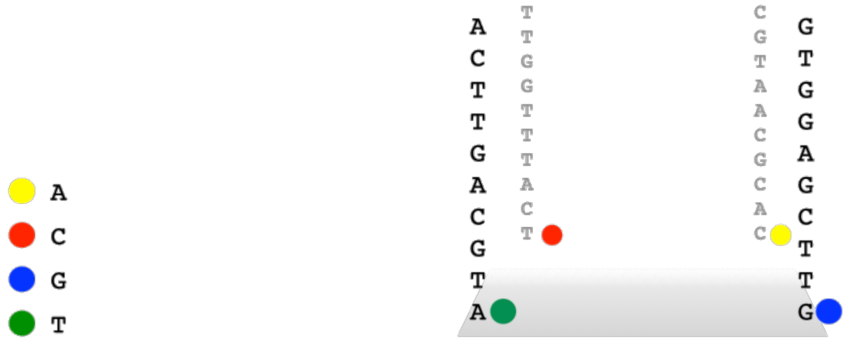


Sequencing by synthesis



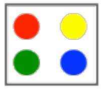
The bases are incorporated to the first base of each fragment

Sequencing by synthesis

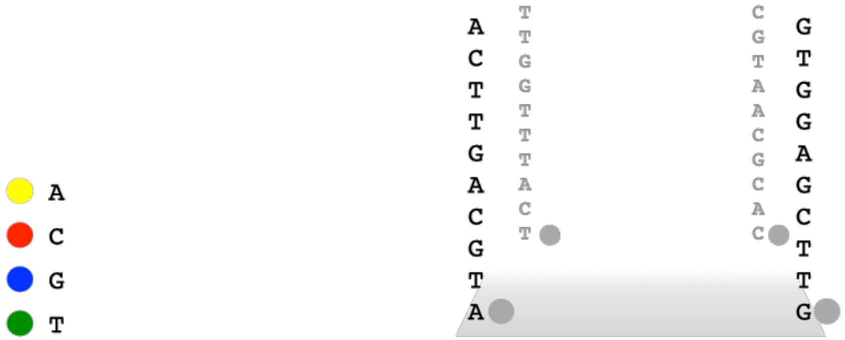


The bases are incorporated to the first base of each fragment

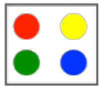
Sequencer takes a picture of the flowcell



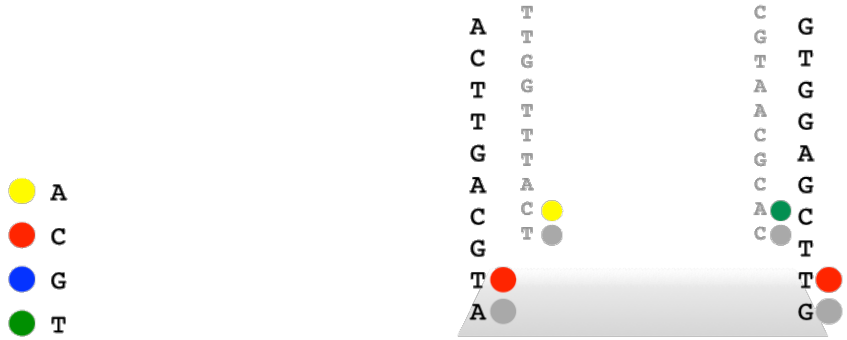
Sequencing by synthesis



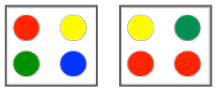
The colours are washed off, and the cycle is repeated



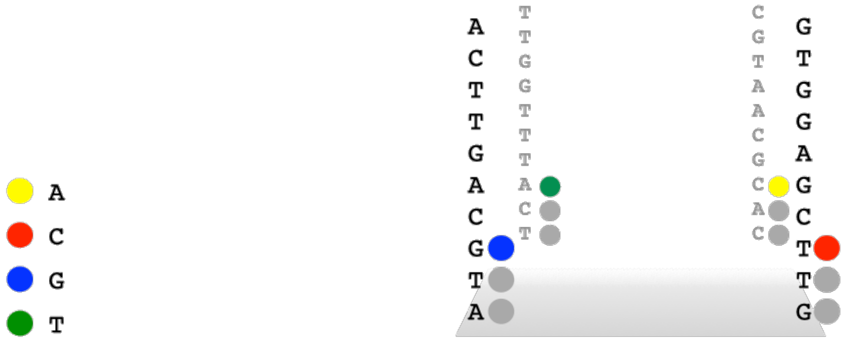
Sequencing by synthesis



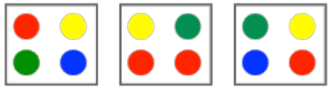
Chemical wash - Base addition - Imaging



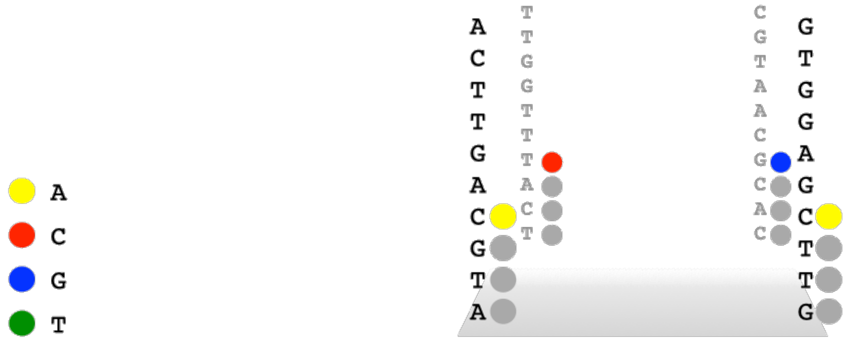
Sequencing by synthesis



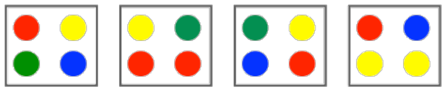
Chemical wash - Base addition - Imaging



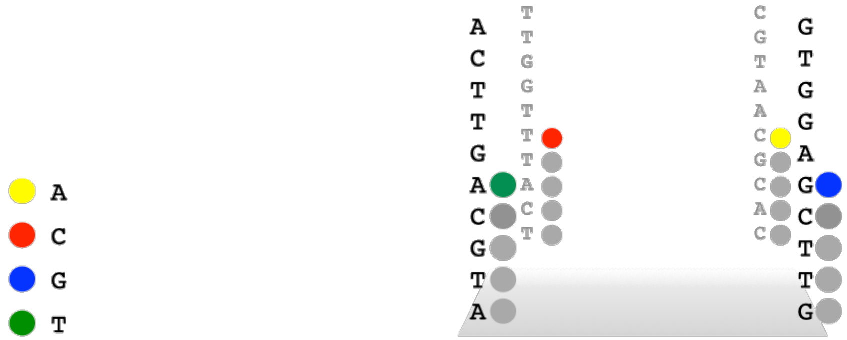
Sequencing by synthesis



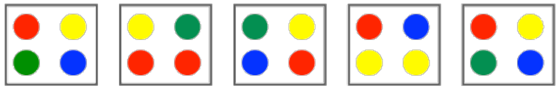
Chemical wash - Base addition - Imaging



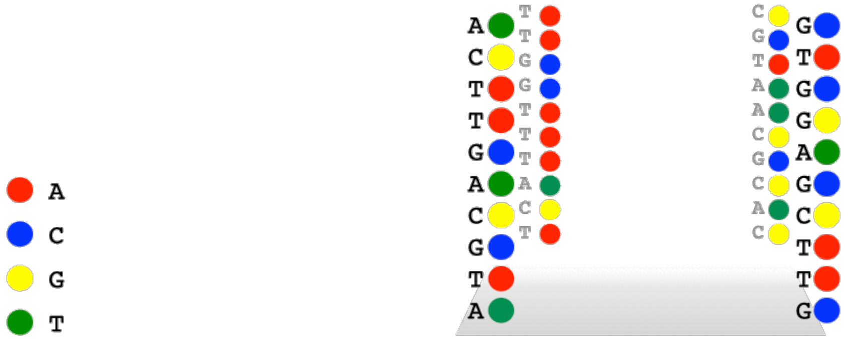
Sequencing by synthesis



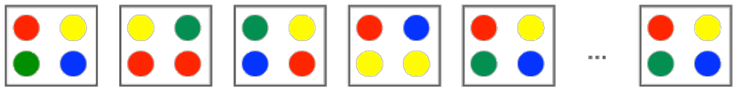
Chemical wash - Base addition - Imaging



Sequencing by synthesis

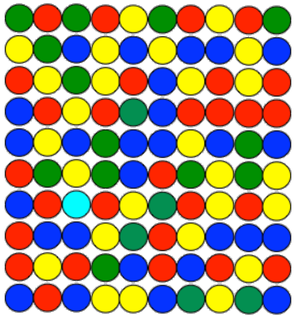


And the process repeats until each fragment is sequenced completely



Sequencing by synthesis

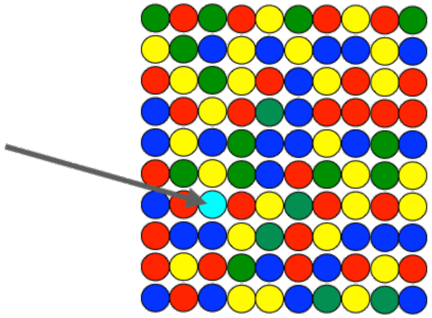
This matrix does not contain 400,000,000 fragments, but illustrates one type of problem that can occur



Sequencing by synthesis

This matrix does not contain 400,000,000 fragments, but illustrates one type of problem that can occur

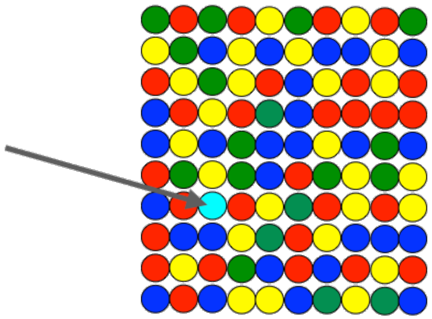
Sometimes a probe does not shine as bright as it should and the sequencer can not be confident that it is calling the correct colour



Sequencing by synthesis

This matrix does not contain 400,000,000 fragments, but illustrates one type of problem that can occur

Sometimes a probe does not shine as bright as it should and the sequencer can not be confident that it is calling the correct colour

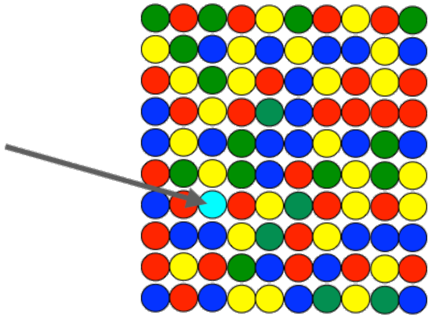


Quality scores, that are part of the output, reflect how confident the machine is that it correctly called a base.

Sequencing by synthesis

This matrix does not contain 400,000,000 fragments, but illustrates one type of problem that can occur

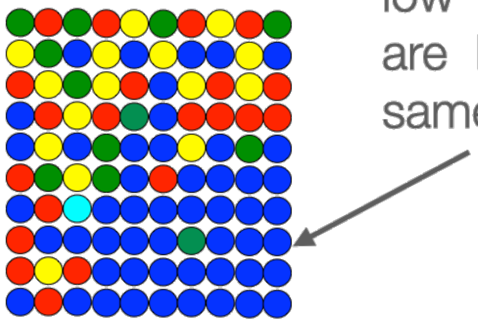
Sometimes a probe does not shine as bright as it should and the sequencer can not be confident that it is calling the correct colour



Quality scores, that are part of the output, reflect how confident the machine is that it correctly called a base.

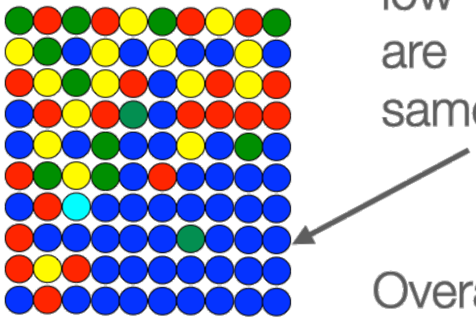
In this case, the faded dot would get a low quality score.

Sequencing by synthesis



Another reason you might get a low quality score is when there are lots of probes that are the same colour in the same region.

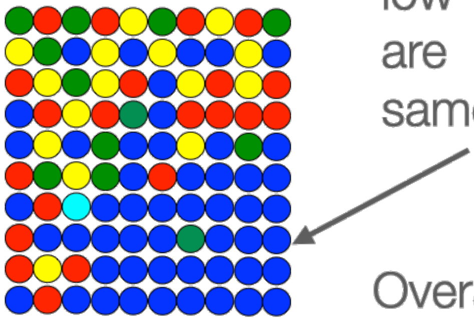
Sequencing by synthesis



Another reason you might get a low quality score is when there are lots of probes that are the same colour in the same region.

Overabundance of a single colour can make it hard to identify individual sequences.

Sequencing by synthesis



Another reason you might get a low quality score is when there are lots of probes that are the same colour in the same region.

Overabundance of a single colour can make it hard to identify individual sequences.

In this case, all the dots in this low complexity region will get a low quality score.

RNAseq Workflow

Experimental Design

Library Preparation

Sequencing

Bioinformatics Analysis

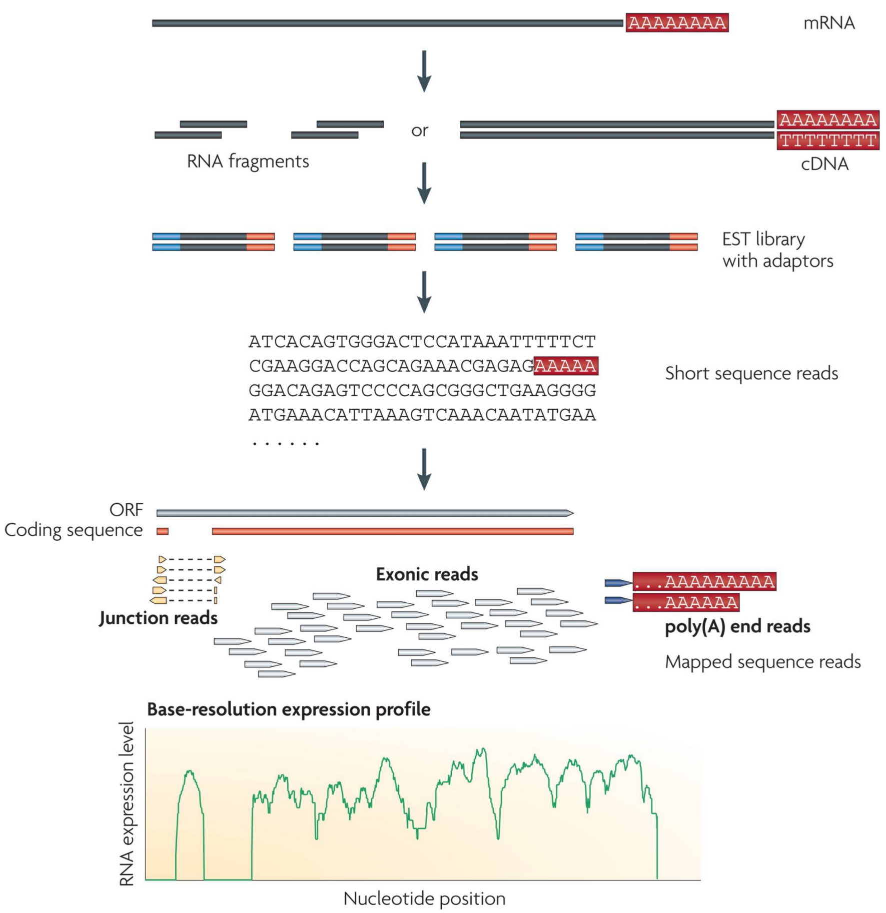
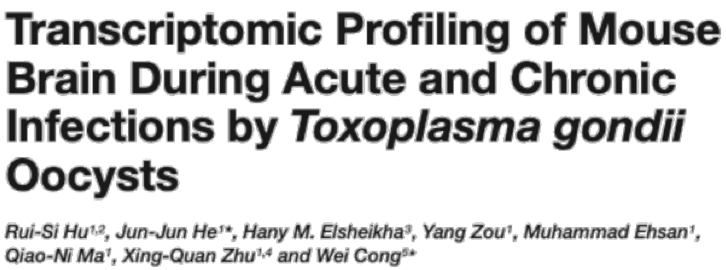


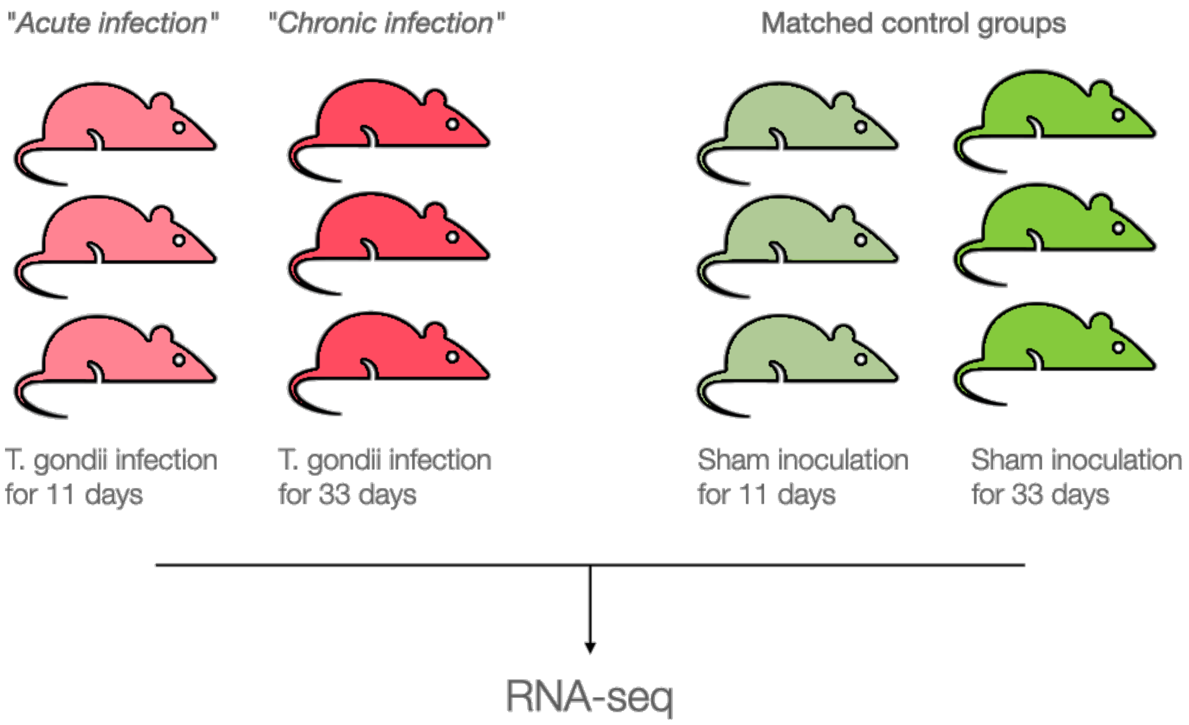
Image adapted from: Wang, Z., et al. (2009), Nature Reviews Genetics, 10, 57–63.

Case Study



Transcriptomic Profiling of Mouse Brain During Acute and Chronic Infections by *Toxoplasma gondii* Oocysts

Rui-Si Hu[1,2], Jun-Jun He[1], Hany M. Elsheikha[3], Yang Zou[1], Muhammad Ehsan[1], Qiao-Ni Ma[1], Xing-Quan Zhu[1,4] and Wei Cong[5]**



"Acute infection"

"Chronic infection"

Matched control groups

T. gondii infection for 11 days

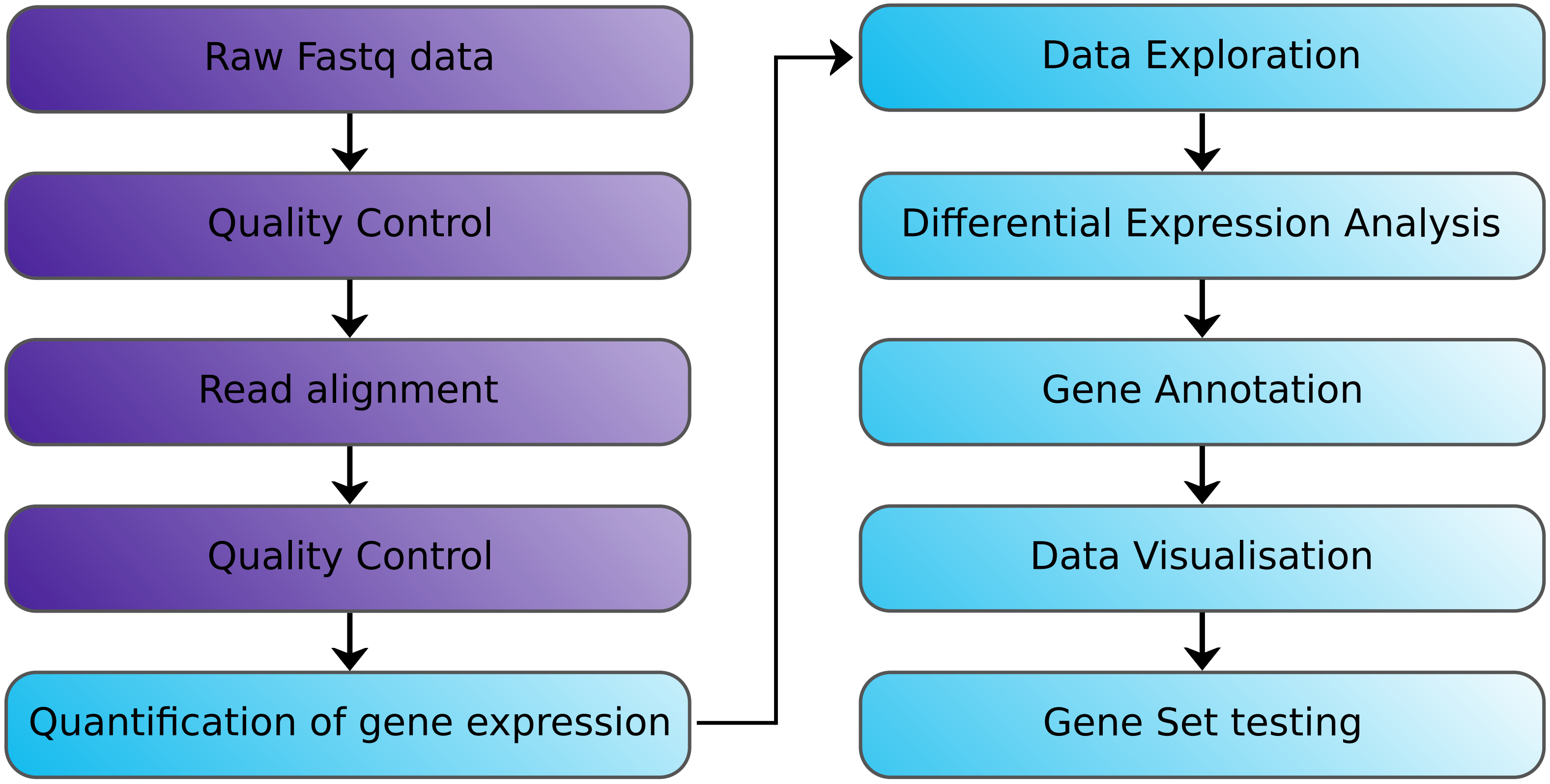
T. gondii infection for 33 days

Sham inoculation for 11 days

Sham inoculation for 33 days

RNA-seq

Differential Gene Expression Analysis Workflow



Raw Fastq data

Quality Control

Read alignment

Quality Control

Quantification of gene expression

Data Exploration

Differential Expression Analysis

Gene Annotation

Data Visualisation

Gene Set testing