Overview of RNA-Seq

Dr. Princess Rodriguez

2025-01-29

Learning Objectives:

- Understand applications of RNA sequencing
- Introduce the overall differential expression workflow
- Understand experimental design concepts such as replicates and batch effects
- Understand different types of library preps, their requirements and uses.

Overview of RNA-seq

RNA-seq is an exciting experimental technique that is utilized to explore and/or quantify gene expression within or between conditions.

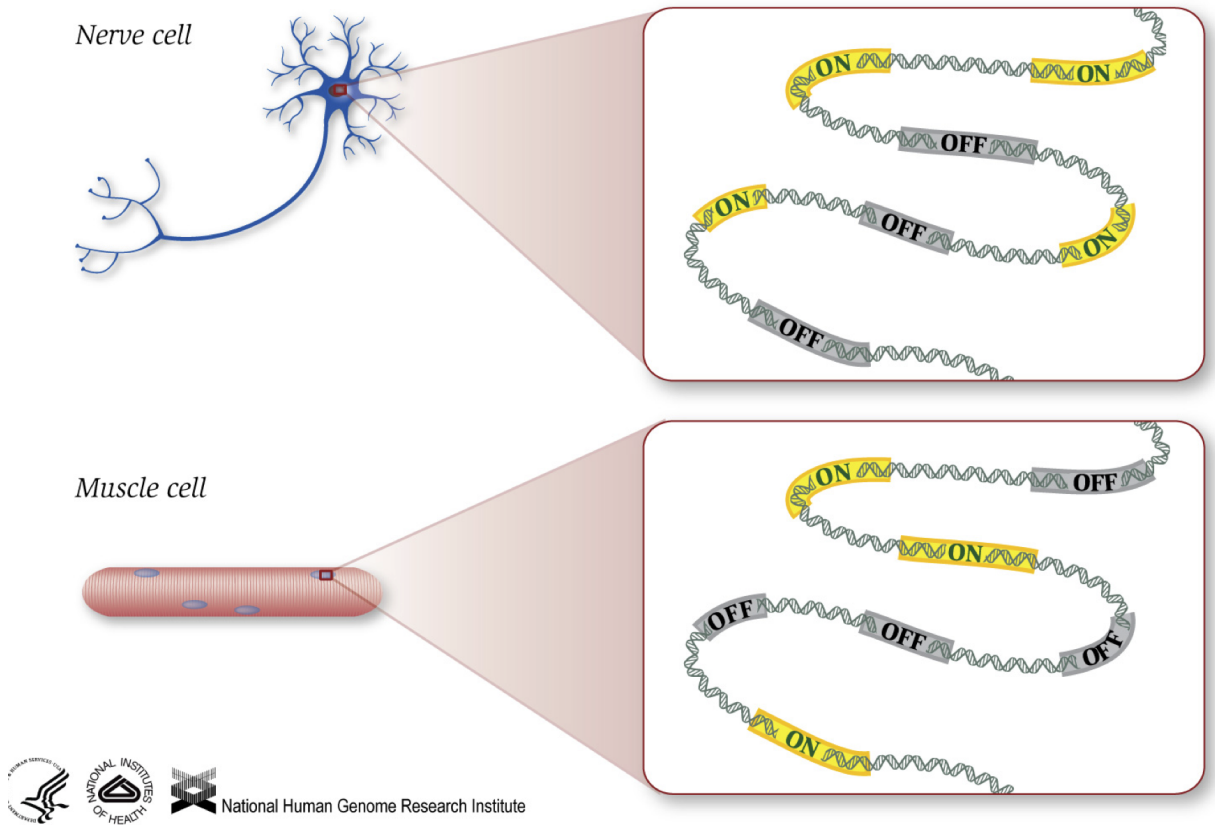
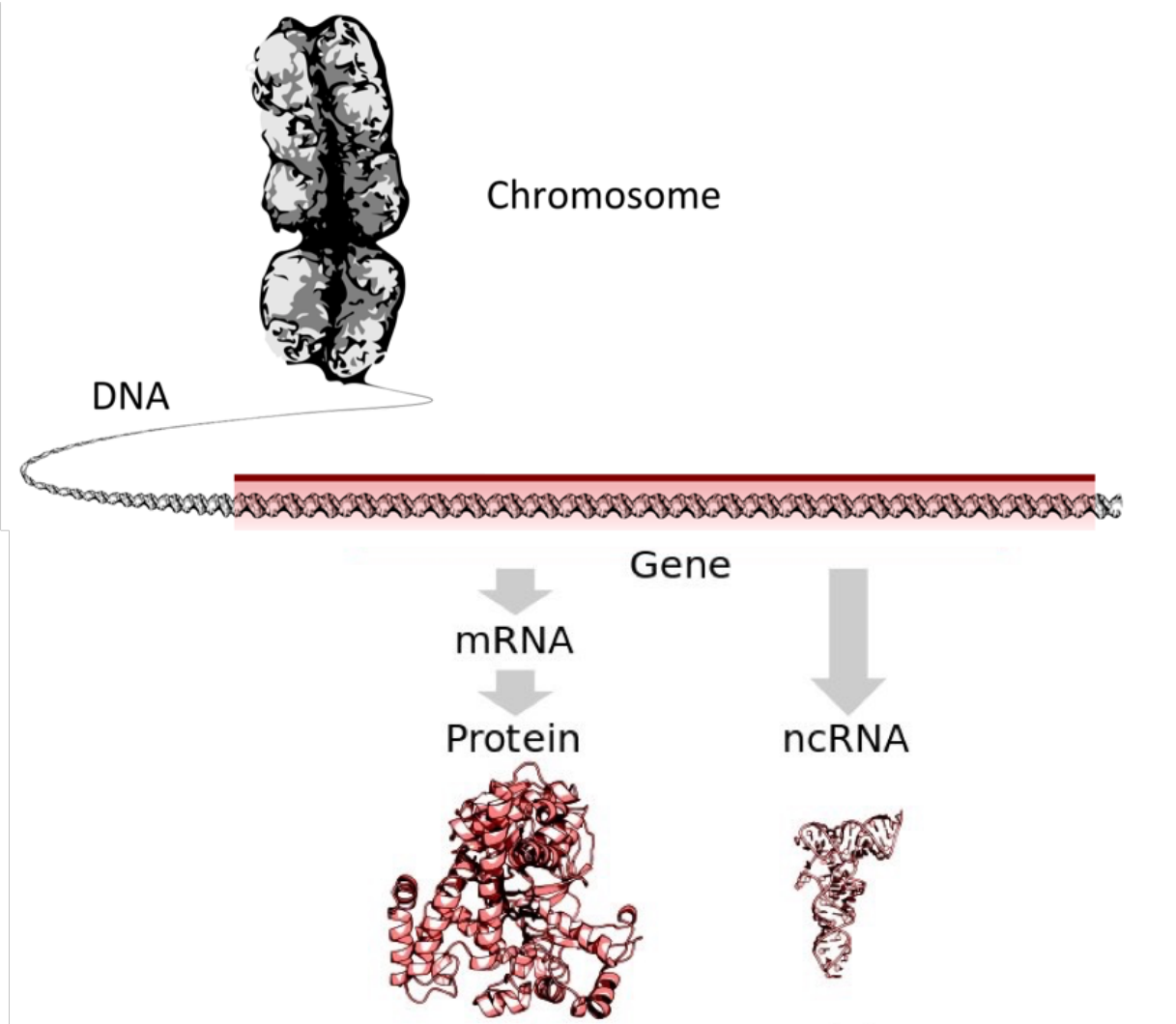


Figure 1: Gene Expression in Cells

Transcription

All RNAs transcribed from genes are called transcripts.



Translation

To be translated into proteins, the RNA must undergo processing to generate the mRNA.

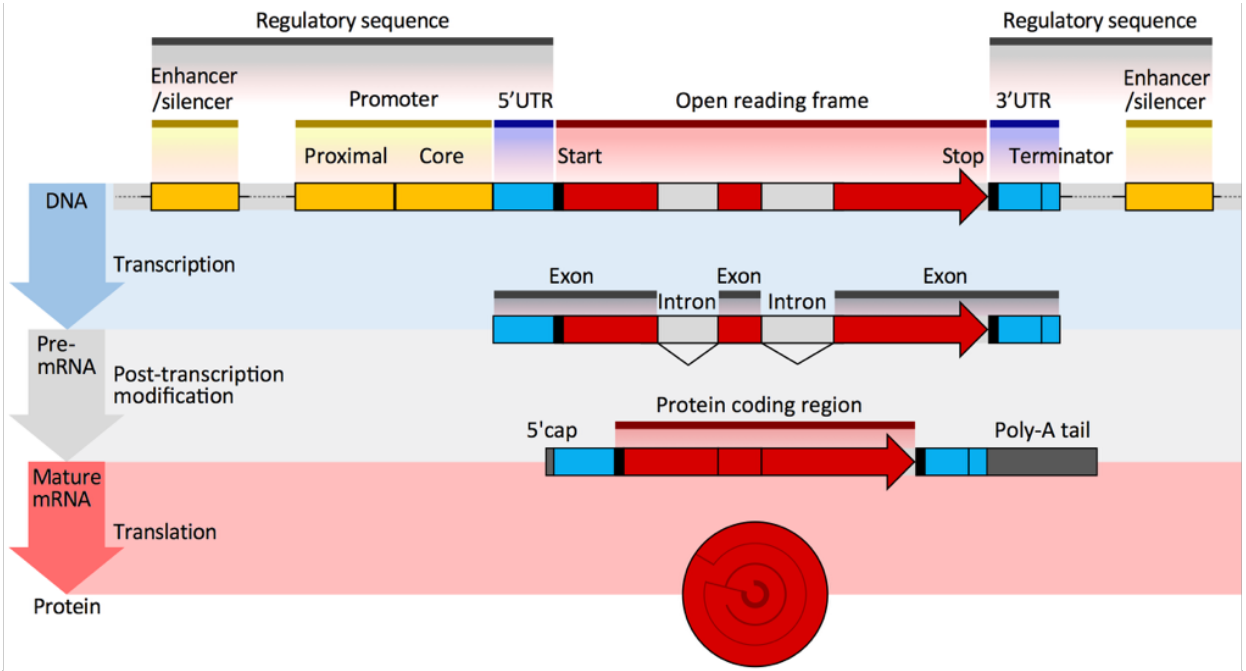


Figure 3: Gene Structure

Transcriptomics

The transcriptome is defined as a collection of all the transcript readouts present in a cell. RNA-seq data can be used to explore and/or quantify the transcriptome of an organism, which can be utilized for the following types of experiments:

- **Differential Gene Expression**: *quantitative* evaluation and comparison of transcript levels between conditions
- **Transcriptome assembly**: building the profile of transcribed regions of the genome, a *qualitative* evaluation
- **Refinement of gene models**: building better gene models and verifying them using transcriptome assembly
- **Metatranscriptomics**: community transcriptome analysis

Illumina Library preparation

Briefly, the RNA is isolated from the sample and contaminating DNA is removed with DNase.

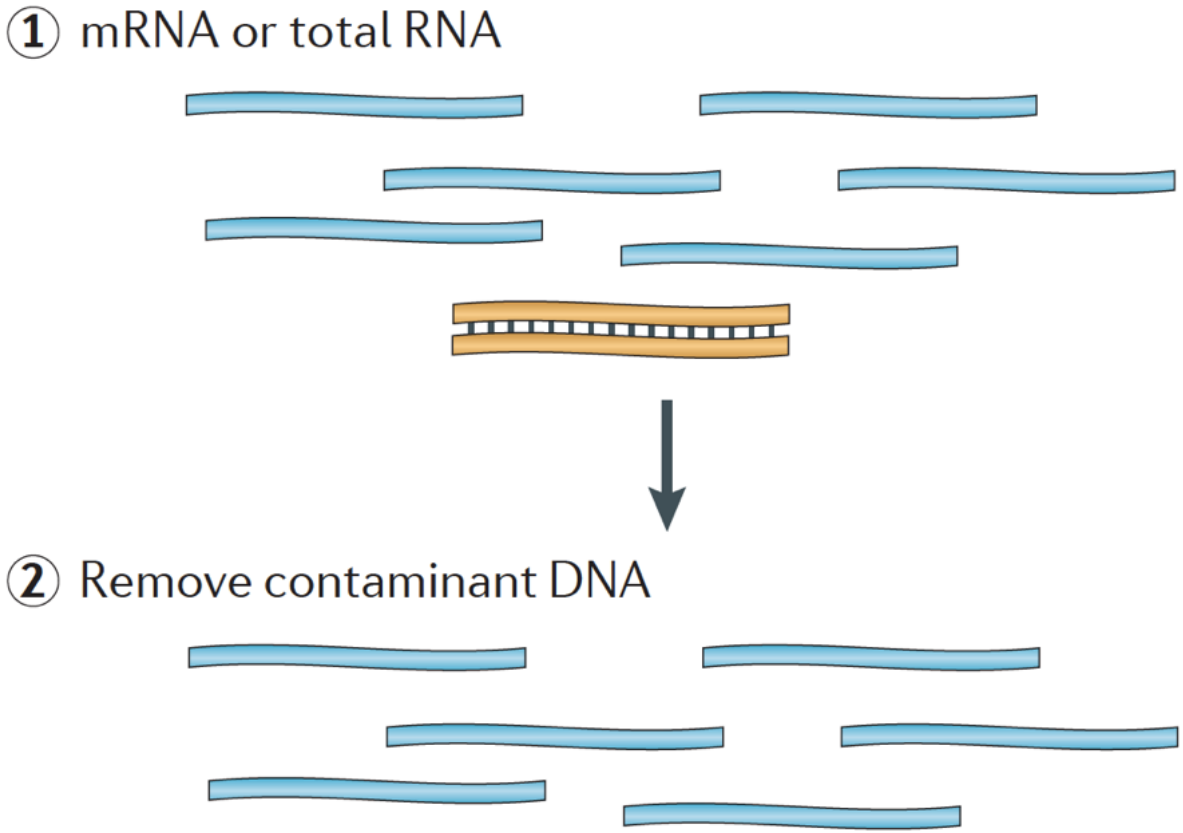


Figure 4: Library Prep

The RNA sample then undergoes either selection of the mRNA (polyA selection) or depletion of the rRNA. The resulting RNA is fragmented.

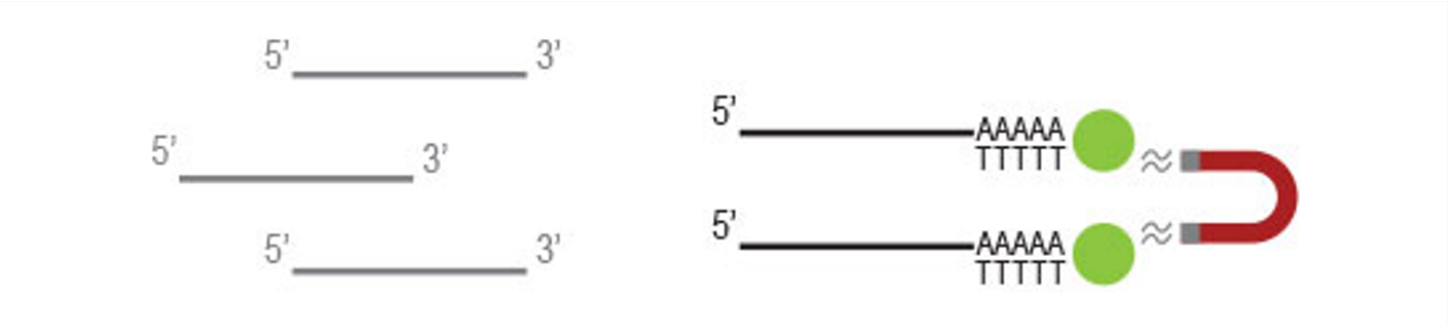


Figure 5: PolyA Tail

The size of the target fragments in the final library is a key parameter for library construction. DNA fragmentation is typically done by physical methods or enzymatic methods

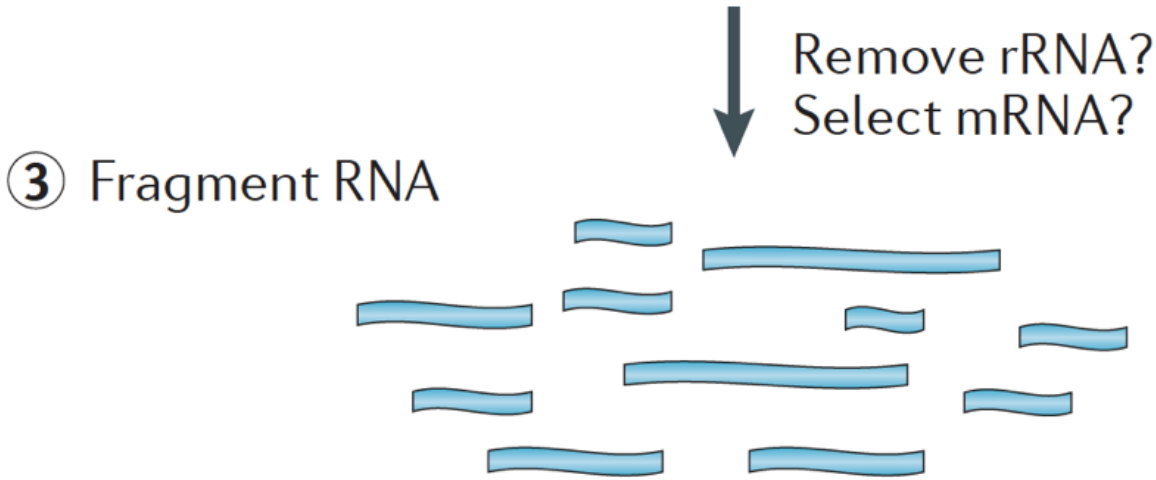


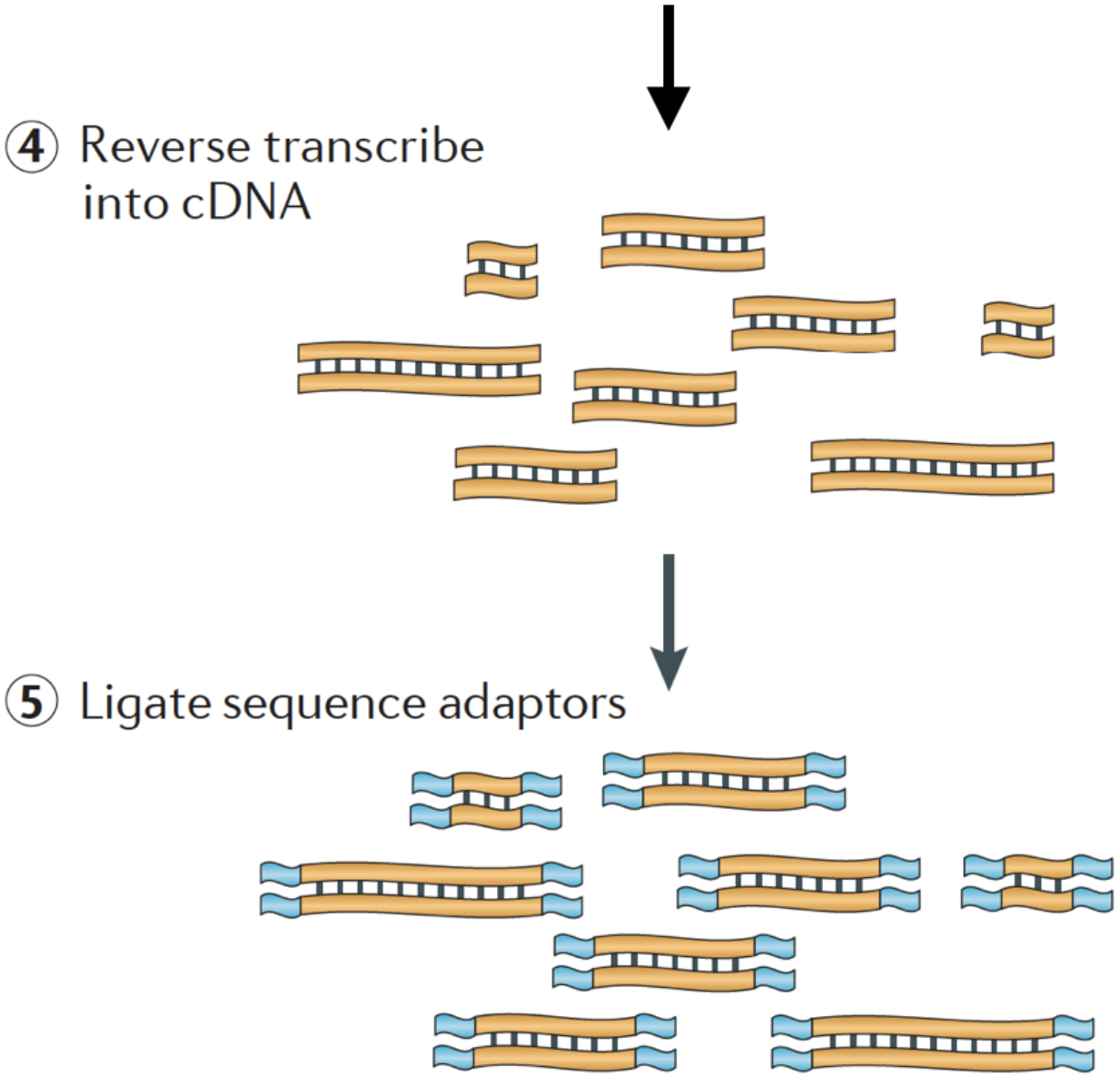
Figure 6: Library Prep

The RNA is then reverse transcribed into double-stranded cDNA and sequence adapters are then added to the ends of the fragments.

The cDNA libraries can be generated in a way to retain information about which strand of DNA the RNA was transcribed from. Libraries that retain this information are called stranded libraries.

There are 3 types of cDNA libraries available:

- *Forward (secondstrand) – reads resemble the gene sequence or the secondstrand cDNA sequence*
- *Reverse (firststrand) – reads resemble the complement of the gene sequence or firststrand cDNA sequence (TruSeq)*
- *Unstranded*


④ Reverse transcribe into cDNA
⑤ Ligate sequence adaptors

Finally, the fragments are PCR amplified if needed, and the fragments are size selected (usually ~300-500bp) to finish the library.

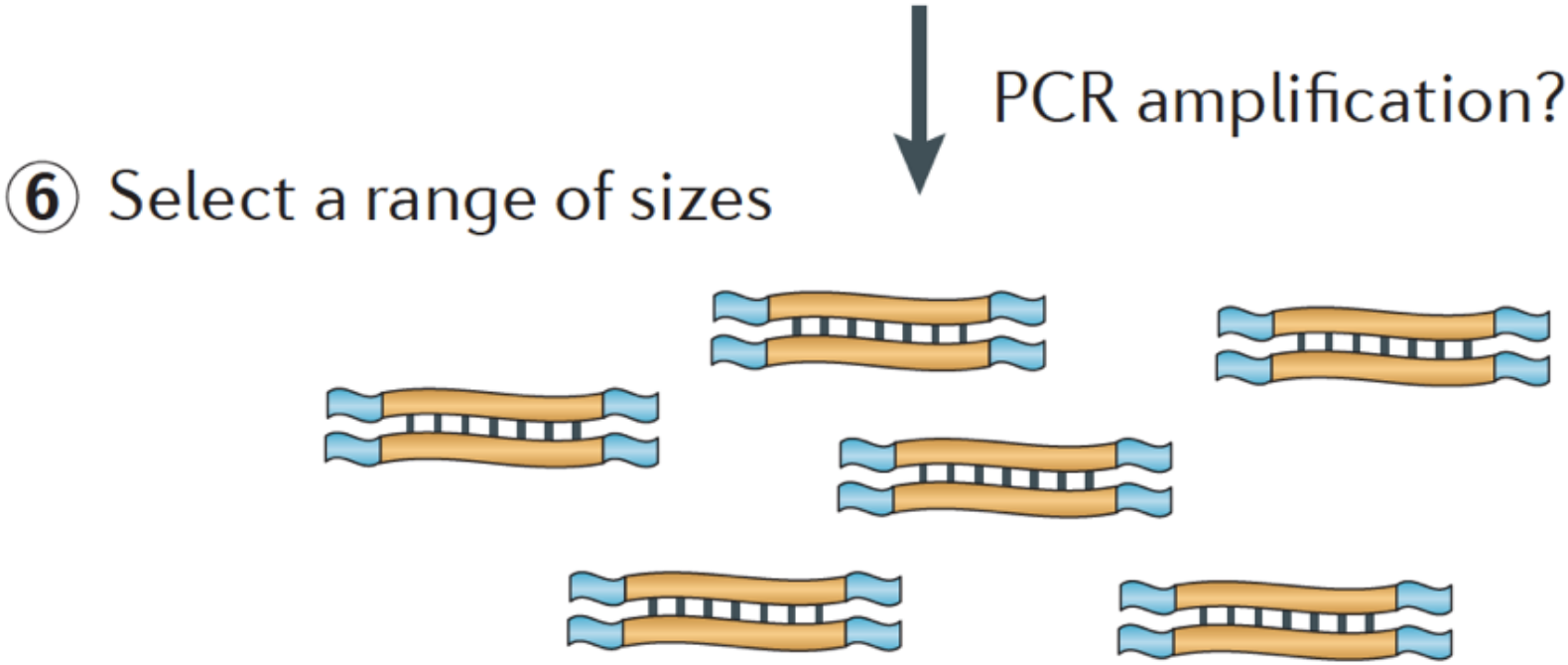


Figure 8: Library Prep Final Step

Image credit: Martin J.A. and Wang Z., Nat. Rev. Genet. (2011)

Strandedness

The implication of **stranded** libraries is that one could distinguish whether the reads are derived from the forward or reverse-encoded transcripts.

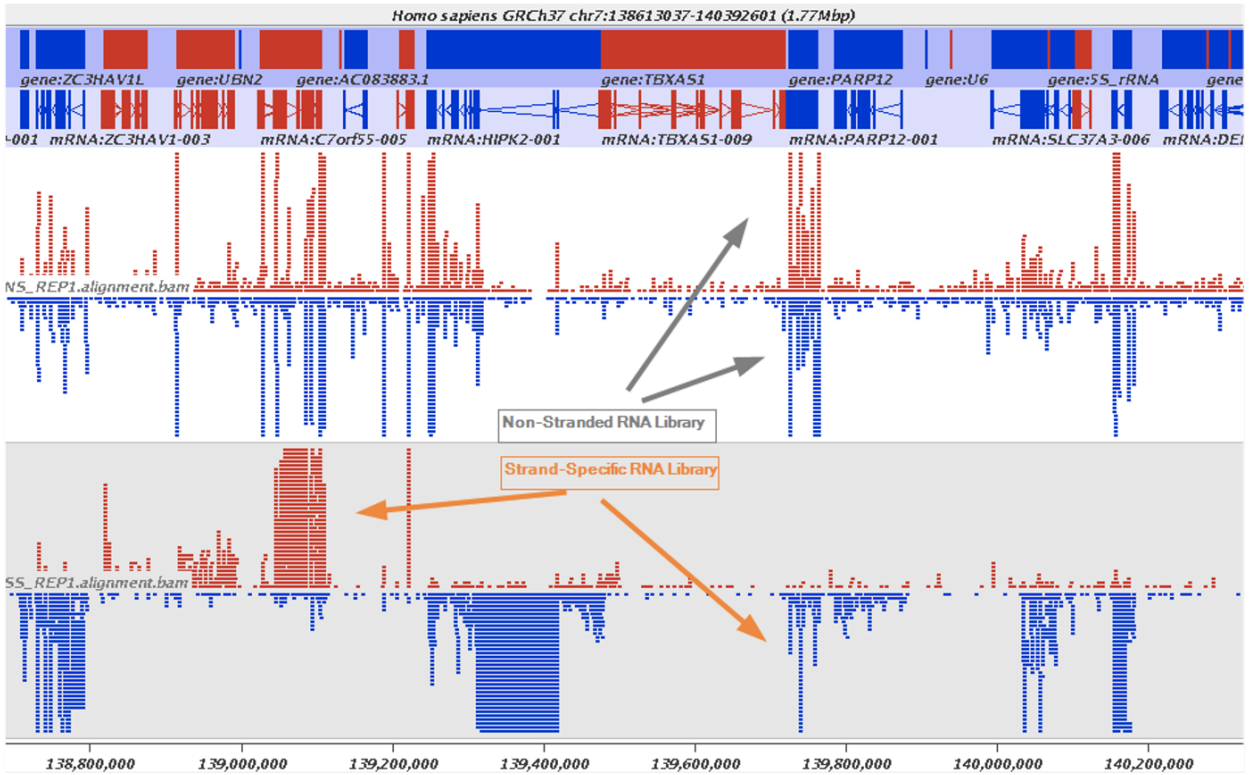


Figure 9: Strandedness

Single-end versus Paired-end

After preparation of the libraries, sequencing can be performed to generate the nucleotide sequences of the ends of the fragments, which are called **reads**. You will have the choice of sequencing a single end of the cDNA fragments (single-end reads) or both ends of the fragments (paired-end reads).

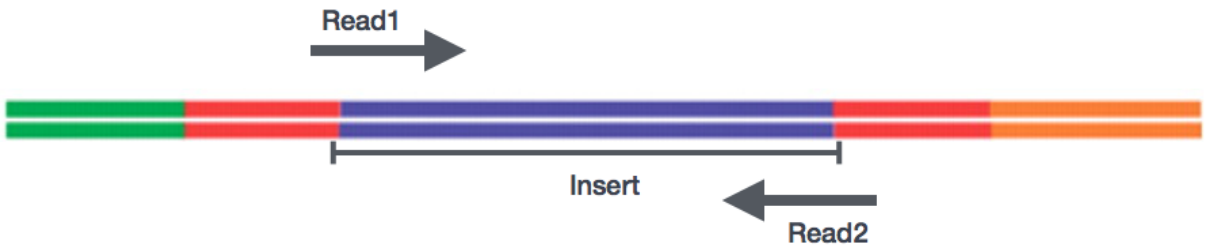


Figure 10: Paired End Reads

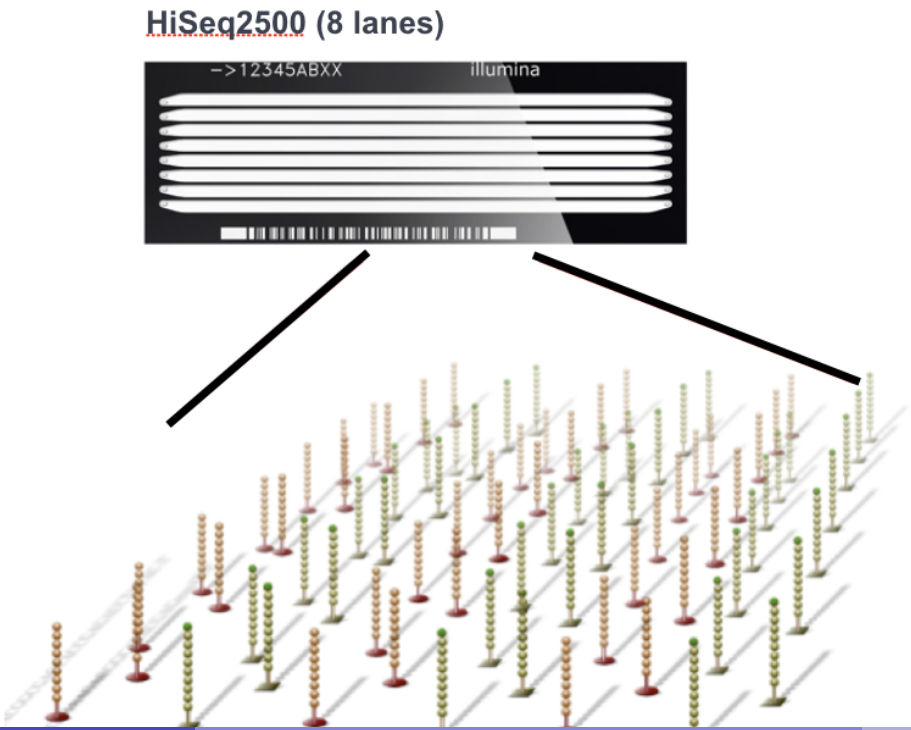
- SE - Single end dataset => Only Read1
- PE - Paired-end dataset => Read1 + Read2

Different sequencing platforms

There are a variety of Illumina platforms to choose from to sequence the cDNA libraries.

Figure 11: Illumina Platforms

Differences in platform can alter the length of reads generated, the quality of reads, as well as the total number of reads sequenced per run and the amount of time required to sequence the libraries. The different platforms each use a different flow cell, which is a glass surface coated with an arrangement of paired oligos that are complementary to the adapters added to your template molecules. **The flow cell is where the sequencing reactions take place**.



Multiplexing

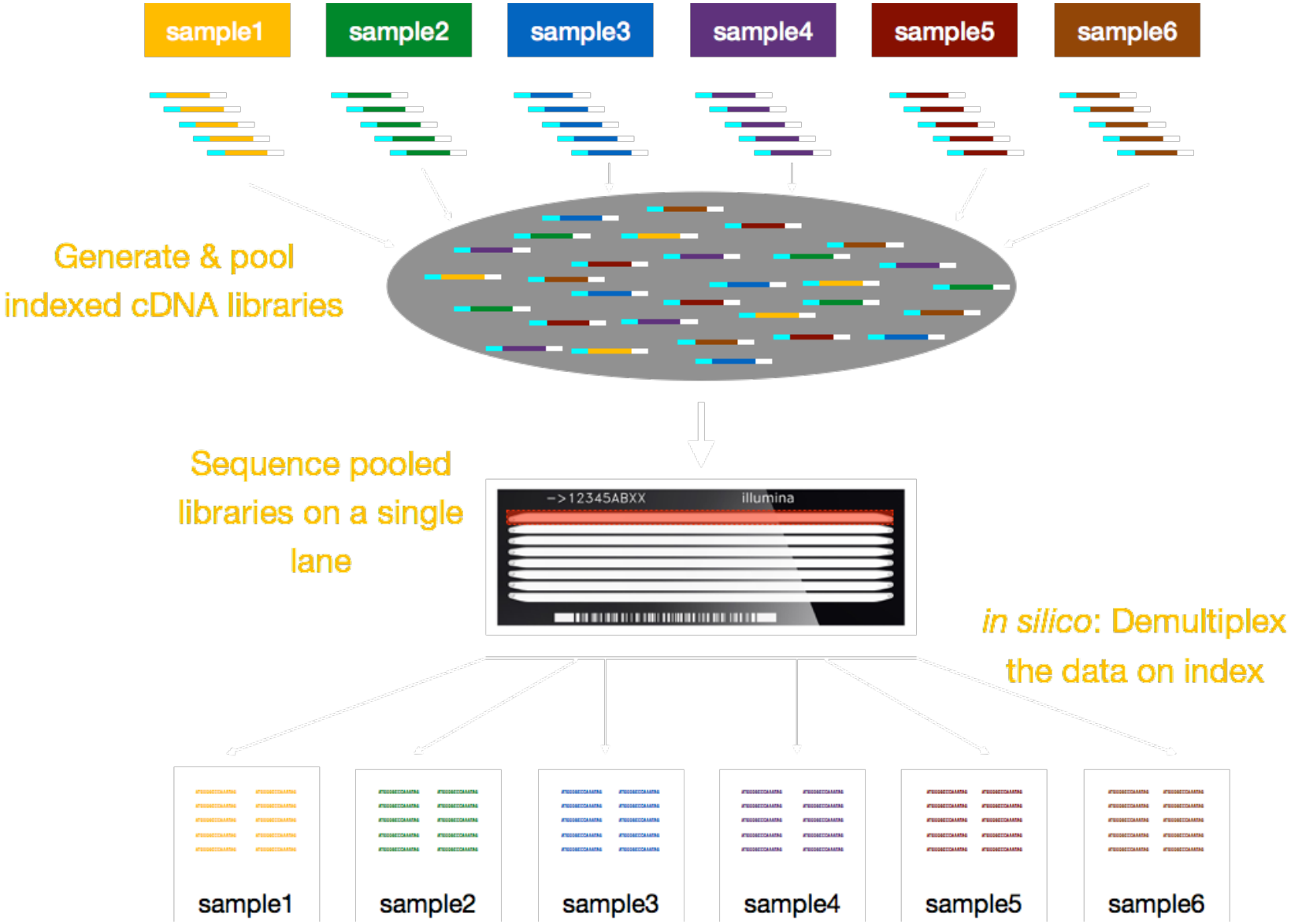
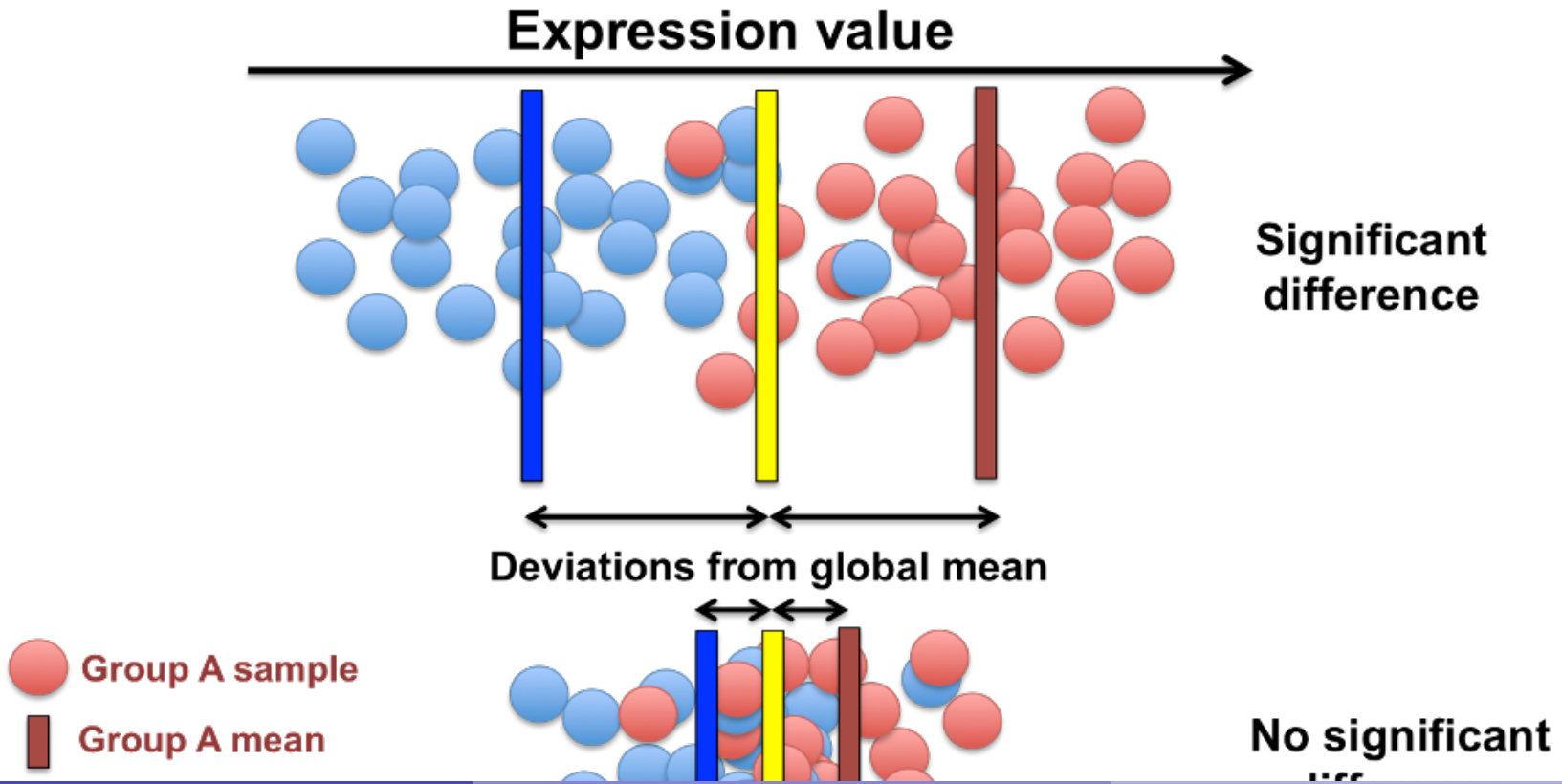


Figure 13: Demultiplexing

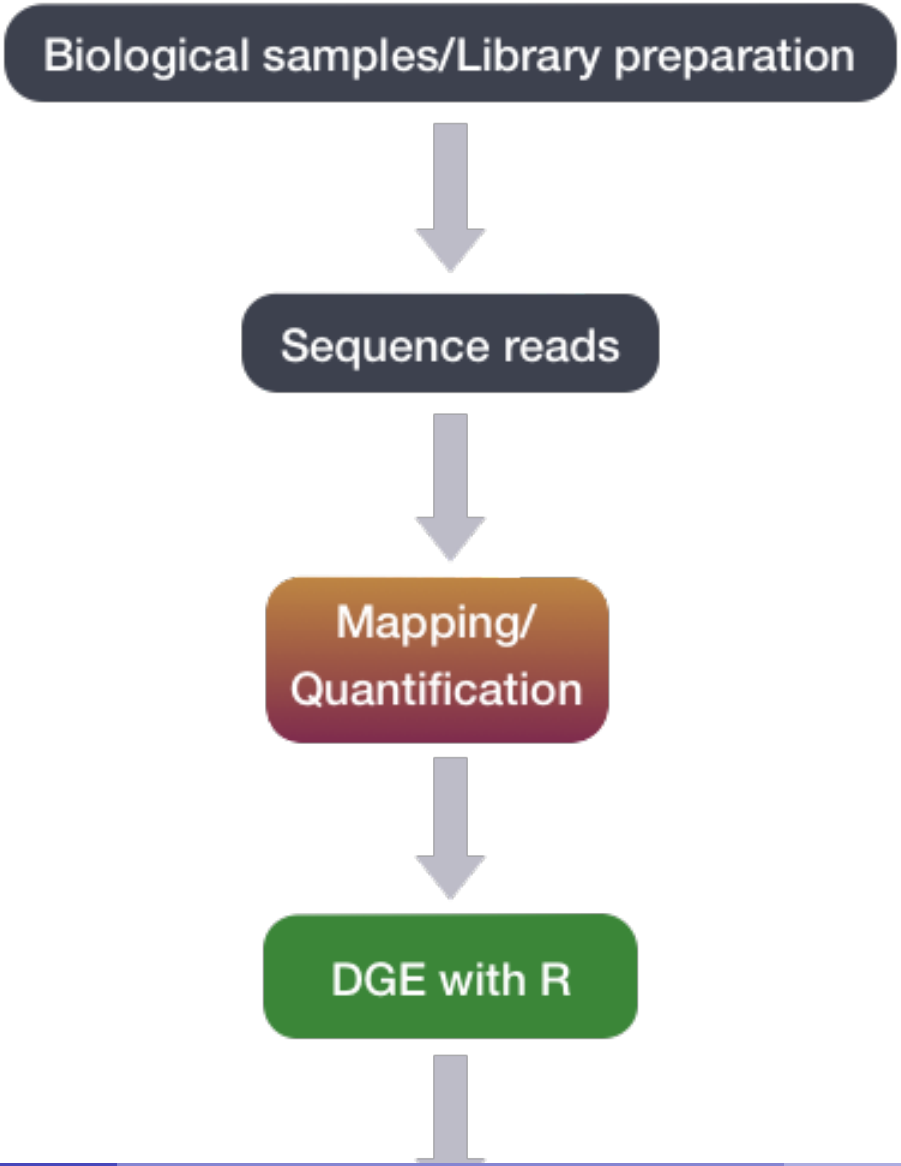
- **General gene-level differential expression:**
  - ENCODE guidelines suggest 30 million SE reads per sample (stranded).
  - 15 million reads per sample is often sufficient, if there are a good number of replicates (>3).
  - Use of an HiSeq or NextSeq, or NovaSeq for sequencing

Differential gene expression

Differential gene expression analysis allows us to explore the gene expression changes that occur in disease or between different conditions, by measuring the quantity of RNA expressed by all genes in each of the different conditions.



To perform differential gene expression analysis, we perform the following steps:



Experimental Design

These important considerations include:

1. Number and type of **replicates**
2. Avoiding **confounding** variables
3. Addressing **batch effects**

Replicates

Experimental replicates can be performed as **technical replicates** or **biological replicates**.

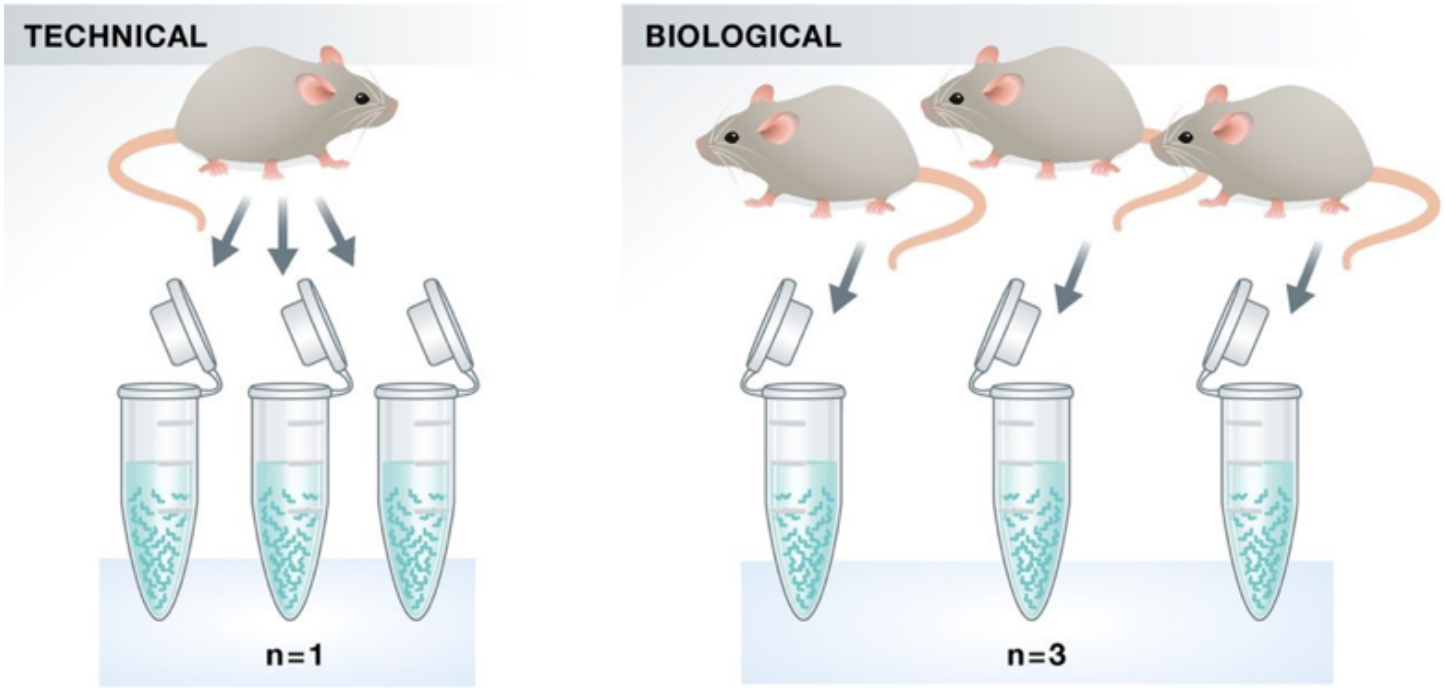


Figure 16: Biological Replicates

Image credit: Klaus B., EMBO J (2015) ***34****: 2727-2730*

- **Technical replicates:** use the same biological sample to repeat the technical or experimental steps in order to accurately measure technical variation and remove it during analysis.
- **Biological replicates** use different biological samples of the same condition to measure the biological variation between samples.

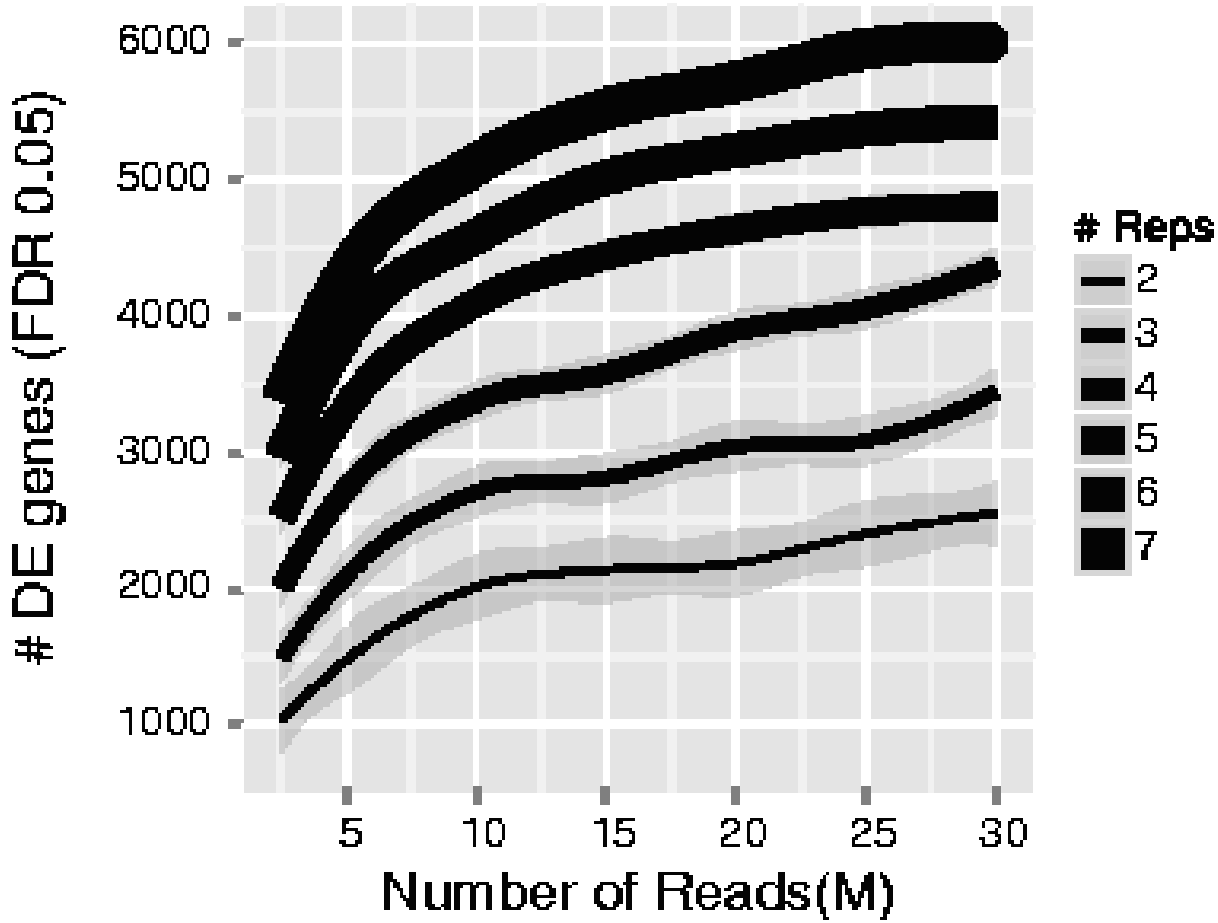


Figure 17: DE Replicates

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

As the figure above illustrates, **biological replicates are of greater**

Confounding

A confounded RNA-Seq experiment is one where you **cannot distinguish the separate effects of two different sources of variation** in the data.

For example, we know that sex has large effects on gene expression, and if all of our *control* mice were female and all of the *treatment* mice were male, then our treatment effect would be confounded by sex. **We could not differentiate the effect of treatment from the effect of sex.**

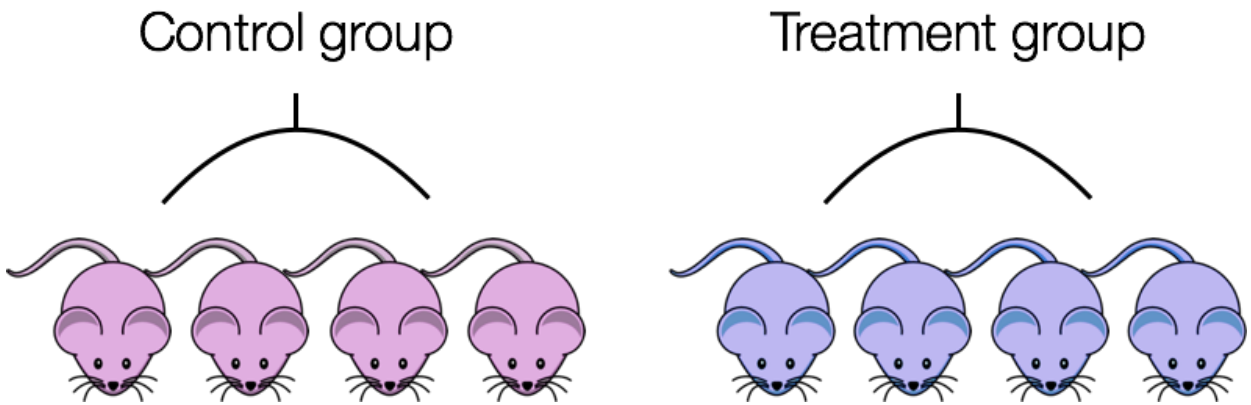


Figure 18: Confounding Variables

To AVOID confounding:

- Ensure animals in each condition are all the **same sex, age, litter, and batch**, if possible.
- If not possible, then ensure to split the animals equally between conditions

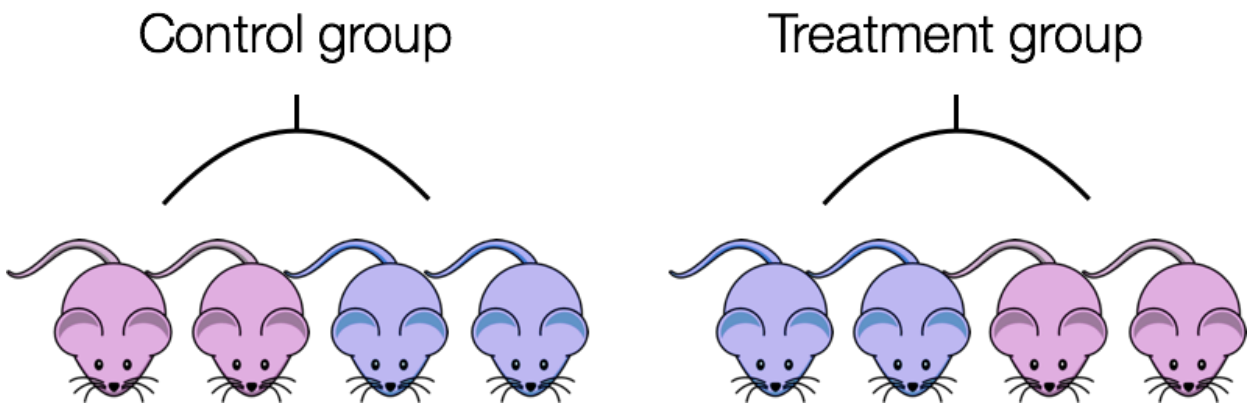


Figure 19: Non Confounded Design

Batch effects

Batch effects are a significant issue for RNA-Seq analyses, since you can see significant differences in expression due solely to the batch effect.

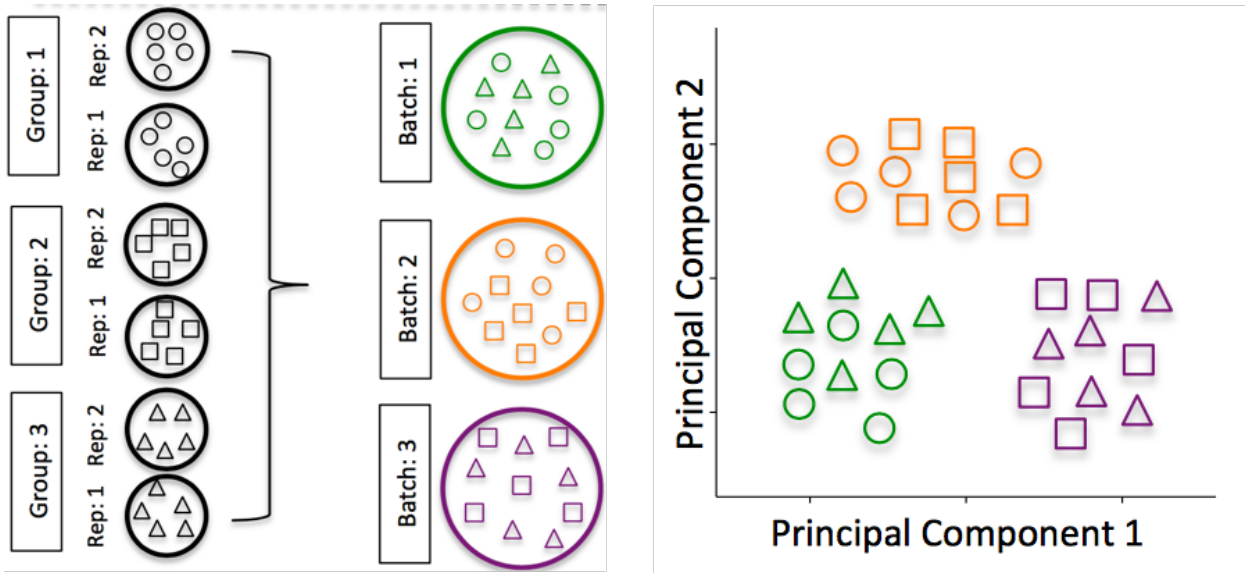


Figure 20: Non Confounded Design

Image credit: Hicks SC, et al., bioRxiv (2015)

Citation

This lesson has been developed by members of the teaching team at the Harvard Chan Bioinformatics Core (HBC). These are open access materials distributed under the terms of the Creative Commons Attribution license (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.
Authors: Mary Piper, Meeta Mistry, Radhika Khetani
Other sources - https://umich-brcf-bioinf.github.io/rnaseq_demystified_workshop/site/Module3a_Design_Prep_Seq#2_Experimental_Design_and_Practicalities