1. **What is RNA-Seq?**

RNA-Seq is a method for transcriptome profiling that uses next generation sequencing technologies. RNA-Seq provides a comprehensive, quantitative, and unbiased view of RNA transcripts within every sample. It is the most powerful tool currently available for analyzing gene expression.

2. **Name a difference between mRNAseq and smallRNAseq.**
Name a difference between bulk sequencing and single cell sequencing.

mRNAseq = focuses on quantifying the expression of mRNAs (protein coding genes).
The long mRNA is excised into small fragments which are then amplified. Due to sequencing bias the coverage across a transcript is not uniform, even if the probability of excision is uniform across the transcript.
smallRNAseq = focuses on small, non-coding RNAs.
The fragments are shorter than the sequenced read i.e. the adapter is present and needs to be trimmed.
The differential expression analysis follows the same principles as for mRNAseq
Single cell RNAseq = the libraries are specific per cell. The sequencing depth is shallower than for bulk mRNAseq, but the number of samples can be several orders of magnitude higher.

3. **Name few applications for RNAseq experiments**
[1] Measure gene expression. Assess the distribution of signal across transcripts
[2] Discover and annotate complete/ new transcripts.
[3] Characterize alternative splicing and polyadenylation.
[4] infer regulatory interactions and build Gene Regulatory Networks

4. **Describe few characteristics of qualitative and quantitative RNAseq experiments.**
Qualitative data includes identifying expressed transcripts, and identifying exon/intron boundaries, transcriptional start sites (TSS), and poly-A sites. We refer to this type of information as "annotation".
Quantitative data includes measuring differences in expression, alternative splicing, alternative TSS, and alternative polyadenylation between two or more treatments or groups. We focus specifically on experiments to measure differential gene expression (DGE).

Discuss biological replicates, transcript coverage, sequencing depth, stranded libraries, read lengths.

5. **State two library preparation methods. Provide two examples illustrating the advantages of using one option or the other.**

Coverage depends on the method used to prepare the library.
[1] **oligo-dT priming** for first-strand synthesis can be used to accurately annotate 3'-ends but often fails to evenly cover the 5'-end.
[2] **Random priming** suffers from uneven coverage due to sequence/structure priming biases, and can result in poor coverage of either of the ends.
Regardless of the method used to prepare the libraries, we observe uneven coverage of individual transcripts. To obtain reads that span the entire transcript more reads are required (i.e. deeper sequencing).

6. **State the sources of variance associated with counts. Describe one of the sources in detail.**
Sources of variances associated with the counts:
Sampling variance: the millions of reads represent only a small fraction of the nucleic acid that is actually present in the library.; also a sample = snapshot in time
Technical variance: Library preparation and sequencing procedures involve a series of complex chemical reactions which all contribute to between-sample variance.

Biological variance: the aim is to measure differences between individuals.
Biological systems are inherently complex and very sensitive to perturbations.
Bio-variance = the nascent variance that is present within a treatment or control group.

Technical replication : sequencing multiple libraries derived from the same biological sample.
Non-biological variation in estimated transcript abundance across these samples can arise from unintended differences during library preparation or sequencing.
Biological replication: sequencing multiple libraries derived from synonymous biological samples.
Assess variation among different individuals or tissues, usually with respect to experimental treatments.
Usually biological variation is large relative to technical variation.

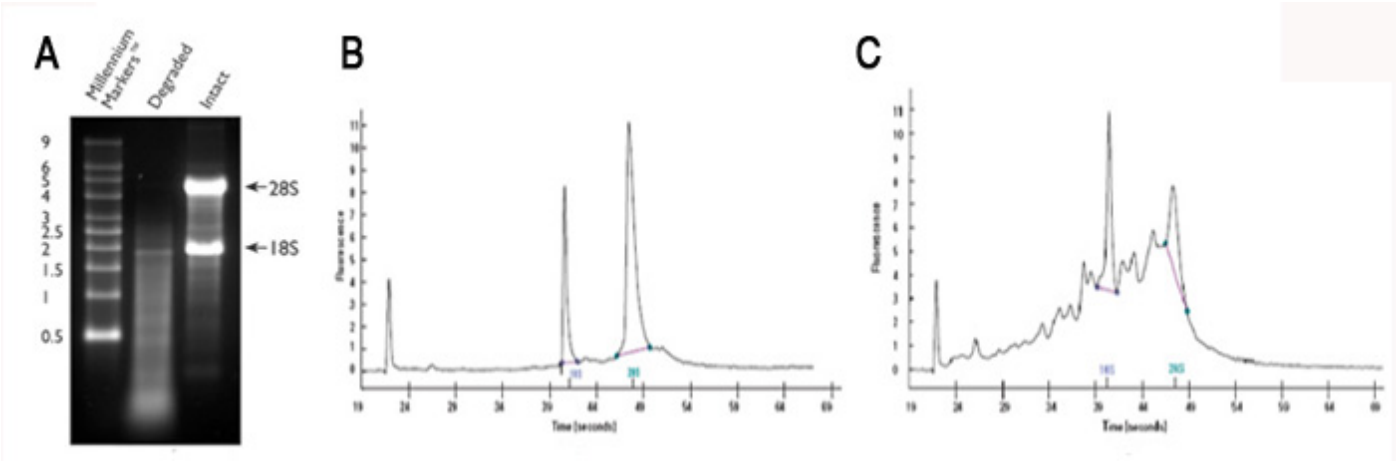
7. **How many reads do you need for an experiment?**

Variation due to the sampling process largely contributes to the total variance among individuals for transcripts represented by few reads. This means that identifying a treatment effect on genes with shallow coverage is not likely amidst the high sampling noise

Discuss size of the genome, transcriptome, observed variability between individuals, size of the measured effect. Also mention the presence of technical noise (high overview description is fine).

The number of reads required depends upon the genome size, the number of known genes, and transcripts. Generally, the recommendation is 5-10 million reads per sample for small genomes (e.g. bacteria) and 20-30 million reads per sample for large genomes (e.g. human, mouse). Medium genomes often depend on the project, but it is generally recommended between 15-20 million reads per sample. For de novo transcriptome assembly projects, the recommendation is ~100 million reads per sample.

8. **RNA preparation – describe the information provided in the figure below**

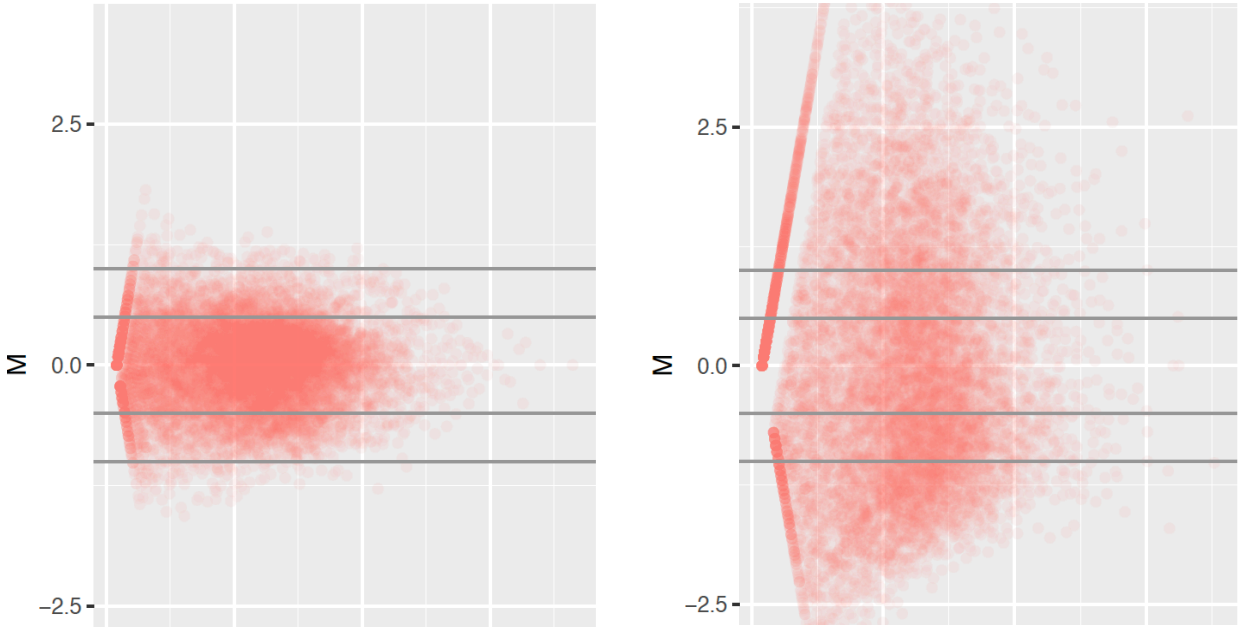


9. **Define the concepts of multiplexing and demultiplexing**

10. **Define randomisation and batch effects**

1. **What is an MA plot? What information can you obtain from it?**

2. **Discuss the following MA plots. What type of technical issues do these illustrate?**



3. **List three standard quality checks and describe one of them in detail.**

Density plots (focus on the localisation of signal and noise and similarity across samples)
PCAs (similarity across samples, incremental approach to avoid the noise effect)
Jaccard Similarity index (JSI - similarity across samples, incremental approach to avoid the noise effect)

4. **Define noise. Comment on the consequences of including/ excluding the noise from an analysis (both qualitative and quantitative).**

5. **Define normalisation. What is its role?**

6. **List few examples of parametric and non-parametric nomalisations. Discuss one of each in detail.**

7. **Using a list of differentially expressed genes, describe an approach to link these back to the biology.**

8. **What is a microRNA? How is it linked to mRNAs?**

9. **List examples of public databases containing additional information on coding or non-coding RNAs, or on the interaction between them.**

Ensemble for an overview of both coding and non-coding genes.
MiRbase = database of microRNAs
rFam = RNA database
mirTarBase = database describing the interactions between microRNAs and mRNAs

10. **List the current Gene Ontologies, briefly describe one of them.**

11. **List two pathway databases, briefly describe one of them.**

12. **Describe the hypergeometric test used for gene enrichment analysis. What background set should be used; provide an example to illustrate the importance of the background set.**