Exam 2020/2021

Final Exam Part 1

1. Mention one "Pro" and one "Cons" for each of the six DNA-seq techniques below:

	Template amplification	Sequencing reaction	Most representative feature	Pro	Cons
Sanger sequencing	Cloning/Emulsion PCR	Nucleotide addition	- 1st ever gene sequencing technique - Each newly added ddNTP is labeled with a different fluorescence that, by capillary electrophoresis, can be detected by that particular nt.	- Long reads - ↑ quality reads - ↓ error rate	- ↓ throughput - Expensive - Can just read one sequence at a time
454/Roche	Emulsion PCR	Nucleotide addition	- Pyrosequencing (sequencing by synthesis) - It detects pyrophosphates (light) that are released on nt incorporation.	- Long reads - Very fast (can make many reads at a time)	- Homopolymer problem - Expensive - Sequence result can contain indels
Illumina	Solid-phase bridge amplification	Cyclic reversible terminator (CTR)	- Uses a cell flow where DNA is sequenced. Also uses paired-end sequencing, doubling the amount of sequence can be generated at a time. - Important for some applications (transcript analysis, structural variants)	- ↑ throughput - ↑ quality reads - Large sequences - Can read several sequences at a time - Cheap	- ↑ error rate, with repetitive sequences - Short reads
Ion Torrent	Emulsion PCR	Nucleotide addition	- It counts how many H+ ions are released when a dNTP is added to a DNA in the complementary strand (pH change). - This is detected in a semi-conductor plate, which converts the reactions into digital information.	- Real-time sequencing - No optical machines - Cheap	- Expensive - Homopolymer problem (when bases are repetitive consecutively)
Pacific Biosciences	Not applicable	Real-time long-read sequencing	- Multiple sequencing protocols are possible: standard sequencing (long reads in one go), circular consensus sequencing (same DNA molecule multiple times), strobe sequencing (long and short strobe reads)	- No amplification - Long reads - ↑ accurate long reads (HiFi)	- ↑ error rate - Expensive
Oxford Nanopore	Not applicable	Real-time long-read sequencing	- Nts are passed through a nanopore and bases are identificated through differences in conductance (changes in electrical current) of nucleic acids.	- No amplification - Can sequence RNA and DNA - Portable - Long reads	- ↑ error rate

2. You want to sequence an eukaryotic genome never sequenced before. Your budget is limited and you decide to make a whole-genome shotgun sequencing with a next-generation sequencing technique. If you could choose one sequencing technique, which one would you recommend? Why?

Since my budget is limited, I would recommend Illumina NGS technique because it is a cheap technology and it is the most used. Also, if we are doing WGS sequencing, we will have a big amount of reads that we could use for detecting and fixing any possible errors that Illumina reads may have.

Would it be a good idea to combine two sequencing techniques? Which ones would you combine? Why?

If our budget is limited, it would not be a good idea to combine two sequencing techniques. But if the budget allow us, I think it could be a good idea to combine Illumina, that provides high-quality short reads and is cost-effective, with PacBio, since it offers long reads for resolving complex genomic regions. The combination allows for error correction using Illumina's accuracy and improves the quality and completeness of the genome assembly.

3. It is time for assembly. Which assembly strategy are you going to follow? Why?

I would follow a *De novo* assembly because it is the first time we are sequencing this eukariotic genome, so we don't have any reference genome in order to do the mapping.

4. You get two separate assemblies of your sequencing data, made by two different assembly software. According to the metrics shown in the table below, which assembly looks best? Why?

I would choose ABySS because of the N50 size value. We know that we will have more than 50% (60 000 aprox) of the contigs with size 7 338 bp, which is relatively close to the size of the longest contig if we compare this same values from Trinity. There is less variation of size from 7 338bp to 21 684bp than from 17 425bp to 468 339bp.

5. What do you need to form scaffolds? Briefly explain the process

In order to form scaffolds, first of all we will need several reads that will form contigs. Then, this contigs will be joined by paired-end and we will have finally our scaffold.

6. Paired-end mapping (PEM) is another application of DNA sequencing. Describe the aim and procedure of the PEM technique.

PEM is a DNA sequencing technique used to determine the relative positions of DNA fragments within a genome by creating a genomic library and performing pair-end sequencing.

1. Construction of a genomic library of DNA fragments of a certain size.
2. Pair end sequencing.
3. Mapping to a reference genome.

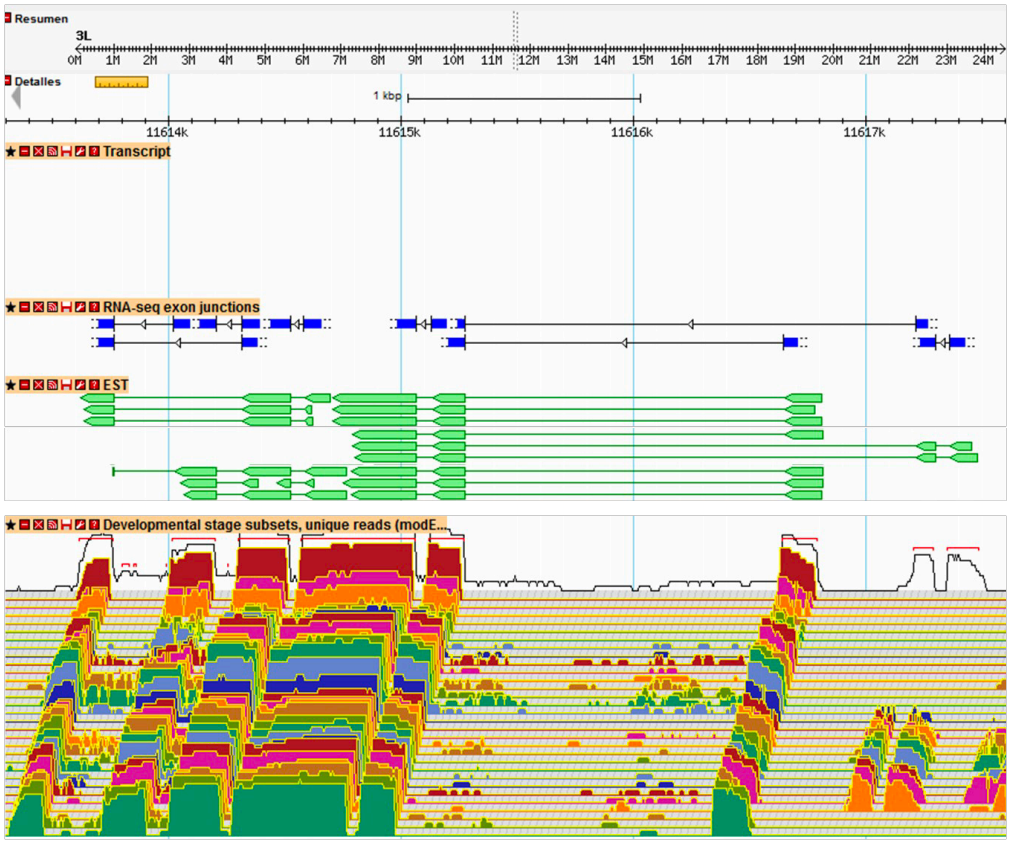
7. Explain what information does RNA-seq transcriptomic data provide you.

From RNA-seq transcriptomic data we can:

- Catalog the different types of RNA: mRNA, non-codingRNA (ncRNA), smallRNA (sRNA)
- Study the structure of the genome: transcription start and end, 5' and 3' UTR, splicing patterns, post-translational modifications
- Analyze the transcript expression over development or under certain physiological conditions.

8. The figure below displays EST and RNA-seq data mapped to a given genomic region.

- How many genes does the genomic region contain? 1
- Do/does the gene(s) show(s) alternative splicing? Draw all the transcripts in the reserved space within the figure. Yes
- What alternative splicing mechanisms are used to generate the different transcripts? Enumerate them and mark the place where they occur in the figure. We only have alternative transcription start and alternative transcription end.
- Do the different transcripts show differential GE throughout development? Yes, transcript 2 starts its expression earlier in the development.
- Are all the proteins encoded by the different transcripts identical? Mark in the figure the beginning and the end of the translation of each transcript. The proteins encoded by the different transcripts are not identical. As some transcripts start earlier/later than others and thus they are not including the same exons. So, the final proteins translated are not the same. But, from this data we cannot know the proteins this gene codify nor the translation start and end.



Final Exam Part 2

Bioconductor, experimental design for OMICS studies and differential expression analysis.

1. The code below illustrates a typical use of Bioconductor packages for microarray analysis. Indeed it is something you have used in class. Explain briefly (not more than 2-4 lines) what does each code chunk do.

```
### chunk 1
    library(GEOquery)
    gse <- getGEO("GSE12345")
    esetFromGEO <- gse[[1]]
```

- First load the Bioconductor package “GEOquery”.
- The `getGEO()` function is used to retrieve the microarray dataset with the accession number "GSE12345" from the GEO database.
- Finally, assign to the esetFromGEO variable the previously obtained dataset but removing its first line, which contained the zip name of the file and can cause problems in the later analysis.

```
### chunk 3
    design_matrix <- model.matrix(~0+targets$groups)
    cont.matrix <- limma::makeContrasts(TreatVSCont, levels=design_matrix)
```

- With the model.matrix() function, a design matrix is created taking in account targets and groups we have assigned.
- Then, from the limma package, call the makeConstast function and create a contrast matrix for the comparison of TreatVSCont (treatment and control groups) using values from the previously created design matrix.

```
### chunk 5
    topTab_AvsB <- topTable (fit.main, number=nrow(fit.main),
    coef="TreatVSCont", adjust="fdr");
    selected <- topTab_AvsB [topTab_AvsB$adj.P.value < 0.05,]
```

- Create a toptab with our fitted values from the TreatVSCont comparisson and only select those with a Pvalue higher than 0.05.
- The argument `number=nrow(fit.main)` specifies to return results for **all** genes and the `adjust="fdr"` option performs multiple testing correction using the false discovery rate (FDR) method.

2. A researcher has done a microarray analysis to compare GE between healthy donors with patients affected from endometriosis in three types of populations. She has done 12 microarrays (or RNseq runs, the design would be the same) 6 from patients with endometriosis and 6 from healthy donors. In each group there are 3 different cell populations call them A, B, and C.

The researcher wishes to make the following comparisons: (i) Compare healthy vs affected in each population. (ii) Compare population A vs B in affected patients.

1. Imagine Samples are named as [H, A][A,B,C][1,2,3]. Write down an appropriate “targets” table to describe the experimental layout for this study

Targets (H, A)

HA1	HA2	HA3
HB1	HB2	HB3
HC1	HC2	HC3

AA1	AA2	AA3
AB1	AB2	AB3
AC1	AC2	AC3

2. Write down the design matrix associated with this study design. HINT: Do as we have done in class and combine the two main factors into a single one

Design matrix
H=0, A=1;
A=0, B=1, C=2;
1=0, 2=1, 3=2

0	0	0
0	0	1
0	0	2
0	1	0
0	1	1
0	1	2
0	2	0
0	2	1
0	2	2
1	0	0
1	0	1
1	0	2
1	1	0
1	1	1
1	1	2
1	2	0
1	2	1
1	2	2

3. Build the contrast matrix needed to do the comparisons described above.

Contrast matrix
(doing everything in the same population)
α1 = E(logHA1) ; α4 = E(logAA1)
α2 = E(logHB1) ; α5= E(logAB1)
α3 = E(logHC1) ; α6 = E(logAC1)

$$\begin{pmatrix} \beta_1^1 \\ \beta_2^1 \\ \beta_3^1 \end{pmatrix} = \begin{pmatrix} -1 & 0 & 0 & 1 & 0 & 0 \\ 0 & -1 & 0 & 0 & 1 & 0 \\ 0 & 0 & -1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \\ \alpha_4 \\ \alpha_5 \\ \alpha_6 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \varepsilon_4 \\ \varepsilon_5 \\ \varepsilon_6 \end{pmatrix}$$

COMPARISSONS 1: Healthy vs Affected
α1 vs α4
α2 vs α5
α3 vs α6

COMPARISSONS 2: A vs B in affected patients
λ1 = E(logAA1)
λ2= E(logAB1)

$$\left(\beta_1^2\right) = \begin{pmatrix} -1 & 1 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix}$$

3. Using the previously defined matrices we have fitted a linear model to the data and have obtained one top tables (one for each comparison). Table below shows the results for a few genes arbitrarily selected

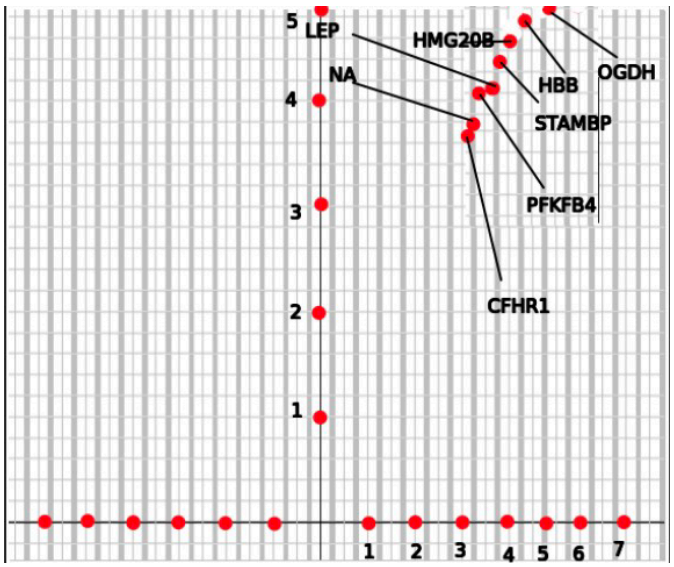
ID	adj.P.Val	P.Value	t	B	logFC	Gene.Symbol
201282_at	0.0074911	1.281663e-05	6.773810	-4.468046	3.031999	OGDH
217232_x_at	0.0088708	5.409429e-05	5.868340	-4.477733	2.969720	HBB
210719_s_at	0.0097080	6.672291e-05	5.742075	-4.479320	3.396428	HMG20B
235361_at	0.0187080	7.161377e-05	5.699805	-4.479866	3.382755	STAMBP
207092_at	0.0339791	1.693045e-04	5.197041	-4.486967	4.011486	LEP
228499_at	0.0459791	1.903590e-04	5.130082	-4.488003	2.510947	PFKFB4
235719_at	0.0907918	2.133572e-04	-5.065256	-4.489028	-2.156150	CYP4V2
230445_at	0.0997918	2.736306e-04	4.924930	-4.491321	2.381681	BTBD17
1569386_at	0.1299791	3.180007e-04	-4.840876	-4.492746	-2.210268	
215388_s_at	0.2397918	3.454539e-04	-4.794777	-4.493543	-3.641796	CFHR1

a. Which genes would you call differentially expressed? Explain the criteria used to make this.

The most differentially expressed genes are the one without name and CFHR1 since are the ones with smaller P-value.

b. b. Draft an approximate Volcano Plot depicting all the genes. Explain what does the Volcano plot represent and why is it interesting (HINT: B ~ -log(P.Value))

Volcano plot represent the P-values fro the expression of our genes. It is interesting because it is a very visual way for knowing with are the most expressed genes in our study.



Final Exam Part 3

1. Write a definition of epigenetics.

Epigenetics is the study of epigenetic variations in the genome of any organism. This genetic variations in DNA or in histones can affect the expression of certain genes without infering in the DNA sequence.

2. How would you expect to find a region of constitutive heterochromatin in terms of nucleosome positioning, DNA methylation and histone modifications?

In terms of nucleosome positioning, a region of constitutive heterochromatin is expected to have tightly packed nucleosomes, forming a condensed and inaccessible chromatin structure. With ↓ nucleosome turnover and ↑ nucleosome occupancy.

In terms of DNA methylation, a region of constitutive heterochromatin is associated with ↑ levels of DNA methylation. DNA methylation is involved in gene silencing and contributes to the stable repression of gene expression in heterochromatic regions.

Regarding histone modifications, a region of constitutive heterochromatin is characterized by specific histone modifications. These modifications include H3K9me3 and H3K27me3, which are associated with gene repression and the maintenance of heterochromatin structure.

3. Which is/are the chromatin state/s most highly associated with the following histone modifications:
   - H3K36me3: transcriptional elongation
   - H3K27me3: repressed state
   - H3K27ac: active promoter and active enhancer
   - H3K4me3: active promoter
   - H3K9me3: heterochromatin

4. What kind of information is provided by ATAC-seq?

ATAC-seq provides information about the accessibility of chromatin regions in a genome. It identifies regions where the chromatin structure is open and accessible to regulatory proteins, such as TFs. ATAC-seq data reveals the locations of accessible chromatin regions, which can indicate active regulatory elements such as promoters, enhancers, and TF binding sites.

5. Rewrite the following terms in hierarchical order:

Nucleosomes, A/B compartments, FIREs, TADs, chromosome territories. (From bigger to small)

Chromosome territories > A/B compartments > TADs > FIREs > Nucleosomes

6. What are UMIs, and what purpose do they serve?

UMIs (Unique Molecular Identifiers) are short, random nt sequences incorporated into individual RNA/DNA molecules during library preparation. They are used for distinguishing and quantifying unique molecules within a sample, even if they have identical sequences. UMIs help to overcome PCR amplification biases and accurately measure the abundance of individual molecules, particularly in ↑-throughput sequencing applications.

7. What are the advantages of using nuclei instead of cells in single-nuclei RNA-seq?

Using nuclei instead of cells in single-nuclei RNA-seq offers several advantages:

- Nuclei are easier to isolate and more robust compared to intact cells, making the process more reliable.
- Using nuclei reduces bias during isolation, ensuring a more representative capture of different cell types.
- Nuclei can be extracted from frozen tissue, allowing researchers to utilize archived samples effectively.
- Analyzing nuclear RNA provides insights into transcriptional processes and splicing events.

8. What is the Louvain algorithm designed for in the context of single-cell?

The Louvain algorithm is designed for community detection in single-cell data. It is used to identify distinct cell populations or clusters based on similarities in GE profiles. It optimizes (maximizes) modularity, aiming to maximize the density of connections within communities and minimize connections between communities. By applying it, researchers can uncover the underlying cellular heterogeneity and classify individual cells into biological groups.

9. Pair:

Ion Source → electrospray, MALDI

Mass filter → TOF, Ion Trap, Quadrupole

10. Describe the purpose of the second MS step in an MS/MS applied to protein identification.

The purpose of the second MS (Mass Spectrometry) step in MS/MS (Tandem Mass Spectrometry) is to fragment specific peptide ions generated from the first MS step. This fragmentation produces smaller peptide fragments, which are then analyzed in the second MS step. By measuring the m/z ratios of these fragmented ions, the second MS step provides information about the sequence and structure of the peptides. This information is important for identifying the proteins present in the sample, as it allows for more accurate peptide sequencing and matching against protein DBs, improving the confidence and specificity of protein identification.

Exam 2021/2022

Part 1

The leafy seadragon (Phycodurus eques) is a marine fish related to the seahorse. It is the only member of the genus Phycodurus. It is native to the waters bordering the Southern and Western coasts of Australia, generally living in template and shallow waters. Your research group wants to collaborate with the Genome 10K project by sequencing the genome of this species for the first time. (total score: 20 points)

1. Your team is considering different technologies for sequencing to decide which one will be applied in the project. Mention one "Pro" and one "Cons" for each of the six DNA-seq techniques below: (+4 points)

2. As the budget is limited and your goal is to achieve a high-quality assembly, equivalent to the quality of the human reference genome, which sequencing technique do you recommend to your group? (Correct answer +1 point, incorrect penalty –0.25)
   - Oxford Nanopore

3. Finally, your group invests a large part of the budget in generating the genomic libraries and sequencing. Now you have in your hand billions of Illumina and Pacific Biosciences sequencing reads. Explain which will be the strengths of Illumina and Pacific Biosciences data, and why it is a good idea to combine both: (+2 points)
   - Illumina: provides high-quality short reads and is cost-effective (cheap)
   - Pacific Biosciences: it offers long reads for resolving complex genomic regions
   - Why combine both: combination of both allows for error correction using Illumina's accuracy, and improves the quality and completeness of the genome assembly.

4. It is time for assembly and you are still discussing which assembly software you are going to use. Which assembly strategy are you going to follow? Why? (Correct answer: +1 point, incorrect: penalty – 0.25)
   - De novo assembly

I would follow a *De novo* assembly because it is the first time we are sequencing this genome, so we don't have any reference genome in order to do the mapping.

5. Finally, you get two separate assemblies of your sequencing data, made by two different assembly software. The first thing you do is compare basic metrics between the two. According to the values shown in the table below, which assembly looks best? Why? (+1 point)

I would choose Velvet because of the N50 size value. We know that we will have more than 50% (60 000 aprox (mitad del numero total de contigs)) of the contigs with size 7 338 bp, which is relatively close to the size of the longest contig if we compare this same values from SOAPdenovo. There is less variation of size from 7 338bp to 21 684bp than from 17 425bp to 468 339bp.

6. Considering that you have assembled 132.13 Gb of sequencing data and that the estimated genome size of the leafy seadragon is 695 Mb, calculate the redundancy (coverage): (+1 point)

Coverage = 132.13*1000 / 695 = 190X

7. You decide to continue with one of the two previous assemblies. Now, to complete the assembly and form scaffolds it is essential to: (Correct answer: +1 point, incorrect: penalty – 0.25)
   - Sequence paired-end reads.

8. The sequencing of a diploid species such as the leafy seadragon reveals sites in the genome where the individual has two different alleles in the form of a polymorphism. How do you think these sites can be detected? (Correct answer: +1 point, incorrect: penalty – 0.25)
   - In the assembly, heterozygous sites have approximately half of the reads with one allele and the other half of the reads with the other allele.

9. A second stage of the genome project of the leafly seadragon is related to RNA-seq. Explain how will you process RNA-seq reads what information does transcriptomic data provide you. (+3 points)

To process RNA-seq reads in the leafy seadragon genome project:

- QC: assess the quality of the reads and perform trimming and filtering to remove ↓-quality or adapter sequences.
- De novo assembly: assemble the reads into longer contiguous sequences (contigs) assembly tools specifically designed for transcriptome assembly.
- Transcript quantification: map the RNA-seq reads back to the assembled contigs/transcripts to estimate their abundance. This can be done with alignment-based approaches (e.g., Bowtie).

- Differential expression analysis: Compare the abundance of transcripts between different conditions or treatments to identify genes that are differentially expressed.

From RNA-seq transcriptomic data we can:
- Catalog the different types of RNA: mRNA, non-codingRNA (ncRNA), smallRNA (sRNA)
- Study the structure of the genome: transcription start and end, 5' and 3' UTR, splicing patterns, post-translational modifications
- Analyze the transcript expression over development or under certain physiological conditions.

10. The figure below displays EST and RNA-seq data mapped to a given genomic region. (+5 points)
    - How many genes does the genomic region contain? 1
    - Do/does the gene(s) show(s) alternative splicing? Draw all the transcripts in the reserved space within the figure. Yes
    - What alternative splicing mechanisms are used to generate the different transcripts? Enumerate them and mark the place where they occur in the figure. We only have alternative transcription start and alternative transcription end.
    - Do the different transcripts show differential GE throughout development? Yes, transcript 2 starts its expression earlier in the development.
    - Are all the proteins encoded by the different transcripts identical? Mark in the figure the beginning and the end of the translation of each transcript. The proteins encoded by the different transcripts are not identical. As some transcripts start earlier/later than others and thus they are not including the same exons. So the final proteins translated are not the same. From this data we cannot know the proteins this gene codify nor the translation start and end.

Part 2

1. Describe the structure of the summarizedExperiment Bioconductor class. Which are its main differences with respect to the expressionSet class?

The summarizedExperiment class is an extension of the expressionSet class in Bioconductor, designed to store and manipulate high-throughput OMIC data. It includes additional slots for storing assay data (such as expression or methylation levels), sample-level metadata (colData), and feature-level metadata (rowData). This class provides a versatile framework for integrating and analyzing diverse types of data in a unified structure.

2. A researcher wants to perform an RNA-seq experiment, followed by a differential GE (DGE) analysis, to compare GE between lung cancer patients and healthy controls. Draw a possible workflow with the steps to carry out such study, starting from the biological question, until the interpretation of the results.
    1. Define biological question
    2. Experimental design
    3. RNA-seq library preparation
    4. Raw data processing: QC, mapping, quantification
    5. Statistical analysis (normalization, exploratory analysis class comparison/discovery/prediction)
    6. Differential GE analysis (e.g., edgeR or DESeq2)
    7. Biological significance interpretation (e.g., GO enrichment, Gene Set Enrichment Analysis, pathway analysis)

3. Among the following metrics: read counts, CPM, RPKM and TPM, select the most appropriate one to compare the expression of a given gene between two technical replicates. Justify your answer.

CPM is the most appropriate metric to compare the expression of a given gene between two technical replicates since it normalizes RNA sequencing data by scaling the read counts of each gene/transcript by the total number of reads in the sample and multiplying by a million.

4. Explain why raw read count data should not be directly modeled using standard (i.e. Normal) linear models. Enumerate two alternative strategies for this purpose.

Raw read count data should not be directly modeled using standard linear models because counts violate the assumptions of normality and homoscedasticity. Two alternative strategies are:

1) Negative Binomial regression, which accounts for overdispersion in count data
2) Using generalized linear models with a Poisson or Quasi-Poisson distribution to model count data.

5. Reason why lowly expressed genes across all samples should be removed prior to differential GE analysis.

Lowly expressed genes across all samples should be removed prior to DGEA to reduce **noise** and focus on genes that exhibit meaningful changes in expression. By excluding genes with low expression levels, the analysis can prioritize genes that are more likely to have biological relevance and provide more robust and interpretable results.

6. Describe the problem of overdispersion of read count data.

Overdispersion in read count data refers to the phenomenon where the observed variation in counts is higher than what can be accounted for by a standard Poisson or normal distribution assumption. It arises due to additional sources of variation or correlation within the data, leading to inflated variability estimates. This can result in incorrect p-values, increased false positives, and reduced statistical power in downstream analyses. Overdispersion in read count data is often addressed by modeling the data using a negative binomial distribution.

7. Which are the main differences between overrepresentation analysis (e.g. GO enrichment) and Gene Set Enrichment Analysis (GSEA)?

Overrepresentation analysis compares gene set presence in a differentially expressed gene list, while GSEA ranks genes by differential expression and assesses gene set enrichment across the ranked list. Moreover, overrepresentation analyzes individual gene significance, while GSEA considers the overall pattern of gene set enrichment.

8. Describe what you could do to identify a batch effect in your expression data.

Unwanted variation removal (UVR) is the process of identifying and removing non-biological sources of variability in RNA sequencing data. Methods like PEER, RUV, and SVA are used to correct for factors such as batch effects. UVR improves downstream analysis accuracy and enables the identification of biologically meaningful variation.

9. A researcher wants to study GE in Alzheimer's disease (AD), mild cognitive impairment (MCI) and healthy (H) conditions. He performs RNA-seq on three individuals per condition, followed by a DGE analysis (voom + limma, models without intercept term) to compare GE between all the conditions pairwise. Draw the corresponding design and contrast matrices.

Design matrix
AD=0, MCI=1, H = 2;
1=0, 2=1, 3=2

AD1	0	0
AD2	0	1
AD3	0	2
MCI1	1	0
MCI2	1	1
MCI3	1	2

H1	2	0
H2	2	1
H3	2	2

Contrast matrix

	AD vs MCI:
AD	-1
MCI	1
H	0

	AD vs H:
AD	-1
MCI	0
H	1

	MCI vs H:
AD	0
MCI	-1
H	1

10. In the previous study, after performing DGE analysis and adjusting for multiple testing via FDR, for the contrast AD – H, the researcher got the following results (only the top 6 genes are shown):

	logFC	AveExpr	t	P.Value	adj.P.Val
ENSG00000179299	-3.031999	3.641797	-6.773810	1.281663e-05	0.0074911
ENSG00000088827	3.396428	4.619954	5.742075	6.672291e-05	0.0097080
ENSG00000134755	4.011486	4.157974	5.197041	1.693045e-04	0.0339791
ENSG00000278195	-2.156150	2.609283	-5.065256	2.133572e-04	0.0907918
ENSG00000111335	2.210268	7.502138	4.840876	3.180007e-04	0.1299791
ENSG00000140443	-3.641796	6.203476	-4.794777	3.454539e-04	0.2397918

How many significant genes are there at 5% FDR? How many significant genes are over- and under-expressed in AD with respect to H? Which plot would you use to summarize the information contained in this table? Justify your answers.

To determine the number of significant genes at a 5% False Discovery Rate (FDR), we need to identify the genes with an adjusted p-value (adj.P.Val) below the threshold of 0.05.

→ Knowing this, we observe 3 significant genes at 5% FDR (first three).

To identify which genes are over- or under-expressed in AD compared to H, we can examine the sign of the log-fold change (logFC). If logFC is positive, it indicates upregulation (over-expression) in AD compared to H, whereas a negative logFC indicates downregulation (under-expression).

→ Knowing this, we observe 3 genes over-expressed and 3 genes under-expressed in AD compared to H.

To summarize the information contained in this table, we could use a volcano plot. It displays the log-fold change on the x-axis and the negative logarithm of the adjusted p-value on the y-axis for each gene. The volcano plot allows for visualization of both the significance (p-value) and magnitude (fold change) of GE differences, making it useful for identifying significant genes and their direction of expression change in a visually manner.

Part 3

1. How would you expect the promoter region of a highly transcribed genes in terms of nucleosome positioning, DNA methylation and histone modifications?

Nucleosome positioning in the promoter region of highly transcribed genes is characterized by an open chromatin structure. This leads to ↓ nucleosome occupancy around the TSS and a ↑ nucleosome turnover. The accessibility of the DNA in this region allows for the efficient binding of TFs and RNA polymerase, facilitating transcriptional activity.

DNA methylation levels in the promoter region of highly transcribed genes are ↓, promoting gene expression. Hypomethylation in these regions enhances accessibility to TFs and facilitates transcription initiation. DNA methylation at CpG sites can impede TF binding, but hypomethylation in highly transcribed genes enables efficient TF recruitment and transcription activation.

Histone modifications in the promoter region of highly transcribed genes reflect active transcriptional activity. These modifications include H3K4me3. Histone acetylation is associated with transcriptional activation and is found in the promoter region of actively transcribed genes. H3K4me3, located around the TSS, is a mark of active transcription and is involved in the recruitment of transcriptional machinery.

2. Methylation at CpG sites:
   - none of the above (occurs at similar extent in all organisms, is always associated to silencing of gene expression, is irreversible)

3. What solution do you know it is used in single-cell genomics to deal with the problem of PCR duplicates?

The problem of PCR duplicates is addressed by incorporating Unique Molecular Identifiers (UMIs). UMIs are short nt sequences that are added to each individual molecule of cDNA during the initial reverse transcription step. These act as molecular barcodes, allowing for the differentiation of PCR duplicates that arise during library amplification.
Therefore, duplicate reads originating from the same molecule can be identified and collapsed into a single read, reducing the impact of PCR amplification bias.

4. Why we digest proteins to peptides before MS instead of running the whole molecule?

Proteins are digested into peptides before Mass Spectrometry (MS) analysis because peptides are smaller and more amenable to ionization and fragmentation compared to intact proteins. Peptides are easier to separate and analyze, allowing for improved sensitivity and specificity in protein identification and characterization.

5. Which protein property is used to separate each dimension in a 2D-SDS-PAGE electrophoresis?

In a 2D-SDS-PAGE (Two-Dimensional Sodium Dodecyl Sulfate Polyacrylamide Gel Electrophoresis), proteins are separated based on their isoelectric point (pI) in the 1st dimension and molecular weight (MW) in the 2nd dimension.
The combination of these two properties allows for the separation of proteins into distinct spots on the gel.