EXAMEN FINAL OMICS 2022 (T_T)

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)

	Pro	Cons
Sanger		
454/Roche		
Illumina		
Ion Torrent		
Pacific Biosciences		
Oxford Nanopore		

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)

- □ 454/Roche
- □ Oxford Nanopore
- □ A combination of the two above
- □ The objective that you propose is not a realistic goal

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 :

Pacific Biosciences:

Why combine both :

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)

- □ Mapping against a reference
- □ De novo assembly
- □ Any of the two above
- □ Expression profiling

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)

	Velvet	SOAPdenovo
Number of contigs	120,479	47,571
N50 size (bp)	7,338	17,425
Longest contig (bp)	21,684	468,339

- □ Velvet
- □ SOAPdenovo

Why?

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)

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.
- □ eliminate repetitive regions.
- □ sequence the transcriptome by RNA-seq.
- □ compare contigs with a database of proteins of a nearby species.

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 Illumina reads, heterozygous sites have an intermediate coloration between the two nucleotides corresponding to the two alleles.
- □ 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.
- □ In the assembly, heterozygous sites have double the redundancy (coverage) than the rest.
- □ The sequencing and assembly of a diploid individual results finally in 2n chromosomes assembled separately, so that heterozygous positions correspond to the differences between the two chromosomes.

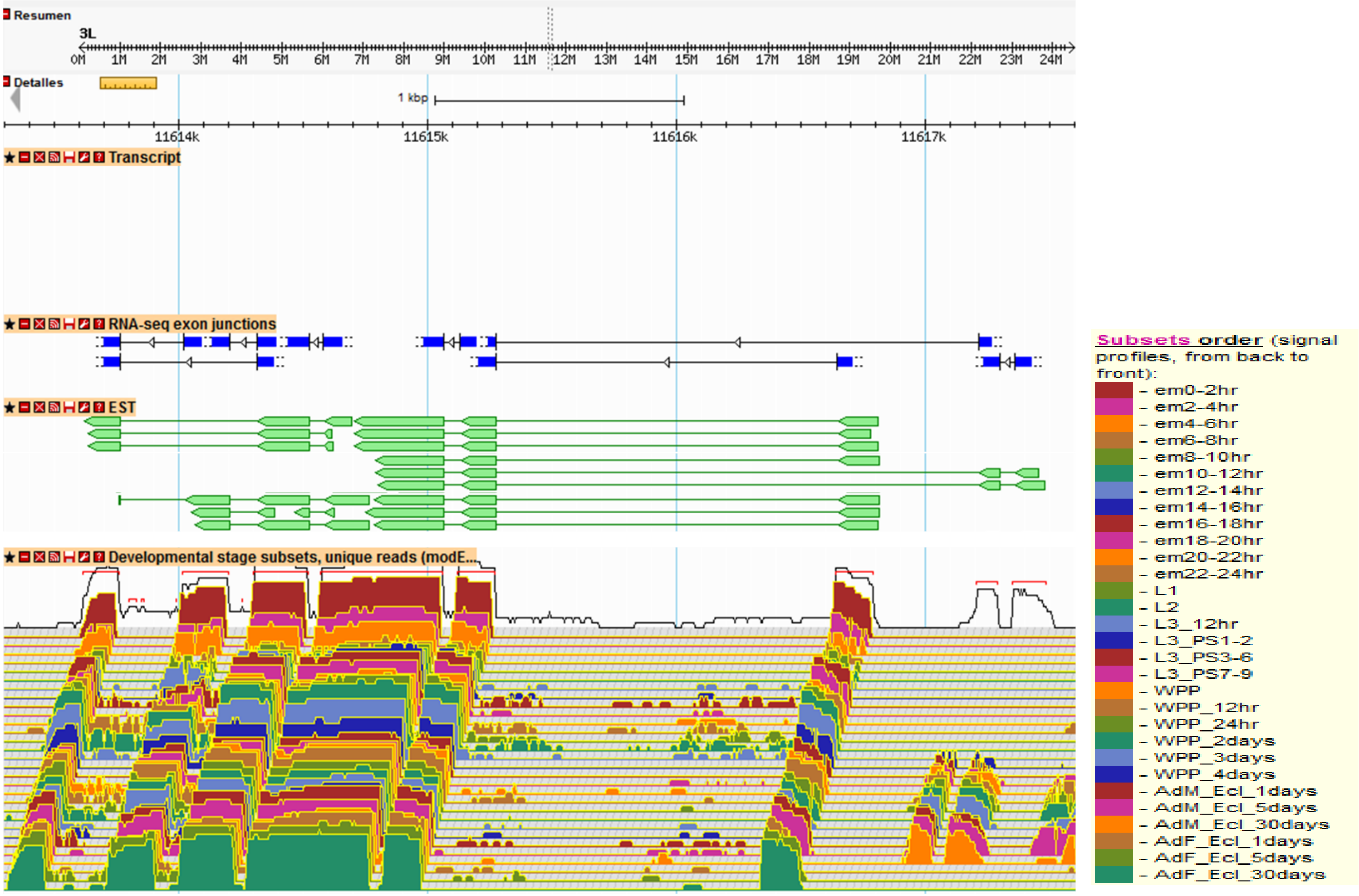
9. A second stage of the genome project of the leafly seadragon is related to RNA-seq. Explain how will you process RNA-seq reads what information does transcriptomic data provide you. (+3 points)

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?
- □ Do/does the gene(s) show(s) alternative splicing? Draw all the transcripts in the reserved space within the figure.
- □ What alternative splicing mechanisms are used to generate the different transcripts? Enumerate them and mark the place where they occur in the figure.

- □ Do the different transcripts show differential gene expression throughout 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.



part .2

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

2. A researcher wants to perform an RNA-seq experiment, followed by a differential gene expression (DGE) analysis, to compare gene expression 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.

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.

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.

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

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

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

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

9. A researcher wants to study gene expression 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 gene expression between all the conditions pairwise. Draw the corresponding design and contrast matrices.

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.

part .3

1 – Write a definition of epigenetics.

2 – How would you expect to find the promoter region of a highly transcribed genes in terms of nucleosome positioning, DNA methylation and histone modifications?

3 – Which is/are the chromatin state/s most highly associated with the following histone modifications:

- H3K36me3 :
- H3K27me3 :
- H3K27ac :
- H3K4me3 :
- H3K9me3 :

4 – Methylation at CpG sites:

a) occurs at similar extent in all organisms.
b) is always associated to silencing of gene expression.
c) is irreversible
d) none of the above

5 – Rewrite the following terms in hierarchical order:
Nucleosomes, A/B compartments, FIREs, TADs, chromosome territories.

6 – What kind of information is provided by ATAC-seq?

7 – What is the Louvain algorithm designed for in the context of single-cell?

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

9 – Why we digest proteins to peptides before MS instead of running the whole molecule?

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