Problem set 4

This problem set covers the content from week 4: RNA sequencing.

Tips and rules:

- You can answer in English or in Thai.
- There can be more than one correct answer. What I am looking for from you is not just the correct answer but the rationale for your answer.
- Please provide evidence of how you think and what sources of information you used.
- AI such as ChatGPT may be used. You can also work together with friends. But you must write the answer in your own words.
- Any incidence of plagiarism and copying of another student's work will be reported to the Graduate Affairs.

RNA sequencing experimental design

How would you design your RNA-seq sample preparation, sequencing, and analysis to best study the following biological systems?

	Which classes of RNA molecules are you interested in: messenger RNA, microRNA, and long-noncoding RNA? Is there a way to enrich RNA molecules?	Would you choose a long-read platform (SMRT-seq or Nanopore) or a short-read platform (Illumina)? Why?	Would you perform single-end or paired-end sequencing (for Illumina)? Why?
Q1: Study gene-level transcriptomic changes in cells treated with a drug.			
Q2: Detect novel alternative splicing events in a cancer tissue.			
Q3: Identify potential post-transcriptional factors in a liver disease			
Q4: Identify cancer-specific RNA transcripts that can be targeted for treatment			

Use the following information to answer **Q5-6**. You read from a paper that microRNA m1 and m2 can bind to the 3'UTR region of gene X and down-regulate its expression. Incidentally, gene X is linked to a rare disease in Thailand, and so you want to use test m1 and m2 as a potential treatment.

First, you collect expression data of gene X and microRNA m1 and m2 from a group of 5 local patients, as shown in the table:

Molecule	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5
Gene X	10.3	15.2	9.7	10.6	16
microRNA m1	7.3	4.8	8	6.9	5.1
microRNA m2	3.5	4.1	4.2	3.8	4.5

Q5: Let's assume that the technician forgot to tell you whether this data has been log-transformed. Based on the observed expression values in the table, what would be your guess?

Q6: Based on the observed expression values, which microRNA is promising (could be both)?

Q7: Explain why it is difficult to quantify the expression levels of individual isoforms of a gene using short read data. *Hint: think about what are the same and what differ among isoforms of a gene.*

Q8: Explain why raw read count (the number of sequencing reads that were mapped to a gene) is inappropriate for comparing expression levels across genes and samples. *Hint: think about what differs across genes and samples that can bias the sequencing read count.*

When analyzing human RNA-seq data, where both reference genome and transcriptome are available, what would be the pros and cons of using each of them, and what are the differences between the two bioinformatics pipelines?

	Using reference genome	Using reference transcriptome
Q9: What are the pros of using such reference database?		
Q10: What are the cons of using such reference database?		
Q11: What are the differences in sequence alignment algorithms?		

Differential expression and functional enrichment analyses

Summarize the concepts and compare the differences between several ways to formulate the statistical tests for differential expression

Q12: Wald test	
Q13: Likelihood ratio test	
Q14: Permutation test	

Q15: In the *kallisto-sleuth* pseudoalignment pipeline, the authors explain that the bootstrapping of RNA-seq data performed by *kallisto* captures the technical variances. Discuss why bootstrapping can capture some technical variances. Also discuss whether you think bootstrapping captures all technical variances.

Q16: When examining differential expression result via a volcano plot, we typically apply cutoff to both the fold change and the p-value. Explain why we need to consider both factors (i.e., why is it inappropriate to consider only one of them)?

Provide an example (or describe a situation) of gene expression result that give good fold change but poor p-value.

Q17: Given the following gene count table for functional enrichment analysis. Calculate the p-value for deciding whether kinases are differentially expressed using the HYPGEOM.DIST function in MS Excel. *Hint: read the online manual, and think carefully about the relationship between p-value and cumulative density*

	Kinase	Non-kinase
Differentially expressed	50	350
Other genes	150	5450

Use the following information for answering **Q18-19**. When performing functional enrichment, most bioinformatic tools will ask for an input "background gene list" – or the list of genes that should be considered as the universe for the analysis. The default background gene list is often the set of all genes in an organism's genome.

Q18: When should you change the background gene list from the default setting? What would you set it as? *Hint: think about the platform that you used to obtain the transcriptomics data.*

Q19: What harm can happen if you set the background gene list incorrectly?

Q20: A benefit of Gene Set Enrichment Analysis (GSEA) technique is that it can distinguish up-regulated and down-regulated pathways. However, we might argue that a differentially expressed pathway may consist of both up-regulated and down-regulated genes (such as when one gene represses another). How would you modify the input for GSEA technique to uncover pathways with mixed up-regulation/down-regulation signals?