Problem set 3

This problem set covers the content from week 3: metagenomics and microarray.

Tips and rules:

- You can answer in English or in Thai.
- There can be more than one correct answer. What I am looking 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.

Metagenomics

Q1: Explain how the **rarefaction** procedure tells you whether you have sequenced enough reads to study a metagenomics sample.

Q2: Explain what **read binning** (grouping of reads derived from the same genomes) is and why it should be performed.

Q3: Explain why DNA *k*-mer profile can serve as a signature for distinguishing reads from different species. What do you think is the limitation of using DNA *k*-mer profile to annotate species in metagenomics data?

Q4: Compare the pros and cons of 16S rRNA amplicon sequencing and shotgun metagenomics for studying microbiome samples.

	16S rRNA sequencing	Shotgun metagenomics
Pros		
Cons		

Q5: Plot the graph of Simpson's index $D = \frac{1}{\sum p_i^2} = \frac{1}{p^2 + (1-p)^2}$ for a system with two taxa, where the abundance of one taxon is p and the other is $1 - p$, with p ranging from 0 to 1. From the graph, where does the function reach its maximum and minimum values? From the graph, explain why Simpson's index is a valid indicator of the diversity of the taxa.

Q6: Calculate the 2-mer profile for the sequence AAAGCTAGGATCAGTCAGACT.

2-mer	**AA**	**AC**	**AG**	**AT**	**CA**	**CC**	**CG**	**CT**
Count								
2-mer	**GA**	**GC**	**GG**	**GT**	**TA**	**TC**	**TG**	**TT**
Count								

Q7: Given what you learned about the limitation of the **Bray-Curtis dissimilarity** for measuring similarity in taxonomic profiles between samples, how would you pre-process your data (in the form of taxonomic profiles) to minimize the impact of the limitation of the **Bray-Curtis dissimilarity**.

Q8: Explain what **Metagenomics Assembled Genomes** (MAGs) are. Explain why taxonomic profiling and functional annotation should be performed on MAGs rather than on individual reads.

Q9 [Extra]: What are the main differences between **Markov Model** and **Hidden Markov Model** (HMM)?

Microarray and Nanostring

Q10: Compare the pros and cons of microarray, Nanostring, and RNA-sequencing. When should you select each of them for your transcriptomics analysis?

	Microarray	Nanostring	RNA-seq
Pros			
Cons			

Q11: Explain why multiple microarray probes were designed for each gene. Explain why some probes with mismatches were intentionally included.

Q12: Multi-channel microarray allows multiple samples to be hybridized to the same chip to reduce technical variances. What specific technical variances were removed?

Q13: Why is Nanostring technique limited to ~800 gene targets? What would you do if you want to query a set of 1,000 specific genes in samples? Propose at least two approaches and discuss how you would decide which one to use.

Answer the following questions using data from https://github.com/cmb-chula/comp-biol-3000788/blob/main/problem-sets/Gene_expression_toy_data.xlsx

Q14: Explore **Dataset A**.

Identify whether the measured gene expression values are distributed following a normal or log-normal distribution. Plot histogram(s) of the distribution of the data to support your answer.

If you were asked to summarize the **average** expression values for Control and Treated conditions, will you use arithmetic mean or geometric mean? Why?

Q15: Explore **Dataset B**.

Identify whether the measured gene expression values are distributed following a normal or log-normal distribution. Plot histogram(s) of the distribution of the data to support your answer.

Perform *t*-tests to check whether Gene A and Gene B are differentially expressed between the Control and Treated conditions. If you believe the data are log-normal, transform them before performing the *t*-tests.

Report how you perform the tests and the resulting p-values.

Q16: The **P-values** sheet contain a collection of raw p-values derived from *t*-tests on 784 genes.

Use Bonferroni and Benjamini-Hochberg procedures to correct the p-values for multiple testing a at the alpha level of 0.05.

How many p-values remain statistically significant after performing each correction procedure?