

Practical Bioinformatics

Quiz-3

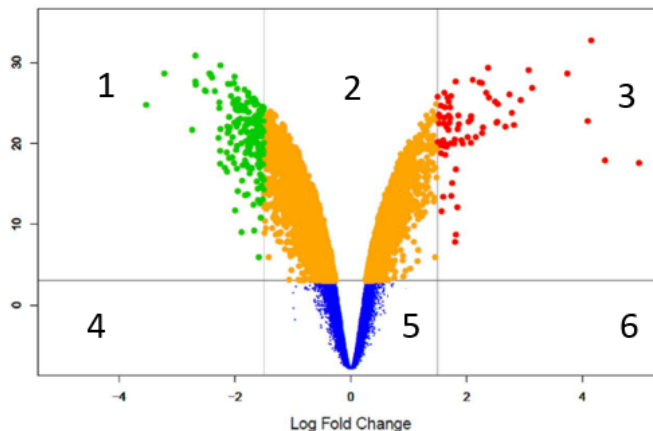
Multiple choice questions:

Total marks: 20

Each question carries 1 mark

1. What is the purpose of normalizing gene expression data?
 - a. **To compare gene expression levels across different samples**
 - b. To amplify gene expression signals
 - c. To sequence genes of interest
 - d. To create knockout mutants
2. Which are the following statements about principal components true?
 - a. **Each principal component is orthogonal to all others**
 - b. Each principal component is a linear combination of the original variables
 - c. Each principal component has the same variance as the original data
 - d. **Each principal component retains all information from the original dataset**
3. A researcher wants to test whether the mean weight of a certain breed of dogs is different from 50 pounds. What would be the null hypothesis (H_0)?
 - a. **The mean weight of the breed is equal to 50 pounds**
 - b. The mean weight of the breed is less than 50 pounds
 - c. The mean weight of the breed is greater than 50 pounds
 - d. There is no need for a null hypothesis in this case
4. How does the binomial distribution differ from the hypergeometric distribution?
 - a. **The binomial distribution models independent trials with replacement, while the hypergeometric distribution models dependent trials without replacement.**
 - b. The binomial distribution models dependent trials with replacement, while the hypergeometric distribution models independent trials without replacement.
 - c. Both distributions model independent trials with replacement, but the binomial distribution uses continuous data while the hypergeometric distribution uses discrete data.
 - d. Both distributions model dependent trials without replacement, but the binomial distribution uses continuous data while the hypergeometric distribution uses discrete data.
5. Basecalls in Sanger sequencing have an offset of 33, which explains:
 - a. Each quality score is represented by a single ASCII character
 - b. **Quality scores ranging from 33 to 126**
 - c. Quality scores ranging from 0 to 33
 - d. Quality scores are encoded using ASCII characters ranging from 0 to 33
6. What is the correct decision when the p-value is less than the significance level (α)?
 - a. **Reject the null hypothesis**
 - b. Fail to reject the null hypothesis
 - c. Accept the null hypothesis
 - d. Further data collection is required to make a decision
7. In a RNA-seq experiment, raw sequence data are in the following file:
 - a. SAM file
 - b. CRAM file
 - c. **FASTQ file**
 - d. BAM file
8. In the context of gene expression data, if you want to find out if the differentially expressed genes are associated with a certain biological process, you will choose the following approach:
 - a. Display the results in a heat map

- b. **Perform a gene set enrichment analysis**
 - c. Use a clustering method
 - d. Order the genes in a rank list according to their differential expression
- 9. What is the correct decision when the p-value is less than the significance level (α)?
 - a. **Reject the null hypothesis**
 - b. Fail to reject the null hypothesis
 - c. Accept the null hypothesis
 - d. Further data collection is required to make a decision
- 10.



Identify up and downregulated in significant region in this plot

- a. **1 & 3**
- b. 1 & 2
- c. 3 only
- d. 1, 3, 2

Close end:

***Question 3 carries 2 marks**

1. **Ontologies** refer to the relationships between concepts in a way that allows computational logical reasoning.
2. The formula for the phred quality score Q are logarithmically related to the base-calling error Probabilities P and defined as formula?

$$Q = -10 \log_{10} P.$$

This relation can be also be written as

$$P = 10^{\frac{-Q}{10}}.$$

3. The below statement is true and false for the FASTQ format when we read the 4 lines of the data?
 - a. The first line starts with & is a unique ID for the sequence. **False**
 - b. Third line is always a "+" character. **True**
4. What is the difference between a one-sample and two-sample hypothesis test?

Answer: A one-sample test compares a sample mean to a known population parameter, while a two-sample test compares the means of two independent samples.

5. Higher Phred scores correspond to lower sequencing quality. **False**

6. What is the primary objective of Principal Component Analysis (PCA) in bioinformatics?

Ans: In bioinformatics, Principal Component Analysis (PCA) primarily aims to reduce the dimensionality of complex biological datasets while retaining the most significant variation, facilitating visualization, clustering, and interpretation of underlying biological patterns or relationships among variables.

7. Explain Type II error in terms of null hypothesis.

Answer: Type II error occurs when the null hypothesis is not rejected when it is false, meaning a significant effect exists but the test fails to detect it.

8. Why is focusing on gene sets enrichment analysis considered a powerful approach in interpreting gene expression data?

Ans: Gene set enrichment analysis provides biological context by focusing on functionally related gene groups, reducing data complexity, and increasing statistical power to detect coordinated gene expression changes, aiding in robust interpretation of biological mechanisms from gene expression data.

9. Name any two methods to perform Gene Set Enrichment Analysis(GSEA).

Ans: GAGE, DAVID, KEGG