**BIM3007 Computational Genomics and Proteomics (2022-2023 Term 1)**

**Assignment #4**

**Deadline:** **23:59:59 25th December**, 2022 (Delayed submission is not allowed for any reason)

**Percentage:** 10%

**Purpose:** to enhance the learning outcomes for the topics in “RNA-Seq”, and “Protein Post-translational Modifications (PTM)”.

**1 [RNA-Seq] (40 points)**

1.1 In RNA-seq analysis, what is the program and significance of data quality control? **(10 points)**

1.2 After performing the data quality control for the RNA-Seq reads, we need to align the reads to the reference genome. What is the alignment program and what is the purpose of the alignment? **(10 points)**

1.3 After alignment step, we obtain bam files. What program help us to summarize the gene counts? **(10 points)**

1.4 What is the input and output of the standard RNA-seq analysis (featureCounts program analysis as the end point), and what kind of analysis can we do on its output? **(10 points)**

**2 [PTM Prediction] (60 points)**

2.1 Please write a python script to extract the positive dataset and the negative dataset **(13-mer)** from the raw dataset (**Blackboard**: Content/[Tutorials (Videos, PPTs, and other course materials)](https://bb.cuhk.edu.cn/webapps/blackboard/content/listContentEditable.jsp?content_id=_196812_1&course_id=_5863_1)/Tutorial Session: Preprocessing of training datasets/Ubiquitination\_sites.txt), and use CD-HIT program with cut-off values of sequence similarity of 80 % to remove the homologous sequences of both datasets.

Please provide a table for the data statistics of positive and negative sequences before and after removing the homologous sequences. **(30 points)**

2.2 After the removal of homologous sequences, please apply WebLogo tool to generate sequence logos for positive sequences and negative sequences, respectively. Then, please employ TwoSampleLogo tool to conduct the comparison of position-specific AAC between positive and negative data. **(15 points)**

2.3 Finally, Please give out a comparison of AAC (Amino acid composition) between positive and negative sequences in terms of bar-chart visualization. **(15 points)**

In RNA-seq analysis, what are programs for data quality control?

There are several programs and tools that are commonly used for quality control (QC) in RNA-seq analysis. These tools can be used to assess the quality of the raw RNA-seq data, as well as the quality of the alignment and the downstream analysis. Some examples of programs and tools for QC in RNA-seq analysis include:

1. FastQC: This tool is used to assess the quality of the raw RNA-seq reads. It generates a report that includes information about the quality of the reads, the sequence content, and potential biases.
2. RSeQC: This tool is used to assess the quality of the RNA-seq alignment. It generates a report that includes information about the mapping rate, the exon-intron structure, and the GC content.
3. MultiQC: This tool is used to combine the results from multiple QC tools into a single report. It allows you to easily compare the results from different samples or different analysis pipelines.
4. RNASeqQC: This tool is used to assess the quality of the RNA-seq data and the downstream analysis. It generates a report that includes information about the expression levels, the differential expression, and the functional annotation.

There are many other tools and programs available for quality control in RNA-seq analysis, and the specific tool you choose will depend on your specific needs and goals.

why should we conduct quality control

Quality control (QC) is an important step in any RNA-seq analysis because it helps to ensure that the data are of high quality and that the downstream analysis is reliable. There are several reasons why it is important to conduct QC in RNA-seq analysis:

1. Data quality: RNA-seq data can be affected by a variety of factors, such as sequencing errors, biases, and contaminants. QC helps to identify and address these issues, which can improve the accuracy and reliability of the analysis.
2. Alignment quality: Proper alignment of the RNA-seq reads is critical for accurate expression estimates. QC tools can help to assess the quality of the alignment, including the mapping rate and the exon-intron structure.
3. Differential expression analysis: Differential expression analysis is used to identify genes that are differentially expressed between two or more conditions. QC can help to ensure that the data are suitable for this type of analysis, and that the results are reliable.
4. Functional annotation: Functional annotation is the process of assigning biological functions to the differentially expressed genes identified in an RNA-seq analysis. QC can help to ensure that the functional annotation is accurate and relevant.

Overall, conducting QC in RNA-seq analysis is important because it helps to ensure that the data and the analysis are of high quality, which can improve the reliability and interpretability of the results.

Comparing between 7,429 positive and 39.291 negative sequences, it is realized that Q is frequently enriched in the flanking region of ubiquitination sites. It is also noticable that K is often depleted in the positive samples