

BIM3007-Assignment4

Junyang Deng

December 24, 2022

1 RNA-Seq

1. In RNA-seq analysis, what is the significance of data quality control? What softwares can be used for quality control? (10 points)

Significance: Quality control is usually the first step in any RNA-seq analysis because it helps ensure the reliability of downstream analysis. Specifically, quality control can improve the accuracy of alignment and make differential analysis or functional annotation more reliable.

Softwares for quality control include FastQC, RSeQC, MultiQC, RNASeqQC etc. In tutorial we learned about **FastQC** and **MultiQC**.

- FastQC: When a .fastq file is input, this software will generate a report for the data in .html format. The report includes the following tests and results (Figure 1). The mark in front of each item indicates the test result. A green tick means 'pass', an exclamation mark means 'warning', and a red cross means 'failure'.

Summary











-  [Basic Statistics](#)
-  [Per base sequence quality](#)
-  [Per sequence quality scores](#)
-  [Per base sequence content](#)
-  [Per sequence GC content](#)
-  [Per base N content](#)
-  [Sequence Length Distribution](#)
-  [Sequence Duplication Levels](#)
-  [Overrepresented sequences](#)
-  [Adapter Content](#)

Figure 1: Fastqc report

- MultiQC: This software can be 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.

- 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)

Significance: Alignment helps people identify the corresponding gene of the transcript. With alignment information, people can use RNA-seq data to quantify the expression level of each gene or identify alternative splicing.

Softwares for alignment can be divided into two types: Splice-aware aligners (STAR, TopHat, HiSat) and non-splice aligners (BWA, Bowtie and HiSat). **Splice-aware aligners** are willing to open gaps rather than mismatches when two sequences cannot coincide. On the contrary, **non-splice-aware aligners** are more likely to generate mismatches, and gaps usually have higher penalties. The splice-aware ones can be used when transcripts are mapped to the whole genome, while the non splice-aware ones can be used to map

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

Program: featureCounts. This program will generate a table of genes by samples with raw sequence abundances.

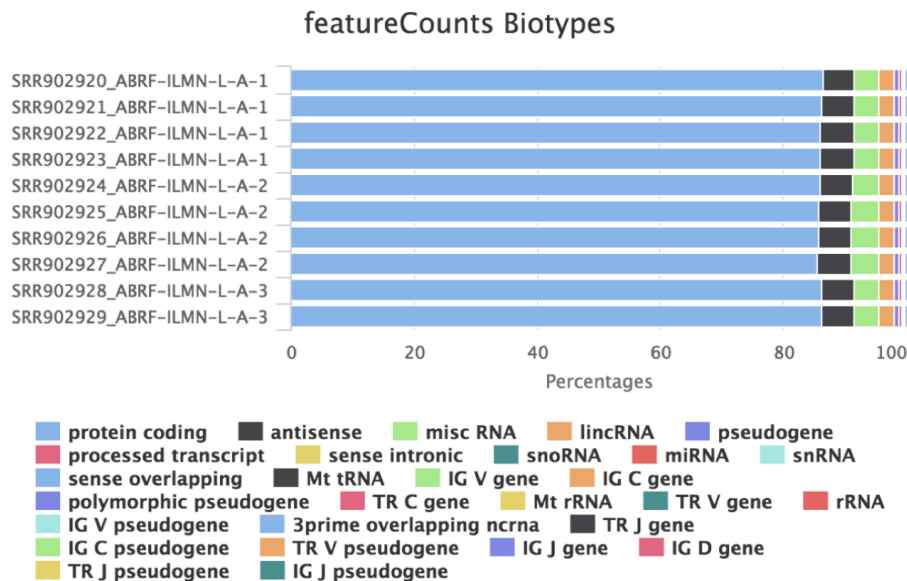


Figure 2: An example of table generated by featureCounts.

- 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)

Input: .fastq file

Output: a read count matrix

Analysis: Many R packages can be used to conduct downstream analysis. For differential expressed gene (DEG) analysis can be done using DESeq2, limma and edgeR. After finding DEGs, we can do functional analysis using EnrichR and GSEA.

2 PTM Prediction Preprocessing

1. Please write a python script to extract the positive dataset and the negative dataset (13-mer) from the raw dataset (`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)

After processing, 8000 positive and 39292 negative sequences with window size 13 are generated.

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)
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)