BIM3007-Assignment4

Junyang Deng

December 25, 2022

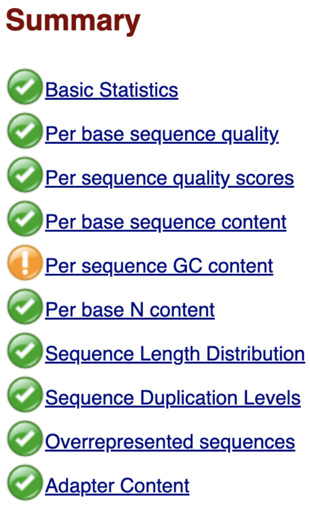
1 RNA-Seq

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.

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)*

   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

3. 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.

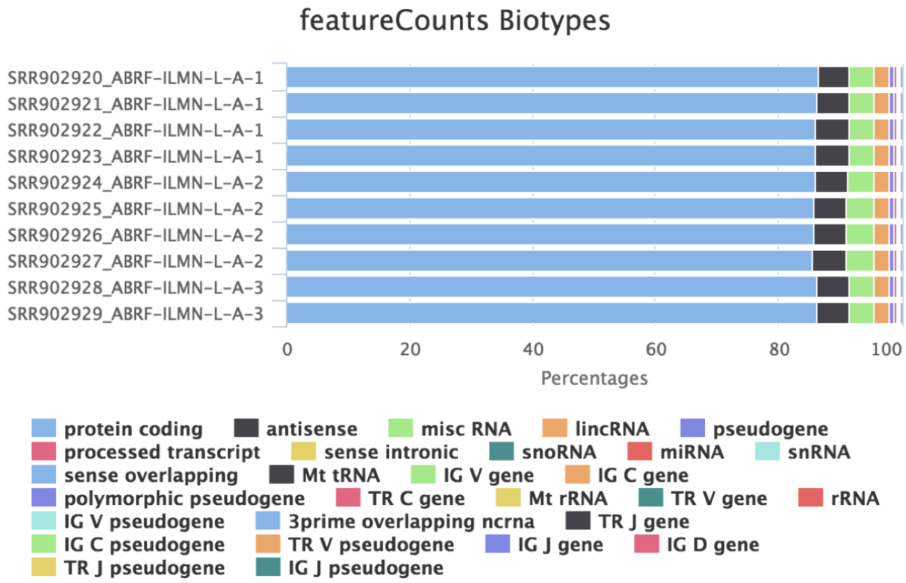


Figure 2: An example of table generated by featureCounts.

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)*

   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)

   The Python script used to extract the (13-mer) positive and negative dataset is named `processing.py`. The program did the following things:

   (a) For a site in a protein sequence, a sequence fragment with $2n+1$ amino acids is constructed by taking n upstream residues and n downstream residues from the site, respectively. The n here is 6;

   (b) When the residues are not enough, e.g. for the sites located in N- or C-terminus, assign a non-existing residue X to fill in the corresponding position;

   (c) Delete the negative sequences where a real ubiquitination site is involved.

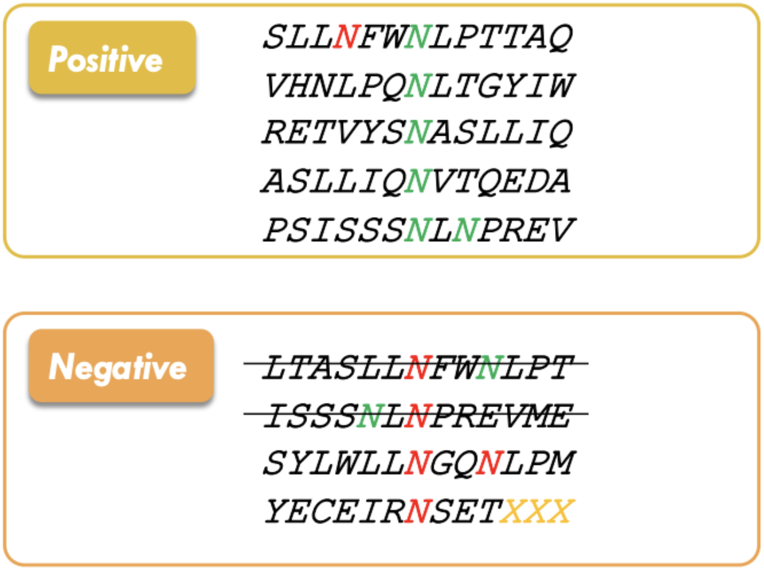


Figure 3: An illustration of what the program does

These steps are illustrated by Figure 3. After running the script, `positive.fasta` and `negative.fasta` were generated since `.fasta` file is needed for CD-HIT. Names for sequences are just numbers. The output of CD-HIT is `positive_80.fasta` and `negative_80.fasta`. Statistics before and after removing homologous sequences is shown in Table 1.

	Positive	Negative
Before	8,000	39,291
After	7,429	32,141

Table 1: Data statistics before and after removing homologous sequences

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)

   One sample logo: The logo of positive and negative sequences are shown in Figure 4 and 5. Two plots look identical with K standing in the middle. Since only K is visible in this plot, and no information is shown here, two zoomed logo plots are also given in Figure 6 and 7. Now, more information is shown. Besides logo plot, probability plots are shown in Figure 8 and 9.

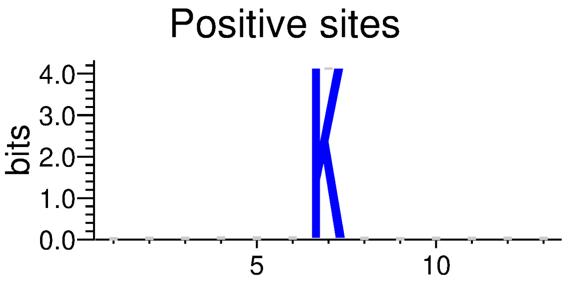


Figure 4: Logo of positive sequences

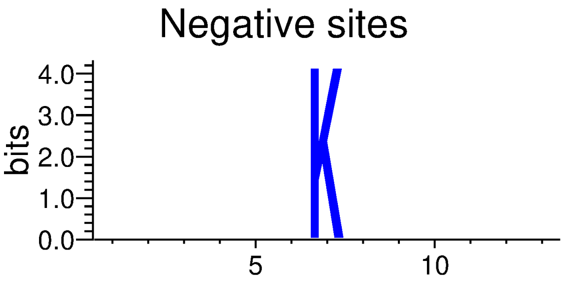


Figure 5: Logo of negative sequences

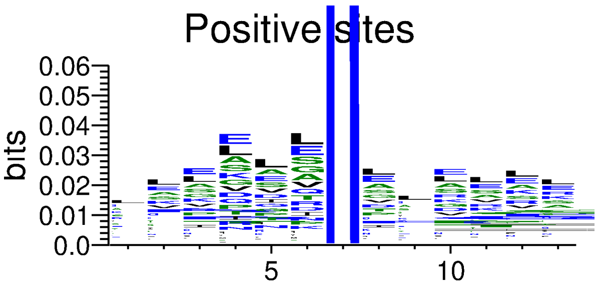


Figure 6: Logo of positive sequences (zoom-in)

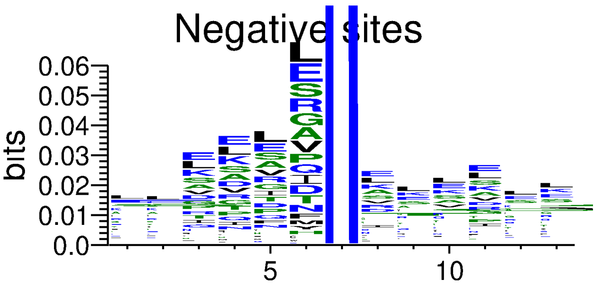


Figure 7: Logo of negative sequences (zoom-in)

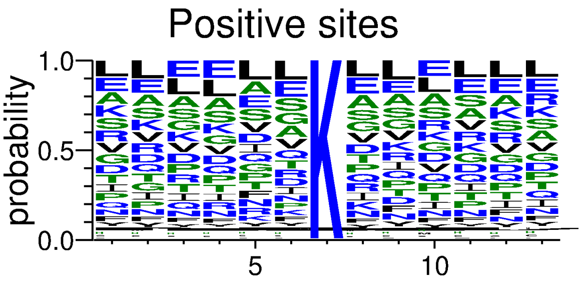


Figure 8: Probability plot of positive sequences

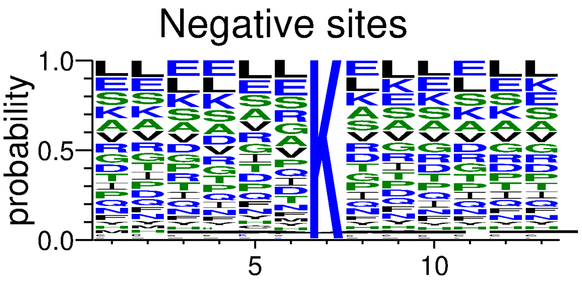


Figure 9: Probability plot of negative sequences

   Two sample logo: Comparing between 7,429 positive and 32,141 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. The two sample logo plot generated by TwoSampleLogo is shown in Figure 10

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)

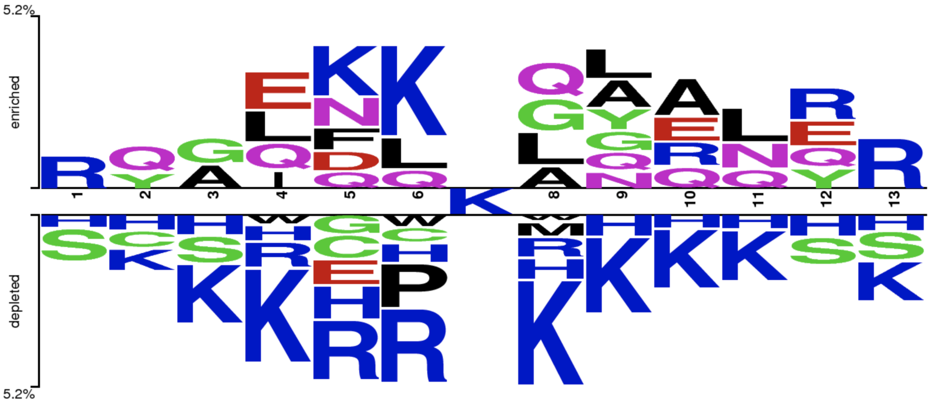


Figure 10: Two sample logo. This figure is generated by TwoSampleLogo by comparing positive and negative samples.

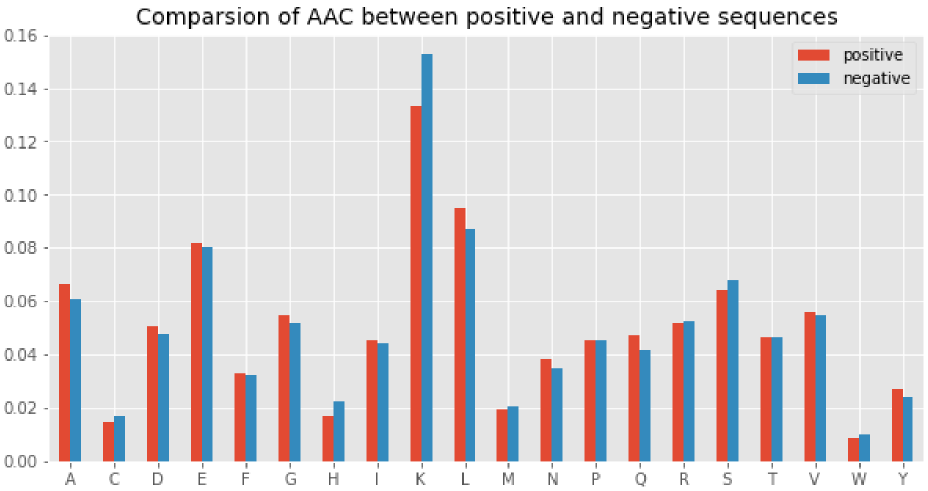


Figure 11: Comparison of AAC

AAC of a sequence is calculated by the following formula:

$$\text{AAC}x = \frac{\text{Occurance of amino acid } x}{\text{Length of the sequence}}$$

Using this formula, the average AAC for positive and negative sequences are calculated. Figure 11 shows the comparison between positive and negative samples. It is quite clear that the percentage of K (lysine) is higher in negative sequences than positive ones. This result aligns with the previous Figure 10 that K is frequently depleted in positive sequences. In addition, A (alanine) and Q (glutamine) appears more frequently in positive samples.