Analyzis of E.coli strains for outbreak investigation via identification pathogenic genes

Ignat Sonets, Kamilla Faizullina

We assemble the genome of deadly E. coli X strain. We find genes which lead to pathogenicity.

1 Introduction

2 Methods

To analyze the E. coli X strains, We use the dataset from the TY2482 sample [1]. To estimate genome size, I use Jellyfish [3]. For estimation the genome size, the following formulas is used:

$$N = \frac{M * L}{L - K + 1}, Genome\_size = \frac{N}{T},$$

where N — Depth of coverage, M — k-mer peak, K — k-mer-size, L — average read length, T — Total bases).

For assembling the genome the SPAdes tool is used [4]. SPAdes uses information about the distances between reads within read-pairs to combine contigs into ordered collections of adjacent contigs called scaffolds.

3 Results

We use Fastqc for [2] for estimation number of reads and quality control. Table 1 represents the number of reads of the sequencing data. We run Jellyfish tool only on the data labeled SRR292678. The length of mer is equal to 31. From the Figure 1, the peak position is ≈ 54. $Genome\_size \approx 5Gb$.

The information related to the quality of the resulting assembly after SPAdes usage is available in lab journal.

The sequence	Reads
SRR292678 forward	5499346
SRR292678 reverse	5499346
SRR292862 forward	5102041
SRR292862 reverse	5102041
SRR292770 forward	5102041
SRR292770 reverse	5102041

Table 1: Number of reads

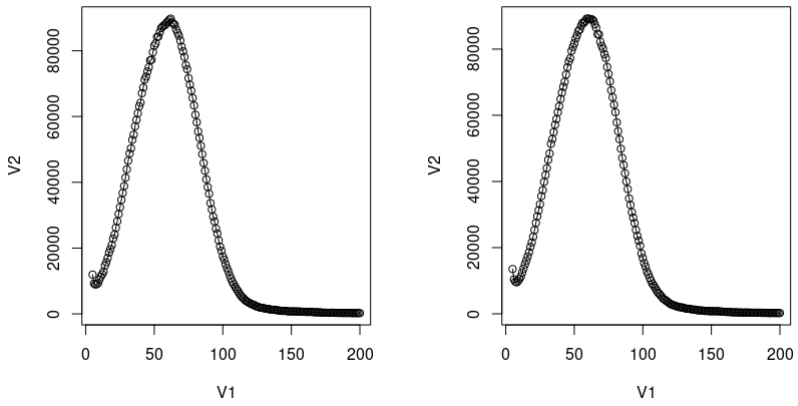

Figure 1: The k-mer distribution in the forward and reverse data among region between 8 and 200

4 Discussion

References

[1] Datasets: (forward and reverse)
SRR292678:
https://d28rh4a8wq0iu5.cloudfront.net/bioinfo/SRR292678sub_S1_L001_R1_001.fastq.g
https://d28rh4a8wq0iu5.cloudfront.net/bioinfo/SRR292678sub_S1_L001_R2_001.fastq.g
SRR292862:
https://d28rh4a8wq0iu5.cloudfront.net/bioinfo/SRR292862_S2_L001_R1_001.fastq.gz
https://d28rh4a8wq0iu5.cloudfront.net/bioinfo/SRR292862_S2_L001_R2_001.fastq.gz
SRR292770:
https://d28rh4a8wq0iu5.cloudfront.net/bioinfo/SRR292770_S1_L001_R1_001.fastq.gz
https://d28rh4a8wq0iu5.cloudfront.net/bioinfo/SRR292770_S1_L001_R2_001.fastq.gz

[2] Fastqc : https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

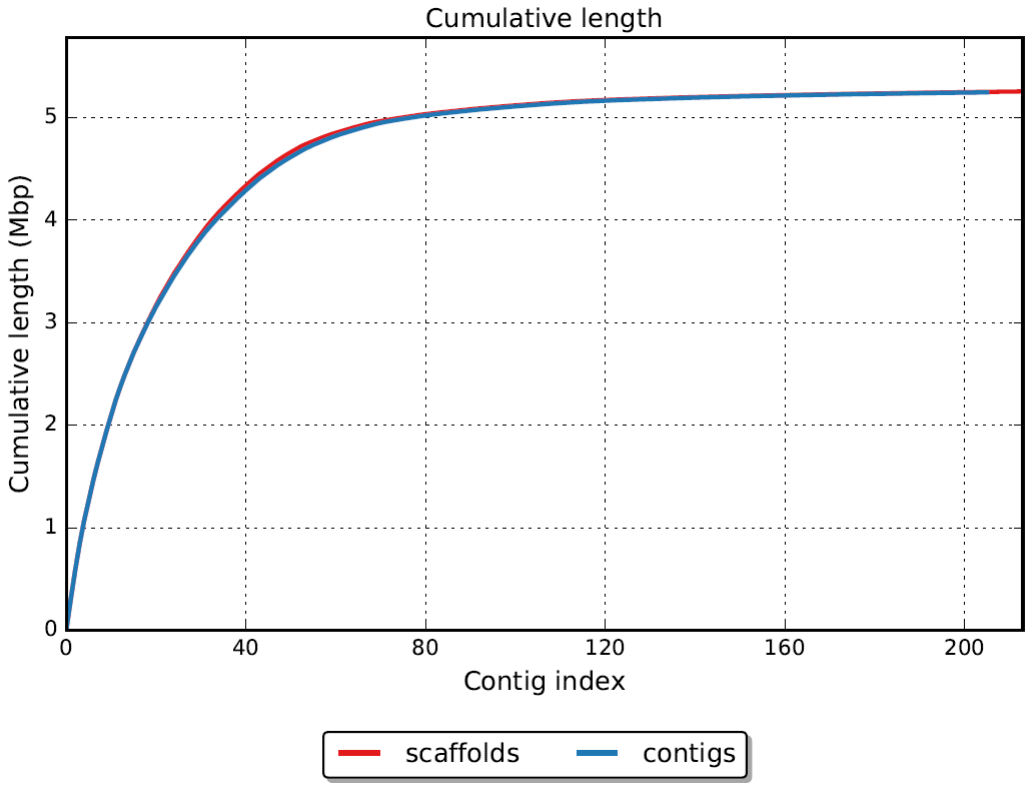

Figure 2: Assessment of the quality of the paired processed data after using SPAdes

[3] Guillaume Marcais and Carl Kingsford, A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics (2011) 27(6): 764-770 (first published online January 7, 2011) doi:10.1093/bioinformatics/btr011

[4] SPAdes: http://cab.spbu.ru/software/spades/

[5] QUAST: http://quast.bioinf.spbau.ru/