Assembly and annotation of WGS of *Klebsiella pneumoniae*

Abstract

The bacterium *Klebsiella Pneumoniae* (SRR32133086) was selected for the work. The data of it's WGS were obtained from searching the SRA database.

Genome assembly identified the total number of nucleotides in the genome - 5,542,871 bases. Genome annotation provided enhanced information on the genome composition and individual genes. In the annotation, we also found 16S rRNA and determined the taxonomic affiliation of the bacterium. This data provides a snapshot of the organism's genome structure and functional elements, useful for understanding its biology and potential applications in research or biotechnology.

Introduction

Klebsiella Pneumoniae is an opportunistic pathogenic microorganism of the saprophytic flora of the human gastrointestinal tract. Some strains serve as opportunistic pathogens that cause severe infectious processes such as pneumonia, urinary tract infections, and circulatory system infections, especially in patients with weakened immune systems [1].

When studying this bacterium using bioinformatic tools, it is possible, for example to predict of proteins that are essential for the survival of bacterial pathogens that can facilitate the development of medicines against it [2].

It is also possible to identify the genes responsible for their multiple antibiotic resistance and virulence of this bacteria that can help to monitor *Klebsiella pneumoniae* outbreaks [3].

Using bioinformatics approaches, researchers can gain a deeper understanding of *Klebsiella pneumoniae* biology, identify novel drug targets, and develop more effective strategies to combat this challenging pathogen.

Methods

On the first stage we obtained raw sequencing data of *Klebsiella pneumoniae* from the NCBI database using SRR accession number SRR32133086. You can do this by entering the number in the NCBI search bar. The data from NCBI has a .sra format so we have to unpack the archive using SRA-tools (fastq-dump) and separate forward and backward reads.

After that we have to check the quality of the reads before proceeding further using FastQC program. As with SRA-tools, we installed fastqc using conda. We ran fastqc for both files and as a result we got the files with the .html extension and opened in the browser. This way we can look at the quality of the reads and decide whether we need to clean them for further steps. We didn't have to do this for our bacteria, but trimming can be done with any convenient tool, for example, Cutadapt version 4.4 can be used. You can install it using conda. After clearing the data, you need to run fastqc again and view the new quality report. If the data quality is acceptable, you can proceed to the next steps.

Filtered reads that can be assembled into longer fragments - contigs and scaffolds using Spades tool. Install spades using conda. To check the quality of the resulting assembly we used QUAST tool, which is also available in conda. As a result we got report.txt file with all the characteristics. If the assembly quality is acceptable, you can proceed with the annotation using special tools.

One of the most popular annotator tools for bacterial genomes is Prokka. It is available from conda. As a result we got files in various extensions that can be used in further research. In our case we decided to find 16S rRNA and determined the taxonomic affiliation of the bacterium. To do this,

using the grep command, we extracted information about 16S rRNA from the annotation file and did the alignment using BLAST.

Results

1 Looking through the qualitative assessment of fastq files, we came to the conclusion that cleaning before assembly is not necessary, since almost all indicators were optimal (Fig. 1-3), and minor increases in GC content is normal for this kind of bacteria (Fig. 4).

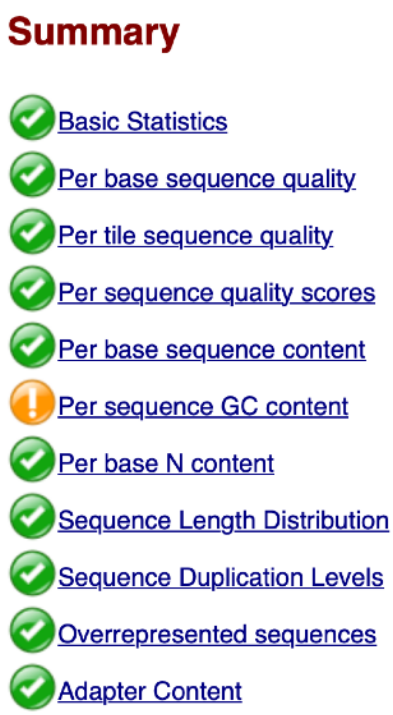

Fig. 1 Assessment of the quality of the reads using FastQC

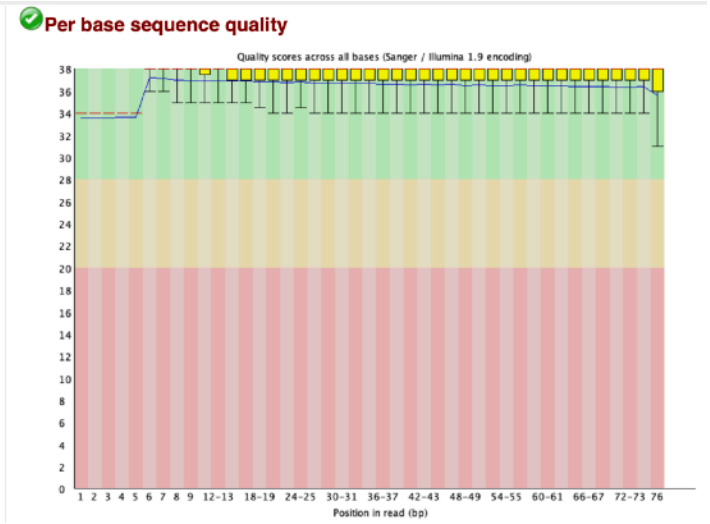

Fig. 2 Per base sequence quality for the forward reads

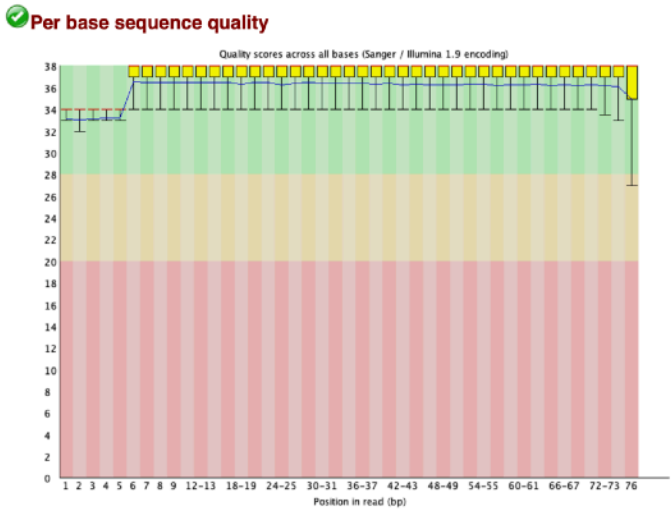

Fig. 3 Per base sequence quality for the backward reads

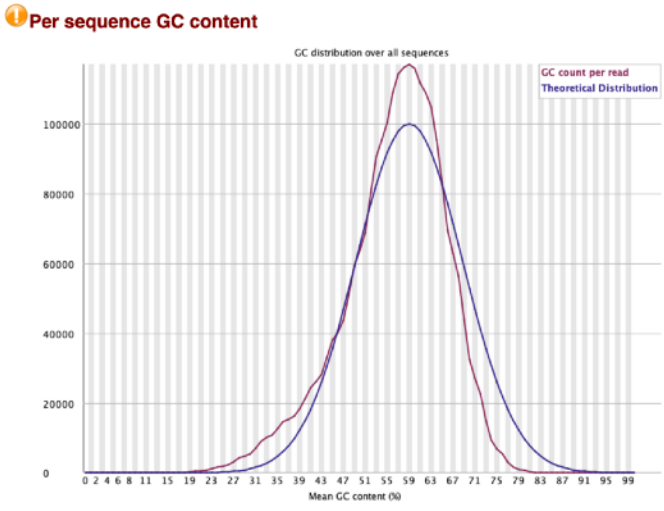

Fig. 4 GC content for the forward reads

2 After assembling the genome, we also looked at the report.txt file to check the characteristics (Fig. 5). Based on the assembly statistics, the assembly appears to be of good quality overall. Total Length (5,502,430 bp) represents the total size of the assembled genome should match the expected size of the genome (~5m), which is a good starting point for assessing completeness. Contig Count (100) is based on contigs of size >= 500 bp and shows the low number of contigs suggests the assembly is reasonably contiguous. The number of all existing contigs is equal to 360. A large contig size (668,545 bp) indicates that the assembler was able to resolve longer stretches of

Assembly	contigs
# contigs (>= 0 bp)	360
# contigs (>= 1000 bp)	87
# contigs (>= 5000 bp)	57
# contigs (>= 10000 bp)	51
# contigs (>= 25000 bp)	43
# contigs (>= 50000 bp)	33
Total length (>= 0 bp)	5542871
Total length (>= 1000 bp)	5494694
Total length (>= 5000 bp)	5422384
Total length (>= 10000 bp)	5377129
Total length (>= 25000 bp)	5237380
Total length (>= 50000 bp)	4910712

# contigs	100
Largest contig	668545
Total length	5502430
GC (%)	57.17
N50	149322
N90	44870
auN	249085.2
L50	10
L90	34
# N's per 100 kbp	0.00

Fig. 5 Assessment of the quality of genome assembly of *K. pneumonia* using QUAST

the genome. The N50 value (149,322 bp) is the length of the shortest contig at which the sum of the lengths of all contigs equal to or greater than this length is at least one-half of the total length of the assembly. the higher the N50 value, the fewer fragmented regions in the genome. N50 values for bacteria should be in the range of hundreds of thousands of nucleotides, so the figure in our assembly is appropriate. The L50 value (10) indicates how many large fragments are required to cover 50% of the genome. The L50 value between 1 and 10 is considered acceptable. Also there is no N's per 100 kbp (0.00) - assembly is highly accurate with no gaps.

From the annotation results we can see some genomic metrics (Fig. 6): 360 contigs suggest the genome is fragmented into 360 pieces, the total number of nucleotides in the genome is 5,542,871 bases. CDS indicates the number of protein-coding genes in the genome, which is 5,125. The number of ribosomal RNA genes, essential for protein synthesis is 6. Repeat_region represents regions with repetitive sequences in the genome - only one is identified. The number of transfer RNA genes involved in translation is 59. The number of transfer-messenger RNA gene, which helps rescue stalled ribosomes during translation, is 1.

PROKKA_03302025.txt

```
organism: Genus species strain
contigs: 360
bases: 5542871
CDS: 5125
rRNA: 6
repeat_region: 1
tRNA: 59
tmRNA: 1
```

Fig. 6 Genomic metrics from the PROKKA report

3 When we did the alignment of 16S rRNA using BLAST, 16S rRNA 100% confirmed that the sequencing data belonged to *Klebsiella pneumoniae.*

Discussion

The genome assembly and annotation of *Klebsiella pneumoniae* allows researchers to obtain a huge amount of information about this bacterium, for example, provide insights into its antimicrobial resistance, virulence, evolutionary dynamics and etc. Based on the data obtained in and as a result of genome assembly and annotation, it is possible, for example, to build machine learning models to predict clinical outcomes based on genomic data, integrating virulence and AMR scores and etc. There are many possible applications of the data, depending on the goals of the researchers.

Citation

1. Shamina, O. V., Samoylova, E. A., Novikova, I. E., & Lazareva, A. V. (2020). Klebsiella pneumoniae: Microbiological characterization, antimicrobial resistance, and virulence. *Russian Pediatric Journal*, *23*(3), 191–197. https://doi.org/10.18821/1560-9561-2020-23-3-191-197
2. Pranavathiyani, G., & Pan, A. (2024). Prediction of essential proteins of *klebsiella pneumoniae* using integrative bioinformatics and Systems Biology Approach: Unveiling new avenues for Drug Discovery. *OMICS: A Journal of Integrative Biology*, *28*(3), 138–147. https://doi.org/10.1089/omi.2024.0001
3. Institut Pasteur. (2022, March 31). *A bioinformatics tool to monitor the resistance and virulence of Klebsiella pneumoniae bacteria*. https://www.pasteur.fr/en/bioinformatics-tool-monitor-resistance-and-virulence-klebsiella-pneumoniae-bacteria