**Lab Report**

Searching for viruses in the sequencing data of a patient's biomaterial.

Darina Bespalaya Applied Genomics, ITMO

03.10.2023

**Abstract**: In this research a bronchoalveolar lavage fluid sample of a 41-year-old patient who was admitted to Central Hospital of Wuhan on December 26 2019 was analyzed. This sample was subjected to deep meta-transcriptome sequencing before. These results were used as the original data in the study. The input data was uploaded to BLAST and the sequence belonging to the pathogen to be detected was identified. Gene annotation was done using Prokka. The results of this research revealed that the sequence sought has a high percentage of overlap with other sequences from the SARS (Severe Acute Respiratory Syndrome) coronavirus family identified in a previous similar MERS outbreak. The importance of this study is to better understand the nature of the origin of the disease and to more effectively detect, treat and prevent it.

**Introduction**: Coronaviruses are a group of respiratory viruses that can infect both humans and a variety of mammal species. They earned their name from the crown-like spikes on their cell surface. Being RNA viruses, coronaviruses have a higher mutation rate compared to DNA-based viruses.

In the 21st century, viruses from this family have been responsible for multiple global epidemics. In 2002, the emergence of SARS in Southern China led to a rapid spread to more than 20 countries. Since coronoviruses exist on our planet and have caused more than one disease outbreak, it is necessary to learn how to identify and treat them quickly and accurately. This study aims to analyze the assembly characteristics of the metatranscriptome of biomaterial from an infected patient.

**Methods**: The original sequence was taken from the NCBI (National Center for Biotechnology Information) website. The quality of the assembly was determined primarily using QUAST (Quality Assessment Tool for Genome Assemblies) and PROKKA (Prokaryotic Genome Annotation System), which helps to annotate and understand the genetic content of organisms more completely. The full results for these analyses will be presented below.

Next, a program was written using Python tools to select the best-fitting contigs based on their length to screen out non-viral contigs. A length of over 1000 base pairs was taken.

The BLAST (Basic Local Alignment Search Tool) tool was also used, necessary to compare and identify sequences included Entrez Query constraints - 1900/01/01:2020/01/01/01 [PDAT] and viruses as organisms.

Results: The results of the QUAST tool are summarized in Table 1. Analyzing the QUAST results became clear that this assembly has a total of 666 contigs, all of which are greater than 1000 bp, the longest contig has 29907 bp. The total length of all contigs was 1042805 bp.

Table 1

|  |  |
| --- | --- |
| Statistics without reference | spades\_scaffolds |
| # contigs | 666 |
| # contigs (>= 0 bp) | 111219 |
| # contigs (>= 1000 bp) | 666 |
| # contigs (>= 5000 bp) | 4 |
| # contigs (>= 10000 bp) | 1 |
| # contigs (>= 25000 bp) | 1 |
| # contigs (>= 50000 bp) | 0 |
| Largest contig | 29907 |
| Total length | 1042805 |
| Total length (>= 0 bp) | 27362468 |
| Total length (>= 1000 bp) | 1042805 |
| Total length (>= 5000 bp) | 48741 |
| Total length (>= 10000 bp) | 29907 |
| Total length (>= 25000 bp) | 29907 |
| Total length (>= 50000 bp) | 0 |
| N50 | 1482 |
| N90 | 1067 |
| auN | 2617.4 |
| L50 | 231 |
| L90 | 566 |
| GC (%) | 44.75 |
| Per base quality |  |
| # N's per 100 kbp | 23.97 |
| # N's | 250 |

In order to find more suitable viral contigs according to their length and coverage, the Scaffold program used by Linux was used. These contigs were subsequently loaded into the BLAST tool.

The result presented by BLAST allowed to know that the selected 5 contigs have a similarity with Bat SARS-like coronavirus isolate bat-SL-CoVZC45 with 89.12%, query coverage 95% and E-value 0.0. in NODE\_1 genome and Enterobacteria phage phiX174 with 100% Query coverage and Identity score in NODE\_3 genome.

By analyzing PROKKA tool results for NODE\_1 and NODE\_3 as the most suitable contigs, information on DNA and nucleotide sequence, genes annotation, location and predicted proteins was obtained. The full report provided by PROKKA is available [1].

**Discussion**: The results of the study showed that the desired assembly obtained by meta-transcriptome sequencing of a bronchoalveolar lavage fluid sample from a patient admitted to the hospital with an unknown disease matched SCARS-coronavirus with a percentage of identity of about 89.12% and 95% of query length for NODE\_1 gene. It can be conclude that the admitted patient was infected with one of SCARS-coronavirus types, which was established later by doctors and scientists. But also NODE\_3 had 100% of querty coverage and Identity score with Enterobacteria phage phiX174. So there is an open question about the presence of this bacterium in the lung fluid sample.

**Data Availability:**

1. All files and reports generated in the course of this study are available in the repository <https://github.com/darina2424/MDA_hw1.git>;
2. Initial data for the study: NCBI Sequence Read Archive (SRA) database with BioProject accession number PRJNA603194.