Student report

**“Searching for viruses in sequencing data”**

Authors: Anton Filev1, Julia Solomennikova1

1ITMO University, St. Petersburg, Russian Federation

**Abstract**

In this report, we demonstrate metagenomic approach, examining matches to the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) in all high throughput sequencing data sets in the NCBI Sequence Read Archive. The results show that a viral genome under accession number PRJNA603194 is mostly related to a group of SARS-like coronaviruses. While analysis of these reads indicates the presence of a similar viral sequence in Enterobacteria phage phiX174. In addition to the implications for SARS-CoV-2 emergence, this report illustrates the utility and limitations of metagenomic search tools in effective and rapid characterization of significant nucleic acid sequences.

**Introduction**

Many of the major human infectious diseases, including presented in humans and absent from animals, are still unexplored, because identifying and measuring the community dynamics of viruses is technically and computationally complicated process (Edwards & Rohwer, 2005). There an example of severe respiratory disease which have been reported in Wuhan, Hubei province, China. (Wu et al., 2020). The origins of this disease is still a source for further research. However, metagenomics offers the promise of unbiased detection of emerging pathogens, without requiring prior knowledge of the identity of the responsible agent or its genomic sequence. Therefore, the objective of this paper is to provide insights into the development stages and to search for similarities between groups, annotate and compare the viral genome.

**Methods**

The data containing the sequences of the target virus was used from the NCBI Sequence Read Archive (SRA) database under BioProject accession number PRJNA603194 published by Wu F. et al (2020). For downloading raw reads a special tool called *fastq-dump* used from SRA-ToolKit. The *fastq-dump* tool extracts data in FASTQ- or FASTA-format from SRA-accessions (Leinonen et al., 2011).

Assembling of metatranscriptome was performed by SPADes assembler (Bankevich et al., 2012). There are many launch parameters, but we used the standard Illumina paired libraries. At the end Illumina produces paired-end reads in two

files: R1.fastq and R2.fastq. with gaps between them. The adapter and/or quality trimming software has been used prior to assembly (Bolger et al., 2014).

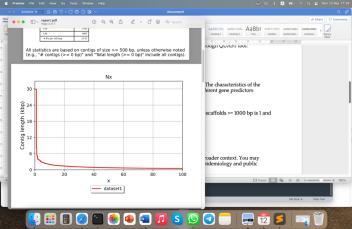
To evaluate he quality of assemblies, the special tool QUAST were used, which is one of the most famous assessment tools available for genome assemblies (Gurevich et al., 2013). QUAST can evaluate assemblies using reference genomes, as well as without reference genomes. The final quality assessment reports in the html file.

For classification of contigs as viral, non-viral or uncertain, based on gene content was used ViralVerify, which predicts genes in the contigs using Prodigal in the metagenomic mode, runs hmmsearch on the predicted proteins and classifies the contig as vrial or non-viral by applying the Naive Bayes classifier (NBC) (Pu & Shamir, 2022).

Indexing of multi-fasta file was performed by Samtools (Danecek et al., 2021). In order to add an options to specify index location or the name of contig, *faidx* option was used for scaffolds.fa files. *Xargs* option were used for selecting a certain length.

To match chosen unknown viral contigs with known viral sequences BLAST was used with the following limitations: viruses (taxid:10239), and entrez query: 1900/01/01:2020/01/01[PDAT].

Annotation features of interest in a set of genomic DNA sequences might be performed by differnet software, but inder the ceration conditions the data from GenBank by checking “Graphics” options (GenBank accession number NC\_045512).

**Results** 

1. Quality assessment of genome assemble

We found 3667 different contigs (fig. 1). Among them only one was >= 1000 bp; its length was 29907 bp

Figure 1. The relation between contig length (kbp) and the number of contigs. More information about the assembled date evaluation - see in Supp.1.

2. Viral detection

Contig analysis showed that 194 of sequences could have viral origin. Other contigs (111076) had chromosomal or undetected origins (supp. 2).

3. Similarity search

We compare chosen unknown viral contigs with known viral sequences using BLAST (See whole report here

https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Get&RID=HZ4SK6ET01R). The longest contig (29997 bp) had the highest identity (89,12%) with Bat SARS-like coronavirus (Taxonomy ID: 1508227) (supp.3)

4. Annotation

The annotation of Severe acute respiratory syndrome coronavirus 2 isolate Wuhan Hu-1 (complete genome) is presented in the Supplementary 4.

**Discussion**

Due to human population which is constantly increasing new lethal microorganisms will appear. Therefore, it is crucial to obtain the data rapidly as much information about new pathogens as we can. In this project we provided the algorithm that could cope with this problem. However, it is necessary to create a universal set of tools which will speed up this process and last but not least, attach detailed instructions for using this set. According to the BLAST report, there is Enterobacteria-hosted phage

on the sequence. Therefore, it is critical to prepare library carefully to exclude contaminations and others interruptions or consider different evolution scenarios.

**References:**

1. Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A., & Pevzner, P. A. (2012). SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology*, *19*(5), 455–477. https://doi.org/10.1089/cmb.2012.0021

2. Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics*, *30*(15), 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

3. Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2), 1–4. https://doi.org/10.1093/gigascience/giab008

4. Edwards, R. A., & Rohwer, F. (2005). Viral metagenomics. *Nature Reviews Microbiology*, *3*(6), 504–510. https://doi.org/10.1038/nrmicro1163

5. Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: Quality assessment tool for genome assemblies. *Bioinformatics*, *29*(8), 1072–1075.

https://doi.org/10.1093/bioinformatics/btt086

6. Leinonen, R., Sugawara, H., & Shumway, M. (2011). The sequence read archive. *Nucleic Acids Research*, *39*(SUPPL. 1), 2010–2012. https://doi.org/10.1093/nar/gkq1019

7. Pu, L., & Shamir, R. (2022). 3CAC: improving the classification of phages and plasmids in metagenomic assemblies using assembly graphs. *BioRxiv*, *30*, 2021.11.05.467408.

https://www.biorxiv.org/content/10.1101/2021.11.05.467408v2%0Ahttps://www.bi orxiv.org/content/10.1101/2021.11.05.467408v2.abstract

8. Wu, F., Zhao, S., Yu, B., Chen, Y. M., Wang, W., Song, Z. G., Hu, Y., Tao, Z. W., Tian, J. H., Pei, Y. Y., Yuan, M. L., Zhang, Y. L., Dai, F. H., Liu, Y., Wang, Q. M., Zheng, J. J., Xu, L., Holmes, E. C., & Zhang, Y. Z. (2020). A new coronavirus associated with human respiratory disease in China. *Nature*, *579*(7798), 265–269. https://doi.org/10.1038/s41586-020-2008-3

**Supplements**

**Supplement 1:** https://drive.google.com/file/d/1COZNL7mp-VJGAPltCweZ0zX YT6uzHBo/view?usp=sharing

**Supplement 2:** https://docs.google.com/spreadsheets/d/1HL7IUDxPefe dsEVWO5UqLX7Wa0Pyu0qmhY2rdenNEE/edit#gid=2141147666 **Supplement 3:** https://drive.google.com/file/d/1POwiOu-cZlrUKnZCILw Vym6S1T4\_VA4/view?usp=sharing

**Supplement 4:**

https://drive.google.com/file/d/164dHMrss3p8fQ9fqQCj5roUy07DIloF0/view?usp=s haring