doi 10.18699/bgrs2024-12.3-22

Millions of SARS-CoV-2 genomes in the RAM of a regular PC: fast and efficient analysis of evolutionary changes

Palyanov A.Yu.[1, 2, 3]*, Palyanova N.V.[2]

[1] *A.P. Ershov Institute of Informatics Systems of the Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia*

[2] *Research Institute of Virology, Federal Research Center for Fundamental and Translational Medicine, Novosibirsk, Russia*

[3] *Novosibirsk State University, Novosibirsk, Russia*

* *palyanov@iis.nsk.su*

Key words: coronavirus; SARS-CoV-2; genome; variants; evolution; analysis; software

Motivation and Aim: In addition to the enormous harm caused to humanity, the SARS-CoV-2 pandemic has also left us, thanks to the efforts of virologists and bioinformaticians, an unprecedented amount of information about how the genome of the virus changed day by day on a planetary scale for more than four years. Analysis of these data made it possible to learn a lot both specifically about SARS-CoV-2 and about respiratory viruses in general. As of May 3, 2024, there are more than $16.7 \cdot 10^6$ genetic sequences of this virus in the GISAID database, and more than $8.6 \cdot 10^6$ in NCBI Genbank. For comparison, the influenza virus is represented in the GISAID database by $5.22 \cdot 10^5$ variants (collected since 1905). Typically, to process such volumes of data, they are read directly from the HDD or SSD, but what if we could place all available SARS-CoV-2 genomes into the RAM of a regular PC and thus be able to access them and perform various analytical calculations at a much higher speed? This approach allowed us to analyze the evolution of SARS-CoV-2 throughout the pandemic.

Methods and Algorithms: One copy of the SARS-CoV-2 genome, a single-stranded RNA(+) virus, has a size of about 29900 nucleotides, and accordingly, it occupies about 29.9 kb in memory. For 16.7 million genomes, 500 GB will be required, while the typical memory capacity of a modern desktop PC is 32 – 64 GB. If alignments are calculated for all genomic variants, considering the initial SARS-CoV-2 genome as a reference sequence, then each genome can be represented as a data structure that describes the differences between considered genomic variants and the reference genome without data loss. These differences consist of a list of point mutations (a position and a nucleotide at that position after replacement), deletions (a list of intervals within which nucleotides were removed from the genome) and insertions (a list of positions where insertions occurred, and fragments of nucleotide sequences that were inserted). At each position of the aligned genome there can be either the same nucleotide as in the reference genome, or a different one (appearing as a result of mutation or insertion), or its absence (as a result of deletion), called gap ("–"). Each position that differs from the reference one increases the editing distance (or the distance between sequences) by one. Even for the variants with maximal distance from the reference genome in most cases it does not exceed 200. This representation of the data allows for compression of 1500 times relative to its original size, i. e. only 330 MB of memory. In addition, calculation of editing distance between any two sequences doesn't require comparison of all positions, only

those which differ from the reference, which provides another advantage – makes it work much faster.

Results: The implementation of the proposed approach was included in the new version (v.1.1) of the ParSeq software package previously developed by our team [1]. It now allows to read multiple SARS-CoV-2 genomes from a given set of FASTA files and align them relative to the reference sequence (using locally installed NextAlign console application [2]), convert alignment results to compact structures containing information about all the differences between a particular genome and the reference one, and also contains an algorithm for quickly and efficiently comparing a pair of any two genomes with each other. Among the mutations, deletions and insertions that determine the differences between the first and second genomes from the reference one, there can be both different and identical ones; when performing calculations, different ones are taken into account, but identical ones, respectively, are not.

Using ParSeq1.1 together with the complete dataset of SARS-CoV-2 genomes from Genbank [3], we obtained the following results:

a) Of the total number of genomes, only 21.3% do not contain unidentified nucleotides ("N") in at least one of the positions. The proportion of genomes with only one "N" is less than 1.2%, two "N"s – 0.95%, three "N"s – 0.85%, etc. (distribution of the number of genomes depending on the number "N" in them was obtained, Fig. 1).

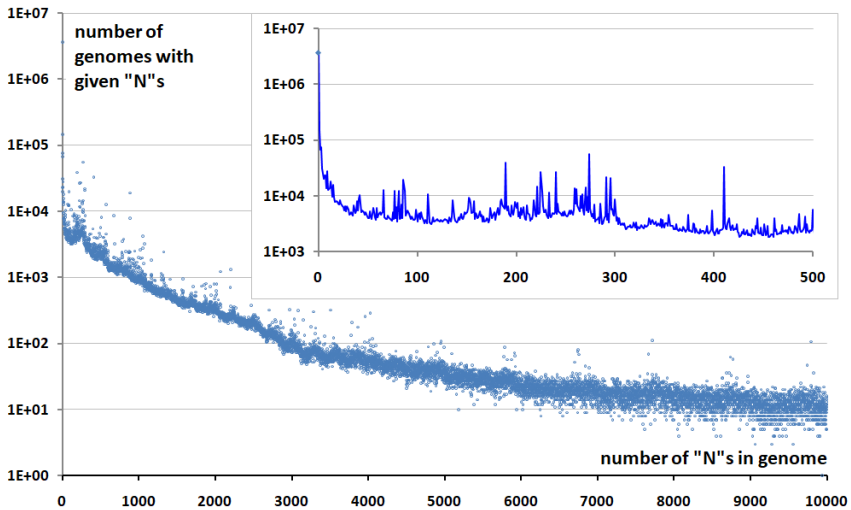

Fig. 1. Distribution of the number of sequences versus the number of "N"s in them

However, if we consider only the coding part of the genome, excluding the 3' UTR (265 nt) and 5' UTR (229 nt), then the percentage of genomes without "N" will increase to 43.3, and this is the dataset that we use in further calculations (their number is about 3.7 million).

b) The number of identical genomes (those in which the nucleotide sequence is completely identical, but the metadata may differ) was calculated. As it turns out, 54.7% of the 3.7 million genome variants are not unique – they can be represented as 459102 clusters, each composed of identical sequences, ranging in size from 2 to 12824. The distribution of cluster sizes depending on their number in the list sorted by size was calculated.

c) In Fig.e 2, the above-mentioned 3.7 million SARS-CoV-2 genomes are sorted by the sample collection date (from December 23, 2019 to May 5, 2024). For each genome

variant, both the total difference from the reference genome and the difference in the number of mutations, deletions and insertions were calculated. Based on these data, changes in the virus population were visualized over time, including the disappearance of old and the emergence of new genetic lines, emerging due to deletions, insertions and recombinations. As a result, without explicitly constructing a phylogenetic tree, an analogue of it was obtained, showing in detail the ongoing evolutionary processes, including periods of slow changes caused by mutations (occurring at almost similar rate, approximately 0.044 mutations per day, for all these years), and rapid jumps over longer distances through recombinations, deletions and insertions.

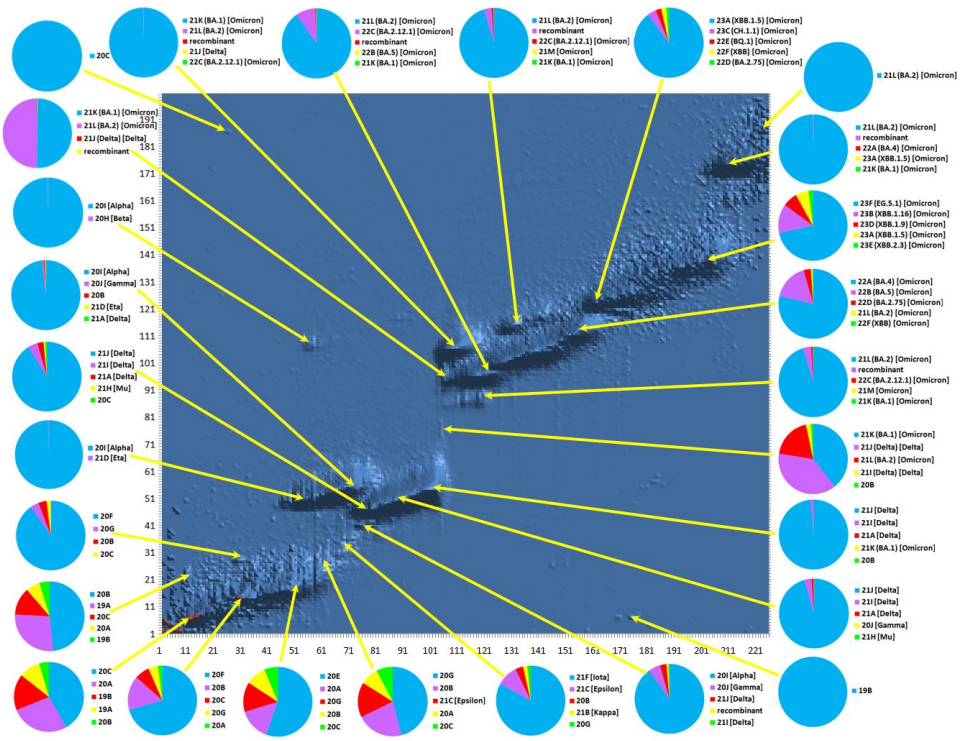

Fig. 2. Landscape of variants of SARS-CoV-2, "visited" by virus since its appearance in the end of 2019 till nowadays. Horizontal axis is time (in weeks) since 23.12.2019, vertical axis – editing distance between considered variant and reference sequence (Wuhan-Hu-1/2019)

Conclusion: A powerful, fast and efficient software tool, ParSeq v.1.1, has been developed for working with large volumes of genetic data of SARS-CoV-2 viral sequences, which has already yielded a number of interesting results and also opens up prospects for new research. In addition, the proposed approach can be applied not only to coronaviruses, but also to many other viral genomes.

Funding: The study is supported by RSF grant (No. 23-64-00005).

References

1. Aksamentov I., Roemer C. et al. Nextclade: clade assignment, mutation calling and quality control for viral genomes. J. Open Source Software. 2021;6(67):3773. doi 10.21105/joss.03773
2. Sayers E.W., Bolton E.E. et al. Database resources of the national center for biotechnology information. *Nucleic Acids Res.* 2022;50(D1):D20-D26. DOI 10.1093/nar/gkab1112
3. Palyanov A.Yu., Palyanova N.V. On the space of SARS-CoV-2 genetic sequence variants. *Vavilov Journal of Genetics and Breeding*. 2023;27(7): 839-850. doi 10.18699/VJGB-23-97