

Analysis of SARS-CoV-2 genomic data using haplotype networks

Samoilov A.E.*, Chudinov I.K.

Research Institute for Systems Biology and Medicine, Moscow, Russia

* andrei.samoilov@gmail.com

Key words: SARS-CoV-2; haplotype networks; viral surveillance

Motivation and Aim: Since the beginning of the COVID-19 pandemic, the world scientific community has faced the problem of analyzing SARS-CoV-2 genomic sequences. One of the methods of visualizing and analyzing genetic information is the construction of haplotype networks, which allow overcoming the limitations of classical methods of phylogenetics. Here we have utilized haplotype networks to detect and characterize imports of SARS-CoV-2 to Russia.

Methods and Algorithms: Genomic sequences of SARS-CoV-2 longer than 29,000 bp and N content lower than 1 % belonging to pangolin [1] lineages BA.1, BA.2 (excluding BA.2.75, XBB and later lineages) and Delta were selected from GISAID [2] and VGARus databases. Filtration was performed with Nextclade [3] and R script. Missing data in non-full Russian genomes was restored using a python script. Haplotype network graphs were built and visualized using R script.

Results: Filtration of genomes was performed as follows: for each lineage among the analyzed (BA.1, BA.2 and Delta), the list of the parent lineage substitutions and deletion was obtained by analyzing the most frequent mutations. Also the list of true reversion substitutions (i.e. mutation occurring at the same position for the second time, effectively reverting the nucleotide to the one present in the Wuhan genome) associated with certain pangolin lineages was created. Genomes with pseudoreversions (or reversion mutations not associated with a certain lineage and most likely caused by bioinformatic error) were filtered out, as well as low quality genomes.

After the filtration, to increase the amount of Russian genomes in the analysis, genomes with N nucleotides were repaired by aligning them against closest full genome sequences collected earlier and replacing Ns with consensus sequence in the N regions or discarded if there was no consensus in the alignment in the N regions.

To create haplotype network graphs, we have determined all of the potential ancestors and descendants of the analyzed Russian genomes (a substitution mutations of a potential ancestor genomes are a subset of a substitution mutations of a potential descendant genome) and created a directed graph (with the direction from the root to the increasing amount of substitutions, in accordance with the direction of the evolution) by connecting each observed substitution pattern to all of its potential ancestors, excluding the potential ancestors of potential ancestors. An example of haplotype network graph for Delta lineage is presented on the Figure 1. Next, the obtained haplotype network graph was trimmed by selecting more likely ancestor in cases where there was more than one using following algorithm: firstly, the closest ancestor was selected, secondly, the ancestor detected in the same country or countries as the descendant was selected, lastly, the earliest ancestor was selected. An example of haplotype network graph for Delta lineage is presented on the Fig. 1. After trimming, we were able to obtain imports of SARS-CoV-2 into Russia

by obtaining every substitution pattern found in Russia and whose closest ancestor was not detected in Russia.

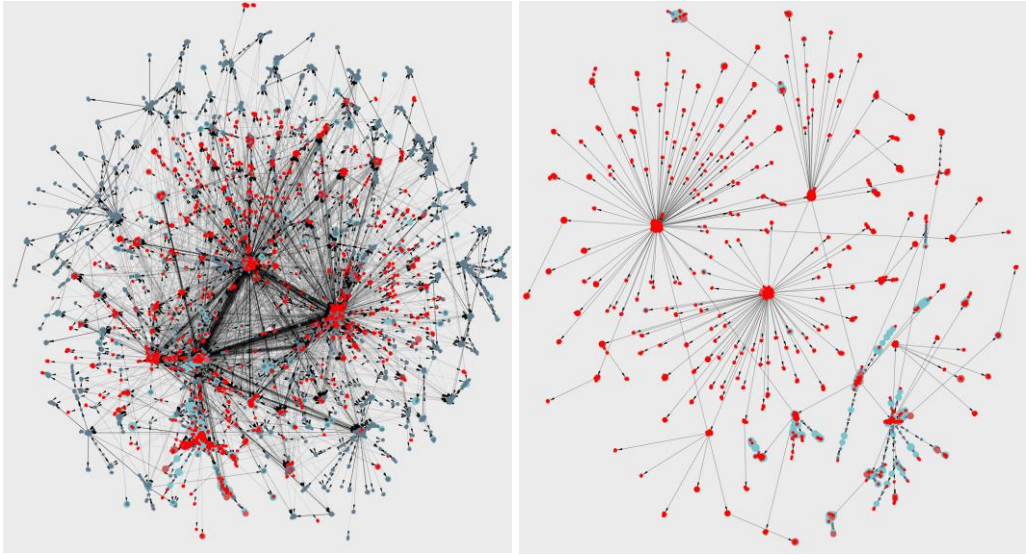


Fig. 1. Haplotype networks of Russian SARS-CoV-2 genomes. Graph vertices represent substitution patterns (a certain combination of substitution mutations), which have been observed in SARS-CoV-2 genomes at least once. Direction of the haplotype network graph represents an increase in the number of mutations from the root (parent genome for the lineage) to the increased amount of substitutions. Russian substitution patterns (detected only in Russia or detected in Russia first) are colored in red, substitution patterns detected in single country outside Russia are colored in dark grey, patterns detected in several countries outside Russia are colored in blue. The size of the vertices is increased with the amount of countries this substitution pattern was observed in. Trimmed haplotype networks are created from untrimmed ones by selecting the most likely path to the root. Left: untrimmed haplotype network graph of Delta lineage in Russia. Right: trimmed haplotype network graph of Delta lineage in Russia

We have detected 1012 imports of Delta lineage, 773 imports of BA.1 lineage and 569 imports of BA.2 to Russia. Since Delta existed for about 10 months, while BA.1 and early BA.2 existed for a few months each, the low amount of Delta imports suggests that anti-epidemic measures which were applied in 2021 were effective way to decrease the number of imports of SARS-CoV-2 in Russia. Another interesting result is the comparison of our data with the paper by Klink et al. [4], which suggests that 90 % of Russian Delta genomes were descendants of one single import. Our data shows that only 60 % of Russian Delta genomes were descendants of that import event, while the other 30 % were re-imported repeatedly over 500 times during all of the course of pandemic. *Conclusion:* Haplotype networks are an efficient method of analysis, which allows to visualize and characterize big genomic data during massive pandemic events.

Funding: The study is supported by the subsidy of the Russian Federal Service for Surveillance on Consumer Rights Protection and Human Wellbeing No. 141-02-2023-208.

References

1. O'Toole Á. et al. Pango lineage designation and assignment using SARS-CoV-2 spike gene nucleotide sequences. *BMC Genomics*. 2022;23(1):121. doi 10.1186/s12864-022-08358-2
2. Khare S. et al. GISAID's role in pandemic response. *China CDC Weekly*. 2021;3(49):1049. doi 10.46234/ccdcw2021.255
3. Aksamentov I., Roemer C., Hodcroft E.B., Neher R.A. Nextclade: clade assignment, mutation calling and quality control for viral genomes. *J Open Source Softw*. 2021;6(67):3773. doi 10.21105/joss.03773
4. Klink G.V., Safina K.R., Nabieva E., Shvyrev N., Garushyants S., Alekseeva E. et al. The rise and spread of the SARS-CoV-2 AY. 122 lineage in Russia. *Virus Evol*. 2022;8(1):veac017. doi 10.1093/ve/veac017