Genetic distance among SARS-CoV-2 VOCs and VBMs

Date of analyses

December 2, 2021.

Data source

The reference SARS-CoV-2 genome was downloaded from NCBI's RefSeq (accession no. NC_045512).

Sequences from 2 Variants of Concern (VOC) and 10 Variants Being Monitored (VBM) were downloaded from GISAID's EpiCov. Please see **Acknowledgements** for a complete list of authors from the originating laboratories responsible for obtaining the specimens, and the Submitting laboratories where the genome data were generated and shared via GISAID.

For each variant, we downloaded the first 100 complete genomes (>29,000 bp) submitted to EpiCoV. The original number of sequences was 1,301. A total of 275 sequences with more than 5% of missing characters ("N" or "?") were removed. Therefore, the final alignment has 1,026 sequences. The 5'-UTR and the 3'-UTR regions were removed from the alignment.

Spike prediction and alignment

The 1,026 sequences from VOCs and VBMs were annotated following the strategy described in Machado et al. (2021). We kept 315 unique spike nucleotide sequences with less than 5% of missing data. We aligned these 315 spike sequences based on their translation using `TranslatorX` (Abascalet al., 2010) and MAFFT v7.475 (Katoh, 2013). The following command line was used for TranslatorX:

```
perl TranslatorX.pl -i variants_spike.fasta -c 1 -p F -o alignment_spike
```

Distance estimation

The pairwise *p*-distances between each pair of sequences were calculated using `MEGA` version 11.0.10 (Stecher et al., 2020).

Here, distances are the proportion (*p*) of nucleotide sites at which two sequences being compared are different. It is obtained by dividing the number of nucleotide differences by the total number of nucleotides compared. It does not make any correction for multiple substitutions at the same site, substitution rate biases (for example, differences in the transitional and transversional rates), or differences in evolutionary rates among sites.

Results

It is interesting to note that, although the Omicron RBM should be categorized by nine mutations (S:N440K, S:G446S, S:S447N, S:T478K, S:E484A, S:Q493R, S:G496S, S:Q298R, S:N501Y), at least some sequences have a tenth mutation (S:Y505H). Also, two other mutations (S:N440K and S:G446S) are not present in all Omicron RBM sequences.

The Omicron variant is the variant more distantly related to the reference in the proportion of shared nucleotides. Also, the Omicron is the variant that is more distantly related to the Gamma variant.

The Omicron variant is the variant more distantly related to the reference in the proportion of shared amino acids in the RBM of the spike. Also, the Omicron is the variant that has the RBM sequence that is the most distantly related to Beta, Epsilon, and Kappa.

References

Abascal F, Zardoya R, Telford MJ. TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. *Nucleic Acids Research*. 2010 Jul 1;38(suppl_2):W7-13. https://doi.org/10.1093/nar/gkq291

Jacob Machado D, Scott R, Guirales S, Janies DA. Fundamental evolution of all Orthocoronavirinae including three deadly lineages descendent from Chiroptera-hosted coronaviruses: SARS-CoV, MERS-CoV and SARS-CoV-2. *Cladistics*. 2021:461-88. https://doi.org/10.1111/cla.12454

Katoh S. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*. 2013 30:772-

780.http://mbe.oxfordjournals.org/content/30/4/772

Kumar S, Stecher G, Tamura K. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Molecular Biology and Evolution*. 2016 Mar 22;33(7):1870-4. https://doi.org/10.1093/molbev/msz312

Description of files

- Whole-genome alignment (MAFFT v7.475, `auto` option) of the original dataset: `alignment_1301.fasta`
- Whole genome alignment after duplications and sequnces with more than 5% of missing data were removed: `alignment_1026.fasta`
- Nucleotide alignments of the spike gene (no duplications, only sequences with less than 5% missing data): `spike_315.fasta`
- Amino acid alignments of the RBM (no duplications, only sequences with less than 5% missing data): `rbm_23.fasta`
- Summary of the *p*-distance tables (spike and RBM): `distances.xlsx`

Acknowledgements

Please see the attached PDF files (organized according to the SARS-CoV-2 variant) for a complete list of authors from the originating laboratories responsible for obtaining the specimens, as well as the Submitting laboratories where the genome data were generated and shared via GISAID.

1. Alpha: `VBM-Alpha.pdf`
2. B.1.617.3: `VBM-B.1.617.3.pdf`
3. Beta: `VBM-Beta.pdf`
4. B.1.427: `VBM-Episilon-B.1.427.pdf`
5. B.1.429: `VBM-Episilon-B.1.429.pdf`
6. Eta: `VBM-Eta.pdf`
7. Gamma: `VBM-Gamma.pdf`
8. Iota: `VBM-Iota.pdf`
9. Kappa: `VBM-Kappa.pdf`
10. Mu: `VBM-Mu.pdf`
11. Zeta: `VBM-Zeta.pdf`
12. Delta: `VOC-Delta.pdf`
13. Omicron: `VOC-Omicron.pdf`