**Phylogenetic analysis of the SARS-CoV-2 spike variants**

**Introduction**

The world has been hit by the COVID-19 pandemic caused by SARS-CoV-2, a beta-coronavirus family member. SARS-CoV-2 is a single positive-stranded RNA virus that uses angiotensin-converting enzyme-2 (ACE2) to enter the cells and infect the host. Infected individuals show symptoms such as fever, cough, shortness of breath, fatigue, and body pain (2). Although SARS-CoV-2 has been reported to have a relatively low mutation rate, it has developed various stable mutant strains (1). It has been more than a year, and the pandemic is still far from being under control. One of the primary reasons why we are struggling to control the pandemic is the adaptive evolution of different variants across the continents. There have been many reported cases of different lineages among the SARS-CoV-2 that is harboring significant mutations rendering the variants more infectious and, thus, the higher risk of mortality among the infected individuals.

SARS-CoV-2 uses its spike glycoprotein (S) to fuse with receptors and enter the host cells. S protein is a trimeric protein consisting of approximately 180 kDa monomers with S1 and S2 subunits (3). S1 is responsible for binding to the host receptors cells while S2 helps viral and cellular membrane fusion. As a class I fusion peptide, spike glycoprotein undergoes two proteolytic cleavages: one at the junction of S1 and S2 for priming and the other at S2 for activation.

Chart, box and whisker chart

Description automatically generatedFig 1. Schematic diagram of spike glycoprotein of SARS-CoV-2.

Spike glycoprotein is a 1237 aa long peptide chain containing the receptor-binding domain (RBD) that interacts with ACE-2 receptor cells of the host (Fig. 1). The N-terminal domain (NTD) follows the signal peptide extending from 1 to 13 aa. The RBD extends from 319 aa to 541 aa, followed by fusion peptide (FP) 786 aa to 806 aa. Two heptapeptide repeat sequences (HR1 and HR2) extend up to 1213aa, followed by the transmembrane domain (1213aa-1237aa) and finally the cytoplasmic domain (1237 aa to 1273 aa). There are numerous critical residues in each domain of the S1 and S2 subunit that take part in interactions with the ACE-2. The mutation in such key residues can significantly enhance the strength of the interactions (4).

Since the spike is significant for the infectivity of the virus and the fusion process, any changes in the amino acid residues bring about significant alteration. As the various studies have shown, SARS-CoV-2 is going through rapid adaptive evolution to create various fit variants with multiple substantial changes in the spike protein. This is the reason why there is a growing concern that the battle against this pandemic might lengthen as our immune system continues to struggle to create neutralizing antibodies against rapidly changing spike glycoprotein. In this study, we will be trying to look at the genetic variation among the SARS-CoV-2 spike sequences worldwide and understand where the variation is heading.

**Material and Methods**

The whole-genome data files of SARS-CoV-2 from various samples were obtained from the GISAID website. Global Initiative on Sharing Avian Influenza Data (GISAID) is a non-profit organization that deals with keeping a record of multiple epidemics and pandemics and facilitates sharing such information across the globe. For this project, we have analyzed the viral genome and produced a phylogenetic tree that would be helpful to understand the evolution of the variant. A Python script was used to extract the spike-protein coding region from the whole genome, and they were stored in FASTA format. Mafft was used to align the sequences. The aligned sequences are then used to create a maximum likelihood tree with the help of RAXML-NG (7). We have used the GTR+G model to obtain the maximum likelihood tree. The tree was run through 1000 bootstraps support, and the final tree was observed using iTOL as well as FigTree. The selected sequences were also viewed in MEGA to look for the variation in the amino acid chain.

**Results and Discussions**

We have taken the 26 spike nucleotide sequences from patients from different parts of the world. There are 3 sequences from the USA, 2 from Australia, Egypt, Nepal, Sweden, Rwanda, South Africa, Brazil, and England. Similarly, there is one sequence from Hongkong, India, Italy, Iceland, Finland, and Taiwan. We have also used the SARS-CoV-2 sequence from Wuhan [NC\_045512.2](https://www.ncbi.nlm.nih.gov/nuccore/NC_045512.2?report=fasta&log$=seqview&from=21563&to=25384), which was reported during the beginning of the pandemic. SARS-Coronavirus sequence ([DQ231462.2](https://www.ncbi.nlm.nih.gov/nuccore/DQ231462.2?report=fasta)) obtained from China during the 2003 outbreak of SARS is used as an outgroup for the analysis.

A picture containing application

Description automatically generated

Fig. 2. Bootstrap support tree (using FigTree) for 26 SARS-CoV-2 spike protein gene sequences from various parts of the globe. [We can also view this on the iTOL](https://itol.embl.de/tree/173162223417901619235148).

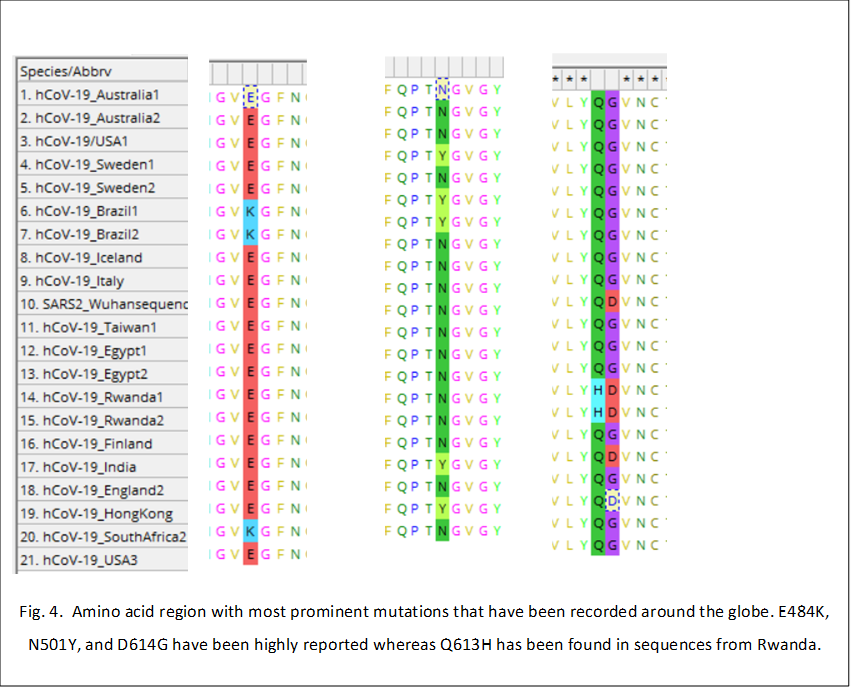
From the tree above in Fig. 2, we can see the variation among the spike sequences. As we have SARS-Coronavirus as an outgroup, we can see that sequences from the different regions have formed different clades with strong bootstrap support values. The bootstrapping converged at 650 replicates with a final loglikelihood value of -8881.281122. The spike sequences from Brazil, Egypt, Rwanda, and South Africa seem to have formed distinct clades with very strong bootstrap support. The bootstrap values for clades, including sequences obtained from Nepal (Nepal1 and Nepal2), Australia (Australia2), and the USA (USA3), seem to have low bootstrap support. This lower support might be the result of the improper alignment of the sequences. Interestingly, the clades of South Africa (SouthAfrica1 and SouthAfrica2) and Brazil (Brazil1 and Brazil2) have been found with significant changes in their coding sequences.

Chart, box and whisker chart

Description automatically generated

Fig 3. Cladogram (using FigTree) showing the relationships among the selected sequences from around the globe.

The cladogram in fig. 3 highlights the clades in a better way. We can see the SARS-CoV-2 is evolving rapidly and forming different clades. It has been documented that a special lineage named B.1.1.7, which is predominant in the UK, has been growing rapidly. It has now spread to 50 countries, and since this lineage shows enhanced transmissibility, the outbreak could be even worse (8). In fig 2 and fig 3, we can clearly see that the sequence obtained from England has formed a separate clade from the rest of the group. The story is the same for the sequences obtained from South Africa and Brazil. The lineage named 2B.1.351 was first detected in South Africa in late 2020 and has become a dominant variant with heightened transmissibility (8).



D614G

Q613H

E484K

N501Y

Similarly, in northern Brazil, a new lineage named P1 has evolved out of nowhere as they saw that 42% of the new samples were found to have such variants (9). In fig. 4, we can observe the critical amino acid change between selected spike sequences. The Brazilian sequences have mutations leading to change in amino acids 484 and 501. It has been found that these variants are highly infectious with a 7-fold increase in binding affinity to ACE2 cells of the human host (5). Similarly, amino acid D614 has mutated to G in most of the sequences, which is found to increase the transmissibility of the virus and increase the infectivity to 9 folds (6).

**Conclusions**

With the help of this study, we can conclude that SARS-Cov-2 is evolving rapidly around the globe, with various mutations in its spike protein sequences giving rise to highly infectious variants. These variants might increase the disease transmission and overcome our effort to control the pandemic. On the other hand, some of the variants might not show any response towards the vaccines that have been developed. The various lineages are emerging and being locally dominant. Such dominant variants then get transferred to other communities exposing thousands of people to the new strain. This study clearly suggests that we need to be proactive in pandemic management and direct out resources to study for another possible outbreak due to evolved mutated variants. A strong network of the scientific community is of utmost importance to find, record, and manage local outbreaks at the local level, preventing them from becoming a massive problem globally.

**References**

1. Ogawa, Junko, et al. *The d614g Mutation in the Sars-cov2 Spike Protein Increases Infectivity in an ace2 Receptor Dependent Manner*. 22 July 2020, [www.ncbi.nlm.nih.gov/pmc/articles/PMC7386486/](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC7386486/)
2. Wang, Dawei, et al. “Clinical Characteristics of 138 Hospitalized Patients With 2019 Novel Coronavirus-Infected Pneumonia in Wuhan, China.” JAMA, American Medical Association, 17 Mar. 2020, [www.ncbi.nlm.nih.gov/pmc/articles/PMC7042881/](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC7042881/).
3. Ou, Xiuyuan, et al. “Characterization of Spike Glycoprotein of SARS-CoV-2 on Virus Entry and Its Immune Cross-Reactivity with SARS-CoV.” Nature News, Nature Publishing Group, 27 Mar. 2020, [www.nature.com/articles/s41467-020-15562-9?report=reader](http://www.nature.com/articles/s41467-020-15562-9?report=reader).
4. Huang, Yuan, et al. “Structural and Functional Properties of SARS-CoV-2 Spike Protein: Potential Antivirus Drug Development for COVID-19.” Nature News, Nature Publishing Group, 3 Aug. 2020, [www.nature.com/articles/s41401-020-0485-4](http://www.nature.com/articles/s41401-020-0485-4).
5. Supasa, Piyuda et al. “Reduced neutralization of SARS-CoV-2 B.1.1.7 variant by convalescent and vaccine sera” Cell, Elsevier Inc. 18 February, 2021, <https://doi.org/10.1016/j.cell.2021.02.033>
6. Zhang, Lizhou et al. “The D614G mutation in the SARS-CoV-2 spike protein reduces S1 shedding and increases infectivity” preprint. <https://www.biorxiv.org/content/10.1101/2020.06.12.148726v1.full.pdf>
7. Katoh, et al. “MAFFT: a Novel Method for Rapid Multiple Sequence Alignment Based on Fast Fourier Transform.” OUP Academic, Oxford University Press, 15 July 2002, academic.oup.com/nar/article/30/14/3059/2904316
8. Wang, Pengfei, et al. Antibody Resistance OF SARS-CoV-2 Variants B.1.351 and B.1.1.7. 8 Mar. 2021, [www.nature.com/articles/s41586-021-03398-2](http://www.nature.com/articles/s41586-021-03398-2).
9. Faria et al “Genomic Characterisation of an Emergent SARS-CoV-2 Lineage in Manaus: Preliminary Findings.” Virological, 12 Jan. 2021, virological.org/t/genomic-characterisation-of-an-emergent-sars-cov-2-lineage-in-manaus-preliminary-findings/586.