International Journal of Advanced Biochemistry Research

ISSN Print: 2617-4693
ISSN Online: 2617-4707
IJABR 2024; SP-8(1): 554-559
www.biochemjournal.com
Received: 15-11-2023
Accepted: 22-12-2023

Manoj Kumar Goud
Ph.D. Scholar, Division of Veterinary Biotechnology, ICAR-IVRI, Izatnagar, Uttar Pradesh, India

Priyanshi Yadav
Ph.D. Scholar, Division of Veterinary Microbiology, ICAR-IVRI, Izatnagar, Uttar Pradesh, India

Ayushi Singh
Ph.D. Scholar, Division of Animal Genetics, ICAR-IVRI, Izatnagar, Uttar Pradesh, India

Rashmi Mishra
Ph.D. Scholar, Division of Veterinary Parasitology, ICAR-IVRI, Izatnagar, Uttar Pradesh, India

Khushboo Panwar
Ph.D. Scholar, Division of Veterinary Microbiology, ICAR-IVRI, Izatnagar, Uttar Pradesh, India

***First two authors contributed equally to the paper**

Corresponding Author:
Priyanshi Yadav
Ph.D. Scholar, Division of Veterinary Microbiology, ICAR-IVRI, Izatnagar, Uttar Pradesh, India

Unveiling SARS-CoV-2 evolution: Insights from nucleotide and protein sequence phylogenetic analysis

Manoj Kumar Goud, Priyanshi Yadav, Ayushi Singh, Rashmi Mishra and Khushboo Panwar

DOI: https://doi.org/10.33545/26174693.2024.v8.i1Sh.396

Abstract
SARS-CoV-2 has evolved since the beginning of the pandemic, generating new virus versions with different characteristics that have replaced pre-existing variants. The study aimed to provide a comprehensive understanding of the evolving landscape of SARS-CoV-2 by analyzing the most recent and diverse nucleotide and amino acid sequences of the surface glycoprotein from the alpha, beta, delta, and omicron variants. The inclusion of sequences from various sources enriched the dataset, offering a more detailed perspective on the virus's genetic diversity and current evolutionary dynamics. The phylogenetic analysis revealed unexpected clustering patterns among the variants, highlighting potential shared genetic characteristics, evolutionary pressures, and convergent evolution. The study's meticulous curation of sequences and exclusion of redundant data aimed to improve the precision of the evolutionary relationships presented. The findings contribute valuable insights into the genetic diversity, adaptation, and global spread of SARS-CoV-2 variants, emphasizing the importance of continued genomic surveillance and in-depth analyses to understand the evolving dynamics of the virus.

Keywords: SARS-CoV-2, spike (S) protein, phylogenetic tree

1. Introduction

Corona Virus Disease 2019 (COVID-19) is still being spread by a novel coronavirus known as severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2). This coronavirus first appeared in December 2019. Coronaviruses of SARS-CoV and MERS-CoV responsible for outbreaks in previous years are also known to result in serious sickness in people, in addition to COVID-19. Since full-length sequence of genome studies have recognized all three of the aforementioned viruses as beta-coronaviruses, SARS-CoV-2 diverges from SARS-CoV and MERS-CoV to form a separate clade on the phylogenic tree. The SARS-CoV-2 genome is a single-stranded, positive-sense RNA of about 30 kb in length, comprising 14 open reading frames (ORFs) and other genes linked to the viral envelope encoding canonical structural proteins (SPs) like spike (S), nucleocapsid, envelope and membrane proteins are found in the remaining downstream segment [1, 2]. The SARS-CoV-2 spike (S) protein, a trimeric class I transmembrane glycoprotein, is crucial for viral entry into host cells. It consists of a large ectodomain with a receptor-binding domain (RBD) in the S1 subunit and a membrane-fusion subunit (S2), which mediates fusion with the host cell membrane. The S protein undergoes extensive structural rearrangement upon interacting with the host cell, allowing viral fusion and entry. The S2 subunit includes a fusion peptide and other machinery for mediating membrane fusion, ultimately enabling the viral genome to enter the host cell cytoplasm [3].

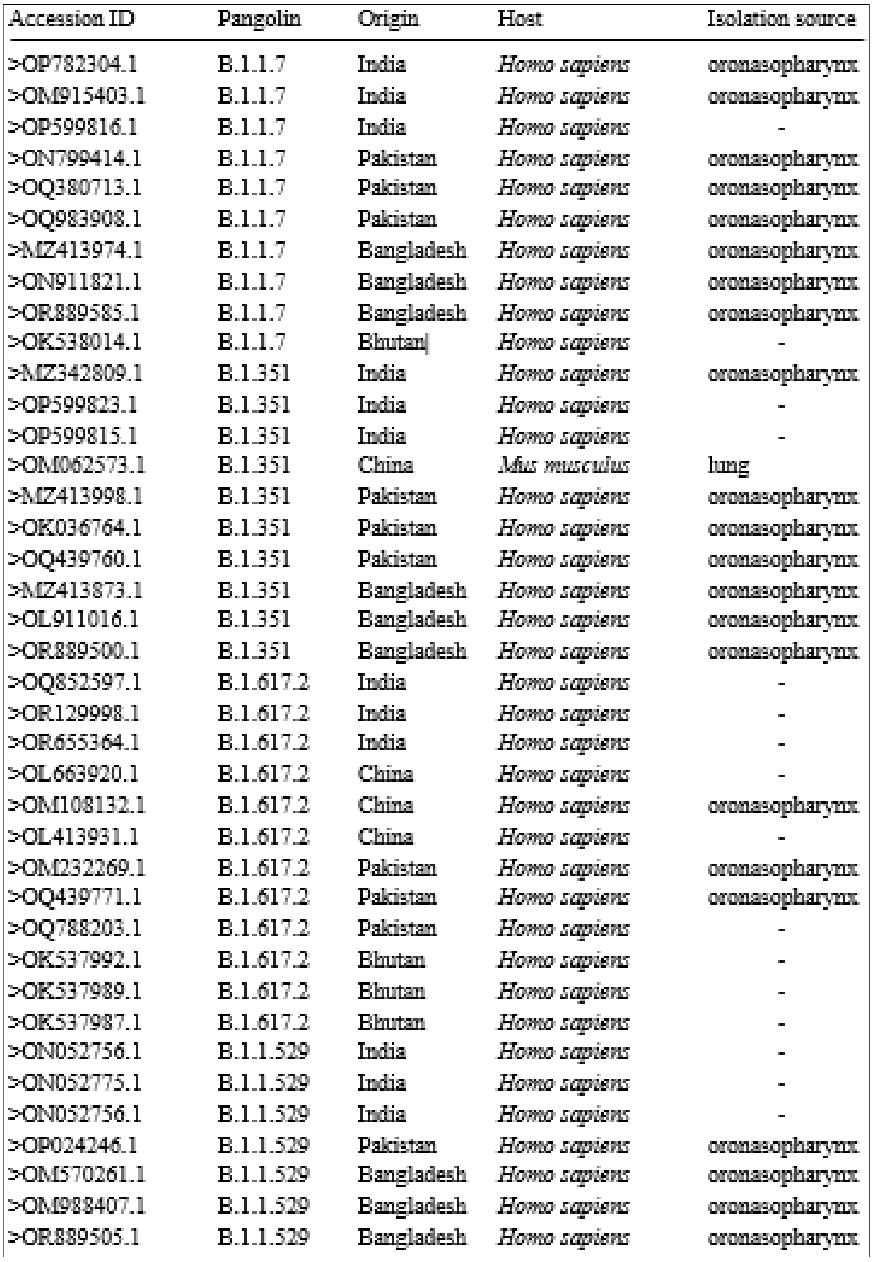
The SARS-CoV-2 virus has been rapidly evolving, leading to the generation of Variants of Concern (VoC), which have shown increased fitness. The key mutations are present in These VoC viruses of the spike protein, allowing better survival and evasion of host defense mechanisms. The D614G mutation in the spike domain is found in the majority of VoC, including Alpha, Delta, Omicron, and its current variants [4]. The development of vaccines and antibody treatments has mostly targeted the SARS-CoV-2 spike protein, which is an essential part of the viral entry process. Effective vaccines and antiviral medications against. SARS-CoV-2 need a thorough understanding of the structural, molecular, and mutational characteristics of the S protein.

2. Materials and Methods

2.1 SARS-Cov-2 isolates

The isolates of SARS-Cov-2 were retrieved from the NCBI virus [8] (National Centre of Biotechnology Information, USA) database. (Table 1).

Table 1: SARS-Cov-2 isolates downloaded from NCBI Virus database



Accession ID	Pangolin	Origin	Host	Isolation source
>OP782304.1	B.1.1.7	India	*Homo sapiens*	oronasopharynx
>OM915403.1	B.1.1.7	India	*Homo sapiens*	oronasopharynx
>OP599816.1	B.1.1.7	India	*Homo sapiens*	-
>ON799414.1	B.1.1.7	Pakistan	*Homo sapiens*	oronasopharynx
>OQ380713.1	B.1.1.7	Pakistan	*Homo sapiens*	oronasopharynx
>OQ983908.1	B.1.1.7	Pakistan	*Homo sapiens*	oronasopharynx
>MZ413974.1	B.1.1.7	Bangladesh	*Homo sapiens*	oronasopharynx
>ON911821.1	B.1.1.7	Bangladesh	*Homo sapiens*	oronasopharynx
>OR889585.1	B.1.1.7	Bangladesh	*Homo sapiens*	oronasopharynx
>OK538014.1	B.1.1.7	Bhutan	*Homo sapiens*	-
>MZ342809.1	B.1.351	India	*Homo sapiens*	oronasopharynx
>OP599823.1	B.1.351	India	*Homo sapiens*	-
>OP599815.1	B.1.351	India	*Homo sapiens*	-
>OM062573.1	B.1.351	China	*Mus musculus*	lung
>MZ413998.1	B.1.351	Pakistan	*Homo sapiens*	oronasopharynx
>OK036764.1	B.1.351	Pakistan	*Homo sapiens*	oronasopharynx
>OQ439760.1	B.1.351	Pakistan	*Homo sapiens*	oronasopharynx
>MZ413873.1	B.1.351	Bangladesh	*Homo sapiens*	oronasopharynx
>OL911016.1	B.1.351	Bangladesh	*Homo sapiens*	oronasopharynx
>OR889500.1	B.1.351	Bangladesh	*Homo sapiens*	oronasopharynx
>OQ852597.1	B.1.617.2	India	*Homo sapiens*	-
>OR129998.1	B.1.617.2	India	*Homo sapiens*	-
>OR655364.1	B.1.617.2	India	*Homo sapiens*	-
>OL663920.1	B.1.617.2	China	*Homo sapiens*	-
>OM108132.1	B.1.617.2	China	*Homo sapiens*	oronasopharynx
>OL413931.1	B.1.617.2	China	*Homo sapiens*	-
>OM232269.1	B.1.617.2	Pakistan	*Homo sapiens*	oronasopharynx
>OQ439771.1	B.1.617.2	Pakistan	*Homo sapiens*	oronasopharynx
>OQ788203.1	B.1.617.2	Pakistan	*Homo sapiens*	-
>OK537992.1	B.1.617.2	Bhutan	*Homo sapiens*	-
>OK537989.1	B.1.617.2	Bhutan	*Homo sapiens*	-
>OK537987.1	B.1.617.2	Bhutan	*Homo sapiens*	-
>ON052756.1	B.1.1.529	India	*Homo sapiens*	-
>ON052775.1	B.1.1.529	India	*Homo sapiens*	-
>ON052756.1	B.1.1.529	India	*Homo sapiens*	-
>OP024246.1	B.1.1.529	Pakistan	*Homo sapiens*	oronasopharynx
>OM570261.1	B.1.1.529	Bangladesh	*Homo sapiens*	oronasopharynx
>OM988407.1	B.1.1.529	Bangladesh	*Homo sapiens*	oronasopharynx
>OR889505.1	B.1.1.529	Bangladesh	*Homo sapiens*	oronasopharynx

2.2 Nucleotide sequence preparation

SARS-CoV-2 nucleotide sequences (surface glycoprotein gene) of the above isolates were retrieved from the GenBank® Database. Multiple sequence alignment of the sequences was done by Multiple Alignment using Fast Fourier Transform (MAFFT) by European Molecular Biology Laboratory-European Bioinformatics Institute (EMBL-EBI) on the command line.

2.3 Amino acid sequence preparation

SARS-CoV-2 amino acid sequence (surface glycoprotein protein) of the above isolates were retrieved from the GenBank® Database. Multiple sequence alignment of the amino acid sequence was done by Multiple Alignment using the Fast Fourier Transform (MAFFT) by the European Molecular Biology Laboratory-European Bioinformatics Institute (EMBL-EBI) [5] on the command line

2.4 Molecular phylogenetic analysis

The purpose of a phylogenetic tree in evolutionary biology is to illustrate the evolutionary history and relatedness of different organisms or sequences and to provide insights into their common ancestry, diversification, and adaptation over time [10]. Evolutionary Genetics Analysis 11 (MEGA 11) (6) was employed to find out the best model fit for nucleotide and amino acid sequences with the maximum likelihood method [11]. The analysis is performed on 1000 bootstrapped [9] input datasets to validate the phylogenetic tree and cross-referencing is done with the Tamura-Nei substitution model for nucleotide sequences, and for amino acid sequences the analysis is performed on 1000 bootstrapped input datasets to validate the phylogenetic tree and cross-referencing is done with the Jones-Taylor-Thornton model by using Randomized Axelerated Maximum Likelihood (RAxML) [7].

3. Results and Discussion

3.1 Results

We wanted to capture a thorough representation of the constantly changing landscape of the coronavirus by combining the most recent and diverse sequences. In the present study, we have retrieved the most recent nucleotide sequences and amino acid sequences of the surface glycoprotein of the four major strains alpha, beta, delta, and omicron from different submitters from five different countries in South Asia. The inclusion of sequences from different sources seeks to enrich our dataset, providing a larger and more nuanced perspective on the virus's genetic diversity. By adding recent submissions, we tried to depict the virus's current evolutionary dynamics, considering any emergent lineages or variants. This approach also helps in a more accurate depiction of the ongoing molecular evolution of the virus.

The phylogenetic analysis of alpha and beta variants from diverse geographical locations has revealed a noteworthy and unexpected clustering, as sequences from disparate regions consistently group within the same clade. Interestingly, in stark contrast, a distinct clustering pattern emerges for sequences associated with the delta variant, with a subset forming a cohesive clade. Intriguingly, within the intricate tapestry of the phylogenetic tree, omicron variant sequences appear to bridge between the disparate clades formed by delta variant sequences. The omicron variants from different geo-locations are clustered indicating the evolutionary convergence of these sequences. Whereas, two omicron sequences from India analyzed in this study show different clades displayed disparate placements within the phylogenetic tree, each forming distinct clades. This divergence in clade formation among Omicron variants from the same geographical origin suggests intricate genetic heterogeneity and potential sub-lineages within the local viral population. The observation prompts questions about localized evolutionary dynamics, selective pressures, or introduction events that may have influenced the genetic divergence of the Omicron variants. In the analysis, one of the delta clades is positioned proximately to the alpha clade within the evolutionary tree This close phylogenetic proximity suggests potential genetic affinities or shared evolutionary traits between these prominent SARS-CoV-2 variants, shedding light on intricate patterns of viral evolution.

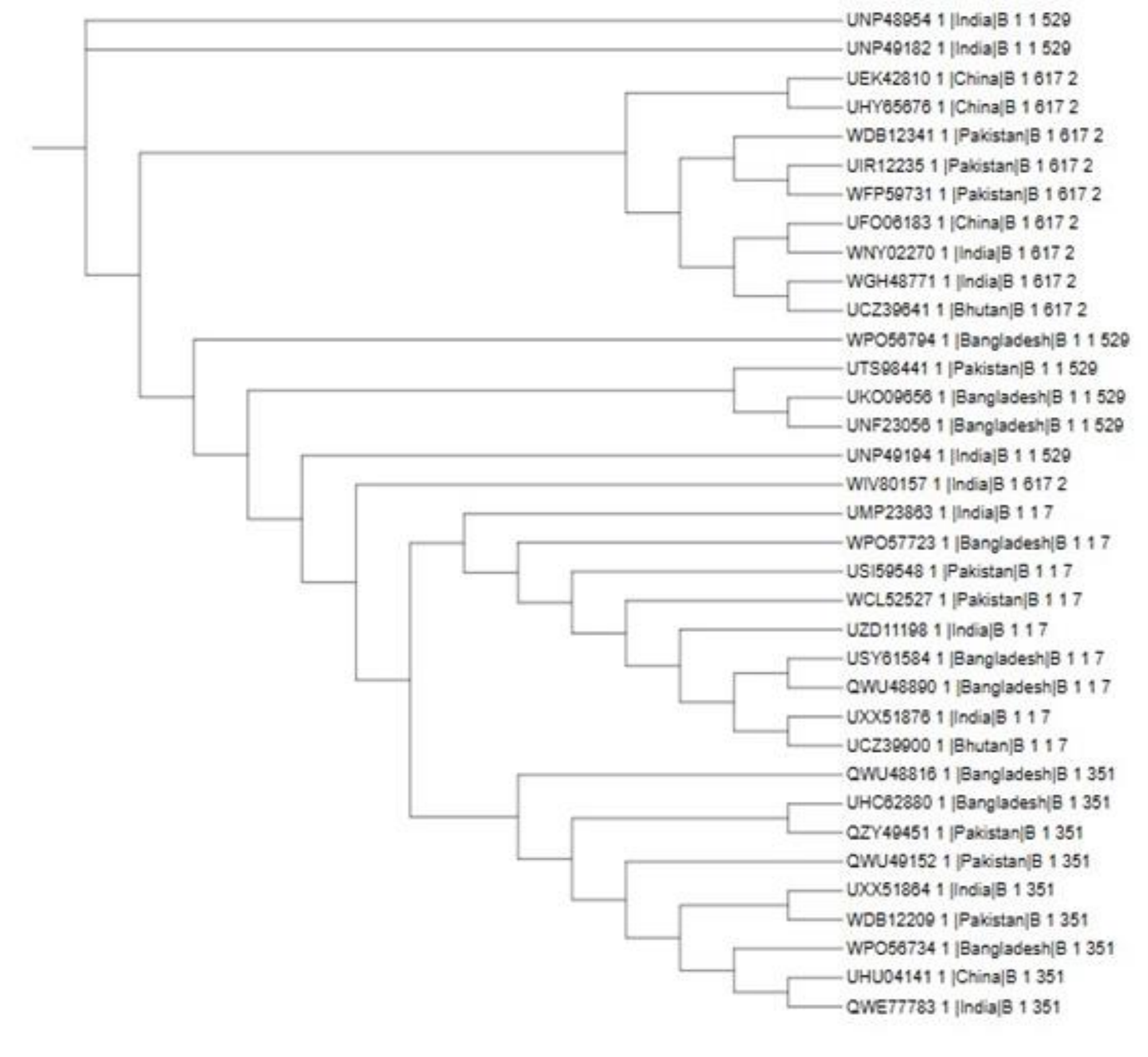

UNP48954 1 |India|B 1 1 529
UNP49182 1 |India|B 1 1 529
UEK42810 1 |China|B 1 617 2
UHY65676 1 |China|B 1 617 2
WDB12341 1 |Pakistan|B 1 617 2
UIR12235 1 |Pakistan|B 1 617 2
WFP59731 1 |Pakistan|B 1 617 2
UFO06183 1 |China|B 1 617 2
WNY02270 1 |India|B 1 617 2
WGH48771 1 |India|B 1 617 2
UCZ39641 1 |Bhutan|B 1 617 2
WPO56794 1 |Bangladesh|B 1 1 529
UTS98441 1 |Pakistan|B 1 1 529
UKO09656 1 |Bangladesh|B 1 1 529
UNF23056 1 |Bangladesh|B 1 1 529
UNP49194 1 |India|B 1 1 529
WIV80157 1 |India|B 1 617 2
UMP23863 1 |India|B 1 1 7
WPO57723 1 |Bangladesh|B 1 1 7
USI59548 1 |Pakistan|B 1 1 7
WCL52527 1 |Pakistan|B 1 1 7
UZD11198 1 |India|B 1 1 7
USY61584 1 |Bangladesh|B 1 1 7
QWU48890 1 |Bangladesh|B 1 1 7
UXX51876 1 |India|B 1 1 7
UCZ39900 1 |Bhutan|B 1 1 7
QWU48816 1 |Bangladesh|B 1 351
UHC62880 1 |Bangladesh|B 1 351
QZY49451 1 |Pakistan|B 1 351
QWU49152 1 |Pakistan|B 1 351
UXX51864 1 |India|B 1 351
WDB12209 1 |Pakistan|B 1 351
WPO56734 1 |Bangladesh|B 1 351
UHU04141 1 |China|B 1 351
QWE77783 1 |India|B 1 351

Fig 1: Phylogenetic tree of nucleotide sequences

Our protein sequence-based phylogenetic analysis revealed a consistent and cohesive tree structure; however, a nuanced technique was used to improve the precision of the evolutionary relationships presented. Notably, a group of sequences shared a significant degree of similarity, prompting a strategic decision to eliminate redundant or closely related sequences to reduce potential biases in the tree topology. This meticulous curation sought to improve the representation of evolutionary diversity and highlight distinct genetic markers within the dataset. By deleting sequences with significant similarity, our technique prioritized the inclusion of distinct and useful data points, contributing to a more accurate portrayal of the evolutionary relationships among the examined protein sequences. This method ensures that the resulting phylogenetic tree reflects true genetic distinctions and evolutionary dynamics, allowing for a more accurate understanding of the complex relationships between the analyzed sequences.

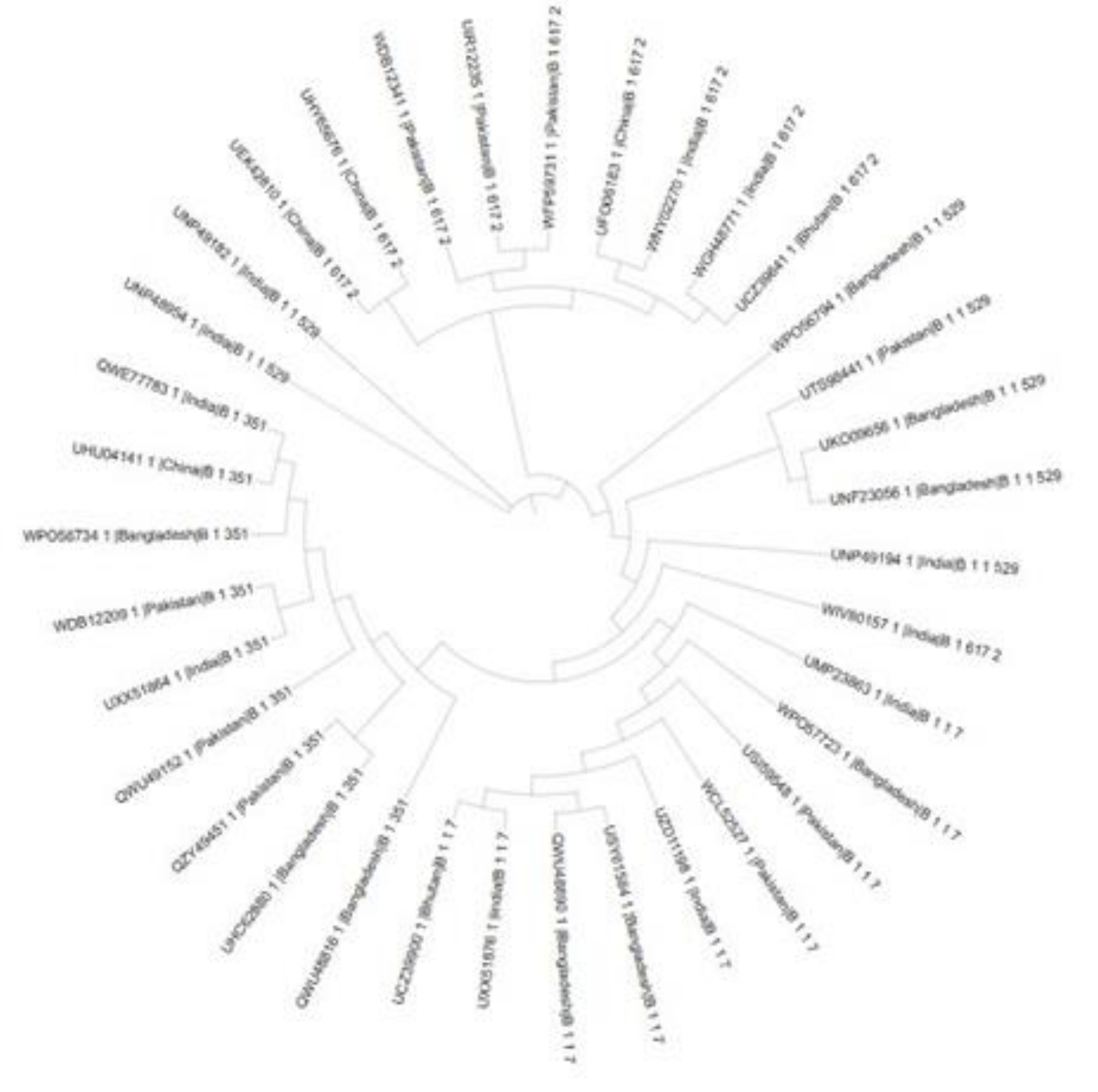

WFP49731 1 |Pakistan|B 1 617 2
UFO06183 1 |China|B 1 617 2
WNY02270 1 |India|B 1 617 2
WGH48771 1 |India|B 1 617 2
UCZ39841 1 |Bhutan|B 1 617 2
WPO56794 1 |Bangladesh|B 1 1 529
UTS98441 1 |Pakistan|B 1 1 529
UKO00656 1 |Bangladesh|B 1 1 529
UNF23056 1 |Bangladesh|B 1 1 529
UNP49194 1 |India|B 1 1 529
WIV80157 1 |India|B 1 617 2
UMP23863 1 |India|B 1 1 7
WPO57723 1 |Bangladesh|B 1 1 7
USI50648 1 |Pakistan|B 1 1 7
WCL52527 1 |Pakistan|B 1 1 7
UZD11196 1 |India|B 1 1 7
USY61584 1 |Bangladesh|B 1 1 7
QWU48890 1 |Bangladesh|B 1 1 7
UXX51676 1 |India|B 1 1 7
UCZ39900 1 |Bhutan|B 1 1 7
QWU48816 1 |Bangladesh|B 1 351
UHC62880 1 |Bangladesh|B 1 351
QZY45451 1 |Pakistan|B 1 351
QWU49152 1 |Pakistan|B 1 351
UXX51864 1 |India|B 1 351
WDB12209 1 |Pakistan|B 1 351
WPO56734 1 |Bangladesh|B 1 351
UHU04141 1 |China|B 1 351
QWE77783 1 |India|B 1 351
UNP48954 1 |India|B 1 1 529
UNP49182 1 |India|B 1 1 529
UEK42610 1 |China|B 1 617 2
UHY65676 1 |China|B 1 617 2
WDB12341 1 |Pakistan|B 1 617 2
USR12235 1 |Pakistan|B 1 617 2

Fig 2: Phylogenetic tree of the protein sequences

3.2 Discussion

Our nucleotide and protein sequence-based phylogenetic analyses have shown several key observations and methodological considerations. The nucleotide sequence analysis provided insights into the genetic relationships among different variants, notably highlighting the distinct evolutionary trajectories of the alpha, beta, and delta variants. The unexpected clustering of alpha and beta variants from diverse geographical locations within the same clade suggested potential shared genetic characteristics or evolutionary pressures.

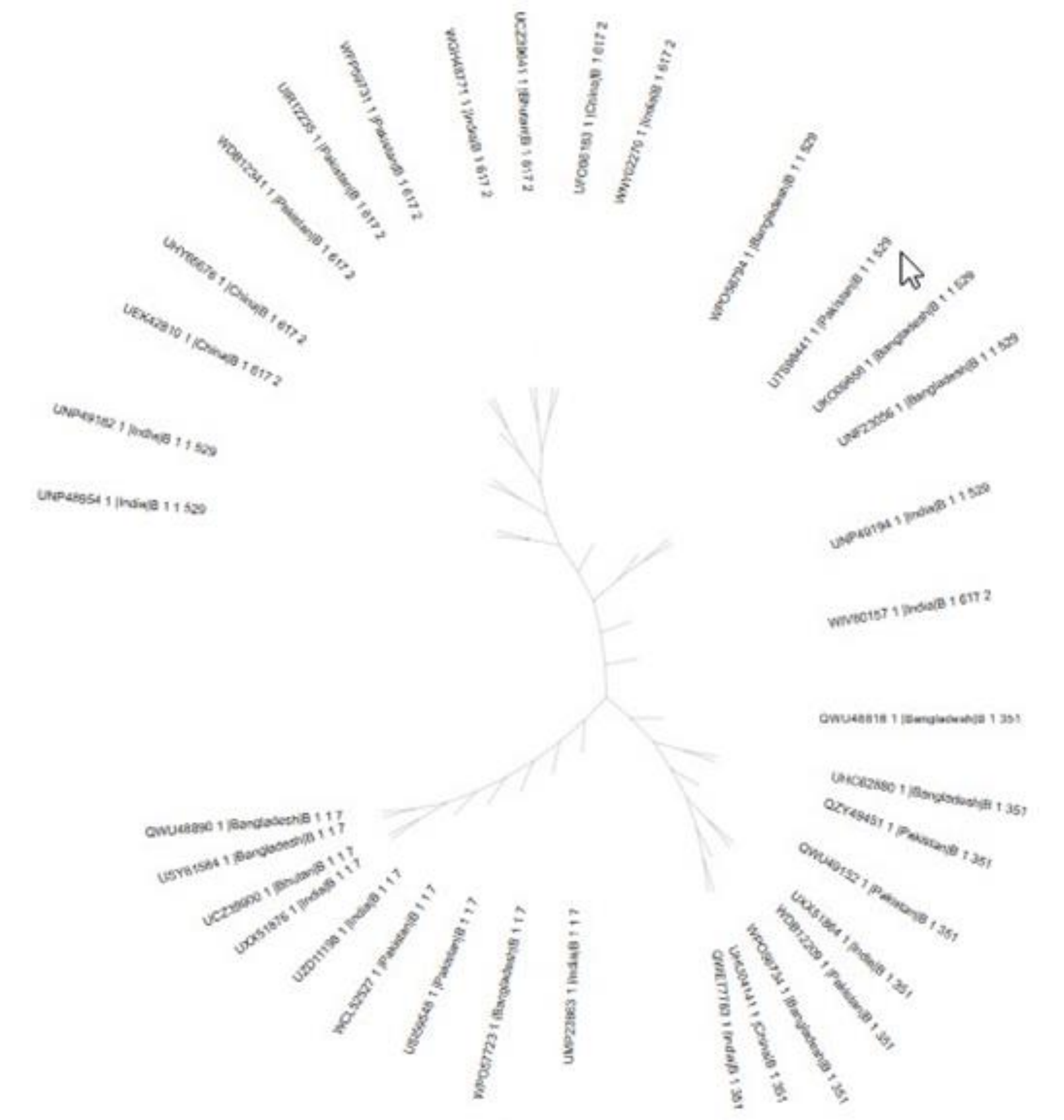

UNP48954 1 |India|B 1 1 529
UNP49182 1 |India|B 1 1 529
UEK42810 1 |China|B 1 617 2
UHY66676 1 |China|B 1 617 2
WDB12341 1 |Pakistan|B 1 617 2
UBI12Z25 1 |Pakistan|B 1 617 2
WFP59731 1 |Pakistan|B 1 617 2
WGH48771 1 |India|B 1 617 2
UCZ39841 1 |Bhutan|B 1 617 2
UFO66183 1 |China|B 1 617 2
WWY02270 1 |India|B 1 617 2
WPO58794 1 |Bangladesh|B 1 1 529
UTS98441 1 |Pakistan|B 1 1 529
UKO09650 1 |Bangladesh|B 1 1 529
UNF23056 1 |Bangladesh|B 1 1 529
UNP49194 1 |India|B 1 1 529
WIV60157 1 |India|B 1 617 2
QWU48818 1 |Bangladesh|B 1 351
UHC82880 1 |Bangladesh|B 1 351
QZY49451 1 |Pakistan|B 1 351
QWU49152 1 |Pakistan|B 1 351
UXX51864 1 |India|B 1 351
WDB12209 1 |Pakistan|B 1 351
WPO58734 1 |Bangladesh|B 1 351
UHS04141 1 |China|B 1 351
QWE77183 1 |India|B 1 351
UMP22863 1 |India|B 1 1 7
WPO57723 1 |Bangladesh|B 1 1 7
USS56548 1 |Pakistan|B 1 1 7
WCL52527 1 |Pakistan|B 1 1 7
UZD11198 1 |India|B 1 1 7
UXX51876 1 |India|B 1 1 7
UCZ39900 1 |Bhutan|B 1 1 7
USY61584 1 |Bangladesh|B 1 1 7
QWU48890 1 |Bangladesh|B 1 1 7

Fig 3: Unrooted tree of the protein sequences

On the other hand, the protein sequence analysis corroborated the general tree structure observed in the nucleotide analysis, reaffirming the evolutionary relationships among variants. The decision to exclude closely related protein sequences enhanced the precision of the analysis, reducing redundancy and potential bias in the tree topology. The identification of convergent evolution among a subset of sequences further emphasized the dynamic interplay between genetic determinants and environmental pressures, contributing to the cohesive clustering observed.

Notably, the distinct phylogenetic placements of Omicron variants from the same geographic location underscored the complex evolutionary dynamics within this emerging lineage. The surprising divergence in clade formation among Omicron variants prompts further investigation into localized evolutionary pressures or introduction events that may have influenced their genetic divergence.

The close phylogenetic proximity of one delta clade to the alpha clade raises intriguing questions about potential genetic affinities or shared evolutionary traits between these prominent SARS-CoV-2 variants. This proximity suggests a complex interplay of genetic factors that may influence the adaptive landscape of the virus.

In summary, our nucleotide and protein sequence analyses have provided a nuanced understanding of the evolutionary relationships among different variants. The careful curation of sequences, exclusion of redundant data, and identification of convergent evolution have refined our interpretation of the phylogenetic trees. These findings contribute valuable insights into the genetic diversity, adaptation, and global spread of SARS-CoV-2 variants, emphasizing the importance of continued genomic surveillance and in-depth analyses to stay abreast of the evolving dynamics of the virus.

4. Conclusion

The use of the phylogenetic tree construction in the analysis of coronaviruses has provided detailed insights into the evolution and mutation of SARS-CoV-2. This approach has challenged the assumption of mutation of nucleotide and protein gene of SARS-CoV-2 and revealed various common mutations in SARS-CoV-2. The timely monitoring of the variation and evolution of SARS-CoV-2 is crucial for the treatment, control, and prevention of COVID-19 and its future outbreaks. Therefore, the application of the phylogenetic tree method has significantly contributed to our understanding of the evolutionary dynamics of SARS-CoV-2, which is essential for informing public health strategies and interventions.

5. References

1. Gao Y, Yan L, Huang Y, Liu F, Zhao Y, Cao L, *et al*. Structure of the RNA-dependent RNA polymerase from COVID-19 virus. Science. 2020 May 15;368(6492):779-82.
2. Michel CJ, Mayer C, Poch O, Thompson JD. Characterization of accessory genes in coronavirus genomes. Virology journal. 2020 Dec;17(1):1-3.
3. Wrapp D, Wang N, Corbett KS, Goldsmith JA, Hsieh CL, Abiona O, *et al*. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. Science. 2020 Mar 13;367(6483):1260-3.
4. Khatri R, Siddqui G, Sadhu S, Maithil V, Vishwakarma P, Lohiya B, *et al*. Intrinsic D614G and P681R/H mutations in SARS-CoV-2 VoCs Alpha, Delta, Omicron and viruses with D614G plus key signature mutations in spike protein alters fusogenicity and infectivity. Med Microbiol Immunol. 2023 Feb;212(1):103-122. DOI: 10.1007/s00430-022-00760-7. Epub 2022 Dec 30. PMID: 36583790; PMCID: PMC9801140.
5. EMBL-EBI. Search and sequence analysis tools services from EMBL-EBI in 2022.
6. Tamura K, Stecher G, Kumar S. MEGA11: Molecular Evolutionary Genetics Analysis version 11 (Tamura, Stecher, and Kumar 2021).
7. Stamatakis A. RAxML Version 8: A tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies. Bioinformatics. 2014;30(9):1312-1313.
8. NCBI Virus (nih.gov).
9. Felsenstein J. Confidence limits on phylogenies: An approach using the bootstrap. Evolution. 1985;39(4):783-791.
10. Singh A, Yadav P, Singh D, Vempadapu V. Phylogenetic tree analysis of HSP4A gene across important livestock taxa. The Pharma Innovation Journal. 2023;12(11S):495-498.
11. Felsenstein J. Evolutionary trees from DNA sequences: A maximum likelihood approach. Journal of Molecular Evolution. 1981;17(6):368-376.