Title: Temporal Pattern of Mutation Accumulation in SARS-CoV-2 Proteins: Insights from Whole Genome Sequences Pan-India Using Data Mining Approach

Author:
Chakrakodi N Varun [1]

Affiliation:

[1] Department of Neurovirology
National Institute of Mental Health & Neuro Sciences (NIMHANS)
Bangalore, India

Email: varuncn.micro@gmail.com
ORCID ID: 0000-0003-2450-4601

Abstract

Mutation is a fundamental factor that affects host-pathogen biology and consequently viral survival and spread. Close monitoring and observation of such mutation help decipher essential changes in the SARS Cov2 genome. A plethora of mutations have been documented owing to increased whole genomic sequencing. Understanding how conserved the specific mutations are and the temporal pattern of mutation accumulation is of paramount interest. Using an in-house data mining approach, pan- India data was mined and analysed for 26 proteins expressed by SARS-CoV-2 to understand the spread of mutations over 28 months (January 2021- April 2023). It was observed that proteins such as Nsp3, Nsp4, ORF9b, among others, acquired mutations over the period. In contrast, proteins such as Nsp6- 10 were highly stable, with no detectable conserved mutations. Further, it was observed that many of the mutations that were highly prevalent in the delta variants were not observed in the omicron variants, which probably influenced the host-pathogen relationship. The study attempts to catalogue and focus on well-conserved mutations across all the SARS-CoV-2 proteins, highlighting the importance of understanding non-spike mutations.

Keywords: Molecular evolution; SARS-CoV-2 proteins; Mutation;

1. Introduction:

Mutation and evolution are fundamental processes that influence host-pathogen biology and, thereby, virus survival and spread. Constant surveillance and monitoring to decipher new and high-frequency mutations in the SARS-CoV-2 genome are thus essential to understand their adaptation. SARS-CoV-2 which was previously declared as a public health emergency (1), has recently been de-escalated from its pandemic status by the WHO (2). Understanding the dynamics of these mutations over a more extended period is essential and provides significant leads to understanding virus evolution. With the availability of a rapidly increasing number of open-access whole genome sequence data for SARS-CoV-2, there has been a growing interest in data mining and exploring genetic variations (3,4).

The 30 Kbp SARS-CoV-2 genome consists of a single-stranded + sense RNA. The genome encodes 12 open reading frames (ORFs), supporting the expression of 22 non- structural proteins (NSPs) and four structural proteins (spike, envelope, membrane and nucleocapsid). Further, novel coding transcripts have also been proposed (5–7). The SARS-CoV-2 mutation rate is estimated at 10^{-6} mutations per nucleotide per replication cycle (8). However, various parts of the genome accumulate mutations at various degrees, probably due to selection pressure. For example, spike protein, the primary target of vaccines, has been rapidly evolving over the period.

Though a vast volume of published literature catalogues various mutations in various genes of SARS-CoV-2, many of these mutations have occurred in selective subvariants or at low frequencies, thereby losing focus on conserved mutations over a longer period across multiple proteins. The analysis presented herein provides a catalogue of mutations on the SARS-CoV-2 evolution at an amino acid sequence level. In summary, pan-India data was mined and analysed for 26 proteins expressed by SARS-CoV-2 to understand the breadth of conserved mutations over 28 months (January 2021- April 2023). As expected, the spike shows a rapidly accumulating number of mutations; other proteins, such as nsp3 and nsp4, show a slow accumulation of mutations over the period. Parallelly, Nsp’s 7-10 were found to be highly stable.

2. Methodology:

2.1 Data Collection and raw data availability:

All the genomics and metadata reported in the analysis herein were downloaded from the GISAID's EpiCoV database. In brief, the database was searched for SARS-CoV-2 original, high-coverage sequences for samples collected from January 2021 to April 2023. A total of 71,878 sequences were downloaded for screening. The relevant GISAID Identifiers and their doi are provided in Supplementary data 1.

2.2 Data pre-processing for downstream analysis:

The data were screened manually, and partial genome sequences were removed from further analysis. The whole genome data was then subject to analysis using the Nextclade tool (v2.13.0) (9) using Wuhan-Hu-1/2019 (MN908947) as a reference sequence. Month-wise protein sequences were derived into a multi- fasta file. The overall qc (quality control) status for each sequence, based on the qc score, was obtained in a .tsv file. Only those sequences assessed as "good" by the nextclade algorithm were taken for further downstream analysis. The workflow summary is presented in Figure 1.

2.3 Mutation analysis:

The aligned protein sequences for all the proteins, month-wise, were analysed using an in-house developed program. In brief, the program was written using Jupyter notebook (v 6.4.12; Anaconda Navigator- 2.3.2). The individual sequences were parsed using biopython (v1.81) (10). The codes were further optimised using the "pandas" library. Each sequence that scored as "good" in the nextclade was piped for further analysis. For nsp1- 16, the ORF1ab preprotein was computationally spliced into individual proteins based on the published coordinates (11). Nsp11 and ORF8 were ignored for analysis. Nextclade output-derived multi-fasta files were directly used as an input for all the protein mutation analysis. Each amino acid site between the sample and reference sequence was checked for mutation using a mutation function as given below. Wherever the amino acid was unknown (X), the code ignored them as a mutation. The total mutation for each protein sequence was imported into a .xls file.

```
for reference, test in zip (Reference, test):
        if reference! = test and test! = "X":
```

mutations+ = 1.

2.4 Identification of signature substitution mutations

The pipeline for studying substitution mutations of interest was developed using python and biopython scripts and further optimised using “pandas” library. In essence, the program looked for differences between the reference and the test sequence and reported amino acid change by position, which occurred in at least 80% of the analysed sequences.

3. Results & Discussion:

Of the total 71,878 sequences that were retrieved- 66,405 sequences had a full-length genome sequence suitable for downstream analysis. Amongst these, 49,158 sequences with sample collection dates ranging from January 2021- April 2023 had a good qc status and were taken for further analysis.

3.1 S protein

A highly mutating protein generally indicates an active evolutionary process. As already reported by several groups, the S- protein accumulated multiple mutations over the period. (Figure 2A). Since there is no international defining standard cut-off for the frequency of mutations to designate them as “conserved”, I chose 80% as calculative cut- off for determining “highly conserved” mutations. A set of mutations observed in some of the critical lineages reported from India is depicted in Figure 2B. For instance, more recently found strains, XBB 1.16 and XBB 1.15 were found to carry F486P and F490S mutations. F486P enhances SARS-CoV-2 binding to the ACE2 receptor (12). F490S, previously noted in the lambda variant, confers the ability to escape vaccination immunity (13). One of the first mutations, D614G was consistently observed across all the studied sequences (14).

3.2 N, E and M protein

Though there was a rapid increase in the number of mutations in N protein at the initial period, the number of mutations remained stable over a longer time (Figure 3A). However, the specific amino acid mutations differed between delta and omicron variants. The D63G mutation, highly prevalent in delta variants, was not observed in omicron variants. In

contrast, the mutation G204R not observed in delta was commonly noted in omicron (Table 1).

E protein showed random mutations in the earlier period and later acquired a stable T9I mutation. The T9I mutation is expected to reduce the apoptotic effects on the cell (15). Recently, an additional T11A mutation was observed, which is proposed to reduce cell lethality and lung damage in mouse models (16). M protein showed mutation accumulation similar to E protein. I82T mutation acquired very early, and enhances glucose uptake during viral replication (17), was not observed in omicron. Together, the data suggest that the omicron variant mutations are favoured to be milder than their delta counterpart.

3.3 Non- Structural proteins (Nsp's)

SARS-CoV-2 ORF1ab encodes a total of 16 non-structural proteins. SARS-CoV-2 Nsps are produced biologically by proteolytic cleavage of larger polyproteins. The individual protein sequences were similarly derived using computational slicing as detailed under methodology. Nsp11 encodes a short 13 aa protein, similar to the first segment of Nsp12 and hence was ignored for mutation analysis. It has been previously recognised that ORF1ab accumulated significant mutations, many of which have co-evolved with spike mutations (18,19). It was observed that Nsp1, Nsp3, Nsp4, Nsp5, Nsp12, Nsp13, Nsp14 and Nsp15 accumulated a few significant sets of mutations (Table 2). Other Nsp's, including Nsp2, Nsp 7- 10 and Nsp16, had an extremely low rate of mutation and no recognisable conserved mutations.

Notably, the G671S mutation in the RNA-dependent RNA polymerase (RdRp; Nsp12) prevalent in delta is not observed in omicron. G671S mutation facilitates enhanced replication and increased transmission (20). Further, a three amino acid deletion (ΔSGF) mutation in Nsp6 (106- 108) was observed in more than 80% of omicron variants. Omicron-derived Nsp6 and spike together have shown a preferential upper respiratory tract infection (21). The evidence of these mutation patterns indicates a propensity of omicron towards causing a milder infection. It should be noted that with new emerging variants, this scenario is changing further. For instance, additional mutations such as K47R (Nsp1), L260F (Nsp6) and S36P (Nsp13) were observed in XBB1.16 for which significance is currently not well understood.

3.4 Other accessory proteins

ORF3a, a large accessory protein, over the time period has shown several mutations occurring at a low frequency. Previous studies have suggested several mutations (22), most of which were observed at a low frequency. S26L was found to be a well-conserved mutation earlier, while omicron variants carry T223I, which is predicted to be a destabilising mutation though its functional relevance is unclear (23). ORF6 has acquired a stable mutation- D61L, which probably reduces the ability of the virus to counteract innate immune response (24). ORF7a showed two conserved mutations in delta variants- V82A and T120I. Both these mutations were not present in the omicron variants. Since these mutations occur within the functional domain, the mutational drop is expected to reduce virulence (25). ORF7b has been largely stable, with no identifiable conserved mutations.

ORF9b has recently accumulated a few significant mutations. The accessory protein is known to localise on the mitochondrial membrane and modulate the IFN-I response (26). Currently, the effects of these mutations are unclear. N55S mutation maps to the region that interacts with TOM70, while the I5T mutation is located in the N terminal. Hence, it is speculated that these mutations will modulate the TOM70-ORF9b structure (18).

A detailed mutation analysis for ORF8 protein was not performed since many sequences showed the appearance of stop codons leading to nonsense mutations. It has been suggested that ORF8 is an immunomodulator but not an essential protein (27); hence, loss of protein expression is likely to be of little or no consequence. It was observed that the number of sequences that showed a premature stop codon increased over the period. Indeed, about 91% of the XBB.1.16 sequences showed a stop codon at the 8th position (G8*).

4. Limitations:

The data mining relies mainly on the open data available through the GISAID database and the nextclade analysis tool. Firstly, the GISAID database probably does not host all the SARS CoV2 whole genome sequences (28,29). Second, the nextclade analysis tool assigns a qc status based on certain assumptions in analysis, which may not always be true (9). Third, the analysis pipeline was designed to omit amino acid sequences designated “X” since they represent an unknown amino acid prediction. Though QC scoring and filtering ensures to remove a majority of those sequences with a large number of “X” in them which may bias the overall analysis, it is acknowledged that those with a good score can still have a few

unknown amino acids. Since there is an equal probability that the amino acid is the wild type or a mutation, as a benefit of the doubt, they have not been accounted for in the analysis. However, this represents a fractional number and is thus unlikely to bias the analysis. Further, assuming that the GISAID holds most of the data and nextclade QC scoring helps avoid analysis of sequences with lesser confidence, the limitations does not significantly affect the data presented. Further, as already discussed there are currently no standards available to define a mutation occurrence as “highly conserved”. Since 80% was taken as a cut-off, it is possible that some of the mutations are not captured in the resulting datasets. For example, a spike mutation- T95I, was present in about 74% of the BA.1, but was not captured due to the set cut-off, and hence the data should be interpreted accordingly.

5. Conclusion:

A massive number of mutations have been reported across various proteins of SARS-CoV-2 thereby loosing focus on which mutations are otherwise important. The strength of the analysis presented herein lies in the breadth of data analysed across the period allowing for the confident identification of significantly conserved mutations. An overall theme that also emerged during the analysis was that various mutations which probably conferred higher virulence to the delta variant were not observed in omicron variants. However, mutations that render higher transmissibility were observed more commonly in omicron variants. It should also be noted that omicron lineage arose independently of delta lineage; hence, they do not likely share many mutations of interest. Additionally, mutations in various other proteins including Nsp1, Nsp13 and ORF9b are seen in recent sequences such as XBB.1.16 which needs to be further evaluated. This analysis highlights the importance of ongoing data mining to observe the acquired and conserved mutations and ignore otherwise seemingly random sets of mutations.

Acknowledgement:

I wish to acknowledge the efforts by the Indian SARS-CoV-2 Genomics consortium (INSACOG) which is largely responsible for SARS-CoV-2 whole genome sequencing in India that has generated the sequences available at GISAID database.

Role of funding source:

This work is not linked to any specific grant from public, commercial, or not-for-profit funding agencies.

Conflict of interest:

The author declares no conflict of interest, financial or otherwise

Table 1: Conserved substitution mutations (> 80% sequences carrying the mutation) in N, E and M protein delta, omicron variants and XBB 1.16 subvariant.

SARS-CoV-2 Protein	Delta and all the subvariants	Omicron and all the subvariants	XBB.1.16 subvariant
N protein	D63G, R203M, D377Y	P13L, R203K, G204R, S413R	P13L, R203K, G204R, S413R
E protein	-	T9I	T9I, **T11A**
M protein	I82T	Q19E, A63T	Q19E, A63T

Table 2: Conserved substitution mutations (> 80% sequences carrying the mutation) in various SARS-CoV-2 Non-structural proteins that were observed in the studied sequences.

SARS-CoV-2 Protein	Delta and all the subvariants	Omicron and all the subvariants	XBB.1.16 subvariant
Nsp1	-	S135R	**K47R**, S135R
Nsp3	-	T24I, G489S	T24I, G489S
Nsp4	-	L264F, T327I, L438F, T492I	L264F, T327I, L438F, T492I
Nsp5	-	P132H	P132H
Nsp6	-	-	L260F
Nsp12	P323L, G671S	P323L	P323L
Nsp13	P77L	R392C	**S36P**, R392C
Nsp14	-	I42V	I42V, **D222Y**
Nsp15	-	T112I	T112I

Table 3: Conserved mutations (> 80% sequences carrying the mutation) in various accessory proteins.

SARS-CoV-2 Protein	Delta and all the subvariants	Omicron and all the subvariants	XBB.1.16 subvariant
ORF3a	S26L	T223I	T223I
ORF6	-	D61L	D61L
ORF7a	V82A, T120I	-	-
ORF9b	T60A	P10S	27-ENA-29 del **I5T**, P10S, **N55S**

Figure 1

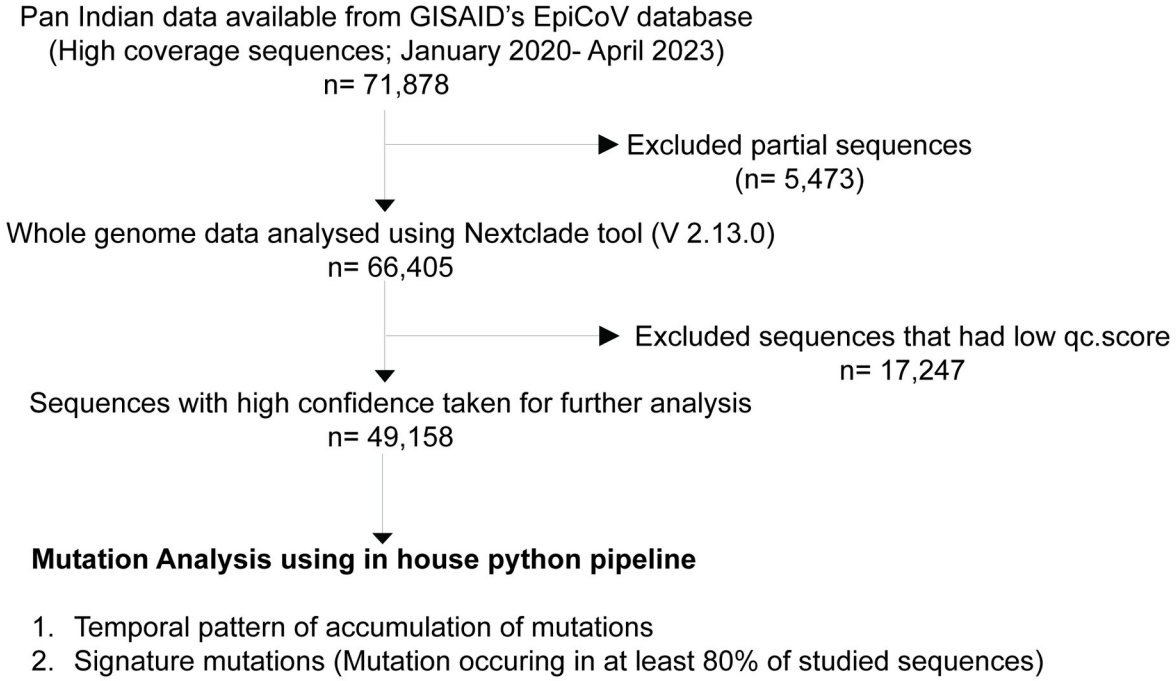

Figure 1: Summary of the data mining and analysis methodology.

Figure 2: Accumulation of Mutation in SARS-CoV-2 spike protein over the period

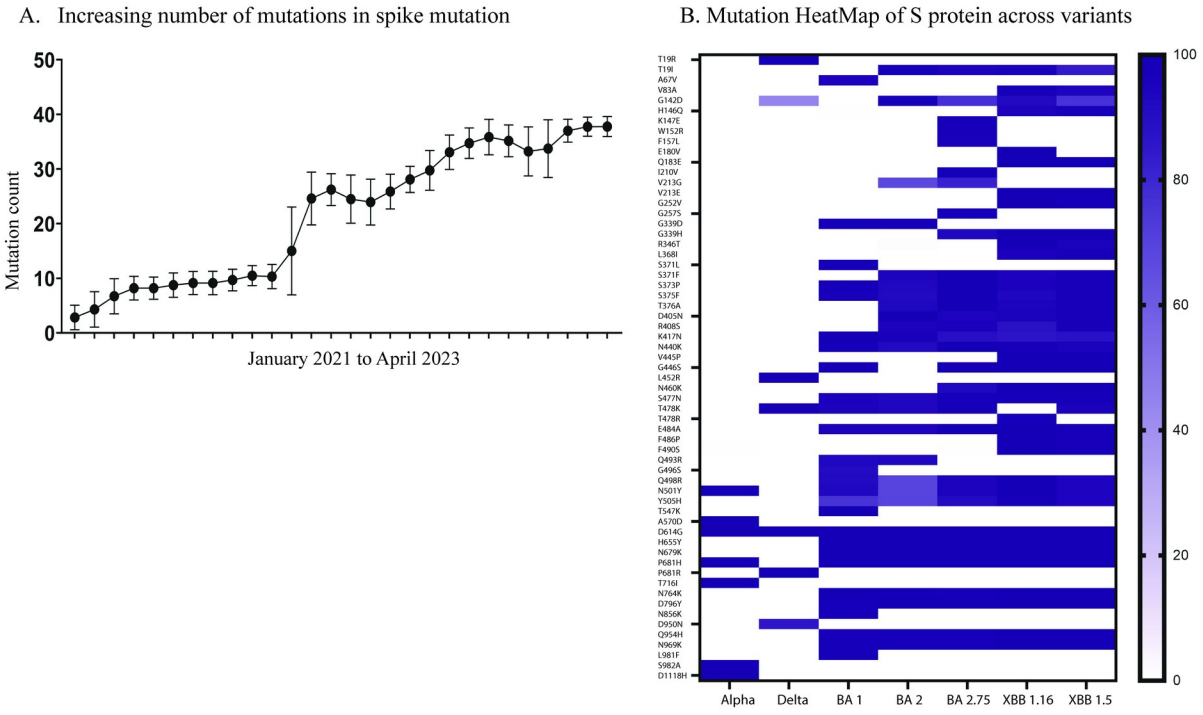

Figure 2: A. Accumulation of mutations in SARS-CoV-2 spike protein over time. X-axis represents by Month; Y-axis represents the number of mutations. Mean and standard deviation are depicted. **B.** SARS-CoV-2 Spike protein mutations across critical variants. The scale bar represents percentage of sequences showing the mutation.

Figure 3: Accumulation of Mutations in N, E and M proteins.

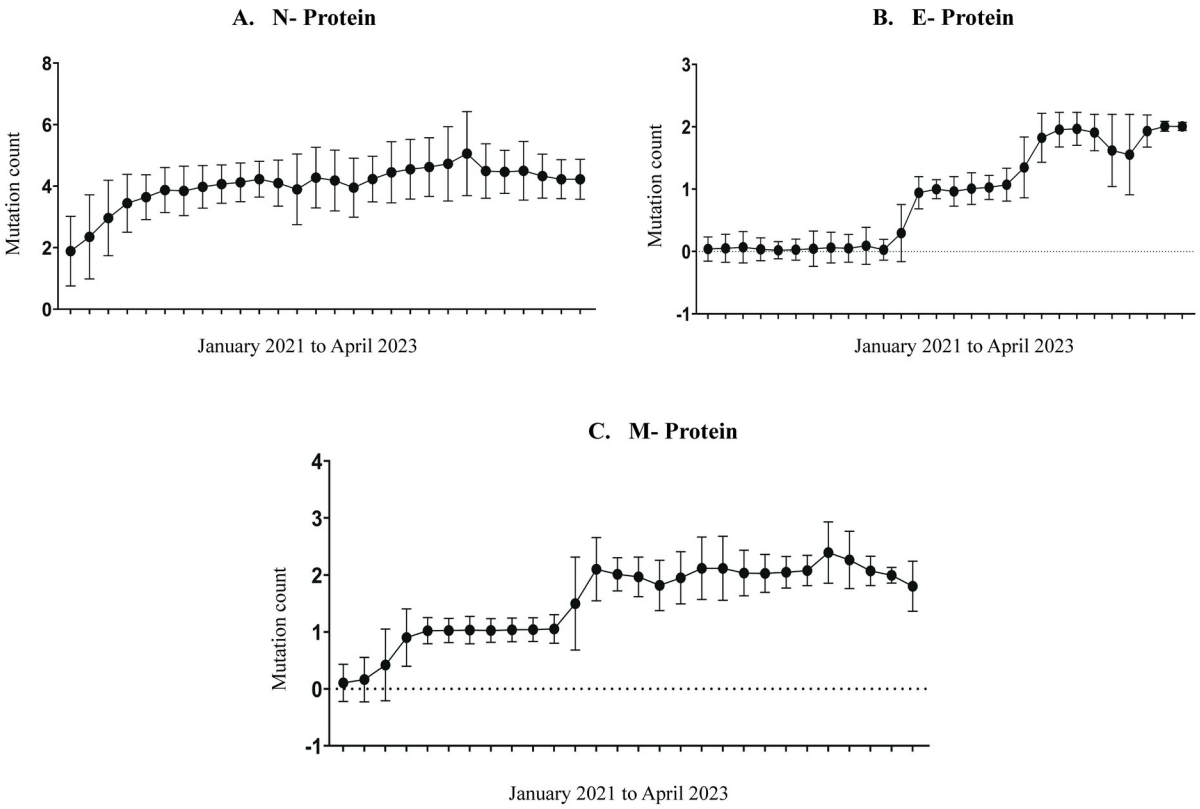

Figure 3: Accumulation of mutations in N, E and M proteins (A- C) over time. X-axis represents by Month; Y-axis represents the number of mutations. Mean and standard deviation are depicted.

References:

1. The Lancet. The COVID-19 pandemic in 2023: far from over. The Lancet. 2023 Jan;401(10371):79.

2. Statement on the fifteenth meeting of the IHR (2005) Emergency Committee on the COVID-19 pandemic [Internet]. [cited 2023 Jul 1]. Available from: https://www.who.int/news/item/05-05-2023-statement-on-the-fifteenth-meeting-of-the-international-health-regulations-(2005)-emergency-committee-regarding-the-coronavirus-disease-(covid-19)-pandemic

3. Shu Y, McCauley J. GISAID: Global initiative on sharing all influenza data - from vision to reality. Euro Surveill Bull Eur Sur Mal Transm Eur Commun Dis Bull. 2017 Mar 30;22(13):30494.

4. Chen Z, Azman AS, Chen X, Zou J, Tian Y, Sun R, et al. Global landscape of SARS-CoV-2 genomic surveillance and data sharing. Nat Genet. 2022 Apr;54(4):499–507.

5. Chan JFW, Kok KH, Zhu Z, Chu H, To KKW, Yuan S, et al. Genomic characterization of the 2019 novel human-pathogenic coronavirus isolated from a patient with atypical pneumonia after visiting Wuhan. Emerg Microbes Infect. 2020 Jan 1;9(1):221–36.

6. Pavesi A. Prediction of two novel overlapping ORFs in the genome of SARS-CoV-2. Virology. 2021 Oct;562:149–57.

7. Finkel Y, Mizrahi O, Nachshon A, Weingarten-Gabbay S, Morgenstern D, Yahalom-Ronen Y, et al. The coding capacity of SARS-CoV-2. Nature. 2021 Jan 7;589(7840):125–30.

8. Markov PV, Ghafari M, Beer M, Lythgoe K, Simmonds P, Stilianakis NI, et al. The evolution of SARS-CoV-2. Nat Rev Microbiol. 2023 Jun;21(6):361–79.

9. Aksamentov I, Roemer C, Hodcroft E, Neher R. Nextclade: clade assignment, mutation calling and quality control for viral genomes. J Open Source Softw. 2021 Nov 30;6(67):3773.

10. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009 Jun 1;25(11):1422–3.

11. Naqvi AAT, Fatima K, Mohammad T, Fatima U, Singh IK, Singh A, et al. Insights into SARS-CoV-2 genome, structure, evolution, pathogenesis and therapies: Structural genomics approach. Biochim Biophys Acta BBA - Mol Basis Dis. 2020 Oct;1866(10):165878.

12. Callaway E. Coronavirus variant XBB.1.5 rises in the United States — is it a global threat? Nature. 2023 Jan 12;613(7943):222–3.

13. Grabowski F, Preibisch G, Giziński S, Kochańczyk M, Lipniacki T. SARS-CoV-2 Variant of Concern 202012/01 Has about Twofold Replicative Advantage and Acquires Concerning Mutations. Viruses. 2021 Mar 1;13(3):392.

14. Korber B, Fischer WM, Gnanakaran S, Yoon H, Theiler J, Abfalterer W, et al. Tracking Changes in SARS-CoV-2 Spike: Evidence that D614G Increases Infectivity of the COVID-19 Virus. Cell. 2020 Aug 20;182(4):812-827.e19.

15. Xia B, Wang Y, Pan X, Cheng X, Ji H, Zuo X, et al. Why is the SARS-CoV-2 Omicron variant milder? The Innovation. 2022 Jul;3(4):100251.

16. Wang Y, Ji H, Zuo X, Xia B, Gao Z. Inspiration of SARS-CoV-2 envelope protein mutations on pathogenicity of Omicron XBB [Internet]. Microbiology; 2023 Jan [cited 2023 Jun 12]. Available from: http://biorxiv.org/lookup/doi/10.1101/2023.01.09.523338

17. Shen L, Bard JD, Triche TJ, Judkins AR, Biegel JA, Gai X. Emerging variants of concern in SARS-CoV-2 membrane protein: a highly conserved target with potential pathological and therapeutic implications. Emerg Microbes Infect. 2021 Dec;10(1):885–93.

18. Hossain A, Akter S, Rashid AA, Khair S, Alam ASMRU. Unique mutations in SARS-CoV-2 Omicron subvariants' non-spike proteins: Potential impacts on viral pathogenesis and host immune evasion. Microb Pathog. 2022 Sep;170:105699.

19. Kannan SR, Spratt AN, Sharma K, Chand HS, Byrareddy SN, Singh K. Omicron SARS-CoV-2 variant: Unique features and their impact on pre-existing antibodies. J Autoimmun. 2022 Jan;126:102779.

20. Kim SM, Kim EH, Casel MAB, Kim YI, Sun R, Kwack MJ, et al. SARS-CoV-2 variants show temperature-dependent enhanced polymerase activity in the upper respiratory tract and

high transmissibility [Internet]. Microbiology; 2022 Sep [cited 2023 May 31]. Available from: http://biorxiv.org/lookup/doi/10.1101/2022.09.27.509689

21. Chen DY, Chin CV, Kenney D, Tavares AH, Khan N, Conway HL, et al. Spike and nsp6 are key determinants of SARS-CoV-2 Omicron BA.1 attenuation. Nature. 2023 Mar 2;615(7950):143–50.

22. Azad GK, Khan PK. Variations in Orf3a protein of SARS-CoV-2 alter its structure and function. Biochem Biophys Rep. 2021 Jul;26:100933.

23. Bianchi M, Borsetti A, Ciccozzi M, Pascarella S. SARS-Cov-2 ORF3a: Mutability and function. Int J Biol Macromol. 2021 Feb 15;170:820–6.

24. Kehrer T, Cupic A, Ye C, Yildiz S, Bouhhadou M, Crossland NA, et al. Impact of SARS-CoV-2 ORF6 and its variant polymorphisms on host responses and viral pathogenesis [Internet]. Microbiology; 2022 Oct [cited 2023 May 31]. Available from: http://biorxiv.org/lookup/doi/10.1101/2022.10.18.512708

25. Cruz CAK, Medina PMB. Temporal changes in the accessory protein mutations of SARS-CoV-2 variants and their predicted structural and functional effects. J Med Virol. 2022 Nov;94(11):5189–200.

26. Jiang H wei, Zhang H nan, Meng Q feng, Xie J, Li Y, Chen H, et al. SARS-CoV-2 Orf9b suppresses type I interferon responses by targeting TOM70. Cell Mol Immunol. 2020 Sep;17(9):998–1000.

27. Vinjamuri S, Li L, Bouvier M. SARS-CoV-2 ORF8: One protein, seemingly one structure, and many functions. Front Immunol. 2022 Oct 24;13:1035559.

28. Kalia K, Saberwal G, Sharma G. The lag in SARS-CoV-2 genome submissions to GISAID. Nat Biotechnol. 2021 Sep;39(9):1058–60.

29. Brito AF, Semenova E, Dudas G, Hassler GW, Kalinich CC, Kraemer MUG, et al. Global disparities in SARS-CoV-2 genomic surveillance. Nat Commun. 2022 Nov 16;13(1):7003.