Dynamic tracking of variant frequencies depicts the evolution of mutation sites amongst SARS-CoV-2 genomes from India

Gaurav Sharma*, Vaishnavi Kolte, Shenu Hudson B., Azra Khan

* Corresponding author

Affiliation: Institute of Bioinformatics and Applied Biotechnology (IBAB), Bengaluru, Karnataka, India-560100

E-mail: gauravsharma@ibab.ac.in

Keywords: COVID-19; SARS; Coronavirus; Genomics; Bioinformatics; Epidemiology

Abstract:

With the exponential spread of COVID-19 pandemic across the world within the last six months, SARS-CoV-2 strains are continuously trying to adapt themselves in their host environment by random mutations. While doing so, some variants with evolutionary advantages such as better human to human transmissibility potential should get naturally selected. This short communication demonstrates how the mutation probability patterns are evolving in 864 SAR-CoV-2 strains isolated from COVID-19 patients across diverse Indian states. We have identified 30 such variants showing contrasting mutational probabilities in the span of four months. Out of these, the mutational probabilities of 25 variants including C14408T (in RdRp gene), A23403G (in spike gene), C6312A (nsp3 gene) are continuously increasing suggesting that these mutations are being propagated with time due to their unexplored evolutionary advantages. In contrast, the mutational probabilities of five variants including C6312A (nsp3 gene), G11083T (nsp6 gene), C28311T (N gene) have significantly decreased in May-June as compared to March-April, suggesting these mutations are being terminated with time. Further in-depth investigation of these identified variants will provide valuable knowledge about the evolution, infection strategies, transmission rates, and epidemiology of SARS-CoV-2 in India.

Letter:

Since the emergence of SARS-CoV-2 in Wuhan, China in December 2019, the infection has spread at a menacing rate throughout the world. As of today, over 10 million active cases of COVID-19 and 0.5 million deaths are telling the horror story of this debilitating virus. Since the whole genome sequencing of SARS-CoV-2 Wuhan Hu-1 strain earlier this year, more than 60,000 genome sequences (1–3) have been deposited in open-source platforms such as GISAID, NCBI Virus, etc. Consortiums such as Nextstrain are providing valuable, unprecedented insights into the demography and epidemiology of SARS-CoV-2 strains along with information to supervise drug/vaccine design and discovery.

Coronaviruses encode a 3′-5′ exoribonuclease (Nsp14) that proofreads RNA (4), therefore maintaining genome fidelity and controlling variations. However, during the few short months of the COVID-19 pandemic, worldwide human-to-human transmission has enabled SARS-CoV-2 to accumulate numerous genetic variations (5). Several mutations among these are non-significant while other mutations could be beneficial for the survival of the virus, its infecting capability, and transmission. As a result, we are learning about newly surfaced symptoms and associated mutations at a rapid pace (6–9). Given that beneficial mutations could be naturally selected in wider populations, studying SARS-CoV-2 genomic variants might help us in understanding viral evolution, behavior, and infection trajectory.

In this study, we have tracked and identified several mutational sites for month-wise separated SARS-CoV-2 genome datasets isolated from Indian COVID-19 patients to distinguish which variants are getting naturally selected to propagate further and the ones which are being dismissed from the population.

Methods: 864 genomes of SARS-CoV-2 isolated from Indian patients till June 30, 2020, were extracted from GISAID **(Table S1)**. Based on the sample collection date of each strain, we distributed them into four categories: 'March' (this included the strains collected in January, February, and March), April, May, and June. The genome of the hCoV-19/Wuhan/WH01/2019 strain (EPI_ISL_406798) was used as a reference throughout this analysis (3). We aligned all genomes using MUSCLE v3.8.1551 (10) and generated their maximum-likelihood phylogeny using RAxML v8.2.12 (11) using GTRGAMMA as nucleotide substitution model with 100 bootstrap values. For each category, we generated multiple sequence alignment and calculated the mutation probability of each site in the alignment as compared to the reference. A hierarchical clustering based heatmap of single-nucleotide loci was generated using mutational probabilities within each category using hclust function in R followed by trend plot generation for each identified cluster using ggplot.

Results: The maximum-likelihood phylogenetic tree depicts the evolutionary divergence amongst SARS-CoV-2 strains isolated and sequenced from the diverse landscape of India during March-June 2020 (**Figure 1; Table S1 and S2)**. The phylogeny was able to categorize all strains into two broad clades (A and B) and eight subclades as demonstrated in GISAID and Nextstrain too. Considering low phylogenetic diversity within SARS-CoV-2 isolates, the branches exhibit low genetic distances. The phylogeny also portrays the sample collection month and location diversity in our dataset suggesting that our working dataset is not biased to any location or time. After alignment, for calculation of mutational probability, we considered only those sites where gaps or non-standard nucleotides were present with a probability of <2%. This analysis identified 200-427 sites with mutation probability ranging from 0.3 to 97.99%. As our interest was in the most relevant mutations, we considered a cutoff of 2% (~17 out of 864 strains) for downstream analysis. Accordingly, we identified 35, 31, 54, and 47 highly mutating sites amongst the March, April, May, and June categories, respectively. Of these, 13 sites were present in all categories, 11 in three categories, 9 in two, and 63 in a single category, resulting in a total of 96 unique sites. Finally, the mutational probabilities of these 96 sites across all time points were visualized using a clustered heatmap where five clusters were obtained with 4, 5, 9, 12, and 66 sites per cluster **(Figure 2A, Table S2).** The largest cluster with 66 sites did not show an upward or downward longitudinal trend, *i.e.*, a "neutral" cluster **(Figure 2B),** while the other four clusters showed significant variations, which are discussed below.

Cluster one included four loci where the mutational probabilities were lowest in March (~35%) with a progressive increase to ~90% in June, indicative of fixation/positive selection in the population (**Figure 2B, Table S2**). Two sites (C3037T and C14408T) are within the orf1ab gene, the third site (A23403G) is in the spike protein-coding gene, and the fourth site (C241T) is in the noncoding 5' UTR region. C3037T is a synonymous mutation within the nsp3 protein-coding gene (F924F at protein level) and may not lead to any major changes in protein function. On the other hand, we noticed that C14408T mutation causes a nonsynonymous P4715L change within the ORF1ab polyprotein (corresponding to P323L in RdRp) which has also been observed in isolates from different continents suggested to change the rigidify of RdRp structure (12). The fourth site, A23403G, is well known to cause a D614G mutation in the spike protein, which interferes in domain S1-S2 interaction. This has been suggested to cause substantial

conformational shifts in spike protein that in turn enhance SARS-CoV-2 infectivity while retaining sensitivity to antibodies that target the receptor-binding domain (6, 7).

The **second cluster** consists of five sites that had high mutational probabilities during early infection months, *i.e.*, March and April, and decreased mutational probabilities in May and June, indicating that these variants are not being selected for further propagation. Two out of these sites, C13730T [which results in an A4489V change in ORF1ab (A97V in RdRp protein)] and C23929T [which results in a Y789Y change in spike protein], either maintain the general chemical nature of the amino acid or are synonymous. However, the other three sites, C6312A, G11083T, and C28311T, result in non-synonymous changes in the nsp3 region (T1198K) of the ORF1a (T2016K), the nsp6 region (L37F) of ORF1a (L3606F), and the nucleocapsid (P13L) protein, respectively. Based on trends, we can hypothesize that the later three non-synonymous mutations were not selected by the virus population for further propagation on account of their lesser efficiency in infection or other types of fitness disadvantages.

Sites identified in the **third cluster** started showing up in April and now in June, they are presenting 35% of the total population. This cluster consists of nine sites, of which six are synonymous mutations (Table S2) while three are non-synonymous mutations and worth exploring further: G25563T site is a part of the orf3a gene and causes Q57H mutation, and G28881A and G28883C are in the nucleocapsid gene and lead to changes in amino acid R203K and G204R respectively. Other loci include synonymous mutations at C2836T [nsp3], C18877T [nsp14], C22444T [S], C26735T [M] and C28854T and G28882A [N] sites, which cause none or insignificant variation as indicated by L6205L, D294D, Y71Y, S194L, and R203R respectively. The **fourth cluster** shows a trend similar to that of cluster three; however, the highest mutational probability is within the range of 10-20% as compared to 30-35% in cluster 3. It consists of 12 sites with six synonymous (C313T [nsp1; L16L], C3634T [nsp3; N1123N], G4300T [nsp3; V1345V], C8917T [nsp4; F2884F], C15324T [RdRp; N5020N], A16512G [nsp13; L5416L]) (Table S2) and six non-synonymous mutations (A2292C [ORF1a; Q676P] [nsp2; Q496P], C5700A [ORF1a; A1812D] [nsp3; A994D], G9389A [ORF1a; D3042N] [nsp4; D279N], C18568T [ORF1ab; L6102F] [nsp14; L177F], C19154T [ORF1ab; T6297I] [nsp14; T372I], G21724T [S; L54F]). Two non-synonymous mutations identified in 3′-5′ exoribonuclease (nsp14 protein) might influence the proof-reading capacity of the virus, however further studies are

needed to test these hypotheses. We hypothesize that these sites in clusters 1, 3, and 5 are being selected for propagation owing to some unknown fitness advantages. Considering the ongoing trends of these variants, we can also predict that their probability will keep increasing. The **fifth cluster** consists of neutral sites where the mutational probabilities are in the range of 2-5% without any significant change in any dataset.

To counter evolutionary pressure, viruses, akin to other living beings, continuously mutate their genetic material to improve their infection strategies, resistance potential to antiviral therapies, and their transmission rate (13–15). Most of these random mutations are synonymous or functionally insignificant. However, a few non-synonymous mutations might give an extra advantage to the virus in its faster transmission, additional infection severity, or higher resistance against antiviral vaccines/treatments along with other fitness advantages. Overall, this study suggests that random-occurring mutations are not evenly distributed across the virus population during different timeframes; however, some loci are more prone to propagate and some get terminated with time. As several virus variants in clusters 1, 3, and 4 have higher mutational probability in June as compared to March, understanding the consequences of those propagating variants in terms of infection and epidemiology will be of great importance. Exploring and recognizing this information might prove helpful for drug and vaccine development. Some reports have already shown the rapid increase of a non-synonymous (D614G) variant across the globe that might have facilitated increased human-to-human transmission (6, 7). In-depth investigations of variants identified in this study will provide newer insights into the evolution and fitness advantages acquired by SARS-CoV-2.

Acknowledgments

GS would like to thank the Department of Science and Technology (DST) and IBAB, Bengaluru for financial and infrastructure support. GS would like to thank his colleague Dr. Shruthi S. Vembar for constructive criticism of the manuscript. We would also like to thank the GISAID database for allowing us to access the genome sequences for this scientific research.

Competing interests

Authors declare that they have no conflicts of interest.

Funding

This work was supported by the DST-INSPIRE (Department of Science and Technology) Faculty Award to GS from the Government of India.

Legends for Figures:

Figure 1: Maximum likelihood phylogenetic tree of 864 Indian SARS-CoV-2 strains isolated from different Indian states (innermost circle) at different times (second circle from inside; month-level). Based on information extracted from Nextstrain, we have also mapped the Clade information on the tree in the outermost circle.

Figure 2: Dynamic tracking of mutational frequencies within genomes extracted from Indian samples in March, April, May, and June. **A) Heatmap of mutational probabilities. B) Trend plots of average mutational probabilities within each cluster.** For each cluster, n represents the number of sites.

Legends for Supplementary Tables:

Table S1: Location-based distribution of 864 Indian SARS-CoV-2 strains analyzed in this study.

Table S2: Metadata for 96 identified mutation sites at Genome, Codon, and protein level. Mutation probability and heatmap-based cluster information are also provided.

References:

1. Lescure FX, Bouadma L, Nguyen D, Parisey M, Wicky PH, Behillil S, Gaymard A, Bouscambert-Duchamp M, Donati F, Le Hingrat Q, Enouf V, Houhou-Fidouh N, Valette M, Mailles A, Lucet JC, Mentre F, Duval X, Descamps D, Malvy D, Timsit JF, Lina B, van-der-Werf S, Yazdanpanah Y. 2020. Clinical and virological data of the first cases of COVID-19 in Europe: a case series. Lancet Infect Dis 0.

2. Zhu N, Zhang D, Wang W, Li X, Yang B, Song J, Zhao X, Huang B, Shi W, Lu R, Niu P, Zhan F, Ma X, Wang D, Xu W, Wu G, Gao GF, Tan W. 2020. A novel coronavirus from patients with pneumonia in China, 2019. N Engl J Med 382:727–733.

3. Fuk-Woo Chan J, Kok K-H, Zhu Z, Chu H, Kai-Wang To K, Yuan S, Yuen K-Y. 2020. Genomic characterization of the 2019 novel human-pathogenic coronavirus isolated from a patient with atypical pneumonia after visiting Wuhan.

4. Sevajol M, Subissi L, Decroly E, Canard B, Imbert I. 2014. Insights into RNA synthesis, capping, and proofreading mechanisms of SARS-coronavirus. Virus Res 194:90–99.

5. Koyama T, Weeraratne D, Snowdon JL, Parida L. 2020. Emergence of Drift Variants That May Affect COVID-19 Vaccine Development and Antibody Treatment. Pathogens 9:324.

6. Yurkovetskiy L, Pascal KE, Tomkins-Tinch C, Nyalile T, Wang Y, Baum A, Diehl WE, Dauphin A, Carbone C, Veinotte K, Egri SB, Schaffner SF, Lemieux JE, Munro J, Sabeti PC, Kyratsous CA, Shen K, Luban J. 2020. SARS-CoV-2 Spike protein variant D614G increases infectivity and retains sensitivity to antibodies that target the receptor binding domain. bioRxiv 2020.07.04.187757.

7. Korber B, Fischer WM, Gnanakaran S, Yoon H, Theiler J, Abfalterer W, Hengartner N, Giorgi EE, Bhattacharya T, Foley B, Hastie KM, Parker MD, Partridge DG, Evans CM, Freeman TM, de Silva TI, McDanal C, Perez LG, Tang H, Moon-Walker A, Whelan SP, LaBranche CC, Saphire EO, Montefiori DC, Angyal A, Brown RL, Carrilero L, Green LR, Groves DC, Johnson KJ, Keeley AJ, Lindsey BB, Parsons PJ, Raza M, Rowland-Jones S, Smith N, Tucker RM, Wang D, Wyles MD. 2020. Tracking changes in SARS-CoV-2 Spike: evidence that D614G increases infectivity of the COVID-19 virus. Cell.

8. Ryder SP. 2020. Analysis of Rapidly Emerging Variants in Structured Regions of the SARS-CoV-2 Genome. bioRxiv 2020.05.27.120105.

9. Simmonds P. 2020. Rampant C→U Hypermutation in the Genomes of SARS-CoV-2 and Other Coronaviruses: Causes and Consequences for Their Short- and Long-Term Evolutionary Trajectories. mSphere 5.

10. Edgar RC. 2004. MUSCLE: A multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5:113.

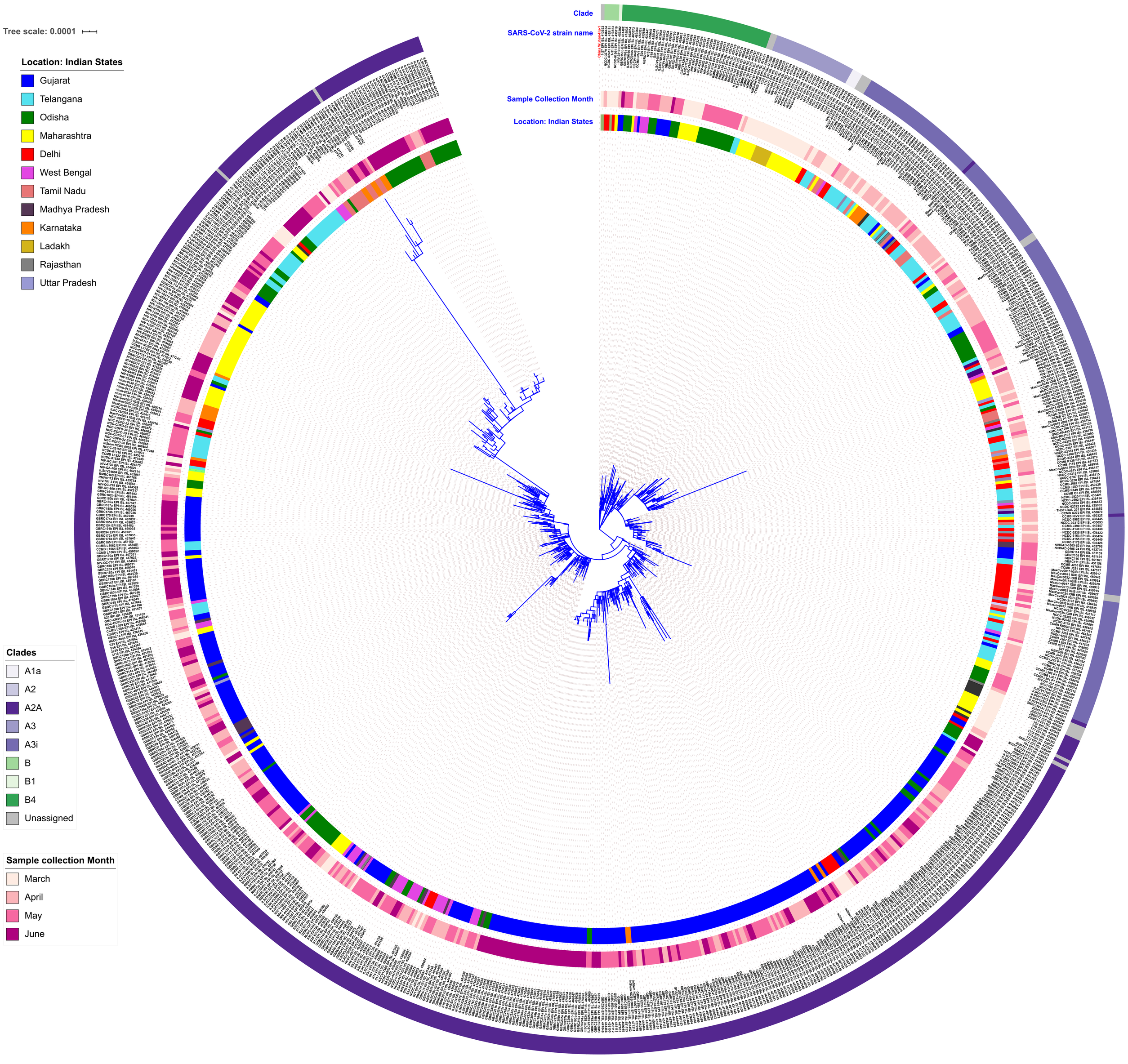
11. Stamatakis A. 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313.

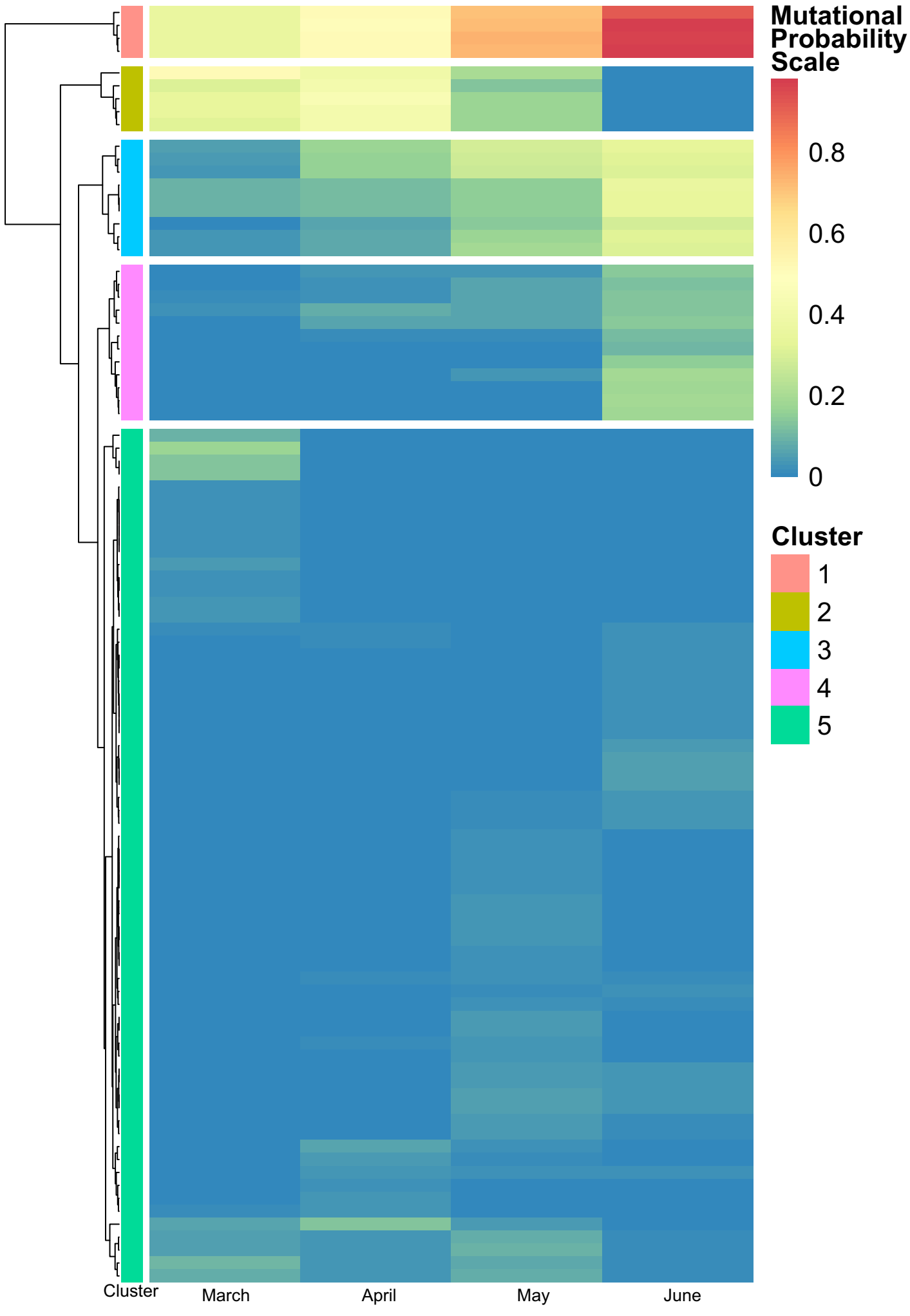
12. Eskier D, Karakülah G, Suner A, Oktay Y. 2020. RdRp mutations are associated with SARS-CoV-2 genome evolution. bioRxiv 2020.05.20.104885.

13. Peck KM, Lauring AS. 2018. Complexities of Viral Mutation Rates. J Virol 92.

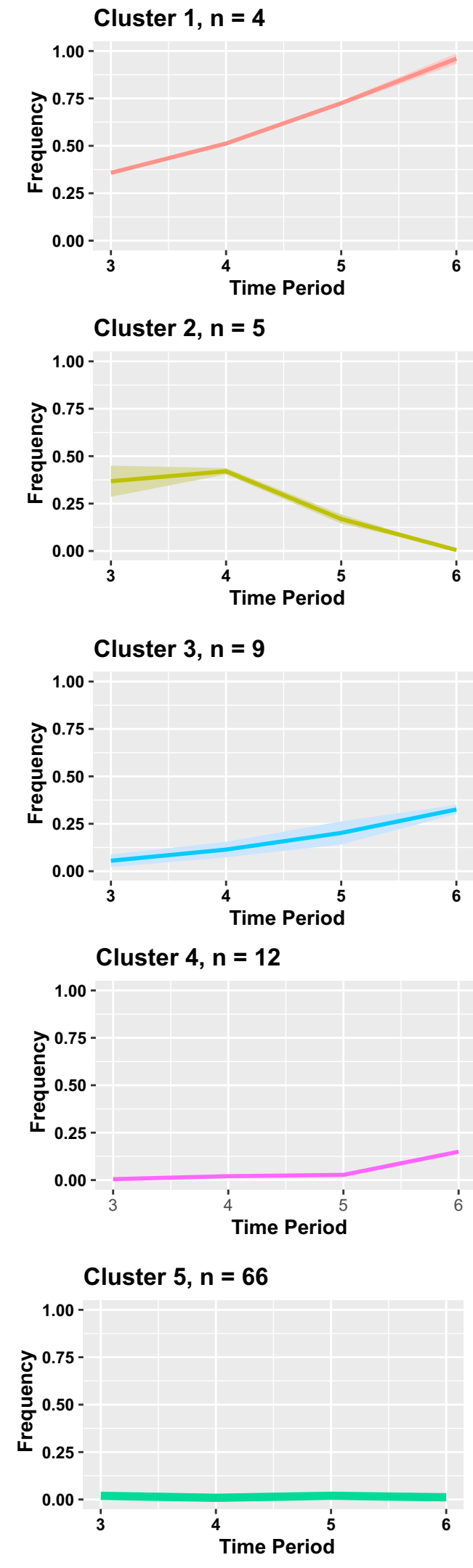
14. Gorbalenya AE, Enjuanes L, Ziebuhr J, Snijder EJ. 2006. Nidovirales: Evolving the largest RNA virus genome. Virus Res 117:17–37.

15. Lynch M, Ackerman MS, Gout JF, Long H, Sung W, Thomas WK, Foster PL. 2016. Genetic drift, selection and the evolution of the mutation rate. Nat Rev Genet. Nature Publishing Group.


Tree scale: 0.0001
Location: Indian States
Gujarat
Telangana
Odisha
Maharashtra
Delhi
West Bengal
Tamil Nadu
Madhya Pradesh
Karnataka
Ladakh
Rajasthan
Uttar Pradesh
Clade
SARS-CoV-2 strain name
Sample Collection Month
Location: Indian States
Clades
A1a
A2
A2A
A3
A3i
B
B1
B4
Unassigned
Sample collection Month
March
April
May
June


Mutational Probability Scale
0.8
0.6
0.4
0.2
0
Cluster
1
2
3
4
5
Cluster
March
April
May
June

A


Cluster 1, n = 4
Cluster 2, n = 5
Cluster 3, n = 9
Cluster 4, n = 12
Cluster 5, n = 66
Frequency
1.00
0.75
0.50
0.25
0.00
3
4
5
6
Time Period

B