Volume 64, Issue 3, 2020

Journal of Scientific Research

Institute of Science,

Banaras Hindu University, Varanasi, India.

National Conference on Frontiers in Biotechnology & Bioengineering (NCFBB 2020), JNTU Hyderabad, India

Viral Phylodynamic Analysis of SARS-COV2 for the Identification of the Clades Prevalent in India

B. Rajesh Abhinav[1], A. Uma[1], Suresh Babu Bastipati[1]

[1]Center for Biotechnology, Institute of Science and Technology,
Jawaharlal Nehru Technological University, Hyderabad. vedavathi1@jntuh.ac.in

Abstract: **Severe Acute Respiratory Syndrome – Coronavirus 2 (SARS-COV2) caused the COVID-19 global pandemic. In a few months, the virus has spread throughout the world and mutated into several variants or strains. These new variants are spreading rapidly overtaking the ancestral variant. Across the world, Different variants have become prevalent. The effect of these mutations on the virulence of the virus is still unknown. However, the mutations can be used to trace the transmission and genetic variations of the virus by phylodynamic analysis of the viral genome. Viral phylodynamics is the study of how different aspects such as evolution, immunology, and epidemiology interact and shape the virus and its phylogeny. The study presented in this paper helps in better management of pandemics and design better vaccines and drugs against the virus. Nextstrain Project helps us to develop phylodynamic studies of different viruses and track their evolution and transmission. A vast collection of genome samples from diverse populations and the metadata connected to the samples helps in constructing a better phylodynamic analysis. It can also help to understand the origin of a virus by tracking zoonotic transfers and genetic variations. Here, we use the community standard Nextstrain pipeline to analyze the SARS-CoV2 phylodynamics and track its transmission and evolution into different strains as it spreads through the population of India.**

Index Terms: **COVID-19, evolution, mutations,Nextstrain,SARS-COV2 (Severe Acute Respiratory Syndrome – Coronavirus 2),transmission, and viral phylodynamics.**

I. INTRODUCTION

Covid-19 is an infectious disease caused by the novel Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-COV2). The first reported outbreak was in Wuhan in December 2019. The first case of COVID-19 in India came to light on 30 January 2020 and since then it has rapidly spread in the country. At the time of writing this paper, there are around 12 lakh confirmed cases of COVID-19 in India and around 29,000 deaths related to this disease i.e., the mortality rate is 2.4%. The R_0 value of the virus in India was found to be 1.471(Kanagarathinam & Sekar, 2020).

Viral Clades Classification

The old classification of clades according to the mutations in the early stages of the pandemic can be seen in *Table 4*. In this classification, the clades were broadly defined as the number of mutations the virus went through was less. Later, the Nextstrain team classified the clades according to the mutations as shown in *Table 3*. In the later classification, the A2a clade was divided into three clades namely 20A, 20B, and 20C. This made the clades more specific and well defined. The GISAID organization on the other hand, classified the clades as shown in *Table 5*.Previously, work has been done in the epidemiology of COVID-19 in India which suggests the prevalence of A2a clade in the country and the rise of a new clade (Banu et al., 2020). But as the pandemic progresses, the distribution of the novel coronavirus clades across the country change, and new clades are being identified with increasing mutations in the virus.

The present research aims to gain insights into the latest composition of viral clades of the novel coronavirus in India, its transmission dynamics, mutations, genetic variants, and the prevalent clades in different parts of India.

II. MATERIALS & METHODOLOGY

The analysis was performed using a community standard Nextstrain pipeline with changes made to run a country-specific analysis.

Nextstrain (Hadfield et al., 2018)

Nextstrain is an open-source project. It makes use of two tools, Augur, a bioinformatics toolkit, and Auspice, a visualization tool. The pipeline consists of individual modules in an order as represented in DAG (Directed Acyclic Graph) which finally produces the results.

*Corresponding Author

DOI: 10.37398/JSR.2020.640308
Selection and Peer-review under responsibility of the Program Chairs.

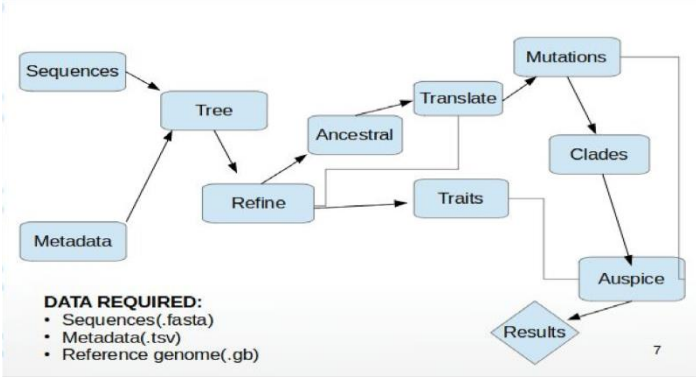

Figure 1: Figure showing DAG followed by Nextstrain pipeline.

Sources of the samples

Samples (genome sequences) along with their metadata were obtained from the GISAID website (Shu & Mccauley, 2017) and NCBI's Genbank (Clark et al., 2015). The reference genome is obtained from Genbank with theaccession ID 'MN908947' (Wu et al., 2020). 1300 samples were collected from both the sources out of which 849 are Indian samples and the rest are from around the world.

Subsampling criteria:

Table 1: Subsampling criteria

Parameter	Number of samples
Location	300
Country	300
Division (States)	300
Date	100

Steps involved in the Nextstrain pipeline:

Filtering and Masking

The sequence data along with metadata were filtered according to the parameters set by the Nextstrain/ncov pipeline. 100 bases from the beginning and 50 bases from the end along with bases at positions 13402, 24389, and 24390as per the reference genome are masked as they are prone to sequencing error according to Nextstrain.

Alignment

The multiple sequence alignment was carried out using MAFFT (Katoh, 2002) and visualized by UGENE (Okonechnikov et al., 2012).

Building a phylogenetic tree

The aligned sequences were used to construct a maximum likelihood phylogenetic tree using IQ-TREE (Nguyen et al, 2014). Branch lengths of the tree depend on the nucleotide divergence. Gene positions and site numbers are derived from the reference genome.

Refining the tree to get a time-resolved tree

The phylogenetic tree is re-rooted with the least-squares method which produces a Timetree (Sagulenko et al, 2018) i.e. it plots the phylogenetic tree against the time. The Timetree is annotated with metadata provided with Augur/Traits subcommand.

Identifying the mutations

Nucleotide mutations in the genome are identified by Augur/Ancestral subcommand. The nucleotide sequences are converted into Amino acids sequences according to the reference sequence and amino acid mutations are identified through Augur/Translate subcommand. All mutations, node data, and annotations are written to a .JSON file.

Visualizing the results

All the required files are collected and written into a .JSON file by Augur/Export v2 subcommand. Auspice visualizes the created .JSON file.

III. RESULTS & DISCUSSION

In the analysis, a total of 1205 genomes out of 1300 were used which were submitted to GISAID between December 2019 and July 2020. In India,the A2a clade was found to be dominant (Banu et al., 2020).But according to the new Nextstrain clade classification A2a is divided into three clades,of which 20B is found to be prevalent in India. Mutations occurring frequently in Indian samples are shown in the table. The mutation rate specific to Indian samples of the virus is found to be 27.619 substitutions per year (~$8*10^{-4}$substitutions/nucleotide/year).

Table 2: Most frequently occurring mutations in the Indiansamples.

Mutated Codons	Gene	Nucleotide Site	Entropy
P314L	ORF1b	314	0.678
D614G	S	378	0.673
F3606L	ORF1a	3606	0.636
P10S	ORF9b	514	0.626
V88A	ORF1b	614	0.612
L13P	N	3606	0.612
T2016K	ORF1a	251	0.598
G204R	N	3220	0.466
G50N	ORF14	84	0.466

Figure 2: Graph representing the most frequently occurring mutations in the Indian samples.

Initially, in India, many of the samples were found to be of an ancestral variant (Wuhan-01). Later, different variants of the virus introduced from around the world started transmitting among the local population and became prevalent overtaking the ancestral type. This can be observed in the time-resolved phylogenetic tree Fig.3.

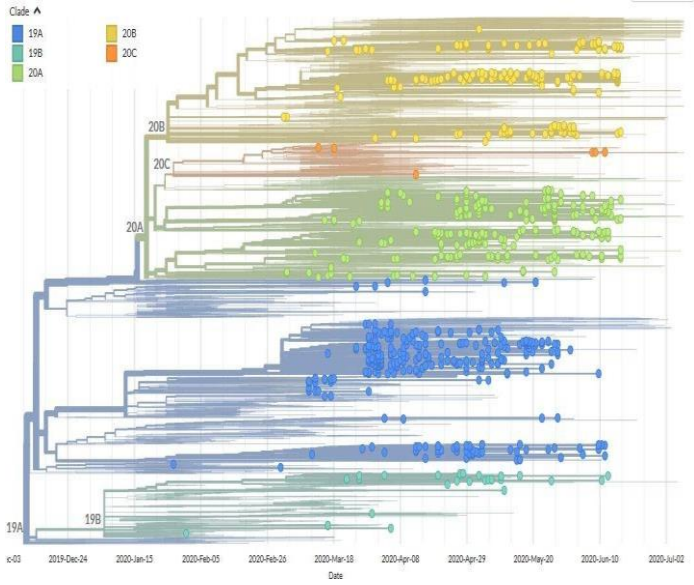

Figure 3: Time-resolved phylogenetic tree of SARS-CoV2 in Indian population

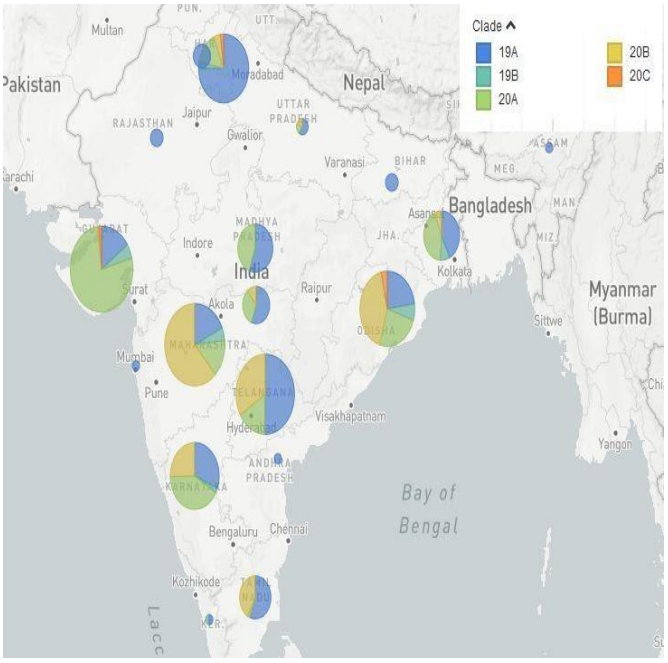

Figure 4: Composition of new Nextstrain clades of the virus in India.

Particularly a strain called '20B' has become prevalent in India. Of the 170 samples collected in June 2020, 84 samples were of 20B clade, 60 of 20A clade, 14 of 19A, 8 of 19B, and 4 of 20C clade. This analysis is also conducted at the state level selecting a few states for which ample numbers of samples were available. The observations are as follows.

A. *Telangana (Mar20-Jun20):* All the new cases emerging are of 'A2a/20B' clade of the Nextstrain clade nomenclature and 'GR' clade of the GISAID clade nomenclature. The most frequent mutations are 'ORF1a: K2016T' & 'ORF1a: F3606L'.

B. *Gujarat (Mar20-Jun20):* 'A2a/20A' clade according to the Nextstrain clade nomenclature and two clades 'G' and 'GH'clade of the GISAID clade nomenclatureare on the rise. The most frequently occurring mutation is 'ORF3a: Q57H'.

C. *West Bengal (Mar20-Jun20):* In West Bengal 'A2a/20A' clade of the Nextstrain clade nomenclature and 'G' clade of the GISAID is prevalent. The most frequent mutation is 'ORF1b:V88A'.

D. *Maharashtra (Mar20-Jun20): M*any of the new cases belong to the 'A2a/20A' clade of the Nextstrain clade nomenclature and 'GR'clade of the GISAID clade nomenclature. The frequent mutation is 'ORF1b: P314L'.

E. *Odisha (Mar20-Jun20):* 'A2a/20A' clade of the Nextstrain clade nomenclature and 'G' & 'O' clade of the GISAID clade nomenclature are prevalent. The most frequent mutation is 'S D614G'.

F. *Karnataka (Mar20-Jun20): M*ost of the new cases belong to 'A2a/20A' clade of the Nextstrain clade nomenclature andGR clade of the GISAID clade nomenclature. The most frequent mutation is 'E*V75F'*.

CONCLUSION

'20B' and '20A' clades are becoming dominant in India with a mutation rate of 27.619 subs per year (~$8*10^{-4}$ subs/nuc/year) and the most frequently occurring mutations are given in *Table 2*.As the pandemic progresses more data will emerge and newly emerging clades can be identified using this analysis. This analysis along with optimum data can help to track the transmission of the virus thus leading to better management of the COVID-19 in India and also identifying any mutations which might have a positive effect on the virulence of the virus.

ACKNOWLEDGMENT

We acknowledge the contribution of all the authors of SARS-COV2 genomes used for this study andthe GISAIDfor making the genome data available without which this study would not have been possible. We are also grateful to the Nextstrain team for providing support.

REFERENCES

Pattabiraman, C., Habib Farhat., Tamma, Krishnapriya. (2020). Perspectives on SARS-CoV-2 strains. ORF ISSUE BRIEF, (365). https://www.orfonline.org/research/perspectives-on-sars-cov-2-strains-67003/.

#IndiaFightsCorona COVID-19. (2020, April 3). https://www.mygov.in/covid-19/.

Andrew Rambaut, Edward C. Holmes, Verity Hill, Áine O'Toole, JT McCrone, Chris Ruis, Louis du Plessis, Oliver G. Pybus bioRxiv 2020.04.17.046086; doi: https://doi.org/10.1101/2020.04.17.046086

Banu, S., Jolly, B., Mukherjee, P., Singh, P., Khan, S., Zaveri, L., . . . Sowpati, D. T. (2020). A distinct phylogenetic cluster of Indian SARS-CoV-2 isolates. doi:10.1101/2020.05.31.126136

Chatterjee, P., Nagi, N., Agarwal, A., Das, B., Banerjee, S., Sarkar, S., … Gangakhedkar, R. (2020). The 2019 novel coronavirus disease (COVID-19) pandemic: A review of the current evidence. Indian Journal of Medical Research, 151(2), 147–159. https://doi.org/10.4103/ijmr.ijmr_519_20

Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2015). GenBank. Nucleic Acids Research, 44(D1). https://doi.org/10.1093/nar/gkv1276

Continued evolution of highly pathogenic avian influenza A (H5N1): updated nomenclature. (2011). Influenza and Other Respiratory Viruses, 6(1), 1–5. https://doi.org/10.1111/j.1750-2659.2011.00298.x

Dorp, L. V., Acman, M., Richard, D., Shaw, L. P., Ford, C. E., Ormond, L., … Balloux, F. (2020). Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. Infection, Genetics and Evolution, 83. https://doi.org/10.1016/j.meegid.2020.104351

Grenfell, B. T. (2004). Unifying the Epidemiological and Evolutionary Dynamics of Pathogens. Science, 303(5656), 327–332. https://doi.org/10.1126/science.1090727

Hadfield, J., Megill, C., Bell, S. M., Huddleston, J., Potter, B., Callender, C., … Neher, R. A. (2018). Nextstrain: real-time tracking of pathogen evolution. Bioinformatics, 34(23), 4121–4123. https://doi.org/10.1093/bioinformatics/bty407

Kanagarathinam, K., & Sekar, K. (2020). Estimation of Reproduction Number (Ro) and Early Prediction of 2019 Novel Coronavirus Disease (COVID-19) Outbreak in India Using Statistical Computing Approach. Epidemiology and Health, 42. https://doi.org/10.4178/epih.e2020028

Katoh, K. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Research, 30(14), 3059–3066. https://doi.org/10.1093/nar/gkf436

Mercatelli, D., & Giorgi, F. M. (2020). Geographic and Genomic Distribution of SARS-CoV-2 Mutations. Preprints. https://doi.org/10.20944/preprints202004.0529.v1

Nguyen, L.-T., Schmidt, H. A., Haeseler, A. V., & Minh, B. Q. (2014). IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. Molecular Biology and Evolution, 32(1), 268–274. https://doi.org/10.1093/molbev/msu300

Okonechnikov, K., Golosova, O., & Fursov, M. (2012). Unipro UGENE: a unified bioinformatics toolkit. Bioinformatics, 28(8), 1166–1167. https://doi.org/10.1093/bioinformatics/bts091

Reid, D. (2020, January 30). India confirms its first coronavirus case. CNBC.com. https://www.cnbc.com/2020/01/30/india-confirms-first-case-of-the-coronavirus.html.

Sagulenko, P., Puller, V., & Neher, R. A. (2018, January 8). TreeTime: Maximum-likelihood phylodynamic analysis. https://academic.oup.com/ve/article/4/1/vex042/4794731.

Shu, Y., & Mccauley, J. (2017). GISAID: Global initiative on sharing all influenza data – from vision to reality. Eurosurveillance, 22(13). https://doi.org/10.2807/1560-7917.es.2017.22.13.30494

Stack, J. C., Welch, J. D., Ferrari, M. J., Shapiro, B. U., & Grenfell, B. T. (2010). Protocols for sampling viral sequences to study epidemic dynamics. Journal of The Royal Society Interface, 7(48), 1119–1127. https://doi.org/10.1098/rsif.2009.0530

Wu, F., Zhao, S., Yu, B., Chen, Y.-M., Wang, W., Song, Z.-G., … Zhang, Y.-Z. (2020). A new coronavirus associated with human respiratory disease in China. Nature, 579(7798), 265–269. https://doi.org/10.1038/s41586-020-2008-3

Zika Tutorial. https://nextstrain.org/docs/tutorials/zika.

APPENDIX

***Table 3*:** Table defining new Nextstrain clades according to mutations on the reference genome.

Clades	Sites on genome	Mutated to
19A	8782	C
19A	14408	C
19B	8782	T
19B	28144	C
20A	8782	C
20A	14408	T
20A	23403	G
20B	8782	C
20B	14408	T
20B	23403	G
20B	28881	A
20B	28882	A
20C	1059	T
20C	8782	C
20C	14408	T
20C	23403	G
20C	25563	T

Source: Nextstrain

***Table 4*:** Table defining old clades according to mutations on the reference genome.

Clades	Gene	Site	Mutated to
A1a	ORF3a	251	V
A1a	ORF1a	3606	F
A2	S	614	G
A2a	ORF1b	314	L
A3	ORF1a	378	I
A3	ORF1a	3606	F
A6	nucleotide	514	C
A7	ORF1a	3220	V
B	ORF8	84	S
B1	ORF8	84	S
B1	nucleotide	18060	T
B2	ORF8	84	S
B2	nucleotide	29095	T
B4	ORF8	84	S
B4	N	202	N
B4	N	202	N

Source: Nextstrain

Table 5: Table defining GISAID clades according to mutations on the genome.

Clade	Gene	Nucleotide Mutations	Site and Mutation
O	ORF1a & ORF1b	8782,14408	C
G	S	C241T, C3037T, A23403G	D614G
GH	ORF3a & S	C241T, C3037T, A23403G, G25563T	D614G, Q57H
GR	S & N	C241T, C3037T, A23403G, G28882A	D614G, G204R
V	ORF3a	G11083T, G26144T	NSP6-L37F, NS3-G251V
S	ORF8	C8782T, T28144C	L84S
L	-	-	-

Source: GISAID.org/epicov.org

***