EMERGING MINDS
JOURNAL FOR STUDENT RESEARCH

ISSN 2976-9094
EMJSR 2024,Vol 2 (1):55-68
https://doi.org/10.59973/emjsr.66

Received: 2024-03-22
Accepted: 2024-04-10
Published: 2024-05-01

Phylodynamics and Phylogenetic Analysis of Recent SARS-Cov-2 Viral Strains from Pune, Maharashtra

Nourin Shamnad[1,*]

[1] School of Biological Sciences, University of Portsmouth, King Henry Building, Portsmouth PO1 2DY, UK

*Corresponding author: anourinshamnad@gmail.com

Abstract - In December 2019, medical practitioners from China identified a novel strain of severe acute respiratory syndrome-CoV (SARS-CoV). The disease was allotted a zoonotic origin and the spillover event is associated with Huanan Wholesale Seafood Market in Wuhan City, Hubei Province, China. On 30 January 2020, the World Health Organization (WHO) declared COVID- 19 as an international public-health emergency concern. In India, the index case was reported on 30th January in the district of Thrissur, state of Kerala and by 25th March 2020, the country was lockdown. This research aims to analyze the viral phylogenetic and phylodynamics of strains from Pune, Maharashtra, followed by a comparison against real time data and an analysis of the potency of government prevention strategies. Publicly available SARS-CoV-2 sequences, specific to Pune, Maharashtra, were downloaded from GISAID during the time frame of the epidemics. Phylogenetic analysis of the sequences, including Maximum Likelihood trees and the nucleotide substitution model, were done using IQ-TREE software. Phylodynamic tools available as part of the BEAST2 software was used to estimate the evolution of R number through time. GTR+F+I was found to be the best-fit nucleotide substitution model for the data acquired. ML trees constructed reported a log likelihood of 44842.9931. Analysis of trace estimates reported R number with an average of 1, indicating transmission of virus. The data contains controversial elements which can only be clarified upon further sequencing of the SARS- CoV-2 strains, however, the results obtained are parallel to real time statistics.

Keywords - SARS-Cov-2; Phylodynamics; Phylogenetics.

1 Introduction

Phylogenetic analysis of viruses takes into consideration the epidemiological, immunological, and evolutionary processes that lead to their phylogenies, which in turn allow researchers to determine the risk factors contributing to particular events using phylogenetic comparative models. Coronavirus are single-stranded enveloped RNA virus. They are currently the largest RNA viruses known to mankind that causes respiratory, gastro-intestinal, hepatic and neurological diseases [1–3]. SARS-CoV-2 is an enveloped RNA virus that has a genomic length of 29.9 kilobases, with 11 encoding regions that encodes ORF1ab polyproteins, Spike (S) glycoproteins, Envelope (E) proteins, Membrane (M) glycoprotein, Nucleocapsid (N) protein and accessory proteins including ORF3a, ORF6, ORF7b, ORF8 and ORF10, along with non-structural proteins (NSP) in the Open Reading Frame (ORF) [4,5]. Genome sequences of

contemporary organisms contain conserved DNA sequences that encode proteins and RNA in related species, from the last common ancestor, as such, the genomic analysis results of 2019 n-CoV has revealed that it is of the subgenus Sarbecovirus, of genus Betacoronavirus of the Coronavirus family [6,7]. Patients with COVID-19 can be asymptomatic or symptomatic. While symptoms of COVID-19 range from mild fever to severe pneumonia and even death, the immune system of asymptomatic carriers show no reactions [1,2]. The analysis of an individual's sequences determines their infection time, in addition to, compensating for the imprecise estimations provided by clinics and health care services, sequence data can also be used to compare strain dynamics [8–11]. The infection time refers to the time period during which an individual stays infectious which was found to range from 1.0×10^{-3} to 8.3×10^{-4} substitutions/site/year for SARS-CoV-2.

Human-to-human transmission of the n-CoV was confirmed along a chain of four generations via contact, respiratory and aerial droplets, in addition to, fecal-oral transmission [1,2]. Viral transmission dynamics is used to contain infections, and phylodynamics allows the study of these processes in conjunction with their effects on virus phylogeny [8,11–13]. An integrated analysis of pathogen genetics and host demographics is essential for the accurate and reliable estimation of factors potentially affecting population dynamics and geographic spread of the virus. By analyzing specific universal factors researchers can determine whether or not global transmission routes are related to particular factors or whether specific prevention and control strategies targeting the pathogen, regardless of the vector, are required to combat it effectively [8,10]. Molecular clock models developed for serially sampled data helps to uncover the timing of transmission events and epidemiological origins. The clock models are estimated from genetic sequences which allows the inference of the date of most common recent ancestor of all sequences, the reproduction rate of the pandemic and the efficacy of viral control [8,11].

Bioinformatics is the storage, manipulation, interpretation, and application of biological data, pertaining to nucleic acids, using computational methods and genomics which is used to study an organism's complete DNA sequence [14]. Bioinformatics uses web-based and command-line tools to gather genomic data which can be accessed via public databases. As a result, DNA sequencing employs Sanger sequencing and next generation sequencing (NGS) technology to determine an organism's genome, including genes encoding RNA molecules and proteins. Comparative analysis of rRNA sequences has revealed that the three major branches on a basic universal phylogenetic tree are Archaea, Bacteria, and Eukaryotes this can be interpreted as that the existing and extinct species have been evolved from these primary branches. At the molecular level, evolution is a process of mutation and selection, however, molecular evolution, in essence, involves comparing genes and proteins. Scientists use molecular clocks, a set of methods and models that indicate how each chromosome is evolving regarding the tree of life, to identify the most recent common ancestor of an organism in order to build the phylogenetic tree, in addition to using this data to allocate a definite timescale to the tree. Phylogeny refers to all relationships among organisms via ancestors and descendants' associations, and it is typically modeled as a tree where the branch lengths define the degree of relationships between the objects, and the topology reveals the relationships or homology among the proteins represented in the tree via a protein alignment. The basis of all phylogenetic analyses is the acquisition of genome sequences. In May 2008, during the sixty first world health assembly, a public database for sharing influenza data in a transparent and verifiable manner was established for public use, named the Global Initiative on Sharing Avian Influenza Data (GISAID) [15]. As of GISAID's launch, WHO Global Influenza Surveillance and Response System (GISRS) collaborates with WHO Collaborating Centers and National Influenza Centers in order to share data for the biannual influenza vaccine virus recommendations [15]. The GISAID Initiative encourages sharing data from all influenza viruses and COVID-19. Researchers can use this genetic sequence, clinical and epidemiological data to determine how viruses evolve and spread [15]. The GISAID users agreed to adhere to a basic etiquette by noting the originators of specimen, and

the submitters of sequences and other metadata, encouraging collaborative research based on open sharing of data and respect for all rights and interests of data collected by GISAID and its users [15]. As part of its work to ensure the integrity of GISAID data, GISAID helps developers integrate or connect their tools with GISAID to facilitate analysis of the data [15]. GISAID was recognized as a member of the PREDEMICS consortium in 2013, a project on the Preparedness, Prediction and Prevention of Emerging Zoonotic Viruses with Pandemic Potential using multidisciplinary approaches [15]. Due to its high-quality genome sequence the coronavirus strain hCoV-19/Wuhan/WI04/2019 (WIV04) is the official reference sequence of GISAID [15]. As of 27th August 2021, 3,085,327 viral sequences from across the globe has been submitted to the GISAID platform [15].
The National Supermodel Committee of India, on establishing the presence of virus in the Indian population, has improvised on the existing standard Susceptible Infected Recovered (SIR) pandemic model in order to differentiate between asymptomatic and symptomatic patients, termed asymptomatic (A) in the category of I in the SIR model [16]. Depending on the physiology of the individual they can be A or I, thereby splitting the S and R categories into SA and SI, RA and R [16]. Daily data is accounted as active infections (I), cumulative removed (Ri) and cumulative deaths (D) [16]. As of 25th August 2021, India had reported 32.6 million cases out of which, 32512366 are active infections, 31754281 recovered patients and 435758 deaths against 595,504,593 vaccination doses in a 1.366 billion population [17,18]. The state of Maharashtra was the worst hit state with 6432699 confirmed cases, 6243034 recovered individuals and a staggering 136355 deaths [17,18].
Maharashtra is a state in India located in central India with a population of 1.26 million, making it one of the most populous states in the country. The index case in Maharashtra was reported on 9th February 2020 in Pune. In the ongoing battle against COVID-19, Maharashtra is the worst-hit state with the death toll overtaking even countries like Germany, and amongst the various districts of Maharashtra, Pune has the most casualties caused by COVID-19. The research titled 'Phylodynamics and Phylogenomics of Recent SARS-CoV-2 Viral Strains from Pune, Maharashtra', as its name suggests, aims to understand the viral phylodynamics and phylogenomic of SARS-CoV-2 in the city of Pune, Maharashtra including, but not limited to, the time to the most recent common ancestor and the effective reproductive number of the virus in the specified geographic location, to compare and contrast the results against real time data, to understand the efficacy of government prevention strategies in the state and across the nation.

2 Materials and Methods

2.1 Data acquisition

SARS-CoV-2 sequences from Pune, Maharashtra were downloaded from GISAID. The obtained sequences were further filtered based on host and low coverage sequences were excluded.

2.2 Phylogenetic and Phylodynamic Analysis

Bioinformatics is a vast subject with branches extending to all possible disciplines in science, but for the purpose of this research statistical bioinformatics is used to derive at the results. Statistical methods allow us to extract information from data and infer relevant patterns from it, understanding and quantifying uncertainties in uncertain observations as well as in inference procedures related to natural phenomena [14,19]. It aims at developing and applying novel data mining methods for studying biology and evaluating results by deriving asymptotic distributions, calculating significance levels, and designing simulation studies to compare results [14,19,20]. Statistics deals with probability and its inferences, which can be understood via either frequentist or Bayesian methods. Bayesian statistical methods were used to analyse the data obtained for the research as by eliminating uncertainty, Bayesian

statistics can do joint processing of all the evidence and quantify the uncertainty explicitly [14]. In addition to using Monte Carlo-based computational tools to analyse its principled organization of data and information which allow a greater emphasis upon the scientific problem rather than on mathematical convenience, the admissibility theorem suggests that sensible answers are always Bayesian in nature [14]. All Bayesian procedures are involved in the creation of a complete probability model, construction of the posterior distribution for a specific parameter, and evaluating the fit and conducting an appraisal of the selected model [14]. Bayesian statistical methods are centered on the Bayes theorem, which states that the probability of an event A occurring if another event B occurs is equal to the probability that the event B occurs if A has already occurred, multiplied by A's probability and divided by B's probability. The mathematical formula for the Bayes theorem is as follows:

$$P(A \mid B) = P(B \mid A)\, P(A) / P(B)$$

Where A and B are the events and, P(A) and P(B) are the independent probabilities of events A and B. P(A | B) is the probability of event A given event B is true and P(B | A) is the probability of event B given event A is true. Phylogenetic analysis of molecular sequences using Bayesian methods can be carried out using BEAST which is a cross-platform tool providing a graphical user interface and a suite of programs for analyzing the results. BEAST contains molecular clock models which are used to estimate roots of time-measured phylogenies. In addition to testing evolutionary hypotheses without considering a particular tree topology, BEAST also reconstructs phylogenies. Each tree in BEAST is weighted proportionally to its posterior probability which is procured by Markov chain Monte Carlo (MCMC) methods [21–24]. MCMC methods is a random sampling method used to determine posterior probability [25].

Multiple Sequence Alignment (MSA) was performed against the ancestral sequence Wuhan-Hu-1 genome, using MAFFTv.7.453 webserver, with the "-addfragment" option [26]. MAFFT v.7.4453 finds the homologous sequences in the genome using Fast Fourier Transformation method. Identifying homologous regions in the given sequences is a central procedure in identifying similar or measuring the similarity between sequences, and thereby mapping the evolutionary pathway [26,27]. Phylogenetic inference of Maximum Likelihood trees and best-fit nucleotide substitution model was performed using IQ-TREE via the Cyberinfrastructure for Phylogenetic Research (CIPRES) web portal [28,29]. Maximum Likelihood methods process the probability of a phylogenetic tree, constructed using the SARS-CoV-2 sequences, given a specific model of evolution [30]. Phylogenetic inferences pivot statistical studies, as such in the absence of a model of probability, the best fit nucleotide model substitute for estimating phylogenies between sequences [31]. The Nexus file containing the sequences was imported to Bayesian Evolutionary Analysis Utility (BEAUti), which is a graphical user-interface that generates the XML file containing specifications for BEAST v.2.6.4.0 [23].

Birth Death Skyline (BDSKY) Serial v.2.4 with a strict molecular clock and tip date calibration with the clock rate at 0.001 per substitutions/site/ year was used [32,33]. The BDSKY-Serial model is the optimal tree prior for pandemic studies with temporal sequences, though contrasting, a strict molecular clock was used for the data set as the pandemic is relatively new [34]. Tip dating uses molecular techniques to infer time-calibrated tree structures. The evolutionary tree is then analyzed using the sample ages provided by the sample. Since the data is contemporaneous, it was pivotal to tip date for the estimation of tree branch length. This was achieved by using data samples with collection date information. Strict Clock Model has only one parameter called the clock rate, which is the substitution rate, measured in substitutions per site per year (s/s/y). This molecular clock model was used to estimate divergence as the observed sequence data is from a single, relatively new epidemic of SARS-CoV-2 virus in humans in at a specific location, thus it was unnecessary to assume different substitution rates for different lineages. Site heterogeneity in the data is accounted for by the Gamma model, which estimates the variability in substitution rates from site to site. Gamma

Site model with gamma category count 4, and empirical frequencies was selected. Gamma distribution is discretized to a small numb of bins (4- 6) to make the analysis traceable. The average of each bin functions as a multiplier for calculating transition probabilities for each scaled substitution rate. The log normal distributed rate heterogeneity for the sites and substitution model were fixed as shown in Table 1. The model selections made in the Site and Clock Model tabs determine which parameters are included in the model.
In the initialization panel of BEAUti, the become uninfectious rate parameter was fixed at 36.5, the dimension of the reproductive number parameter was changed to 7, and the upper value for the origin parameter was fixed at 10. In addition to being a mutation-prone virus, the viral dynamics of SARS-CoV-2 are highly susceptible to change. This change in viral dynamics was accommodated by fixing the dimensions of Reproductive number (Re) at 7 which allowed the parameter to change 7 equally spaced times between the origin of the epidemic and the present time [4,22]. Become Uninfectious rate can be understood as the time period an individual takes to recover or become uninfectious from the disease. A time period of 10 days is the generally accepted time interval from infection to recovery as such the rate was fixed at 365/10 = 36.5 [22].
The Priors tab allow prior distributions to be specified for each model parameter. Birth Death Skyline Serial was the tree prior preferred for the research as it assumes that the collected data are heterochronous. It allows for the assumption that the become uninfectious rate is constant while allowing reproductive number to change. A log normal prior was used which allows for positive rates and high estimates for Re [34]. Uniform become uninfectious rate and a beta sampling proportion of 0.01 was fixed. The sampling proportion refers to the samples of infected individuals collected since the origin of the epidemic and the date of the last sample collected according to the data. Since it is evident that the sampling proportion of the obtained data does not contain all the sequences of tested individuals, a prior between 0 and 1 was selected as a result [34]. Therefore, sampling proportion was hence given a beta distribution which defines values between 0 and 1 [34]. MCMC chains were run for a chain length of 50 million with storing every 10,000. Table 1 summarizes the specifications for the BEAST run, and the remaining parameters were kept at their default values.

Parameters	Values
Clock rate	0.001 substitutions/site/year
Origin	Uniform distribution, lower = 0, upper=10
Sampling proportion	Beta distribution, Alpha =2, beta =10
Become uninfectious rate	Uniform distribution, fixed at 36.5
Reproductive Number	Log normal distribution, dimension =7
MCMC Chain length	50,000,000
Log Every	10,000

Table 1: Table showing the modified parameters and their corresponding values.

The output log file generated by BEAST v.2 was exported to Tracer to analyse the results35. Tracer reads the trace files and estimated the data on a Susceptible, Infectious and Recovered individuals in a population, in addition to evaluating the TMRCA [35]. The results of Tracer were fixed with a burn-in of 10%, which means that the first 500000 iterations generated by the BEAST run was discarded, to remove the presence of any low-probability portions of the posterior distribution. Fluctuations in the Reproductive numbers (Re) are indicators of change in factors that impact viral transmission. In the event that Re is higher than 1, the epidemic will continue to spread and prevention efforts will attempt to push Re below 124. Treatment, vaccinations, quarantines and behavioral changes can all contribute to reducing Re more rapidly after more people have been infected and the population becoming susceptible decreases. If an epidemic is neither growing or declining it has a value of 1 [24].

TreeAnnotator v.2.6.4.0 was used to create the target tree that summarizes the tree file generated by BEAST v.2 [36]. TreeAnnotator, as its name suggests, annotates the single target tree with the summary information from the entire sample. Each node or clade in the 'target' tree is evaluated according to the values of average nodes ages along with HPD intervals, the posterior support and the average evolution rate on each branch which would be included in the tree files generated by the BEAST run [24,36]. Maximum Clade Credibility (MCC) tree was chosen as the target tree type with 10% Burnin and median node heights which gives an average node height of the trees [24]. Maximum clade credibility examines all the trees in the data input and selects the tree with the highest sum of the posterior probabilities of all its nodes [24].
Posterior Probability Limit is defined as the minimum posterior probability for a node in order for TreeAnnotator to store the annotated information [24,36]. The value was set to 0.5 i.e., only nodes with this posterior probability or greater will have information summarized. The target tree thus generated was visualized using Fig Tree v.1.4.4 [37].

3 Results

3.1 Data acquisition

Out of 81 sequences submitted on the database, 68 completed and high coverage sequences were downloaded from the GISAID portal. The first sequence was sampled on 4th of April 2020 and the most recent sequence was sampled on 15th of July 2021.

3.2 Phylogenetic Analysis by IQ-TREE

Best-fit nucleotide substitution model using Model Finder by IQ-TREE generated the results shown in Table 2. Meanwhile, the result of Maximum Likelihood method of Tree inference by IQ- TREE produced is shown in Table 3, followed by a visual representation of the ML Tree using FigTree v1.4.4 software in Figure 1 [37].

Parameters	IQ-TREE Result
Akaike Information Criterion (AIC)	GTR+F+I
Akaike Information Criterion (AIC) Score	89984.8624
Corrected Akaike Information Criterion (AICc)	GTR+F+I
Corrected Akaike Information Criterion (AICc) Score	89986.2270
Bayesian Information Criterion (BIC)	GTR+F+I
Bayesian Information Criterion (BIC) Score	91164.2738
Best-fit Model	GTR+F+I chosen according to BIC
Rate parameters	A-C: 0.2029, A-G: 0.6972, A-T: 0.0637, C-G: 0.1951, C-T: 2.0902, G-T: 1.0000
Proportion of Invariable Sites	0.8865
Base Frequencies	(A) = 0.2987, (C) = 0.1835, (G) = 0.1963 (T) = 0.3215

Table 2: Output generated by IQ-TREE model selection.

Log-likelihood of the tree	-44842.9931
Unconstrained Log-Likelihood (without tree)	-55204.8847
Number of free parameters (#branches + #model parameters)	142
Akaike information criterion (AIC) score	89969.9862
Corrected Akaike information criterion (AICc) score	89971.3508
Bayesian information criterion (BIC) score	91149.3976
Total tree length (sum of branch lengths)	0.0147
Sum of internal branch lengths	0.0030 (20.2727% of tree length)

Table 3: Table showing the output of Maximum Likelihood Tree Inference by IQ-TREE.

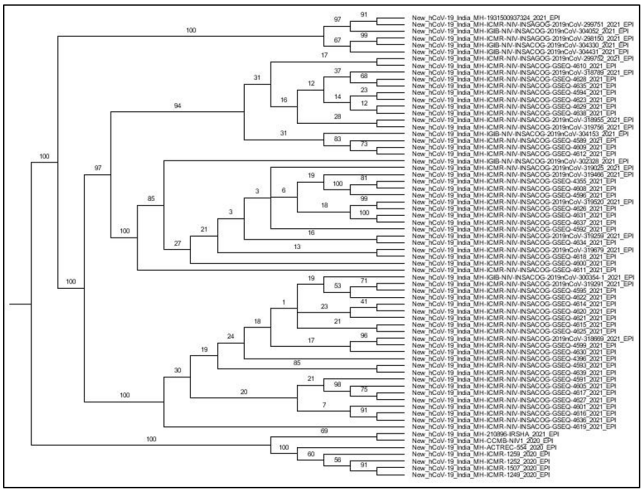

Figure 1: Visualization of Maximum Likelihood Trees with their bootstrap values generated by IQ-TREE using FigTree.

3.3 Phylodynamic Analysis by BEAST v.2

Log file generated by BEASTv.2 was exported to Tracer to analyse the posterior estimates. The Effective Sample Size (ESS) for all parameters is above 200, showcasing ideal convergence of states. Traces with mean of zero and null ESS corresponds to fixed parameters. A summary of the traces is shown in Table 4.

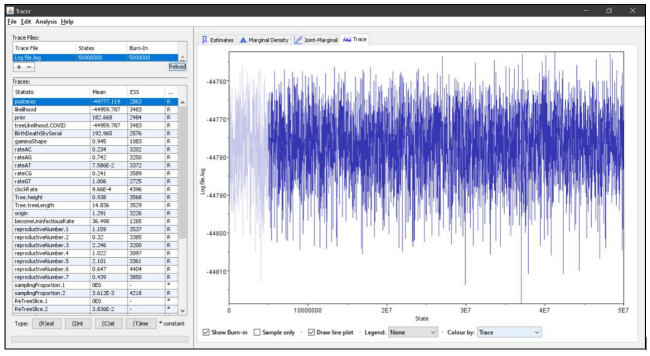

Figure 2: Visualization of Posterior Parameter using Trace Panel in Tracer.

Parameter/ Trace	Median	95% HPD interval
Tree Height	0.9342	[0.7922, 1.0962]
Origin	1.0811	[0.7508, 2.1208]
becomeUninfectiousRate	36.498	[36.4001, 36.5869]
Sampling proportion	3.4381×10^{-3}	[1.3278×10^{-3}, 6.202×10^{-3}]

Table 4: TMedian and 95% HPD interval of BEAST parameters as reported by Tracer. Tracer reported a Tree Height with a median of 0.9342 and 95% HPD interval between [0.7922, 1.0962]

Figure 3 indicates a Kernel Density Estimate (KDE) plot visualization of the Tree Height trace. The origin parameter showcased a median of 1.0811 and 95% HPD interval [0.7508, 2.1208], while the become Uninfectious Rate trace showed a median at 36.498 and [36.4001, 36.5869] 95% HPD interval, visual representations of which are shown in figures 4 and 5 respectively.

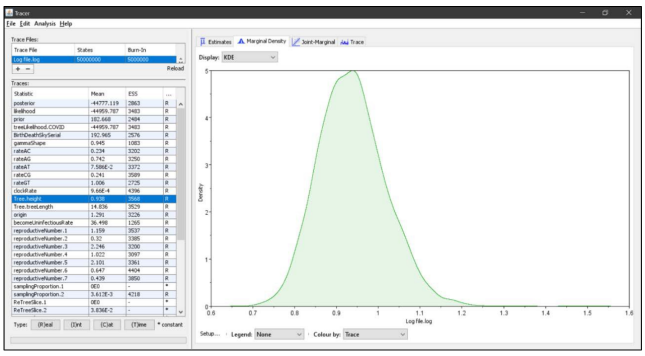

Figure 3: Tracer interface visualizing estimates for Tree Height.

The sampling proportion reported a median of 3.4381×10^{-3} and 95% HPD interval [1.3278×10^{-3}, 6.202×10^{-3}]. A sampling proportion with a mean of zero defines the sequence availability between the origin of the epidemics and the first sequence in the acquired data.
The Reproductive numbers is an average indication of number of secondary transmissions by an infectious individual. Throughout the BEAST run, reproductive number was allowed to change 7 times as shown in Figure 7. The boxplots seen in figure 7, represents the changes in the rate at which the virus spread since the beginning of the epidemic in Pune till 15th July 2021, i.e., from the oldest to the latest time interval from left to right respectively. Upon observation of the results, it is recognized that the values of R number increase for time intervals 3 and 5 thereby stipulating peaks in the epidemics. The tree file generated by BEAST was summarized into a median node height MCC target tree by processing 4501 trees which was visualized with transformed branches in Fig Tree v1.4.4, as represented in figures 8 and 9, respectively.

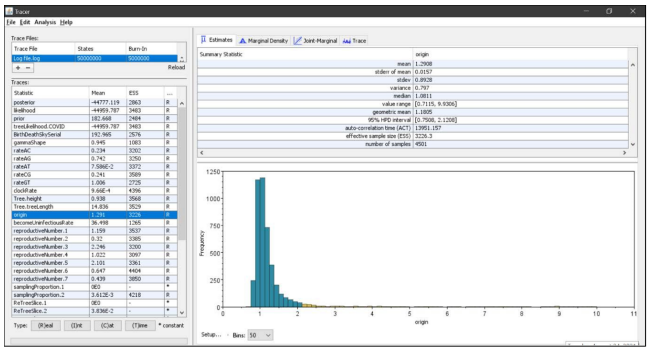

Figure 4: Tracer interface visualizing estimates for Origin.

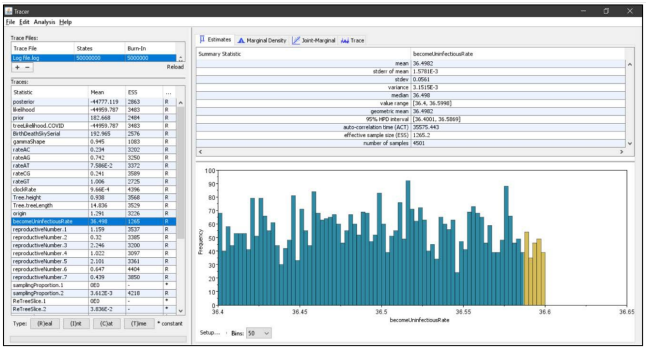

Figure 5: Tracer interface visualizing estimates for become Uninfectious Rate.

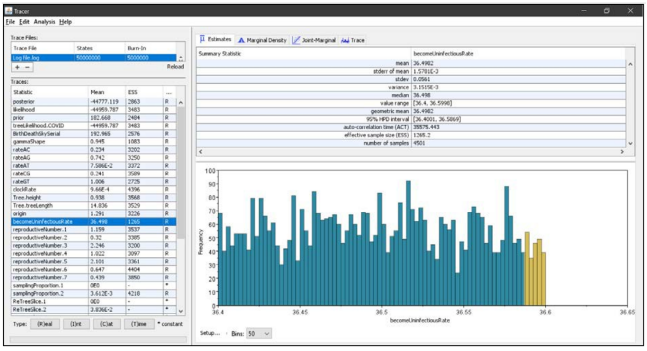

Figure 6: Tracer interface visualizing estimates for Sampling Proportion.

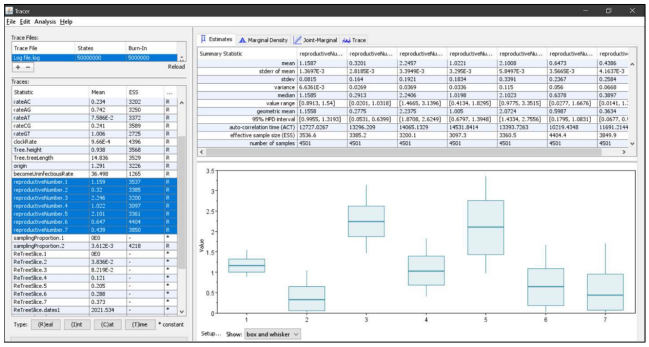

Figure 7: Tracer interface visualizing estimates for Reproductive number.

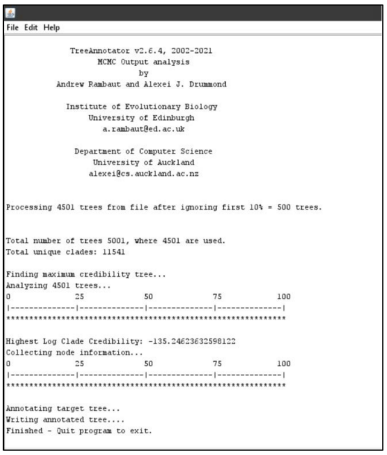

Figure 8: TreeAnnotator v.2.6.4.0 User Interface.

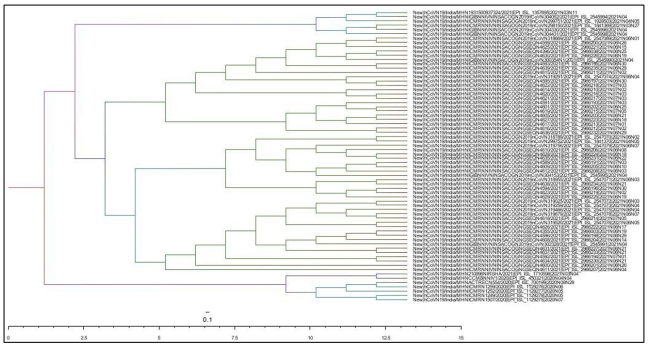

Figure 9: Maximum Clade Credibility (MCC) Tree with Median Node Heights Visualized by FigTree.

4 Discussion

The central axis of any phylodynamic analysis using Bayesian method is the calibration of a molecular clock. Bayesian dating is based on a variety of assumptions about evolution, which can either be realist or empirically supported to varying degrees and analyses of molecular dating can be greatly impacted by the assumptions used [38]. One assumption made is that the sequences are aligned against their homologous regions [38]. This assumption was made to be true by performing MSA using MAFFT v.7.453 [26]. If the data are not aligned, evolutionary divergence estimates are inaccurate and false molecular dating results are observed [38]. Another assumption is the nucleotide substitution model for the data [38]. Molecular dating is heavily based on the presence of homologous regions for the comparison of sequences to attain their evolutionary history, however, different loci are prone to different evolutionary rates [38]. This results in finding the best nucleotide model for the given set of data crucial and for this model selection was performed using IQ-TREE [28]. The assumptions about the priors of the given data, however, are error-prone as there is no concrete evidence to provide the correct values. Priors were fixed based on user discretion and previously generated results.

Bayesian methods are built upon the Bayes theorem which states that the probability distribution is given by the product of the prior distribution and the likelihood score, divided by the probability of the data [38–40]. Likelihood score is generated by the ML trees and substitution model, both of which constitutes the likelihood model [38]. Likelihood is defined as the probability of the given set of data provided by a particular model [28,30,38,41]. The tree model includes the tree topology which defines relationships among nodes, as well as the distances between nodes and their connecting branches [30,38].There are three types of tree nodes, the terminal nodes or tips which indicates the sampled taxa, the internal nodes which are the hypothetical ancestors and the root node which represents the common ancestor of all terminal nodes38. Analysis of the tree topology involves positioning the terminal nodes at the same time point which is always the present, and estimating the positions of the internal nodes38. As a result of taking into account substitutions in the sequences, estimation of the rate of mutation that has occurred since the sequences were copied from their last common ancestor can be calculated [35,38,39]. Based on patterns of sequence variation observed in our alignments, the substitution model explores the probability of different types of substitution events [38,41,42]. The substitution model used and the priors on the parameters of the model affects the estimation of dates, and while model fit is a good indicator of the best model given the available choices, it does not guarantee that the selected model matches the evolutionary process that generated the data [38]. Data priors represent our beliefs about the parameters within the likelihood model, appraising information included in the data. The tree priors can have a wide range of impact on date estimates and since the tree prior contributes to the posterior distribution of all node times, using an inappropriate tree prior or applying unreasonable constraints on a parameter may result in inaccurate dates estimates at each node in the tree [21,22,34,38]. Molecular dating techniques depends on taxon sampling. This

means that the number of samples is proportional to the results, that is to say, the higher the number of samples included in the data, better the inference of topology and accuracy of branch-length estimates due to reduced node-density effect. Failure to sample the recent internal nodes produces an incomplete sampling of taxa which tends to increase the length of terminal branches in phylogenetically challenging trees [38]. The strict clock allows only one rate of substitution per site in the phylogeny. Though a constant- rate model can provide more precise date estimates than a model based on variable rates, using stochastic rates and a relatively small variance between branches, in the presence of significant rate variation, however, constant-rate models have poor accuracy [21,22,34]. In light of the diversity and number of biological factors implicated in influencing substitution rates, especially in the case of viral strains, the assumption that all lineages in the given data set share the same rate of change is to be questioned. Even though the index case in India was reported on 30th January 2020, the sequencing of the viral strain was done only by 6th March 2020. India has reported a total of 32.6 million SARS-CoV-2 infections, however, only 48,065 viral sequences are available on GISAID, out of which 7348 sequences were isolated from Maharashtra and 81 sequences from Pune alone. The first case of COVID-19 in Maharashtra was recorded on 9th March 2020, contradictory to this report however, GISAID cites the first sequence from Maharashtra on 6th March 2020. This may be due to the fact that the laboratories involved, in the collection and submission of the data, was situated in the state of concern, nevertheless, it deems the validity of geographical isolation of the strains to be uncertain. Moreover, the lack of high coverage sequences against the growing infection negatively impacts the phylodynamic analysis of the virus. The Pune Municipal Corporation (PMC) upon learning about the index case in the state sealed all the areas around the accommodation in a 3 km radius in Pune. Government policies to prevent the spread of virus was undertaken starting from 15th March 2020 with initial closing of shopping malls and educational institutions, live orchestra, disco and pubs with a ban to avoid crowds in public places [43]. On 17th of March 2020, guidelines on preventive measure to control the spread of virus was issued, this was followed by the implementation of Janta Curfew on 21st of March 2020 and a state-wide lockdown on 23rd March 2020 [43–45]. Restrictions were eased the following month, but due to escalation in cases at specific regions, containment zones were introduced on 24th April 2020 [43]. In other areas of the state, courts, companies, salons and spas, and government service centres were gradually allowed permission to open from June, under strict COVID-19 regulations43. Another complete lockdown was brought into effect from 14th July 2020 for a week. Maharashtra was the first state to implement fines for not wearing masks in public spaces [43]. By September 2020, the number of infections in Maharashtra had increased exponentially and hospitals were faced with lack of oxygen supply. On 30th March 2021, a Committee for Oxygen Supply was set up in Maharashtra in order to effectively communicate and transport patients and oxygen supply amongst hospitals across the state.

Phylodynamic analysis of the obtained data clearly stipulates the effect of government restrictions and its mismanagement thereof. Reproductive numbers as seen in figure 7 indicates decrease in Re number below 1, which can be the results of the restrictions in place, and the increase in Re number can be assigned to the ease of restrictions. It can be seen from the data that there has been a steep inclination during the months of May and June 2021, with a constant rate at the beginning of the year. The appropriate explanation for this would be the increase in COVID-19 infections from the second wave starting from mid-March. Though government control did bring the curve down, ease of restrictions in April saw a consequent growth in cases starting Mid-April, which peaked in May. Break the Chain order was issued on 4th April 2021 in order to control the second wave of infections in the state [43]. An additional reason as to why the spread of virus was controlled since its outbreak in the state, was that the restrictions in effect were enforced by the State government throughout the entire nation, over-riding the constitutional rights of the local government, it was during the latter part of 2020 that local governmental bodies were allowed to take control of the situation. This negatively impacted the administration of preventive measures as workers

from other states flowed into Maharashtra, as is seen as an increase in June. From this evidence, it is affirmed that restrictions during a surge in infections and their consequent relaxations during decline, is a never- ending cycle and other efforts should be undertaken to effectively control the epidemic. In addition to alleviation of the lockdown, another cause for the increase in Re can be attuned to the celebrations of festivals like Kumbh Mela, Holi and Diwali, which can turn into super spreader events.

The state government's strategies to control the spread of virus also includes vaccinations in two doses. As of 26th August 2021, 54,865,308 vaccinations have been provided to the citizens out of which only 14,594,471 individuals have received the second dose [17]. Despite government interventions, Maharashtra remains the worst affected state in India with 6.44 million number of cases and 137K deaths, and amongst the districts of Maharashtra, Pune has the highest number of casualties [17]. The government has yet to publish the total number of samples tested therefore the effectiveness of the vaccinations provided in the locality remains uncertain, but it is transpicuous that scarcity of vaccines and shortage of oxygen supply has led to a surge of infections and deaths, not just in the state but across the whole nation [46,47].

Research on the phylogenomics and phylodynamics of SARS-CoV-2 in India under a Hasegawa-Kishino-Yano (HKY) Substitution model using Birth Death Serial tree prior and a log normal clock reported Re number at 3.683 [95% HPD: 2.411, 5.401] and infectious rate as 7.44 [95% HPD: 4.51,9.99][48]. Median Joining Network Analysis of 138 haplotypes from GISAID discovered that sequences from Maharashtra might have evolved from cluster B of the virus, which shows linkage with the Wuhan ancestral sequence EPI ISL 406798 [15,48].

These findings, however, are based on a broad spectrum of assumptions and the results, though seemingly consistent with statistical observations, must be utilized under extreme scrutiny. The index case in Maharashtra was a Non-resident Indian (NRI) couple from Dubai, and all the passengers on that flight has been tested positive for COVID-19. The Dubai International Airport is one of the busiest travel routes connecting the Eastern and Western countries. Further analysis of the most recent common ancestor requires that individuals that had arrived in Maharashtra from Dubai and other passengers and personnel that had worked in the airport be tested and samples be collected. The delay caused by the collection and submission of the existing sequences, however, has provided the time interval for the recovery of the patients, thereby making their testing futile. Rather, mapping the travel routes of the ex-carriers is a more practical, though time-consuming, approach. A secondary suggestion would be that the sequence data, if available, of passengers and personnel of Dubai International Airport be combined on a separate platform for a clearer understanding of the transmission dynamics of the pathogen.

Based on the literature and observed results, it is recommended that the restrictions be sustained in the state and that masks, vaccinations and social distancing be made compulsory in order to contain the spread of the virus. Increased testing in the community would be beneficial as it identifies asymptomatic carriers in addition to providing samples for sequencing. This in turn ensures improved results of phylodynamic analysis which can be used to understand viral phylodynamics and potency of preventive measures in place, thus completing the circle.

Abbreviations

2019-n-CoV- 2019 Novel Coronavirus
AIC- Akaike Information Criterion
AICc- Corrected Akaike Information Criterion
BDSKY- Birth Death Skyline
BEAST- Bayesian Evolutionary Analysis by Sampling Trees
BEAUti- Bayesian Evolutionary Analysis Utility
BIC- Bayesian Information Criterion
CIPRES - Cyberinfrastructure for Phylogenetic Research COVID-19- Coronavirus Disease

DNA- Deoxyribonucleic acid
ESS- Effective Sample Size
GISAID- Global Initiative in Sharing Avian Influenza Data
GISRS- Global Influenza Surveillance and Response System
GTR – General Time Reversible
HKY- Hasegawa-Kishino-Yano
HPD – Highest Posterior Density
IQ-TREE- Important Quartet Tree
KDE- Kernel Density Estimate
MAFFT- Multiple Alignment using Fast Fourier Transform
MCC- Maximum Clade Credibility
MCMC- Markov Chain Monte Carlo ML- Maximum Likelihood
NGS- Next Generation Sequencing
NRI- Non-Resident Indian
NSP – Non-structural proteins ORF – Open Reading Frame
PMC- Pune Municipal Corporation
Re / Rt Number - Effective Reproductive number RNA – Ribonucleic Acid
rRNA – Ribosomal Ribonucleic Acid
SARS – severe acute respiratory syndrome
SIR- Susceptible, Infected and Recovered
TMRCA – Time since the Most Recent Common Ancestor
WHO – World Health Organization

References

[1] Rothan HA, Byrareddy SN. The epidemiology and pathogenesis of coronavirus disease (COVID-19) outbreak. J Autoimmun [Internet]. 2020;109(February):102433. Available from: https://doi.org/10.1016/j.jaut.2020.102433.

[2] Wu D, Wu T, Liu Q, Yang Z. The SARS-CoV-2 outbreak: What we know. Int J Infect Dis. 2020;94:44–8

[3] Burki T. The origin of SARS-CoV-2. Lancet Infect Dis. 2020;20(9):1018–9

[4] Khailany RA, Safdar M, Ozaslan M. Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID- 19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information. Gene Reports. 2020;19(January):1–6.

[5] Saha I, Ghosh N, Maity D, Sharma N, Sarkar JP, Mitra K. Genome-wide analysis of Indian SARS-CoV-2 genomes for the identification of genetic mutation and SNP. Infect Genet Evol [Internet].2020;85(July):104457. Available from: https://doi.org/10.1016/j.meegid.2020.104457

[6] Hardison RC. Comparative genomics. PLoS Biol. 2003;1(2):156–60.

[7] Rambaut A, Holmes EC, O'Toole Á, Hill V, McCrone JT, Ruis C, et al. A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. Nat Microbiol [Internet]. 2020;5(11):1403–7. Available from: http://dx.doi.org/10.1038/s41564-020-0770-5

[8] Erik M. Volz, Katia Koelle TB. Viral Phylodynamics. 2013. p. 1–12.

[9] WHO Guidance Note. WHO-convened Global Study of Origins of SARS-CoV-2 : China Part (14 January-10 February 2021). World Heal Organ. 2021;(February):120.

[10] Pybus OG, Rambaut A. Evolutionary analysis of the dynamics of viral infectious disease. Nat Rev Genet. 2009;10(8):540–50.

[11] Rife BD, Mavian C, Chen X, Ciccozzi M, Salemi M, Min J. Phylodynamic applications in 21 st century global infectious disease research. 2017; 1–10.

[12] Park ST, Kim J. Trends in next-generation sequencing and a new era for whole genome sequencing. Int Neurourol J. 2016; 20:76–83.

[13] Houldcroft CJ, Beale MA, Breuer J. Clinical and biological insights from viral genome sequencing. Nat Rev Microbiol. 2017;15(3):183–92.

[14] Jiang R, Zhang X, Zhang MQ. Basics of Bioinformatics. Basics of Bioinformatics. 2013.

[15] Global Initiative on Sharing Avian Flu Data [Internet]. [cited 2021 Aug 1]. Available from: https://www.gisaid.org/.

[16] Department of Science and Technology. Indian Supermodel for Covid-19 Pandemic. Vol.0. National Supermodel Committee.

[17] Government of India COVID website [Internet]. Available from: https://www.mygov.in/covid-19

[18] Ministry of Health and Family Welfare [Internet]. [cited 2021 Aug 25]. Available from: https://www.mohfw.gov.in/

[19] Pevsner J. Bioinformatics and Functional Genomics. Bioinformatics and Functional Genomics. 2005.

[20] Edwards YJK. Bioinformatics and Functional Genomics. Vol. 3, Briefings in Functional Genomics and Proteomics. 2004. 187–190.

[21] Bošková V. Tutorial using BEAST v2.4.2 Introduction to BEAST2.:1–30.

[22] Pečerska J, Bošková V, du Plessis L. Tutorial using BEAST v2.5.0 . 2020;1–39. Available from: http://beast.community/tracer

[23] Bouckaert R, Heled J, Kühnert D, Vaughan T, Wu CH, Xie D, et al. BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. PLoS Comput Biol. 2014;10(4):1–7.

[24] Drummond AJ. Notes Assisting BEAST Tutorials. 2000;1–3.

[25] Jannink J. Likelihood of Bayesian, and MCMC Methods in Quantitative Genetics. Vol. 43, Crop Science. 2003. 1574–1575.

[26] Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7:Improvements in performance and usability. Mol Biol Evol. 2013; 30 (4):772–80.

[27] Daugelaite J, O' Driscoll A, Sleator RD. An Overview of Multiple Sequence Alignments and Cloud Computing in Bioinformatics. ISRN Biomath. 2013: 1–14.

[28] Nguyen LT, Schmidt HA, Von Haeseler A, Minh BQ. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015;32(1):268–74.

[29] Miller MA, Pfeiffer W, Schwartz T. Creating the CIPRES Science Gateway for inference of large phylogenetic trees. 2010 Gatew Comput Environ Work GCE 2010. 2010.

[30] Cho A. Constructing Phylogenetic Trees Using Maximum Likelihood Constructing. Scripps Sr Theses. 2012; 46: 1–49.

[31] Posada D, Crandall KA. Selecting the Best-Fit Model of Nucleotide Substitution. Syst. Biol. 2001;50(4):580–601.

[32] MacLean OA, Orton RJ, Singer JB, Robertson DL. No evidence for distinct types in the evolution of SARS-CoV-2. Virus Evol. 2020;6(1):1–6.

[33] Danesh G, Elie B, Michalakis Y, Sofonea MT, Bal A, Behillil S, et al. Early phylodynamics analysis of the COVID-19 epidemic in France. medRxiv. 2020;1–26.

[34] Müller NF, Plessis L. Tutorial using BEAST v2.4.2 Skyline plots. 2013;1–30.

[35] Rambaut A, Drummond AJ, Xie D, Baele G, Suchard MA. Posterior summarization in Bayesian phylogenetics using Tracer 1.7. Syst Biol. 2018;67(5):901–4.

[36] Helfrich P, Rieb E, Abrami G, Lücking A, Mehler A. Treeannotator: Versatile visual annotation of hierarchical text relations. Lr 2018 - 11th Int Conf Lang Resour Eval. 2019;(February):1958–63.

[37] Rambaut A. FigTree READ ME [Internet]. 2018. Available from: https://beast.community/figtree

[38] Bromham L, Duchêne S, Hua X, Ritchie AM, Duchêne DA, Ho SYW. Bayesian molecular dating: opening up the black box. Biol Rev. 2018;93(2):1165–91.

[39] Inference B. Bayesian Inference Chapter 12. Stat Mach Learn [Internet]. 2014;299–351. Available from: http://www.stat.cmu.edu/ larry/=sml/Bayes.pdf

[40] Moore PG. The Bayesian Approach to Statistics. J Inst Actuar. 1966;92(3):326–39.

[41] Hall BG. Building phylogenetic trees from molecular data with MEGA. Mol Biol Evol. 2013;30(5):1229–35.

[42] Drummond AJ, Nicholls GK, Rodrigo AG, Solomon W. Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. Genetics. 2002;161(3):1307–20.

[43] District of Pune, Government of Maharashtra Website [Internet]. [cited 2021 Aug 3]. Available from: https://pune.gov.in/corona-virus-updates/

[44] Implementation of Janta Curfew [Internet]. Available from: https://pune.gov.in/corona- virus-updates/

[45] Lockdown orders Maharashtra [Internet]. Available from: https://pune.gov.in/corona-virus- updates/

[46] COVID-19: India Outrage Over "No Oxygen Shortage Death Data" Claim. 2021.

[47] Vaccination Declines by 60% as States Say They Have No Doses. The Hindu [Internet]. Jul; Available from: https://www.thehindu.com/news/national/coronavirus-vaccination- declines-by-60-as-states-say-they-have-no-doses/article35303316.ece

[48] Farah, Sameera; Atkulwar, Ashwin; Praharaj, Manas Ranjan; Khan, Raja; Gandham, Ravikumar; Baig M. Phylodynamics of SARS-CoV-2 with Reference to India. medRxiv. 2020;(165):1–13.