Environmental Sustainability (2021) 4:533–541
https://doi.org/10.1007/s42398-020-00123-z

ORIGINAL ARTICLE

Transmission of SARS-CoV-2 in South Asian countries: molecular evolutionary model based phylogenetic and mutation analysis

Anand Prakash Maurya[1] · Rupesh V. Chikhale[2] · Piyush Pandey[1]

Received: 13 July 2020 / Revised: 29 August 2020 / Accepted: 31 August 2020 / Published online: 18 September 2020

Abstract

The on-going coronavirus disease 19 (COVID-19) pandemic has caused a very high number of infections and deaths around the globe. The absence of vaccine/drugs to counter COVID-19 has scrambled scientific communities to repurpose available medicines/vaccines. As the virus is known to mutate, using the whole genome sequences, the transmission dynamics and molecular evolutionary models were evaluated for South Asian countries to determine the evolutionary rate of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Phylogenetic analyses were done using the data available on National Center for Biotechnology Information (NCBI). Different nucleotide substitution models and molecular evolutionary models were analyzed to see how SARS-CoV-2 was transmitted in the populations. Models for the viral 'S' and 'N' protein from selected strains were constructed, validated, and analyzed to determine the mutations and discover the potential therapeutics against this deadly viral disease. We found that the Hasegawa-Kishino-Yano (HKY) nucleotide substitution model was the best model with the lowest Bayesian information criterion (BIC) scores. Molecular clock RelTime analysis showed the evolutionary rate of SARS-CoV-2 substitutions in the genome was at 95% confidence interval, and heterogeneity was observed. Several mutations in the viral S-protein were found with one in the receptor-binding domain concerning SARS-CoV-2/Wuhan-1/S-Protein. Nucleocapsid protein also showed mutations in the strains from India and Sri Lanka. Our analysis suggests that SARS-CoV-2 is evolving at a diverse rate. The mutation leading to substitution in the nucleotide sequence occurred in the genome during the transmission of COVID-19 among individuals in the South Asian countries.

Keywords COVID-19 · Pandemic · Coronavirus · Genetic variability · Receptor binding domain (RBD) · Human mobility

Introduction

In the past two decades, outbreaks of Severe Acute Respiratory Syndrome coronavirus (SARS-CoV) and Middle East respiratory syndrome coronavirus (MERS-CoV) had caused deadly human diseases (Cheng et al. 2007; Chan et al. 2015). The current pandemic of pandemic coronavirus disease 19 (COVID-19) is caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), which spreads through human-to-human contact via droplets due to sneezing or coughing or by touching the contaminated surface or coming in touch with an infected person (Drosten et al. 2003; Gralinski and Baric 2015; Guarner 2020). Coronaviruses cause fever and upper respiratory diseases, but some of these viruses, such as HCoV-KHU1, HCoV-229E, HCoV-OC43, and HCoVNL63 cause only minor respiratory infections (Zumla et al. 2016; Guarner 2020). SARS-CoV-2 is the 7th member of the coronavirus family which has spread in humans with symptoms like fever, and dry cough similar to SARS and MERS-CoV (Ceraolo and Giorgi 2020; Liu et al. 2020). SARS-CoV-2 was first identified in December-2019 in Wuhan, China, from where human-to-human transmission spread throughout the world (Wu et al. 2020; Wang et al. 2020). The investigation pointed out that the main source of this virus were animals sold in the seafood market, but

Anand Prakash Maurya and Rupesh V. Chikhale contributed equally.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s42398-020-00123-z) contains supplementary material, which is available to authorized users.

✉ Piyush Pandey
ppmicroaus@gmail.com; piyushddn@gmail.com

[1] Department of Microbiology, Assam University, Silchar 788011, Assam, India

[2] School of Pharmacy, University of East Anglia, Norwich Research Park, Norwich NR4 7TJ, UK

no specific linkage with SARS-CoV-2 has been identified (Zhou et al. 2020). Based on the genomic arrangements and sequences, it was speculated that this virus originally came from bats (Zhou et al. 2020). However, the intermediate host between bat and humans is still a mystery and a matter of investigation (Li et al. 2020). The SARS-CoV-2 is a unique member of coronavirus, which is enveloped and single-stranded RNA virus (Zumla et al. 2016). The virus consists of enveloped spike (S) protein (Du et al. 2009), and this assists the receptor binding and fusion with the membrane, which is essential for shaping host tropism and the ability to spread the virus (Li 2016).

The SARS-CoV-2 infection came to highlight in January 2020, and the declaration of a COVID-19 pandemic by WHO led to a panic worldwide. Most of the countries have already reported COVID-19 cases by the end of March 2020. In this article, we put focus on the COVID-19 situation in the South Asian countries, which include India, Pakistan, Bangladesh, Afghanistan, Nepal, Bhutan, Sri Lanka, and the Maldives. These countries are represented by the association known as the South Asian Association for Regional Cooperation (SAARC). In this research, we have focused on the COVID-19 situation in the region, from the countries where the genomes of SAR-CoV-2 have being reported, and so, their phylogeny and the mutations have been analyzed. The history of the first cases from different SAARC countries has been given in Table S1. A significant number of SARS-CoV-2 infections have been reported from the South Asian region [COVID-19 dashboard by Center for Systems Science and Engineering, 2020; accessed as on 10th August 2020].

Several SARS-CoV-2 sequences were accessed from India, Pakistan, Sri Lanka, and Nepal, available on the International Nucleotide Sequence Database Collaboration (NCBI gene bank). However, sequences from other South Asian countries like Afghanistan, Bangladesh, and Bhutan were not available on the NCBI gene bank. In this study, we have studied nine sequences of SARS-CoV-2 to construct the molecular evolutionary model based on their phylogenetic analysis and to explore the possible mutations while the spread of the infections in the SAARC regions.

Methods

Bioinformatics analysis, phylogenetic analysis, and molecular evolutionary modeling

The nucleotide sequences were accessed from the gene bank NCBI database (Sayers et al. 2019). A total of 9 genome sequences of countries, including India, Sri Lanka, Nepal, and Pakistan, were analyzed with the Wuhan reference genome sequence. In addition to this, two sequences for analysis were taken from Italy, as some of the infected people traveled from Itlay to India. Multiple sequence alignment (MSA) was done using the MUSCLE program version 3.8.31, and further percent identity matrixes were created by Clustal v2.1 (Edgar 2004)]. Genome sequences were aligned using MAFT v7.42 for further analysis (Katoh et al. 2019).

The phylogenetic tree was constructed using MEGA-X v10.1.8, and Molecular evolutionary (ME) history was inferred using the Maximum likelihood (ML) method, Tamura-Nei model, and Minimum Evolution (ME) method (Rzhetsky and Nei 1992; Tamura and Nei 1993; Kumar et al. 2018). The tree was further authenticated by investigation on 100 bootstrapped input datasets. ML fits the model for nucleotide substitutions was evaluated to find the best nucleotide substitution, models. The tree topology was computed with 11 nucleotide sequences for the estimation of ML values. 1st + 2nd + 3rd + Noncoding codon positions have been included (Nei and Kumar 2000).

Substitution patterns and rates in the sequences were estimated under the Tamura-Nei (1993) model (+G), which is essential for modernizing the evolutionary models. A discrete Gamma distribution was used to model evolutionary rate (Tamura and Nei 1993). Relative values of instantaneous *r* were considered during evaluation. The user-specified tree topology was analyzed by using the ML method and Tamura-Nei model, and the Maximum Parsimony method (Tamura and Nei 1993; Nei and Kumar 2000).

The strict and flexible uncorrelated relaxed Molecular clock test was done among 11 genome sequences to see the equality of evolutionary rate between 3 sequences; Tajima's relative rate test was performed. Genome sequences: **A** (MT012098.1 SARS-CoV-2 human IND 29 2020) and **B** (NC_045512.2 Wuhan Hu 1), and **C** (MT066156.1 SARS-CoV-2 human ITA INMI1 2020) were used as an outgroup. The analysis involved 3 nucleotide sequences, and 1st + 2nd + 3rd + Noncoding codon positions were included (Tajima 1993).

For detecting correlation of evolutionary rates in a phylogenetic tree of all 11 taxa; Corrtest analysis was performed using the branch lengths to compute evolutionary rates (Tao et al. 2018), and the tree topology (branch lengths) was analyzed using the ML method, based on the general time-reversible model (Nei and Kumar 2000).

Modeling of S and N protein

The MSA FASTA files were processed on the T-Coffee server (Paolo et al. 2011), and images were developed on the Boxshade v3.2 [https://embnet.vital-it.ch/software/BOX_doc.html] (ExPASy, 2019). The S-protein sequences were modeled on the locally installed Robetta (Walls et al. 2020), model validation was performed by Qualitative model energy analysis (Qmean) and MolProbity method (Angira et al. 2019). Ramachandran plot and images were generated

in the Molecular Operating Environment (MOE) (Kerzare et al. 2016). A score towards 1 is considered good; in this case, all models showed a score of 0.74, which is considered as highly acceptable. QMEAN estimates the native-ness of the protein model, and the score is based on the Z-score. Any homology model with a value between 0 and − 4 is considered a high-quality model. The Nucleocapsid N-protein for the genome sequences was modeled similarly with the crystal structure of the SARS-CoV-2 nucleocapsid protein RNA binding domain (PDB: 6M3M) (Kang et al. 2020) as the reference template. The rest of the validation parameters and protocols were similar, as mentioned above.

Results

Phylogenetic analysis

We collected 11 full genomic SARS-CoV-2 sequences from the NCBI database, and reference genome of SARS-Cov-2. Selected genome was: MT358637.1 (SARS-CoV-2/human/IND/GBRC1/2020); MT012098.1 (SARS-CoV-2/human/IND/29/2020); MT050493.1 (SARS-CoV-2/human/IND/166/2020); MT077125.1 (SARS-CoV-2/human/ITA/INMI1/2020); MT066156.1 (SARS-CoV-2/human/ITA/INMI1/2020); MT072688.1 (SARS-CoV-2/human/NPL/61-TW/2020); MT371047.1 (SARS-CoV-2/human/LKA/COV38/2020); MT371048.1 (SARS-CoV-2/human/LKA/COV53/2020); MT262993.1 (SARS-Cov-2/human/PAK/Manga1/2020); MT240479.1 (SARS-CoV-2/human/PAK/Gilgit1/2020), and a reference genome: NC_045512.2 Wuhan-Hu-1. MSA using MUSCLE (v3.8.31) shown that all were highly similar (> 99.95%) to the reference genome (Percent identity matrix: Supplementary data 1). Mutations in the nucleotide bases were found by MSA using CLUSTAL O (1.2.4), which shows there were mutations in SARS-CoV-2. Based on the percentage identity phylogenetic tree was constructed (Figure S1), and all 11 genomes show maximum similarity to the reference genome, suggesting that they were spread from Wuhan, China.

The phylogenetic tree with the highest log-likelihood is shown in Fig. 1. The evolutionary history was inferred using heuristic search, that was found by using Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using tamura-Nei model, and then selecting the topology with superior log likelihood value. The result presented that genomes were highly conserved, and phylogenetic reconstruction suggested that transmission of COVID-19 happened from Wuhan, China.

The Close-Neighbor-Interchange (CNI) algorithm has been used for exploring the ME tree at a search level of 1. The molecular clock RelTime method has been used to estimate the divergence times for all branching points in the topology. Also, the bars around each node represent 95% confidence intervals, signifying that the evolutionary rate of nucleotide substitutions per year in SARS-CoV-2. The ME time tree is shown in Fig. 2. All the 11 genome sequences implied that they were originated and spread through Wuhan SARS-CoV-2.

Molecular evolutionary model analysis of SARS-CoV-2

ML fits the model for nucleotide substitutions were evaluated. A total of 24 different nucleotide substitution models were found for ML fits. The model HKY with the lowest Bayesian information criterion (BIC) scores was found to be the best nucleotide substitution pattern model among all (Table S2). This model confirmed that nucleotides arose at diverse rates, and transitions and transversions happened at different degrees in transmission.

The heterogeneity rate among the genomes was determined using the ML estimate of the nucleotide substitution matrix and shown in Table 1. For simplicity, the sum of r values is made equal to 100. The total nucleotide frequencies were A = 29.89%, T/U = 32.11%, C = 18.37%, and G = 19.62%. Computed tree topology of the maximum Log-likelihood was − 40969.372. The results suggested that matrix found significant as the estimated Transition/Transversion bias (R) was 2.90 and confirming that there was heterogeneity among the genomes, which is also supported by MSA using CLUSTAL O v1.2.4.

ML has been estimated for site rates, with the estimated value of 0.0500 for the shape parameter for the discrete Gamma Distribution. A discrete Gamma distribution was used to model evolutionary rate differences among sites (5 categories, [+G]). Mean evolutionary rates in these categories were 0.00, 0.00, 0.00, 0.03, 4.97 substitutions per site. The nucleotide frequencies are A = 29.89%, T/U = 32.11%, C = 18.37%, and G = 19.62%. A tree topology of maximum Log-likelihood for this computation was − 40976.575, indicating that there are differences in the rate of evolution and transmission of SARS-CoV-2.

Evolutionary divergence analysis of tree topology was presented in Figure S2. Inferred Ancestral phylogeny tree was made. A set of possible nucleotides at each ancestral node was introduced in the tree and set of alternative nucleotides at each node, as shown in Figure S3. The result showed that there were similarities between the genome sequences, with some differences in the nucleotides.

The equality of evolutionary rate difference between the genetic diversity among 3 sequences was seen using Tajima's relative rate test through the molecular clock. The χ^2 test statistic was 5.00 ($P = 0.02535$ with 1 degree[s] of freedom). Since the P value here is less than 0.05, the null hypothesis was rejected for equal rates between lineages.

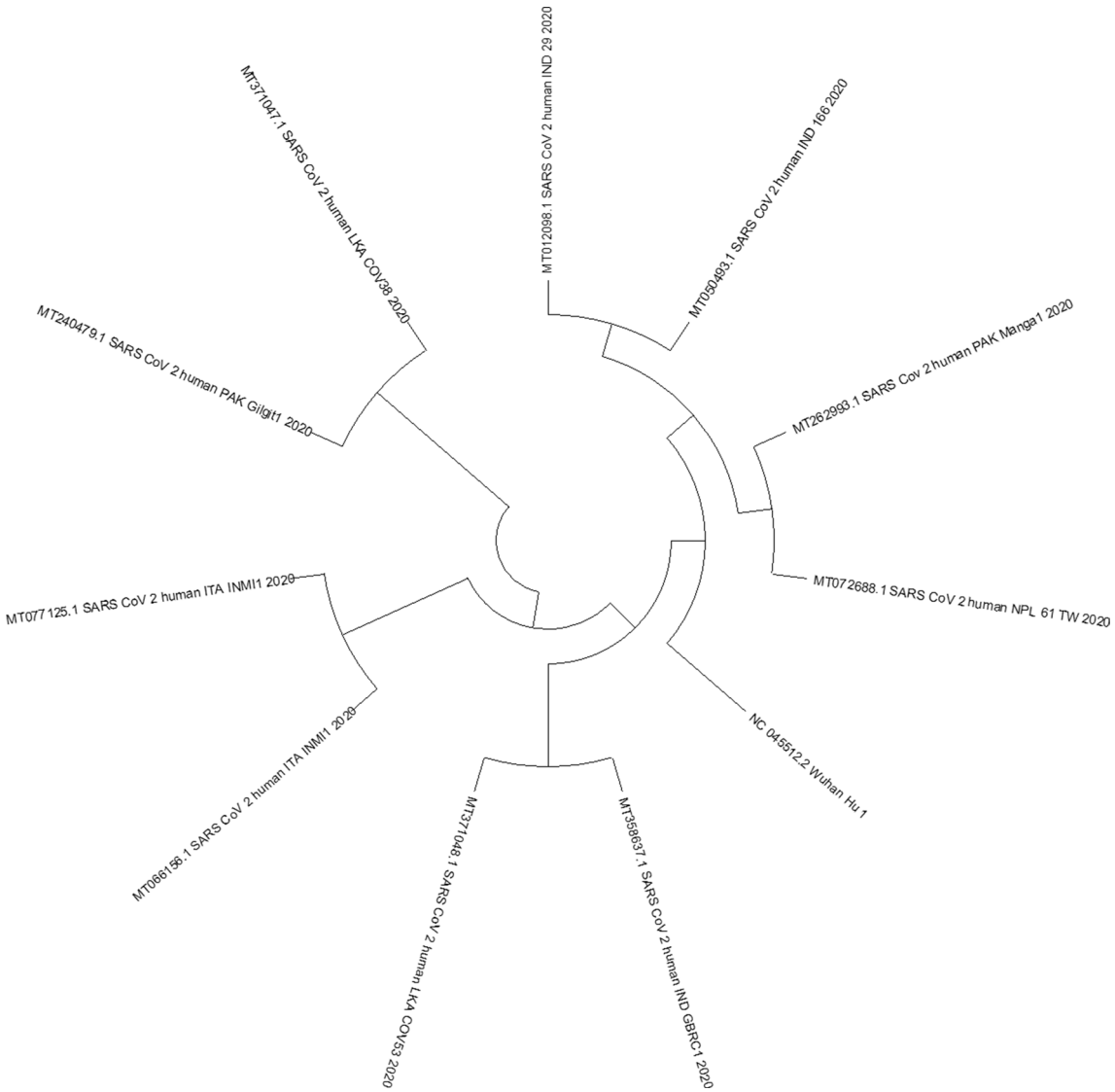

Fig. 1 Phylogenetic tree of SARS-CoV-2 genomes with the highest log likelihood indicating spread of COVID-19. Tree(s) for the heuristic search were obtained automatically by applying Neighbor-Join and BioNJ algorithms to a matrix of pairwise distances estimated using the Tamura-Nei model, and then selecting the topology with superior log likelihood value

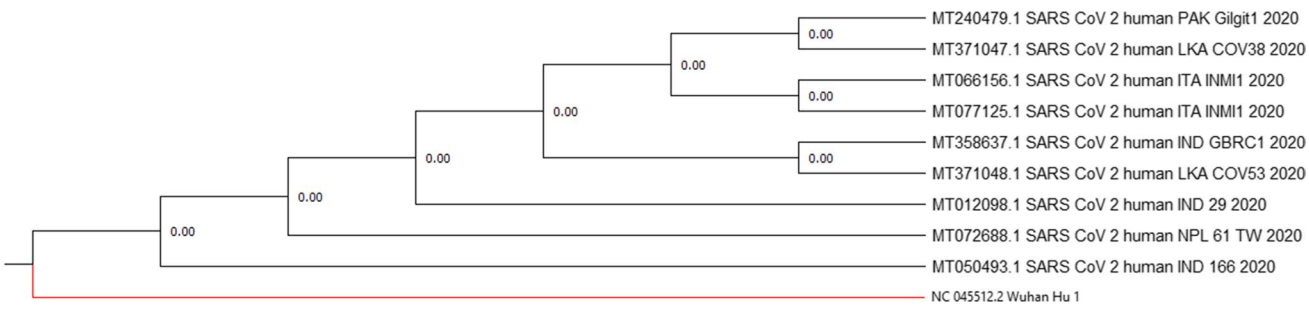

Fig. 2 Minimum evolution (ME) time tree of SARS-CoV-2 genomes: Origin and transmission dynamics at different rate and time intervals. (Each node represents 95% confidence intervals)

Table 1 Maximum likelihood estimate of substitution matrix (transitional substitutions are in bold and transversionsal substitutions are in italics)

Nucleotides	A	T/U	C	G
A	–	*3.92*	*2.25*	**8.18**
T/U	*3.65*	–	**19.99**	*2.40*
C	*3.65*	**34.93**	–	*2.40*
G	**12.46**	*3.92*	*2.25*	–

Each entry is the probability of substitution (r) from one base (row) to another base (column). Rates of different transitional substitutions are shown in bold and those of transversionsal substitutions are shown in italics

Table 2 Tajima's relative rate test (Sequences **A is**: MT012098.1 SARS CoV 2 human IND 29 2020, **B**: NC 045512.2 Wuhan Hu 1, and **C**: MT066156.1 SARS CoV 2 human ITA INMI1 2020)

Configuration	Count
Identical sites in all three sequences	29843
Divergent sites in all three sequences	0
Unique differences in Sequence A	5
Unique differences in Sequence B	0
Unique differences in Sequence C	2
Total	29850

There were a total of 29850 positions in the final dataset (Table 2). Results indicated that there were five mutations in MT012098.1 SARS-CoV-2 human IND 29 2020 and two mutations in MT066156.1 SARS-CoV-2 human ITA INMI1 2020 (Table 2).

The correlation rate among 11 taxa was analyzed to see the evolutionary rates among the phylogenetic tree of the genomes by Corrtest analysis. The tree topology was analyzed using the ML method, and the Corrtest analysis score was 1.0000. Therefore, the null hypothesis was rejected (p value was < 0.001), showing evolutionary rates, and the model is significant. This shows a high correlation among the ML tree and genetic distances in genomes of all taxa.

The MSA of S and N-protein with homology modeling

The multiple sequence alignment of all 11 sequences of spike proteins and the receptor-binding domain (RBD) of the nucleocapsid phosphoprotein was performed on the T-Coffee server, and the Clustal W server and aligned FASTA file were obtained. This file was further formatted in the BoxShade program to obtain the results (Fig. 3a). The MSA of Spike-protein from these 8 sequences shows that four sequences had mutations in reference to the Spike-protein sequence from Wuhan. Each FASTA sequence was further separated and modeled into the Robetta modeler program with the crystal structure of Spike glycoprotein as a template (PDB: 6vyb). This template was used for all 11 SARS-CoV-2 sequences due to its high identity of 99.50%. The models generated were further validated, and structural assessment was performed. The parameters studied were GMQE (Global Model Quality Estimation) score, Qmean score, Molprobity score, and Ramachandran plot analysis. The results of structural validation are provided in Table S3. The GMQE is a quality estimation for the template alignment with the target sequence. All the 11 models generated have a Qmean score between − 1.82 and − 2.06 and thus represents a high-quality model for the spike protein. MolProbity score is based on the analysis of steric problems within the molecular framework of the model; it also checks for H-bonds and Van der Wall's contacts within the generated models. The spike protein models were scored from 1.26 to 1.40, representing high quality and structural stability of the models.

Ramachandran plots were also analyzed, and all models have low outlier percent and a low clash score. The models were visualized and superimposed for structural comparison in MOE. We found five different mutations from the sequences on comparison and superimposition. Out of five mutations in this group, one mutation is observed in the RBD of the spike protein (Fig. 3a, b), and the rest four were in other regions of Spike protein (Figure S4). The SARS-CoV-2 nucleocapsid protein RNA binding domains for all the 11 sequences were modeled, and the models were validated (Table S4). Model validation and assessment reveals high similarity but does not show any mutations in the binding domain (Fig. 4a). The receptor-binding domain of the nucleocapsid protein was only modeled due to a lack of complete protein crystal structure; it shows a good homology (Fig. 4b) with the reference crystal structure (PDB: 6M3M) in structural features and validation results.

Discussion

We collected COVID-19 data of 8 countries; India, Pakistan, Bangladesh, Afghanistan, Nepal, Bhutan, Sri Lanka, and the Maldives. The sequences from India, Pakistan, Nepal, Sri Lanka, and the Maldives were only available at the time of this study. The evolution of death count was calculated using XLSTAT-2020 software since death 1 in all six countries, which indicated that the number of deaths has been increasing day by day in all SAARC countries.

Since the relative analysis of genomic sequences and molecular evolutionary data is essential for evaluating and reconstructing the evolutionary relationships among genomes (Li et al. 2020), we analyzed the different models with data. We found that the HKY model was the best model

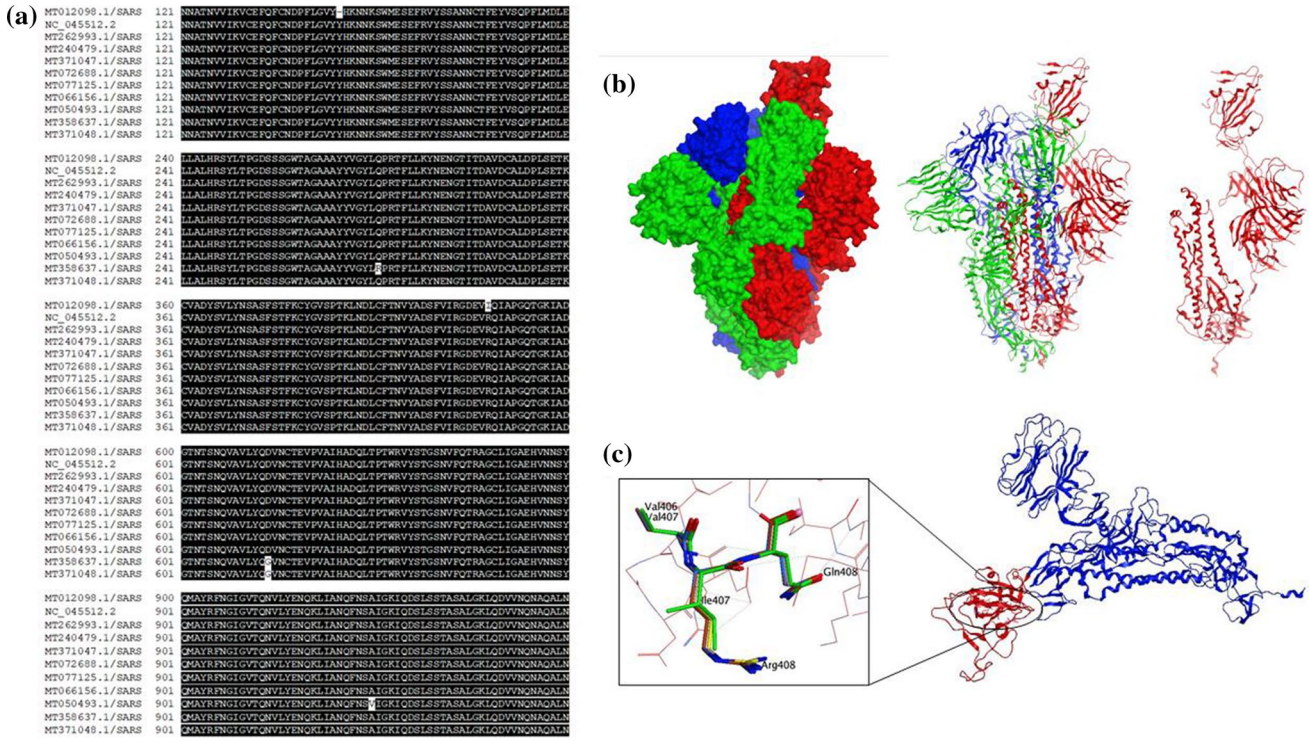

Fig. 3 **a** Multiple Sequence Alignment of 11 SARS-CoV-2 spike protein sequences. **b** The trimeric SARS-CoV-2 spike protein with surface and ribbon representation and one single spike protein with receptor binding domain (RBD) towards the top and tail region at bottom. **c** The highlighted box show homology models of 11 sequences in superposed representation, Ile407 mutation in green colour and the surrounding residues superposed on rest of the modelled spike protein along with the SARS-CoV-2 spike protein crystal structure (PDB: 6vyb)

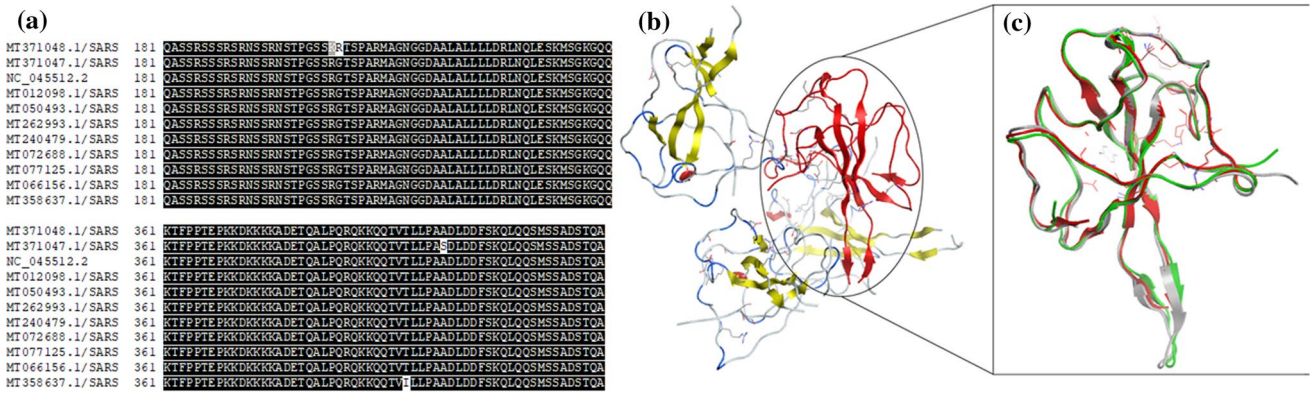

Fig. 4 **a** Multiple sequence alignment of 11 SARS-CoV-2 nucleocapsid protein sequences. **b** The trimeric SARS-CoV-2 nucleocapsid protein receptor binding domain (RBD) in ribbon form. **c** Superposed homology model for 11 SARS-CoV-2 nucleocapsid protein receptor binding domain (RBD) showing high structural similarity

with the lowest BIC scores. This model confirms that SARS-CoV-2 arises at diverse rates, and transitions and transversions happened at different degrees in the genome during transmission of COVID-19 among individuals. Our results indicate that the SARS-CoV-2 genome sequences are highly similar to each other, consistent with earlier records (Ceraolo and Giorgi 2020; Giovanetti et al. 2020). Divergence times in tree topology were at a 95% confidence interval, which supported the evolutionary rate of SARS-CoV-2 substitutions site per year.

ML BIC and evolutionary history indicate that there were mutations among the genomes, but they were highly conserved and similar to Wuhan SARS-CoV-2, which is similar to other reports (Ceraolo and Giorgi 2020; Li

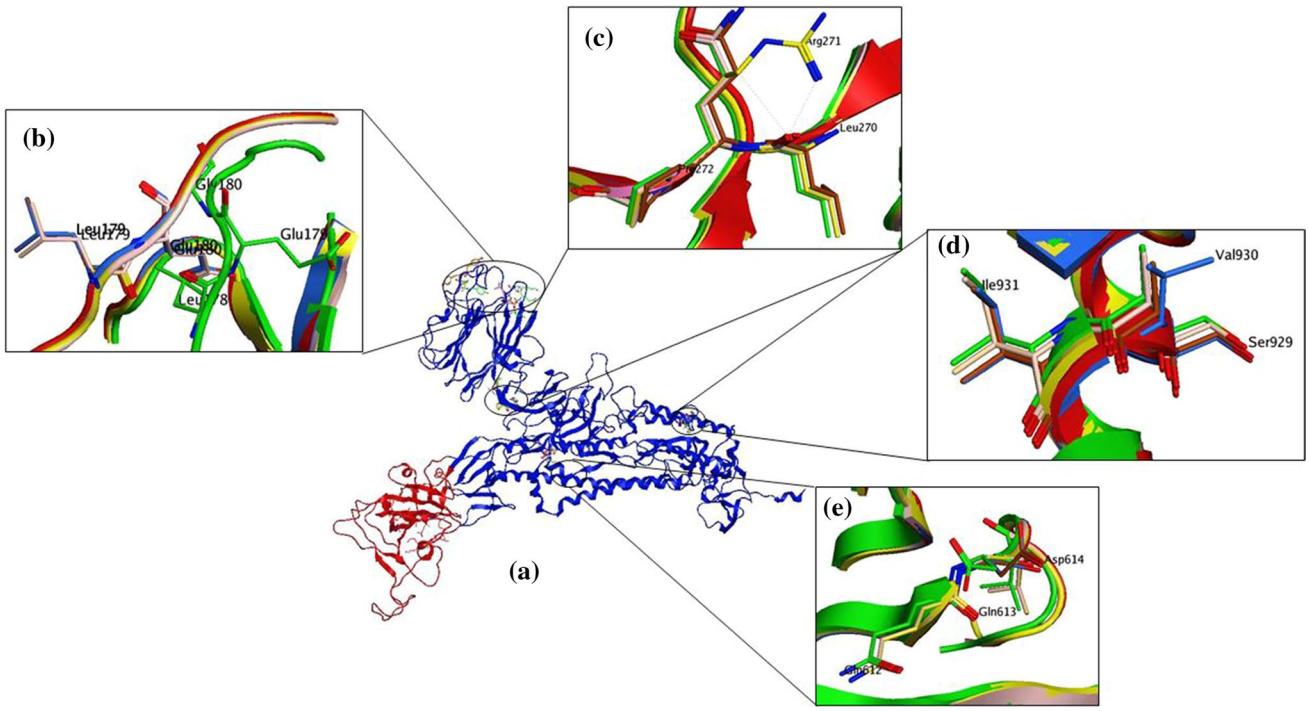

Fig. 5 **a** The SARS-CoV-2 spike protein with receptor binding domain as red coloured region. **b** mutations in the spike proteins SARS-CoV-2/human/IND/29/2020(Gap Val143-Tyr144), **c** SARS-CoV-2/human/IND/GBRC1/2020/S-Protein (Arg271), **d** SARS-CoV-2/human/IND/166/2020 (Val930), **e** SARS-CoV-2/human/LKA/COV53/2020 and SARS-CoV-2/human/IND/GBRC1/2020/S-Protein (Glu614)

et al. 2020). Our results highlighted that there were heterogeneity and evolutionary differences with rates among the genome, which was supported by time tree analysis, Corrtest analysis, and Tajima's relative rate test. The high correlation between ML tree and genetic distances in the genomes was evident that evolutionary rates and molecular models are significant. However, further molecular genetics investigation is needed with these genomes to justify the mutation or polymorphisms.

The multiple differences in the ML estimation of gamma parameter for site rates among the genomes indicate that there is variation in the rate of evolution and transmission. ML evolutionary results indicated that SARS-CoV-2 is spread at different time intervals from Wuhan as the genomes show the highest similarity. Further, our results also confirmed that there is a very low level of variation among all 11 selected genomes, which is supported by the MSA and phylogenetic tree analysis. The Nucleocapsid protein for these sequences showed three mutations compared to the reference genome (NC_045512.2 Wuhan-Hu-1). One of the mutations was found in the MR358637.1 sequence from India, and the other two mutations were found in the MT371047.1, MT371048.1 from Sri Lanka. We found that these data are inconsistent with earlier reports (Ceraolo and Giorgi 2020; Li et al. 2020).

The Spike (S) protein models for sequences had shown mutations as compared to the Wuhan SARS-CoV-2 model. The SARS-CoV-2/human/IND/29/2020 model showing two mutations; out of these two mutations, one is present in the receptor-binding domain of the Spike-protein. RBD is the site that interacts and binds to the ACE-II receptor of the host cell and initiates the viral entry; it extends from residues 331–524 (Fig. 3b, c). The Arg407 has mutated to Ile407 (Fig. 3c) in the RBD domain in the SARS-CoV-2/human/IND/29/2020 sequence when compared to other strains from this region (Yadav et al. 2020). Other mutations were found in the SARS-CoV-2/human/IND/29/2020 (Gap Val143-Tyr144), SARS-CoV-2/human/IND/GBRC1/2020/S-Protein (Arg271), SARS-CoV-2/human/IND/166/2020 (Val930), SARS-CoV-2/human/LKA/COV53/2020 and SARS-CoV-2/human/IND/GBRC1/2020/S-Protein (Glu614) concerning the Wuhan sequence (Fig. 5a–e).

The N -protein is essential for the viral RNA genome packaging, regulation of viral RNA synthesis during replication/transcription, and modulation of infected cell metabolism. The N-terminal RNA-binding domain is responsible for RNA binding, and hence it is vital to target for drug discovery. It also plays an important role in viral self-assembly, and interaction with the multiple host cell proteins such as Smad3, pyruvate kinase, B23 phosphoprotein, and chemokine Ccl16. N protein

is involved in binding to the nucleic acid at many places by the coupled-allostery method [Kang et al., 2020]. The modeled RNA-binding domain does not show any substitutions (Fig. 4b, c). C-Terminal domain plays a significant role in the protein–protein interface, and in our study, we found this C-terminal dimerization domain and intrinsically disordered central Ser/Arg linker region does show few substitutions. The Substitutions observed compared to the Wuhan sequences are as follows; MT371048.1/SARS-CoV-2/human/LKA/COV53/2020 (Lys203, Arg204), MT371047.1/SARS-CoV-2/human/LKA/COV38/2020 (Ser398), MT358637.1/SARS-CoV-2/human/IND/GBRC1/2020 (Ile393) (Fig. 4a). Since N protein is very stable, therefore, it is an excellent diagnostic and therapeutic target for SARS-CoV-2 antiviral therapy. Our observation of this study allows researchers to explore this region for the future exploitation of any possible antiviral therapy.

Conclusion

The phylogenetic and evolutionary models successfully linked the human transmission of the SARS-CoV-2 in the South Asian regions through international travel between various neighboring countries as most of these share land borders. These mutations could also adversely affect the current efforts of drug repurposing because the binding mode of drugs may change in these cases. These observations are important as the mutations discovered in the South Asian region pose a great challenge to the development of a vaccine and or drug that would target all the strains of SARS-CoV-2. High population density and the high volume of national and international travel for business and tourism make this region vulnerable. These findings would also help devise strategies for governments and the public to take all necessary measures to control this pandemic.

Acknowledgements The administration of Assam University, Silchar is acknowledged for providing all necessary facilities.

Author contributions APM and RVC: manuscript writing, data analysis, and compilation; PP: conceptualized the work, suggestions, comments, and final submission.

Funding None.

Compliance with ethical standards

Conflict of interest Authors declare no conflict of interest.

References

Angira D, Chikhale R, Mehta KK, Bryce RA, Thiruvenkatam V (2019) Tracing the GSAP–APP C-99 interaction site in the β-amyloid pathway leading to Alzheimer's Disease. ACS Chem. Neuro 10:3868–3879. https://doi.org/10.1021/acschemneuro.9b00332

Ceraolo C, Giorgi FM (2020) Genomic variance of the 2019-nCoV coronavirus. J Med Virol 92:522–528. https://doi.org/10.1002/jmv.25700

Chan JF, Lau SK, To KK, Cheng VC, Woo PC, Yuen KY (2015) Middle East respiratory syndrome coronavirus: another zoonotic betacoronavirus causing SARS-like disease. Clin Microbiol Rev 28(2):465–522. https://doi.org/10.1128/CMR.00102-14

Cheng VC, Lau SK, Woo PC, Yuen KY (2007) Severe acute respiratory syndrome coronavirus as an agent of emerging and re-emerging infection. Clin Microbiol Rev 20(4):660–694. https://doi.org/10.1128/CMR.00023-07

COVID-19 dashboard by Center for Systems Science and Engineering (2020) https://gisanddata.maps.arcgis.com/apps/opsdashboard/index.html#/bda7594740fd40299423467b48e9ecf6

Di Paolo T, Sebastien M, Ioannis X, Miquel O, Alberto M, Chang J-M, Taly J-F, Notredame C (2011) T-Coffee: a web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension. Nucleic Acids Res 39:W13–W17. https://doi.org/10.1093/nar/gkr245

Drosten C, Günther S, Preiser W et al (2003) Identification of a novel coronavirus in patients with severe acute respiratory syndrome. N Engl J Med 348(20):1967–1976. https://doi.org/10.1056/NEJMoa030747

Du L, He Y, Zhou Y, Liu S, Zheng BJ, Jiang S (2009) The spike protein of SARS-CoV–a target for vaccine and therapeutic development. Nat Rev Microbiol. 7(3):226–236. https://doi.org/10.1038/nrmicro2090

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797. https://doi.org/10.1093/nar/gkh340

ExPASy (2019) Bioinformatics resource portal. BoxShade. https://embnet.vital-it.ch/software/BOX_doc.html

Giovanetti M, Benvenuto D, Angeletti S, Ciccozzi M (2020) The first two cases of 2019-nCoV in Italy: where they come from? J Med Virol. 92:518–521. https://doi.org/10.1002/jmv.25699

Gralinski LE, Baric RS (2015) Molecular pathology of emerging coronavirus infections. J Pathol. 235(2):185–195. https://doi.org/10.1002/path.4454

Guarner J (2020) Three emerging coronaviruses in two decades: the story of SARS, MERS, and now COVID-19. Am J Clin Pathol 153:420–421. https://doi.org/10.1093/AJCP/AQAA029

Kang S, Yang M, Hong Z, Zhang L, Huang Z, Chen X, He S, Zhou Z, Zhou Z, Chen Q, Yan Y, Zhang C, Shan H, Chen S (2020) Crystal structure of SARS-CoV-2 nucleocapsid protein RNA binding domain reveals potential unique drug targeting sites. Acta Pharmaceut Sin B. https://doi.org/10.1016/j.apsb.2020.04.009

Katoh K, Rozewicki J, Yamada KD (2019) MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. Brief Bioinform 20:1160–1166. https://doi.org/10.1093/bib/bbx108

Kerzare D, Chikhale R, Bansode R, Amnerkar N, Karodia N, Paradkar A, Khedekar P (2016) Design, synthesis, pharmacological evaluation and molecular docking studies of substituted oxadiazolyl-2-oxoindolinylidene propane hydrazide derivatives. J Braz Chem Soc 27:1998–2010. https://doi.org/10.5935/0103-5053.20160090

Kumar S, Stecher G, Li M, Knyaz C, Tamura K (2018) MEGA X: molecular evolutionary genetics analysis across computing platforms. Mol Biol Evol 35:1547–1549. https://doi.org/10.1093/molbev/msy096

Li F (2016) Structure, function, and evolution of coronavirus spike Proteins. Annu Rev Virol 3(1):237–261. https://www.annualreviews.org/doi/10.1146/annurev-virology-110615-042301

Li X, Zai J, Zhao Q et al (2020) Evolutionary history, potential intermediate animal host, and cross-species analyses of SARS-CoV-2. J Med Virol. https://doi.org/10.1002/jmv.25731

Liu J, Zheng X, Tong Q et al (2020) Overlapping and discrete aspects of the pathology and pathogenesis of the emerging human pathogenic coronaviruses SARS-CoV, MERS-CoV, and 2019-nCoV. J Med Virol. 92(5):491–494. https://doi.org/10.1002/jmv.25709

Nei M, Kumar S (2000) Molecular evolution and phylogenetics. Oxford University Press, New York

Rzhetsky A, Nei M (1992) A simple method for estimating and testing minimum evolution trees. Mol Biol Evol 9:945–967

Sayers EW, Cavanaugh M, Clark K, Ostell J, Pruitt KD, Karsch-Mizrachi I (2019) GenBank. Nucleic Acids Res. 47(D1):D94–D99. https://doi.org/10.1093/nar/gky989

Tajima F (1993) Simple methods for testing molecular clock hypothesis. Genetics 135:599–607

Tamura K, Nei M (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol 10:512–526. https://doi.org/10.1093/oxfordjournals.molbev.a040023

Tao Q, Tamura K, Battistuzzi F, Kumar S (2018) Corrtest: a new method for detecting correlation of evolutionary rates in a phylogenetic tree. bioRxiv. https://doi.org/10.1101/346635

Walls AC, Park YJ, Tortorici A, Wall A, McGuire AT, Veesler D (2020) Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. Cell 180:281–292. https://doi.org/10.1016/j.cell.2020.02.058

Wang M, Jiang A, Gong L, Luo L, Guo W, Li C, Zheng J, Li C, Yang B, Zeng J (2020) Temperature significant change COVID-19 transmission in 429 cities. medRxiv. https://doi.org/10.1101/2020.02.22.20025791

Wu F, Zhao S, Yu B, Chen Y, Wang W, Song Z, Hu Y, Tao Z, Tian J, Pei Y, Yuan M, Zhang Y, Dai F, Liu Y, Wang Q, Zheng J, Xu L, Holmes EC, Zhang Y (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

Yadav PD, Potdar VA, Choudhary ML, Nyayanit DA, Agrawal M, Jadhav SM, Majumdar TD, Shete-Aich A, Basu A, Abraham P, Cherian SS (2020) Full-genome sequences of the first two SARS-CoV-2 viruses from India. Indian J Med Res. https://doi.org/10.4103/ijmr.IJMR_663_20

Zhou P, Yang X-L, Wang X-G et al (2020) Discovery of a novel coronavirus associated with the recent pneumonia outbreak in humans and its potential bat origin. bioRxiv. https://doi.org/10.1101/2020.01.22.914952

Zumla A, Chan JF, Azhar EI, Hui DS, Yuen KY (2016) Coronaviruses drug discovery and therapeutic options. Nat Rev Drug Discov. 15(5):327–347. https://doi.org/10.1038/nrd.2015.37