Future VIROLOGY

Short Communication

Phylogenomics and phylodynamics of SARS-CoV-2 genomes retrieved from India

Sameera Farah[1,2], Ashwin Atkulwar[1,3], Manas Ranjan Praharaj[4], Raja Khan[5], Ravikumar Gandham[4] & Mumtaz Baig*,[1,2]

[1]Department of Zoology, Laboratory of Molecular & Conservation Genetics, Govt. Vidarbha Institute of Science & Humanities, VMV, Road, Amravati 444604, India
[2]Department of Integrative Biology, University of Guelph, Guelph, 50 Stone Road, East, N1G 2W1, ON, Canada
[3]Department of Zoology, Amolakchand Mahavidyalaya, Godhani Road, Yavatmal 445001, India
[4]National Institutes of Animal Biotechnology, Gowlidoddy, Hyderabad, Telangana 500032, India
[5]Division of Veterinary Biotechnology, ICAR-Indian Veterinary Research Institute (IVRI), Izatnagar, Bareilly, Uttar Pradesh, India
*Author for correspondence: mumtaz.baig@uoguelph.ca

Background: This is the first phylodynamic study attempted on SARS-CoV-2 genomes from India to infer the current state of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) evolution using phylogenetic network and growth trends. **Materials & Methods:** Out of 286 retrieved whole genomes from India, 138 haplotypes were used to build a phylogenetic network. The birth–death serial model (BDSIR) package of BEAST2 was used to calculate the reproduction number of SARS-CoV-2. Population dynamics were investigated using the stamp date method as implemented in BEAST2 and BEAST 1.10.4. **Results:** A median-joining network revealed two ancestral clusters. A high basic reproduction number of SARS-CoV-2 was found. An exponential rise in the effective population size of Indian isolates was detected. **Conclusion:** The phylogenetic network reveals dual ancestry and possibility of community transmission of SARS-CoV-2 in India.

First draft submitted: 17 July 2020; Accepted for publication: 17 November 2020; Published online: 30 November 2020

Keywords: India • phylodynamics • SARS-nCoV-2

Throughout human history, viral outbreaks have been an omnipotent threat to global human health from the spread of the Spanish flu epidemic in 1918 to the recent corona virus (CoV) pandemic in December 2019 [1–3]. CoVs are members of family *Coronaviridae* with a characteristic crown-like structure and have a positive-sense ssRNA genome of 26–32 kb, while SARS-CoV-2 has a genome of 29,844–29,891 nt [4–6]. The prevalence of CoVs has been reported previously in livestock and wild species [7–10]. Their rapidly evolving nature makes them capable of adapting and infecting a range of hosts with broad tissue tropism [11,12]. The first corona viral outbreak – porcine epidemic diarrhea – was reported in pigs in Europe and Asia in 1971 [8]. However, the zoonotic shift of CoVs in humans began with the worldwide outbreak of SARS-CoV-1 in China during 2002–2003 and subsequently with the Middle East respiratory syndrome in 2012 [13–15]. In the current ongoing epidemic, the first case of the novel CoV, caused by the new strain of CoVs, in other words, SARS-CoV-2, was reported from the Wuhan city of China on 31 December 2019 and spread worldwide by January 2020 [16,17]. The WHO declared this outbreak as a global pandemic on 11 March 2020. The novel CoV strain is characterized primarily by its potential to infect the human respiratory tract, causing severe pneumonia in the affected person. Whole-genome sequencing and phylogenetic analysis suggests that bats might be the original hosts [5]. Several studies, however, point toward the *S gene* of SARS-CoV-2 exhibiting high similarity of functional domains with isolates from the pangolin [18–20]. Advances in computational genomics and the availability of genomic data enable us to understand the spatial distribution and epidemiology of pandemics. The whole-genome sequencing and annotation of a novel SARS-CoV-2 reference genome opens up rapid sequencing and assembly of many global SARS-CoV-2 genomes worldwide. Pathogen genomics is an effective tool not only in monitoring the ongoing pandemic but also in effective vaccine development. Genomic data contributed by various laboratories across the world provides a platform to track down

Future Medicine

10.2217/fvl-2020-0243 © 2020 Future Medicine Ltd ISSN 1746-0794

the origin and events of community transmission of the virus in their countries [21]. Genomic studies have already proved crucial in contact tracing of infection and would become more important in the second wave of infection after the release of lockdown in the ongoing covid19 pandemic [22]. In this study, we analyzed 121 whole genomes of SARS-CoV-2 from India to infer the phylogeography of SARS-CoV-2 within India, infection genomics to estimate the reproduction number (R0) and infection rate and the past and present evolutionary trajectory of SARS-CoV-2.

Materials & methods

Whole-genome sequence information on SARS-CoV-2 isolates from India, available in the GISAID database until 4 May 2020, was retrieved. Out of 286 SARS-CoV-2 genomes, samples with missing information and gaps were discarded to arrive at a total of 219 Indian SARS-CoV-2 genomes. These genomes were aligned using Clustal Omega [23] and the number of haplotypes was determined using DnaSP v6 [24]. Based on the emergence of 138 haplotypes, the resulting dataset was further used for downstream analyses. In evolutionary studies, phylogenetic networking is becoming a method of choice for reconstructing evolutionary pathways in many species. The Median-Joining Network (MJN) is one such algorithm developed to reconstruct the unambiguous evolutionary history of species [25]. An MJN was constructed with 138 genomes covering all major states of India (Supplementary Table 1). The birth–death serial model (BDSIR), as implemented in the BEAST2 package, was used to estimate the effective R0 of SARS-CoV-2 in India [26, 27]. In BEAST2, an HKY nucleotide substitution model with a gamma category count of 4, a relaxed lognormal clock with a clock rate of 8.3E-5 subs/site/month corresponding to 1 × 10^{-3} subs/site/year [25] were applied. In Markov Chain Monte Carlo (MCMC) analysis, parameters were sampled every 1000 generations over a total of 10 million generations. The basic R0 and BecomeUinfectiousRate were set with the mean distribution to 18.00, assuming a mean recovery time between 18 and 20 days. BEAST v1.10.4, with similar settings to those of the HKY nucleotide substitution model with a gamma category count of 4, a relaxed lognormal clock, and a clock rate of 8.3E-5 subs/site/month corresponding to 1 × 10^{-3} subs/site/year [28] was utilized to reconstruct the evolutionary dynamics of SARS-CoV-2. The tree prior was set to the coalescent exponential growth to calculate the growth rates of the virus in India. The effective sample sizes and 95% highest posterior density intervals for parameters like the basic R0, BecomeUinfectiousRate, growth rate and the demographic reconstruction of growth rates with exponential growth priors were inspected using Tracer v1.7.0. The trees file was summarized in TreeAnnotator by setting the burnin percentage to ten and the target tree type to maximum clade creditability tree, while the node heights were set to mean heights. The time-scaled Maximum Clade Credibility tree based on MCMC analysis of the 138 SARS-CoV-2 genomes was visualized in the FigTree v1.4.4 tree viewer. The Bayesian Skyline Plot (BSP) method as implemented in BEAST v2.2.0 was used to estimate the effective population size (Ne) for these 138 Indian isolates (Supplementary Table 2). The stamped-date method with HKY nucleotide substitution as a model coupled with 4 gamma category and coalescent Bayesian skyline tree priors were set for the analysis. The MCMC chain length of 10 million steps was applied, the first 10% were discarded as burn-in, and a strict clock rate of 8.33E-5 subs/per/site/per month was used. The log file and tree log file were analyzed to draw the BSP in Tracer v1.7.0.

Results

The MJN showed the occurrence of ancestral clusters alongside their newly mutated daughter clusters and haplotypes. The network was defined by two main clusters marked as ‘A’ and ‘B’ (Figure 1). Both A and B clusters showed linkage with the Wuhan, China outbreak haplotype (EPI_ISL_406798) of 26 December 2019. Cluster A, which was dominant in Gujarat, differed by a single median vector from the Wuhan haplotype as compared with cluster B, dominant in south India, which differed by two median vectors. In a biological network, median vectors are interpreted either as unsampled or extinct individuals. Based on this relationship, cluster A was considered as the ancestral node. Genomes of unknown origin contributed by National Institute of Virology, Pune showed greater affinity with SARS-CoV-2 genomes from Wuhan and Ladakh and were linked to ancestral cluster A. However, genome isolates from West Bengal illustrated affinities with both clusters. Of note, haplotypes from the worst hit Maharashtra state, including Mumbai, exhibited closer relatedness to cluster B. Further, the SARS-CoV-2 haplotypes from Delhi were more widespread in distribution, while those from Madhya Pradesh displayed closer relatedness to ancestral cluster A. Many daughter haplotypes accumulated 1–4 mutations that were derived from clusters A and B. In one instance in the network, a daughter cluster derived from cluster B illustrated sharing of haplotypes from Delhi, Telangana and Assam and can potentially be considered as a third minor subcluster. This subcluster further showed the emergence of newer haplotypes by accumulating mutations in the range of 2–3

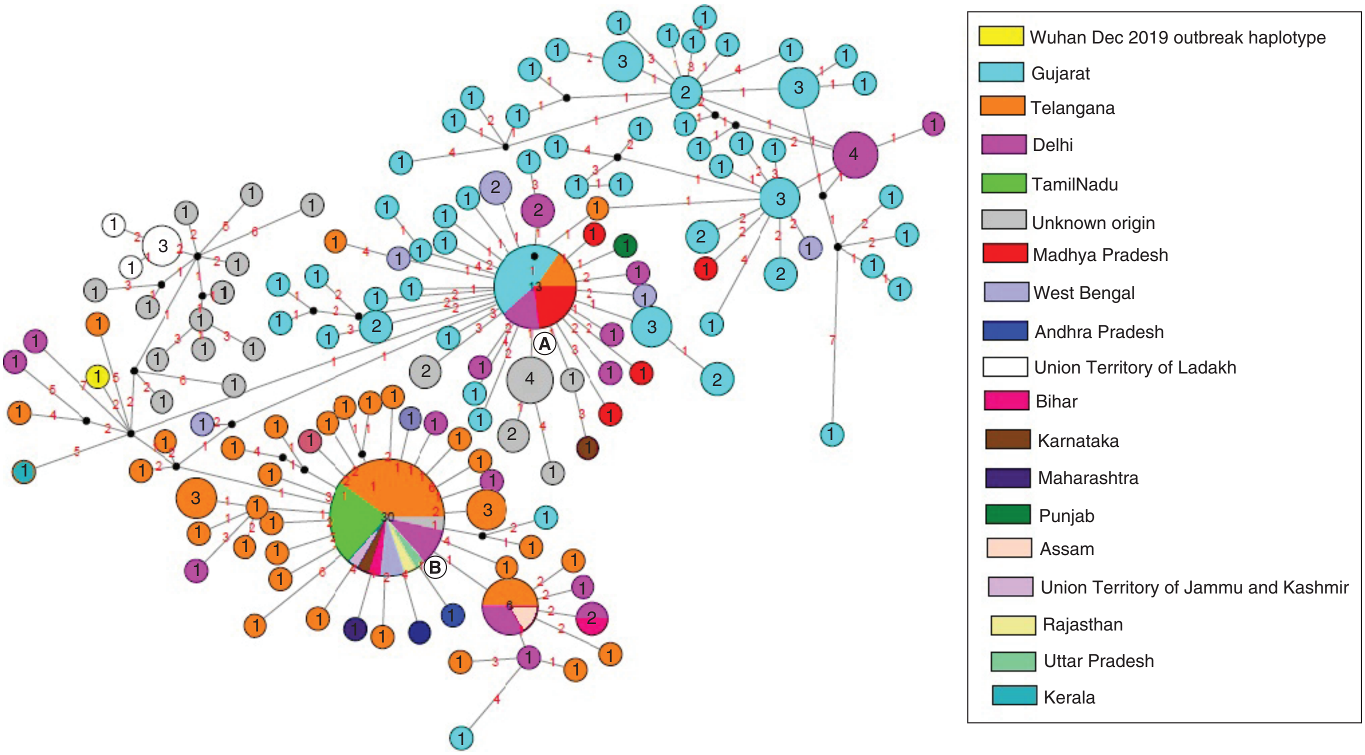

Figure 1. Median-joining network constructed out of 138 Indian haplotypes shows the occurrence of ancestral clusters alongside their newly mutated daughter clusters and haplotypes. The median joining network inferred two main clusters marked as **'A'** and **'B'** showing linkage with the Wuhan, China outbreak haplotype. Circled areas are proportional to the number of shared haplotypes and numbers inside the circle illustrate the number of haplotypes; each number in red on the links represents the position number of the mutated nucleotide. The sequence range under consideration is 56–29,797, with nucleotide position numbering according to the Wuhan, China EPI_ISL_406798 reference sequence. For MJN, Network5011CS (www.fluxusengineering.com/) was used with the parameter epsilon set to zero.
MJN: Median-joining network; np: Nucleotide position.

with the sole representative from Gujarat. Based on this median joining network, autochthonous and community transmission of the virus in India cannot be ruled out. Similarly, the Time-Scaled Maximum Clade Credibility tree based on MCMC analysis with an Exponential Growth tree prior reveals the basal position of the haplotype from Kerala (Figure 2). The tree shows the splitting of haplotypes into two major clades, i.e., Telangana and Gujarat, whereas the samples contributed by the National Institute of Virology, Pune, Maharashtra displayed monophyletic status with the isolates from Telangana (Figure 2). The effective R0 of SARS-CoV-2 in India was found to be 3.683 (95% highest posterior density [HPD] interval: 2.411, 5.401). The estimated mean infectious rate, which is the actual time needed for recovery after the infection, was estimated as 7.44 (95% HPD interval: 4.51, 9.99), which roughly corresponds to approximately 40–50 days. The current population dynamics of SARS-CoV-2 in India from the last week of January (27 January 2020) to the first week of May (4 May 2020) was plotted using a BSP. The population dynamics of Indian isolates exhibited a sigmoidal type of distribution, with exponential growth of the Ne starting from the last week of January to the second week of February (Figure 3A). The pandemic peaked in India around the second week of February and plateaued between the last week of February and 4 May 2020. However, growth rate curve reconstruction using an exponential growth tree prior depicted a continuous increase in the effective population from the last week of January–4 May 2020 (Figure 3B).

Discussion

Our phylodynamic analysis confirms the high effective R0 and deaths recorded in India during this period. Most probably, the occurrence of a plateau phase in the BSP after the last week of February resulted from the nationwide lockdown and social distancing measures (Figure 3A). Likewise, some findings also suggest that exposure to high temperatures also contributed to the lowering of activity and the lifespan of SARS-CoV-2 [29–31]. In another recently

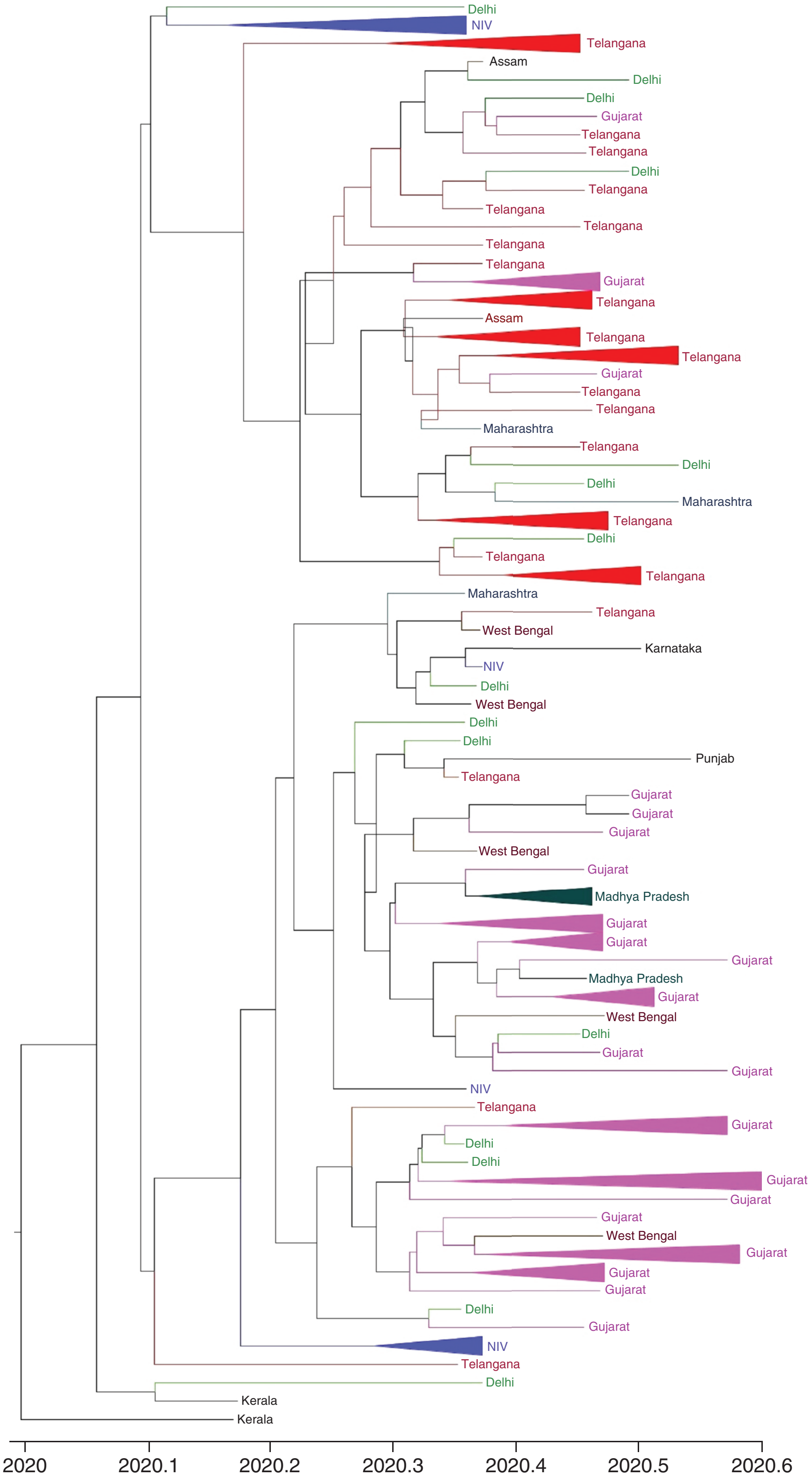

Figure 2. The Maximum clade credibility tree based on Markov Chain Monte Carlo analysis of the 138 Indian SARS-CoV-2 genomes with an exponential growth tree prior.

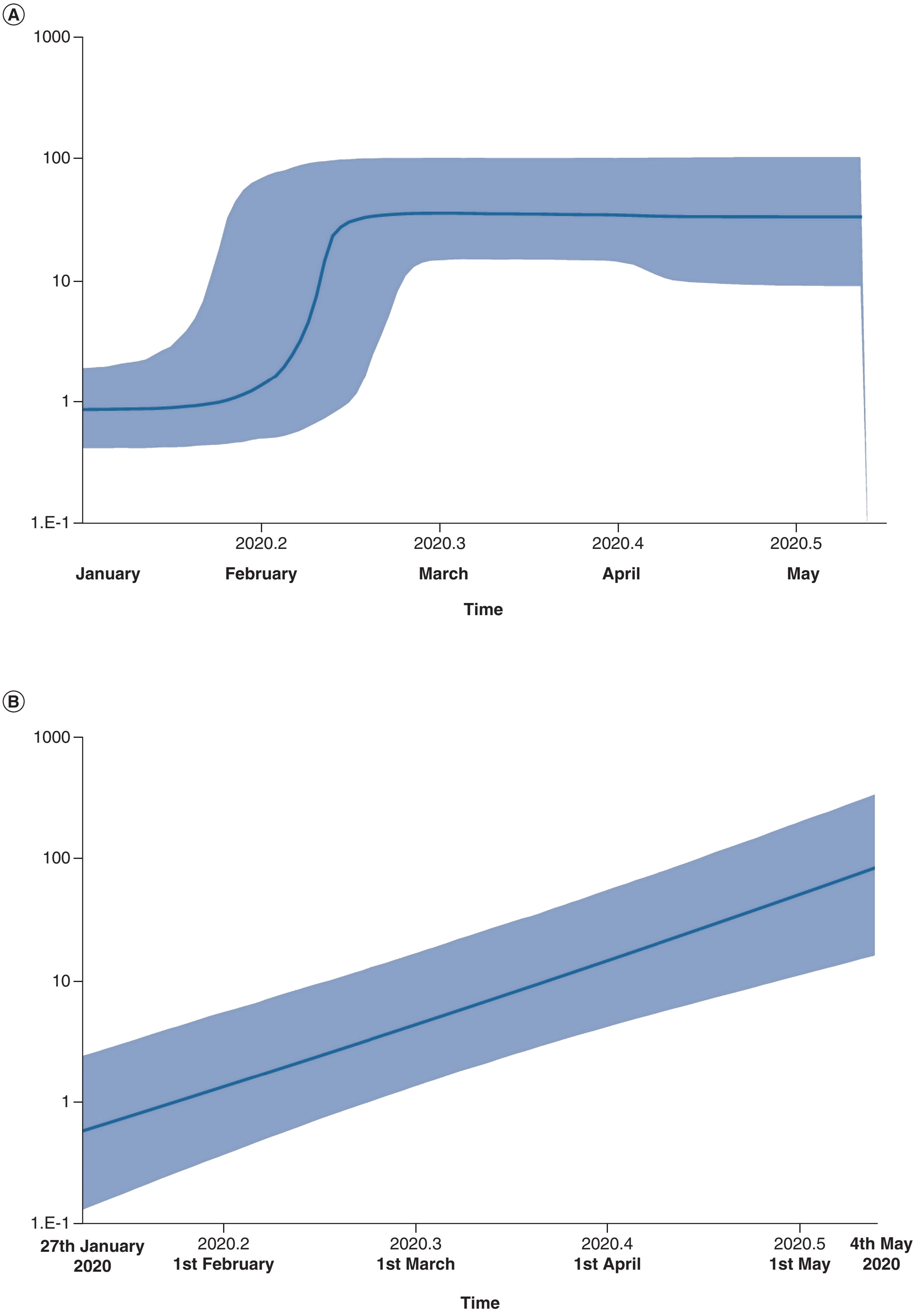

Figure 3. Population dynamics of SARS-CoV-2 in India. Bayesian skyline plots **(A & B)** constructed using 138 Indian isolates by both coalescence Bayesian skyline and exponential growth tree priors depicts the effective population size on the Y-axis, while the X-axis denotes the timeline in months.
Ne: Effective population size.

published study, physical distancing measures taken in Wuhan, China beginning in April played a significant role in lowering the reproductive number of virus [32]. Exponential growth was also confirmed in the population dynamics, which was congruent with the continuous rise of SARS-CoV-2 infections in India.

Conclusion

To conclude, our study provides baseline genome-based phylodynamic information, highlighting genetic affinities between viral isolates sequenced from the major states of India. In the coming days, sequencing and analyses of greater numbers of SARS-CoV-2 genomes from India would help in dealing with the second wave of community transmission after relaxation of the lockdown. At the same time, genomic information produced through such studies can also be utilized to fill the gaps created due to unrealistic assumptions, lack of contact tracing, sampling errors and limited diagnostic testing.

Summary points

- The research presented in this study cast light on the phylogenomics and phylodynamics of SARS-CoV-2 genomes retrieved from India.
- A total of 286 SARS-CoV-2 whole genomes deposited from 26 December 2019 to 4 May 2020, representing all major regions of India, were analyzed.
- Out of 286 retrieved whole genomes, a total of 138 haplotypes were identified and used to build a phylogenetic network using the birth-death serial model (BDSIR) package of BEAST2. The reproduction number (R0) was also calculated using the same dataset.
- The population dynamics were also investigated using the stamp date method of constant coalescence as well as exponential growth models as implemented in BEAST2 and BEAST 1.10.4.
- Our median-joining network analysis confirms the dual ancestry of viral haplotypes in India. The phylodynamic analysis validates the high basic R0 of SARS-CoV-2 and an exponential rise in the effective population size of Indian isolates.
- The high basic R0 and the exponential rise in the effective population size of Indian isolates predicted by the Bayesian model highlight the possibility of community transmission of SARS-CoV-2 in India.
- Following relaxation of the lockdown in India and considering the possibility of a second wave of disease transmission, maintaining strict track of infection using genome analysis would aid in the proper management of SARS-CoV-2.

Supplementary data

To view the supplementary data that accompany this paper please visit the journal website at: www.futuremedicine.com/doi/suppl/10.2217/fvl-2020-0243

Author contributions

S Farah and A Atkulwar retrieved and analyzed the data, MR Praharaj and R Khan assisted in data analysis, R Gandham and M Baig wrote the manuscript, study was conceptualized by M Baig.

Acknowledgments

We gratefully acknowledge all authors and submitting laboratories of the sequences from GISAID. S Farah and M Baig are thankful to Prof. Elizabeth Boulding and Compute Canada for providing access to high computation facilities at www.cedar.computecanada.ca

Financial & competing interests disclosure

The authors have no relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties.

No writing assistance was utilized in the production of this manuscript.

References

1. Patterson KD, Pyle GF. The geography and mortality of the 1918 influenza pandemic. *Bull. Hist. Med.* 65(1), 4–21 (1991).
2. Spreeuwenberg P, Kroneman M, Paget J. Reassessing the Global Mortality Burden of the 1918 Influenza Pandemic. *Am. J. Epidemiol.* 187(12), 2561–2567 (2018).
3. Biondi-Zoccai G, Landoni G, Carnevale R, Cavarretta E, Sciarretta S, Frati G. SARS-CoV-2 and COVID-19: facing the pandemic together as citizens and cardiovascular practitioners. *Minerva Cardioangiol.* 68(2), 61–64 (2020).

4. Woo PC, Huang Y, Lau SK, Yuen KY. Corona virus genomics and bioinformatics analysis. *Viruses* 2(8), 1804–1820 (2010).

5. Lu R, Zhao X, Li J, Niu P *et al.* Genomic characterisation and epidemiology of 2019 novel corona virus: implications for virus origins and receptor binding. *Lancet* 395(10224), P565–574 (2020).

6. Chan JF, Kok KH, Zhu Z *et al.* Genomic characterization of the 2019 novel human-pathogenic corona virus isolated from a patient with atypical pneumonia after visiting Wuhan. *Emerg. Microbes Infect.* 9(1), 221–236 (2020).

7. Lam TT, Shum MH, Zhu H *et al.* Identifying SARS-CoV-2 related corona viruses in Malayan pangolins. *Nature* 583, 282–285 (2020).

8. Pensaert M, Callebaut P, Vergote J. Isolation of a porcine respiratory, non-enteric corona virus related to transmissible gastroenteritis. *Vet. Q* 8, 257–261 (1986).

9. Li W, Shi Z, Yu M *et al.* Bats are natural reservoirs of SARS-like corona viruses. *Science* 310(5748), 676–679 (2005).

10. Hu B, Ge X, Wang L *et al.* Bat origin of human corona viruses. *Virol. J.* 12, 221 (2015).

11. Graham RL, Baric RS. Recombination, reservoirs, and the modular spike: mechanisms of corona virus cross-species transmission. *J. Virol.* 84(7), 3134–3146 (2010).

12. Li F. Structure, function, and evolution of corona virus spike proteins. *Annu. Rev. Virol.* 3(1), 237–261 (2016).

13. Hilgenfeld R, Peiris M. From SARS to MERS: 10 years of research on highly pathogenic human coronaviruses. *Antiviral Res.* 100, 286–295 (2013).

14. Gao H, Yao H, Yang S, Li L. From SARS to MERS: evidence and speculation. *Front. Med.* 10, 377–382 (2016).

15. Coleman CM, Frieman MB. Emergence of the Middle East respiratory syndrome coronavirus. *PLoS Pathog.* 9, e1003595, (2013).

16. Zhou P, Yang X, Wang X *et al.* A pneumonia outbreak associated with a new corona virus of probable bat origin. *Nature* 579, 270–273 (2020).

17. Wu F, Zhao S, Yu B *et al.* A new coronavirus associated with human respiratory disease in China. *Nature* 579, 265–269 (2020).

18 . Wong MC, Cregeen SJJ, Ajami NJ, Petrosino JF. Evidence of recombination in coronaviruses implicating pangolin origins of nCoV-2019. *bioRxiv* doi:10.1101/2020.02.07.939207 (2020) (Epub ahead of print).

19 . Xiao K, Zhai J, Feng Y *et al.* Isolation and characterization of 2019-nCoV-like coronavirus from Malayan pangolins. *bioRxiv* doi:10.1101/2020.02.17.951335 (2020) (Epub ahead of print).

20 . Lam TT-Y, Shum MH-H, Zhu H-C *et al.* Identification of 2019-nCoV related coronaviruses in Malayan pangolins in southern China. *bioRxiv* doi:10.1101/2020.02.13.945485 (2020) (Epub ahead of print).

21. Zhang YZ, Holmes EC. A genomic perspective on the origin and emergence of SARS-CoV-2. *Cell* 181(2), 223–227 (2020).

22. Stevens EL, Timme R, Brown EW *et al.* The public health impact of a publicly available, environmental database of microbial genomes. *Front. Microbiol.* 8, 808 (2017).

23. Sievers F, Wilm A, Dineen D *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* 7, 539 (2011).

24. Rozas J, Ferrer-Mata A, Sánchez-DelBarrio C *et al.* DnaSP 6: DNA sequence polymorphism analysis of large datasets. *Mol. Biol. Evol.* 34, 3299–3302 (2017).

25. Bandelt H-J, Forster P, Röhl A. Median-joining networks for inferring intraspecific phylogenies. *Mol. Biol. Evol.* 16, 37–48 (1999).

26. Kühnert D, Stadler T, Vaughan TG, Drummond AJ. Simultaneous reconstruction of evolutionary history and epidemiological dynamics from viral sequences with the birth–death SIR model. *J. R. Soc. Interface* 11(94), 20131106 (2014).

27. Stadler T, Kühnert D, Bonhoeffer S, Drummond AJ. Birth–death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). *Proc. Natl Acad. Sci. USA* 110(1), 228–233 (2013).

28 . Li R, Pei S, Chen B *et al.* Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (COVID-19). *medRxiv* doi:10.1101/2020.02.14.20023127 (2020) (Epub ahead of print).

29. Kucharski AJ, Russell TW, Diamond C *et al.* Early dynamics of transmission and control COVID-19: a mathematical modelling study. *Lancet Infect. Dis.* 30(20), 144–4 (2020).

30 . Wang M, Jiang A, Gong L *et al.* Temperature significant change COVID-19 Transmission in 429 cities. *medRxiv* doi:10.1101/2020.02.22.20025791 (2020) (Epub ahead of print).

31 . Pawar S, Stanam A, Chaudhari M, Rayudu D. Effects of temperature on COVID-19 transmission. *medRxiv* doi:10.1101/2020.03.29.20044461 (2020) (Epub ahead of print).

32. Prem K, Liu Y, Russell TW, Kucharski AJ, Eggo RM, Davies N. The effect of control strategies to reduce social mixing on outcomes of the COVID-19 epidemic in Wuhan, China: a modelling study. *Lancet Public Health* 5, e261–270 (2020).