Genomics of Indian SARS-CoV-2: Implications in genetic diversity, possible origin and spread of virus

Mainak Mondal[#], Ankita Lawarde[#] and Kumaravel Somasundaram*

Department of Microbiology and Cell Biology, Indian Institute of Science, Bengaluru 560 012, India

World Health Organization (WHO) declared COVID-19 as a pandemic disease on 11 March 2020. Comparison of genome sequences from diverse locations allows us to identify the genetic diversity among viruses which would help in ascertaining viral virulence, disease pathogenicity, origin and spread of the SARS-CoV-2 between countries. The aim of this study is to determine the genetic diversity among Indian SARS-CoV-2 isolates. Initial examination of the phylogenetic data of SARS-CoV-2 genomes (*n* = 3123) from different continents deposited at GISAID (Global Initiative on Sharing All Influenza Data) revealed multiple origin for Indian isolates. An in-depth analysis of 558 viral genomes derived from samples representing countries from USA, Europe, China, East Asia, South Asia, Oceania, Middle East regions and India revealed that most Indian samples are divided into two clusters. A1 sub-cluster showed more similarity to Oceania and Kuwait samples, while A2 sub-cluster grouped with South Asian samples. In contrast, cluster B grouped with countries from Europe, Middle East and South Asia. Viral clade analysis of Indian samples revealed a high occurrence of G clade (D614G in spike protein; 37%), which is a European clade, followed by I clade (V378I in ORF1ab; 12%), which is an Oceania clade with samples having Iran connections. While A1 cluster is enriched with I clade, the cluster B is enriched with G clade type. Thus our study identifies that the Indian SARS-CoV-2 viruses are enriched with G and I clades in addition to 50% samples with unknown genetic variations. The potential origin to be countries mainly from Europe, Middle East Oceania and South Asia regions, which strongly imply the spread of virus through most travelled countries. The study also emphasizes the importance of pathogen genomics through phylogenetic analysis to discover viral genetic diversity and understand the viral transmission dynamics with eventual grasp on viral virulence and disease pathogenesis.

Keywords: COVID-19, genetic diversity, pandemic, SAR-CoV-2, severe acute respiratory syndrome.

A novel corona virus (SARS-CoV-2) causes acute respiratory disease (Coronavirus disease 2019; COVID-19), which was initially found in China but now it is spread all over the world[1]. The total number of COVID-19 cases diagnosed so far exceeds 4.1 million worldwide as on 11 May 2020 with the number almost reaching 68,000 in India[2,3]. SARS-CoV-2 is an enveloped, non-segmented positive-sense RNA virus with a large genome of approximately 30 kb in length[1].

A total of 17,878 viral isolates have been sequenced and deposited online as on 11 May 2020 (ref. 4). Genome sequence analysis of viral genome between countries would help to understand the origin and also the severity of the disease process itself. The sequence information for 173 Indian viral isolates is available in the Global Initiative on Sharing All Influenza Data (GISAID) database[4–6]. In this study, we carried out systematic analysis of genome sequences of Indian SARS-CoV-2 isolates and inferred the possible source of origin and important genetic variants of Indian viruses.

Samples and methods

Sample collection

We have collected 173 genome sequences from Indian clinical samples deposited at GISAID as on 8 May 2020 (ref. 4). In addition, we also collected another 421 representative genome sequences of samples from USA (75), Europe (80), China (75), East Asia (64), South Asia (41), Oceania (75) and Middle East (11). Among Indian viral isolates, seven viral genome sequences that belong to passaged virus through cell lines (Vero CCL81 isolate P1) and one sequence with incomplete genome were excluded in this study. The viral sequences having complete and high coverage (*n* = 137) were alone selected for phylogeny analysis. However, for viral clade analysis, an additional set of 28 Indian viral genomes with low coverage was also used. Genome accession and sample data information can be found in 'SupplymentaryData.xlsx'.

Phylogenetic tree analysis

A total of 558 complete genomes were taken for alignment using MAFFT version 7.402 at CIPRES Science

*For correspondence. (e-mail: skumar1@iisc.ac.in)
[#]Contributed equally.

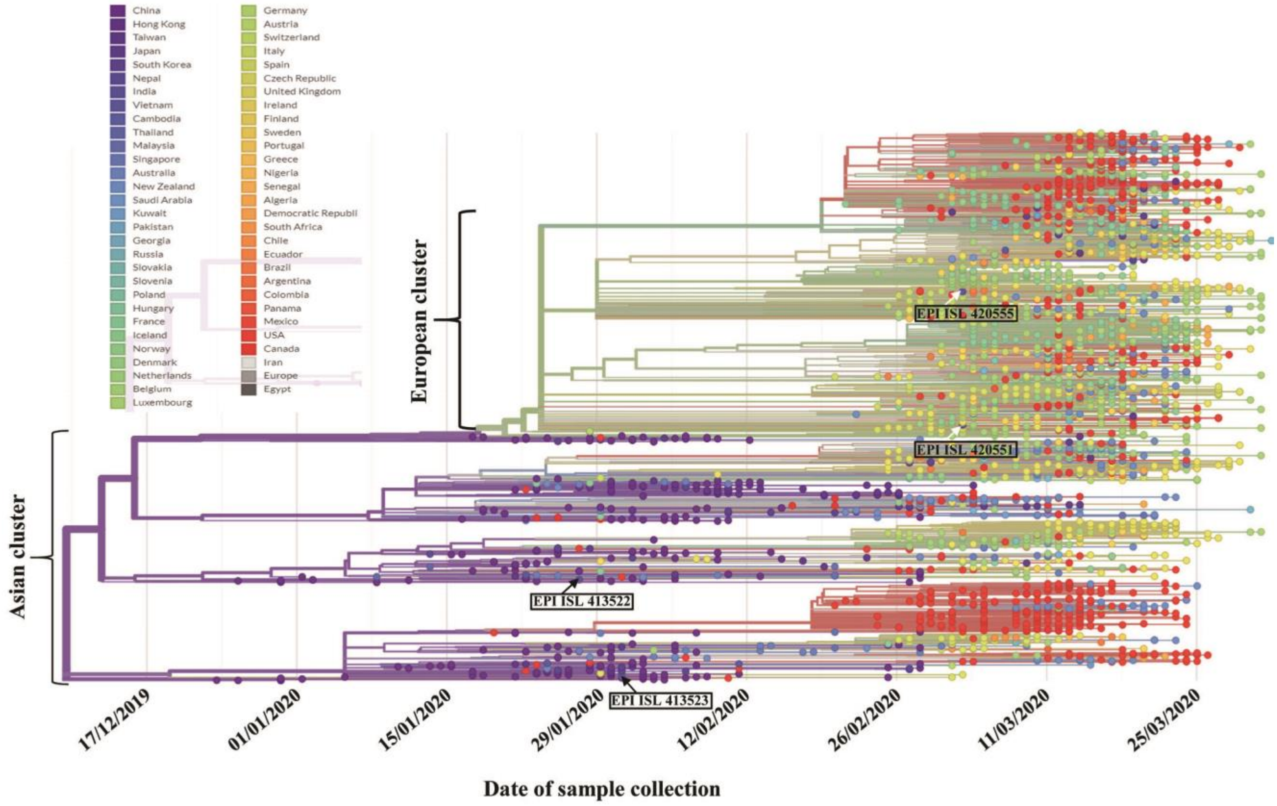

Figure 1. A modified view of phylogenetic analysis (rectangular view) of genome sequences of SARAS-CoV-2 ($n = 3123$) taken from https://www.gisaid.org/epiflu-applications/next-hcov-19-app/. The list of countries from where samples were used is given with their colour code.

Gateway[7]. Phylogenetic analysis by maximum likelihood (ML) method was carried out using IQ tree version 1.6.12 (ref. 8). TIM + F + R2 having lowest BIC score (109220.216) was selected as best substitution model out of 279 substitution model fitted. Analysis was carried out with 10^3 ultrafast bootstrap replicates. The tree file obtained was visualized using Figtree version 1.4.4 (ref. 9).

Viral clade analysis

A reference Wuhan isolate and all the other viral genome sequences ($n = 586$) were obtained and used for this analysis. Individual genes namely *ORF1ab*, *S*, *ORF3* and *ORF8* were extracted from the whole genome. The genes were aligned using CLUSTAL Omega algorithm[10] and translated to amino acid sequences. The aligned protein coding genes was visualized in BioEdit version 7.2.5 (ref. 11).

Results and discussion

To identify the origin of Indian isolates of SAR-CoV-2 virus, we examined the phylogenetic data from GISAID[4]. The phylogenetic data from 3123 samples, which included 4 Indian isolates, available at GISAID website were analysed (Figure 1; Supplementary Figure 1). It is of our interest to note that there are two major clusters – Asian cluster is represented by purple and related colours, while the European cluster is represented by greenish yellow. The Indian samples, represented by the arrows (black and white) clustered with both Asian and European clusters. While the Indian samples with black arrows were isolated during January 2020, the other two samples with white arrows were isolated during March 2020 (more details later).

Further to precisely map the origin of Indian SARS-CoV-2 isolates, we carried out an independent phylogenetic analysis using a selected set of samples representing most regions and countries where the COVID-19 infection rate is high. The samples which were collected earliest during this pandemic in each of the countries were only considered. The set consisted of 558 samples as detailed earlier. The analysis shows interesting features about the possible source of origin of Indian SARS-CoV-2 samples (Figure 2; Supplementary Figure 2). In particular, the Indian samples are located away from China/East Asian samples and are divided between two clusters, A and B. While the sub-cluster A1 consists of mostly Oceania/Kuwait samples besides a large number of Indian samples ($n = 19$), the A2 sub-cluster is enriched with South Asian samples along with Indian samples ($n = 53$). The cluster B consists mostly of European and few numbers of Middle East/South Asian samples besides a large number of

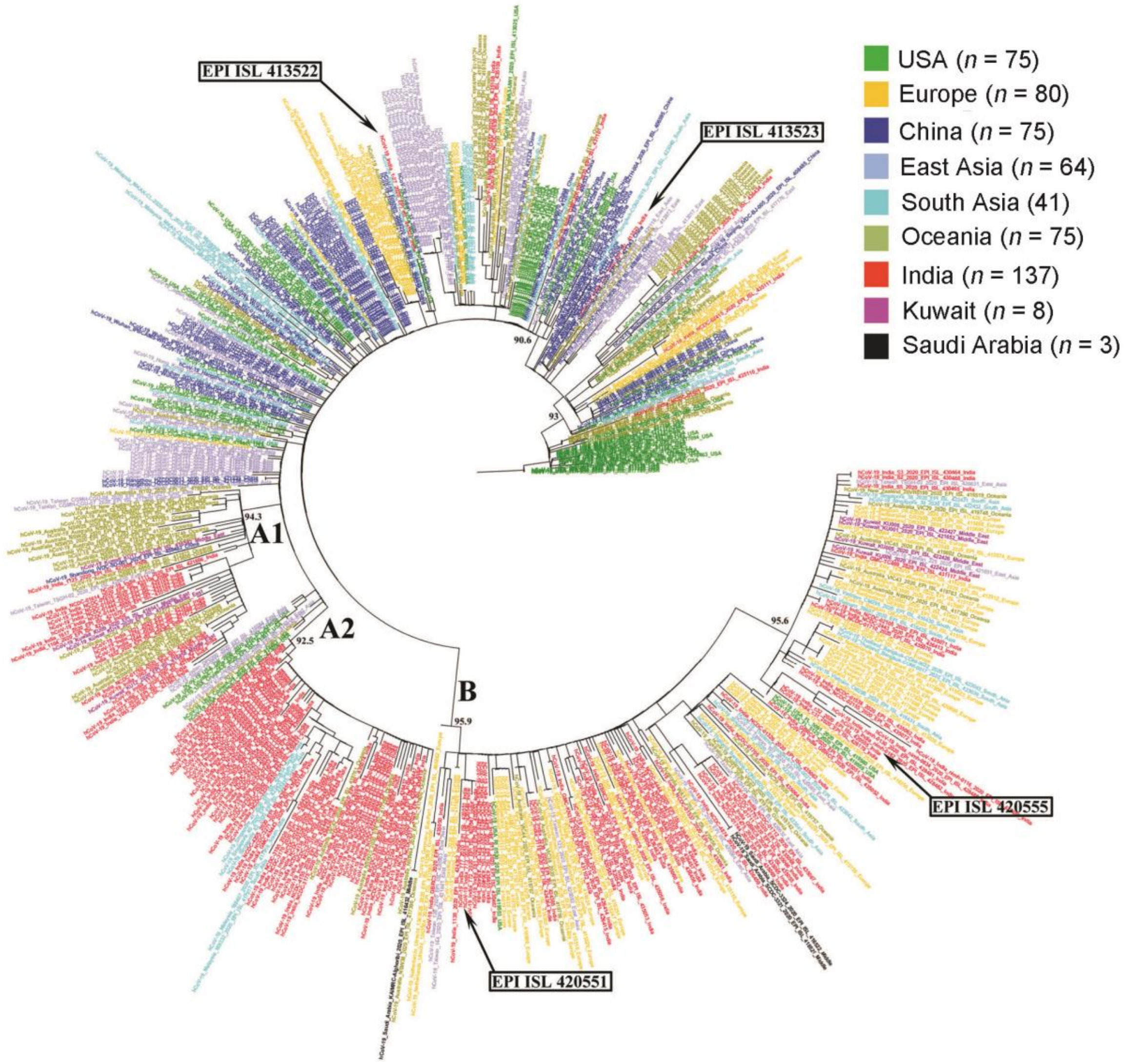

Figure 2. Maximum likelihood (ML) phylogenetic tree was constructed using whole genome sequences obtained from 558 individual SARS-Cov-2 viral isolates. The branch was tipped according to strains with respective country. The nodes represent bootstrap values. The taxa were coloured according to different countries. The colour code for different regions/countries is given.

Indian samples (*n* = 57). The analysis revealed that most Indian SARS-CoV-2 viruses (129 out of 137) show more similarity to that of specific countries. In cluster A, Indian samples show more similarity to the viruses found in Oceania, Kuwait and South Asian samples, while in the cluster B, Indian samples show more similarity to mainly European and few numbers of Middle East/South Asian samples. These results indicate that majority of Indian SARS-CoV-2 viruses have originated from Europe, Middle East, South Asia and Oceania regions. The remaining Indian isolates (*n* = 8) have grouped with other clusters which contained most samples from China and East Asia. This indicates that these viruses might have been introduced by Indian travellers from China and its neighbouring countries as they show close resemblance to ancestral Chinese virus. Indeed, a recent study reported that two viruses (EPI_ISL_413522 and EPI_ISL_413523) from this group were isolated from patients who travelled from Wuhan, China[5]. A similar correlation for the remaining six samples could not be made as their travel information is not available.

Given the fact that the date of sample collection probably coincides with the time of disease occurrence, the collection dates of different Indian samples provide some hint at the origin and spread of virus. The first two Indian viral isolates (EPI_ISL_413522 and EPI_ISL_413523) that were collected during January 2020 from patients who travelled from Wuhan, China showed more similarity with China/East Asia viral isolates. This conclusion is well supported by the fact that the initial outbreak of SARS-CoV-2 virus in Wuhan happened during December 2019 (ref. 12). A large majority of Indian viral isolates (*n* = 129), which were collected during March/April 2020, show more similarity with samples from Europe, Middle East, South Asia and Oceania regions. The delay in the occurrence of majority of Indian COVID-19 cases probably indicates the time taken for the virus to spread from China to other countries from where the Indian travellers would have contracted the virus.

It is interesting to note that two samples (EPI_ISL_420551 and EPI_ISL_420555), that are seen with European cluster in Figure 1, grouped with the cluster B, are enriched with European and few numbers of Middle East/South Asian samples according to Figure 2. Similarly, other two samples (EPI-ISL_413522 and EPI_ISL_413523) (collected from patients who had travel history from Wuhan, China) that grouped with samples from China were also identified to be associated with the major Asian cluster in Figure 1. Thus, the results of our independent phylogenetic analysis (according to Figure 2) match with that of analysis done by GISAID.

According to specific variations in different viral proteins compared to Chinese ancestral SARS-CoV-2 virus, GISAID identified three clades of SARS-CoV-2 namely, G, V and S clades[4,13]. G clade is characterized by D614G (A23403G) in S protein and largely encompasses sequences from Europe. V clade is characterized by G251V (G26144T) in ORF3 and mostly includes Asian and European sequences. S clade is characterized by the presence of L84S (C8782T) in ORF8 and mostly comprises sequences from North America. Recently, a new clade of SARS-CoV-2 carrying V378I (G1397A) in ORF1ab has been linked to travellers returning from Iran to Australia[14]. We then studied the genomes of Indian SARS-CoV-2 viruses to find out the association between Indian samples and different clades. The sample set ($n = 558$) which was used for phylogenetic analysis and an additional set of Indian samples with low coverage ($n = 28$) was subjected to clade analysis which revealed several interesting facts (Supplementary Figures 3 and 4). The samples from China failed to classify with any of the clades except a significant proportion of S clade which signifies the ancestral nature of Chinese viruses. While most groups had significant proportion of unclassified samples, East Asia and South Asia samples were found to split among G, V and S clades. Oceania samples were represented in all clades with a significant high proportion of I and S clades. In contrast to these groups, European samples showed rather a very high proportion of G clade type and USA samples showed a high proportion of S clade type. While 50% of Indian samples do not belong to any of the clades, a significant enrichment in G clade (35.15%) and I clade (12.12%) is seen (Figure 2; Supplementary Figures 3 and 4). The unclassified Indian samples do not appear to be more similar to Chinese ancestral viruses as they are grouped in A and B clusters (according to Figure 2) and found separated from ancestral samples. It is interesting to note that all I clade Indian samples are part of A1 cluster (according to Figure 2) and all G clade Indian samples are part of B cluster (according to Figure 2) (Supplementary Figure 2). Further studies are needed to identify the specific genetic variation(s) unique to this unclassified Indian samples. We conclude from this analysis that there is a higher occurrence of G clade (35.15%) and I clade (12.12%) among Indian samples. The type of unique variations specific to the remaining 50% of Indian samples is yet to be identified.

We have also analysed the type of viral clades in different states of India (Supplementary Table 1; Supplementary Figure 5). This analysis was limited by fewer numbers of viral sequences available for many states. Considering mainly those states where more number of viral sequences are available, G clade is prevalent in many states in particular Delhi, Gujarat, Karnataka, Madhya Pradesh and West Bengal. While a large percentage of I clade samples could not be tied to one state, it is interesting to note that Kargil and Ladakh samples are enriched with I clade.

Our finding that Indian SARS-CoV-2 isolates belong to specific clades may have important consequences with respect to virus transmission rate and virulence; extent of the disease severity and various other aspects of disease pathogenesis. It has been reported that viruses belonging to different clades may differ in their virulence[15]. For example, the G clade viruses carry glycine (G) corresponding to the codon 614 of S protein instead of aspartic acid (D) in other clades. Phylogenetic analysis identified that D614G mutation is originated from ancestral D residue seen in the reference Wuhan virus[16]. This residue is located very close to glycosylation region of the viral spike protein encoded by S gene[17]. It has been proposed that mutations in and around glycosylation region may alter viral spike protein structure and hence the membrane fusion process resulting in varied pathogenicity and transmissibility. Further, the difference in the death rate of COVID-19 patients of East Coast versus West Coast of USA is implicated to their difference in their G clade status[15]. G clade virus has also been identified to be highly transmissible by utilizing multiple mechanisms over its ancestral virus[18,19]. Several other studies also reported that mutations in spike protein of other Corona viruses alter the virulence[20–22].

The presence of multiple clades of SARS-CoV-2 strains in a population may also have serious implications in the accuracy of diagnostic tests that are being employed worldwide. It is not clear whether the diagnostic tests based on detection of antibody or quantifying the viral RNA genome that are in use distinguish these variations. Hence, it is important to develop diagnostic kits based on the type(s) of clades prevalent in an area. Indeed, it is reported that the serology-based rapid tests are ineffective in detecting COVID-19 positive cases in India and elsewhere in the world. The variation created by the presence of different viral clades in a population also needs to be considered seriously in developing vaccines.

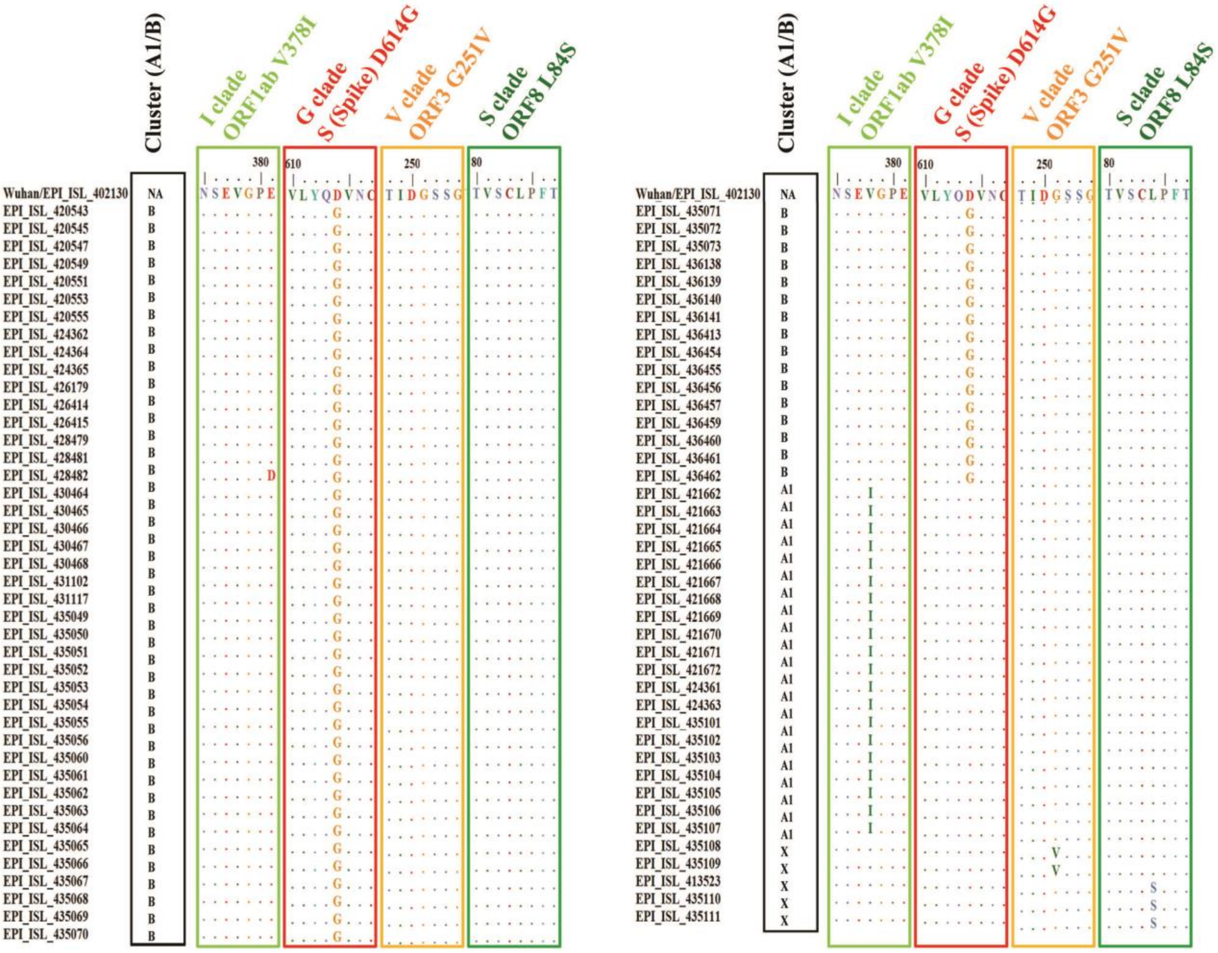

Figure 3. Characterization of clade defining genetic markers of 165viral sequences from the Indian isolates included in this study. Cluster (A1/B) shows the Indian samples belonging to the clusters identified by phylogeny analysis. I clade contains V378I marker in the ORF1ab region; G clade contains D614G marker in the spike protein (S); V clade contains G251V in ORF3 and S clade contains L84S marker in ORF8 region. NA: not applicable; X stands for the samples that do not belong to either A1 or B. Of the analysed Indian samples, 82 of them failed to classify with any of the four clades. The data for the remaining 83 samples is shown.

In particular, mutations in the receptor binding domain (RBD) of S protein are likely to create structural changes thus creating an escape mechanism from antibody recognition[23]. While the structural alterations created by D614G in S protein is not completely understood, it is proposed that the location of this change in the RBD makes it unlikely to affect critical epitopes to be used for vaccine development[23]. It has been also proposed that relaxation models of social distancing should consider the presence of one or more types of viral clades[15].

While this manuscript was under review, a higher occurrence of A2a (45.7%) and A3 (37.1%) types of SARS-CoV-2 in India was reported[24]. Upon comparison, it appears that A2a and A3 types are the same as G and I clades reported in this study. While our study with much higher number of viral genomes also reports the high occurrence of G clade (35.15%) followed by I clade (12.12%), a true picture on the proportion of different viral types in India will emerge only when more number of SARS-CoV-2 genomes are analysed. It is possible that the highly transmissible G clade type may become a dominant form occupying much large proportion in India in the next few months as it was reported in USA[18].

Conclusions

Collectively, we conclude that the G and I clades represent a significant proportion of Indian SARS-CoV-2 viruses based on this limited analysis. The probable source of origin of Indian SARS-CoV-2 viruses is countries from Europe and Oceania regions besides Middle East and South Asian regions. The possible spread of the SARS-CoV-2 virus to India through Middle East countries from Europe and Oceania regions cannot be ruled out. Indeed, both A and B clusters contain samples from Middle East countries. In addition, these samples appear to split between I and G clades (Supplementary Figure 6). In the absence of the information related to travel/contact history of Indian patients, more inference and definite conclusions on the possible source of origin could not be made at present. Thus our result also indicates that there is a close connection between source of virus and the countries that are most travelled by Indians. The study also highlights the power of rapid viral genome sequencing and public data sharing to improve the detection and management of pandemic diseases such as COVID-19. It is important to point out that most countries in America, Europe, Oceania and East Asia were quick in supporting the advanced scientific studies on the virus and disease process itself in suitable containment facilities with appropriate ethical clearance towards developing novel treatment modalities and preventive vaccines. Needless to say that major countries from emerging economies such as Brazil and India should also support experimental research on SARS-CoV-2 pathogenesis.

Certainly, our analysis has clear limitations, the most important one being that we were able to analyse only a small number of Indian SARS-CoV-2 genomes while the number of COVID-19 cases increased beyond 60,000. Further, travel history of the patients and other clinical parameters are needed to make the conclusions definite. Hence, it is required that more number of Indian isolates of SARS-CoV-2 needs to be sequenced. Nevertheless, our study highlights the need for large-scale community surveillance for SARS-CoV-2 introductions and the spread. More importantly, this work underscores the power of pathogen genomics to identify epidemiological understanding of the virus and the disease.

Author contributions. M.M. and A.L. carried out data downloading and analysis. K.S. executed the whole study and wrote the manuscript.

1. Guo, Y. R. *et al.*, The origin, transmission and clinical therapies on coronavirus disease 2019 (COVID-19) outbreak – an update on the status. *Mil. Med. Res.*, 2020, **7**(1), 11; doi:10.1186/s40779-020-00240-0.
2. https://www.arcgis.com/
3. https://www.mohfw.gov.in/
4. https://www.gisaid.org/
5. Yadav, P. D. *et al.*, Full-genome sequences of the first two SARS-CoV-2 viruses from India. *Indian J. Med. Res.*, 2020; doi: 10.4103/ijmr.IJMR_663_20 [Epub ahead of print].
6. Sardar, R., Satish, D., Birla, S. and Gupta, D., Comparative analyses of SAR-CoV2 genomes from different geographical locations and other coronavirus family genomes reveals unique features potentially consequential to host–virus interaction and pathogenesis. bioRxiv: 2020.03.21.001586; doi:https://doi.org/10.1101/2020.03.21.001586.
7. Miller, M. A., Pfeiffer, W. and Schwartz, T., Creating the CIPRES science gateway for inference of large phylogenetic trees. In 2010 Gateway Computing Environments Workshop (GCE), 2010, pp. 1–8.
8. Trifinopoulos, J., Nguyen, L. T., von Haeseler, A. and Minh, B. Q., W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis. *Nucl. Acids Res.*, 2016, **44**(W1), W232–W235; https://doi.org/10.1093/nar/gkw256.
9. http://tree.bio.ed.ac.uk/software/figtree/
10. Madeira, F. *et al.*, The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nuc. Acids Res.*, 2019, **47**(W1), W636–W641; doi:10.1093/nar/gkz268.
11. Hall, T. A., BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. *Nucl. Acids Symp. Ser.*, 1999, **41**, 95–98.
12. Zhou, P. *et al.*, A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*, 2020, **579**(7798), 270–273; doi:10.1038/s41586-020-2012-7. Epub 3 February 2020.
13. Elbe, S. and Buckland-Merrett, G., Data, disease and diplomacy: GISAID's innovative contribution to global health. *Global Chall.*, 2017, **1**(1), 33–46; doi:10.1002/gch2.1018. eCollection.
14. Eden, J. S. *et al.*, An emergent clade of SARS-CoV-2 linked to returned travellers from Iran. bioRxiv: 2020.03.15.992818; doi: https://doi.org/10.1101/2020.03.15.992818.
15. Brufsky, A., Distinct viral clades of SARS-CoV-2: implications for modeling of viral spread. *J. Med. Virol.*, 2020; doi: 10.1002/jmv.25902 [Epub ahead of print].
16. Wu, F. *et al.*, A new coronavirus associated with human respiratory disease in China. *Nature*, 2020, **579**, 265–269; https://doi.org/10.1038/s41586-020-2008-3.
17. Andersen, K. G., Rambaut, A., Lipkin, W. I., Holmes, E. C. and Garry, R. F., The proximal origin of SARS-CoV-2. *Nat. Med.*, 2020, **26**, 450–452; https://doi.org/10.1038/s41591-020-0820-9.
18. Korber, B. *et al.*, Spike mutation pipeline reveals the emergence of a more transmissible form of SARS-CoV-2. bioRxiv: 2020.05.05; https://doi.org/10.1101/2020.04.29.069054.
19. Bhattacharyya, C., Das, C., Ghosh, A., Singh, A. K., Mukherjee, S., Majumder, P. P., Basu, A. and Biswas, N. K., Global spread of SARS-CoV-2 subtype with spike protein mutation D614G is shaped by human genomic variations that regulate expression of TMPRSS2 and MX1 genes. bioRxiv: 2020.05.05; https://doi.org/10.1101/2020.05.04.075911.
20. Krueger, D. K., Kelly, S. M., Lewicki, D. N., Ruffolo, R. and Gallagher, T. M., Variations in disparate regions of the murine coronavirus spike protein impact the initiation of membrane fusion. *J. Virol.*, 2001, **75**(6), 2792–2802; doi:10.1128/JVI.75.6.2792-2802.2001.
21. Geoghegan, J. L. and Holmes, E. C., The phylogenomics of evolving virus virulence. *Nat. Rev. Genet.*, 2018, **19**, 756–769; https://doi.org/10.1038/s41576-018-0055-5.
22. Ontiveros, E., Enhanced virulence mediated by the murine coronavirus, mouse hepatitis virus strain JHM, is associated with a glycine at residue 310 of the spike glycoprotein. *J. Virol.*, 2003, **77**(19), 10260–10269; doi:10.1128/jvi.77.19.10260-10269.2003.
23. Dearlove, B. *et al.*, A SARS-CoV-2 vaccine candidate would likely match all currently circulating strains. bioRxiv: 2020.04.27; https://doi.org/10.1101/2020.04.27.064774.
24. Biswas, N. K. and Majumder, P. P., Analysis of RNA sequences of 3636 SARS-CoV-2 collected from 55 countries reveals selective sweep of one virus type. *Indian J. Med. Res.*, Special issue on COVID-19 (in press, 28 April 2020).

ACKNOWLEDGMENTS. The results reported here are in whole or part based upon data generated by Global Initiative on Sharing All Influenza Data (GISAID) (https://www.gisaid.org/). We gratefully acknowledge the authors, originating and submitting laboratories of the sequences from GISAID's EpiCoV™ Database. K.S. acknowledges DBT, DST-SERB and ICMR, Government of India for research grant. Infrastructure support by funding from DST-FIST, IISc-DBT partnership and UGC (Centre for Advanced Studies in Molecular Microbiology) to M.C.B. is acknowledged. K.S. is a J. C. Bose Fellow of the Department of Science and Technology.

Received 28 April 2020; revised accepted 22 May 2020

doi: 10.18520/cs/v118/i11/1786-1791