Phylogeographic analysis of SARS-CoV-2 variations from Indian isolates

Akshat Joshipura

Manipal Institute of Technology

S Balaji (✉ s.balaji@manipal.edu)

Manipal Institute of Technology

Research Article

Keywords: SARS-CoV-2, Indian isolates, phylogeographic analyses, mutations, amino acid frequency

Posted Date: October 9th, 2020

DOI: <https://doi.org/10.21203/rs.3.rs-90273/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Phylogeographic analysis of SARS-CoV-2 variations from Indian isolates

Akshat Joshipura and S. Balaji*

*Department of Biotechnology, Manipal Institute of Technology, Manipal Academy of
Higher Education, Manipal – 576104, Karnataka, India*

** E-mail: s.balaji@manipal.edu*

Abstract

The aggressive multiplication of coronavirus in the Indian population is probably due to the faster mutation rate. At the time of commencement of this work, India was not present in the list of Top 10 worst-affected countries. However, upon completion of this manuscript, India is ranked No. 3 and during publication of this manuscript it may even elevate to the top two positions due to the pandemic. In this study, SARS-CoV-2 isolates of Indian origin were compared with the Wuhan reference sequence. Phylogenomic, proteomic, and phylogeographic analyses were performed. The genome comparisons indicated that majority of the sequence variations are associated with protein-coding regions, especially Orf1a and spike glycoproteins, while Orf7a had consistent variations, whereas Orfs 6a, 8 and 10 had negligible variations. The terminals of the genomes compared had high sequence entropy. However, the polyadenylation signal was invariant in the analysed dataset. The codon usage frequency indicated that UGU (code for cysteine) is the most frequent codon, while the least frequent was GCG (code for alanine). The amino acid frequency showed that the most abundant was leucine (12.5%), and the least was histidine (2.45%). The phylogeographical patterns were mapped for all the representative states of India, and were supplemented with few representative countries. The unique differences in the sequence of the Kerala isolate (EPI_ISL_413522) were resolved to be sequence errors rather than mutations. Based on the phylogeographic analysis, the high probability of mutations likely to be of Indian origin is attributed to the Gujarat cluster.

1. Introduction

The ongoing pandemic: coronavirus disease (COVID-19), is a respiratory illness caused by severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2), also known as the novel coronavirus (2019-nCoV). This virus surfaced in late 2019, in the Hubei province of Wuhan city, China and is the cause for the ensuing worldwide biological disaster (F. Wu et al., 2020; Zhou et al., 2020). The SARS-CoV-2 belongs to the member of the same coronavirus family that caused the SARS outbreak in 2003 and the Middle Eastern respiratory syndrome (MERS) in 2012. This virus family belongs to an order of non-segmented, enveloped, '+' strand RNA viruses called *Nidovirales*. Coronavirus belongs to the *Coronavirinae* sub-family of *Coronaviridae*, which primarily affects birds and mammals (MacLachlan & Dubovi, 2017).

The virus derives its name due to the glycoprotein projection from its envelope, which appears like a *corona* ('crown' in Latin) and resembles the solar corona or the corona spinarum in art. (Virology: Coronaviruses, 1968; Sturman & Holmes, 1983). It was found that these spike proteins assist in the entry of the virus through the angiotensin-converting enzyme II (ACE2) receptor. The virus enters only those cells that express ACE2, as was found for SARS-CoV-1 and in the initial studies of the SARS-CoV-2 (Zhou et al., 2020). The host-dependent translation of the proteins for this virus, like other coronaviruses, follows canonical methods like ribosome frameshifting (-1 frameshift) as summarised by Maranon et al. (2020) and Rodnina et al. (2020).

The genome of Coronaviruses usually varies in its GC%, ranging from 32 to 43%. There are conserved genes that are observed in the genome (in the same order). From the 5'-start is the ORFa and ORFb, (ORFab) which codes for replicase gene. This is followed by the spike (S), envelope (E), membrane (M) and the nucleocapsid (N) at the 3' end, with a variable number of ORFs in between ORFab and N, depending on the subgroup. (Woo, Huang, Lau, & Yuen, 2010). Extensive phylogenetic analysis has revealed variations in the genome such as L (70%) and S (30%) clusters as well as A, B, and C clusters (Tang et al., 2020; Forster et al., 2020). Although there are several studies on comparison of whole-genome as well as spike proteins of SARS-CoV-2 at global level, there are relatively very few studies published on SARS-CoV-2 sequences from India.

Since the population in India is highly diverse, it was proposed to analyse the genomic variations of SARS-CoV-2 isolates sequenced from different states of India. These variations were compared with the Wuhan reference sequence (NC_045512) of March 2020. While documenting our results, there were few published papers from India emphasising on mutations of available sequences (**Anindita et al., 2020**) and variations in sequences from the state of Gujarat (**Hassan et al., 2020**). However, the mapping of conservations and variations at the strain level, inter-state differences, codon frequency, phylogeographic, phylogenomic and proteomic analyses were not performed exclusively for representative Indian states and union territories. Hence, a complete genomic and proteomic analyses of SARS-CoV-2 isolates of India along with few global representatives are performed and presented here.

2. Materials and Methods

2.1. Sequence retrieval

The SARS-CoV-2 genome sequences were retrieved from the GenBank (**Sayers et al., 2020**) and EpiCoV database of GISAID initiative (<https://www.epicov.org/>). The reference genome (NC_045512) and its corresponding protein sequences (YP_009725318.1) were retrieved from the GenBank database. The genome sequences were collected from different states of India. Partial sequences were ignored, and at least one full sequence from each state of India was collected from the database as on 21st May 2020 (**Fig. 1**). Along with this dataset a representative sequence from the neighboring countries were (available as on 3rd April 2020) also included for comparison.

<... **Figure 1** ...>

2.2. Sequence alignment

The 33 representative genomes were stored in a FASTA formatted file, to which the SARS-CoV-1 genomic sequence, as well as the Wuhan reference sequence (NC_045512), were appended. This brings the total count to 35. After a draft alignment, the sequences were trimmed to a uniform size to ensure better alignment and also to avoid internal gaps that are introduced due to length variations. Multiple sequence alignment was performed using standalone ClustalX 2.1, with default parameters (**Larkin et al., 2007**).

2.3. Shannon Entropy

Shannon entropy was analysed using DAMBE7 (**Xia, 2018**). The ‘Site-specific entropy’ option of the sequence analysis menu was used. Coding regions of the genome were isolated and analysed. Amino acid grouping method was used. A standard text file was generated from which the tables were made. Heatmaps were used to represent data in the form of a matrix with colour gradients to facilitate easier identification of smallest and largest values, which in turn forms the basis of clustering. (**Metsalu & Vilo, 2015**). These 2D matrices with colour gradients make it easier to spot errors (**Yachdav et al., 2014**).

2.4. Codon usage, frequency analysis and Heat maps

DAMBE7 (**Xia, 2018**) was used to generate data on codon usage (standard) and relative synonymous codon usage, nucleotide and amino acid frequencies. Tab-separated value data sets were created that served as input for ClustVis (<https://biit.cs.ut.ee/clustvis/>), an online tool with which heatmaps were created (**Metsalu & Vilo, 2015**). The default parameters were set which used the unit variance method of row scaling and the singular value decomposition method of imputation.

2.5. Proteome analysis

Using YP_009725318.1 as a reference, the corresponding proteins for each sequence were obtained by translating the file containing 34 sequences (using standard codon) in MEGA-X. The same has been described in the results. The sequences of 12 different proteins were aligned using ClustalX 2.1

2.6. Phylogenetic tree construction

The sequence alignment file (*.ALN) obtained from 35 sequences was utilised to construct phylogenetic trees using DAMBE7 (Xia, 2018). Unresolved bases were denoted as 'N'. A phylogenetic tree was made using the distance-based method. The Fast-ME tree was constructed using SARS-CoV-1 genome sequence (NC_004718.3) as the outgroup. Bootstrapping was enabled and set to 100.

2.7. Phylogeographical variations

The first recorded COVID-19 case in India is Kerala (EPI_ISL_413522), which originated in Wuhan, China. Hence, the genomic sequence of the first case was compared with the Wuhan reference sequence using Molecular Evolutionary Genetics Analysis (MEGA X) for Windows (Kumar et al., 2018). Moreover, the mutations in sequences of other Indian states were also compared with the reference mentioned above to determine whether any sequence-specific variations arose in the Indian Subcontinent

3. Results and Discussion

3.1. Shannon Entropy

The sequence variation of the entire genome is depicted in Fig 2. A large number of variations from the base value were observed for Orf1a and Spike glycoprotein, while Orf7a had consistent variations, whereas Orfs 6a, 8 and 10 had negligible variations. This is in accordance with the results obtained by (Hassan et al., 2020). All sequences had the hexamer polyadenylation signal 'AAUUA' (Peng et al., 2016), of which only four sequences Wuhan Reference, Hyderabad (Telangana), South 24 Parganas, and Howrah (West Bengal) (viz., NC_045512, EPI_ISL_431102, 437537, 437537) had a poly-A tail after 57 bases. The remaining sequences were terminated between 53-58 bases downstream, without a poly-A tail. There were no variations in the polyadenylation signal except ambiguous base calls for the two sequences of South 24 Parganas and Howrah (EPI_ISL_437536 and 437537).

<.... Figure 2>

The terminals were highly varying due to length differences and probably due to errors in sequencing. Moreover, there were ambiguous bases which are challenging to deal with, especially while inferring mutations throughout the sequences. The first available SARS-CoV-2 isolate from Kerala (EPI_ISL_413522) is devoid of poly-A tail. This has been misinterpreted as the "absence of a Poly (A) tail might signify a different origin of the Indian isolate when compared to the Chinese isolate, or it may denote an *indel*" (Chatterjee, 2020) as it is well known that the poly-A tail plays a significant role during infection. Truncation of the poly-A tail in Kerala isolate is probably due to sequencing errors or may be due to tail length regulation in coronavirus as the length is not static (Peng et al., 2016). However, it cannot signify a "different origin of the Indian isolate" as reported by Chatterjee (2020).

There are also length variations, especially in the poly-A tail of other isolates, which is vital for the viral translation. It has been shown *in vivo* that the transcriptome with the longer poly-A tail is significantly enhanced over the shorter ones (Wu et al., 2013). The reason behind the poly-A tail length variation is unknown, and it is missing in many isolates except (Wuhan Reference, Hyderabad, South 24 Parganas, and Howrah) This may be due to lack of sequence data. This length variation may lead to misinterpretations in understanding the regulatory mechanism of coronavirus during infection (Wu & Brian, 2010). However,

it was observed that the polyadenylation signal hexamer AAUUA is invariant in the analysed dataset (Wu et al., 2013). This is crucial not only for the stability and translation efficiency of mRNA but also for the export of mature mRNAs as demonstrated by (Preiss et al., 1998; Weiss et al., 1991; Fuke & Ohno, 2008; Wu et al., 2013; Peng et al., 2016)

3.2. Codon usage, Nucleotide and amino acid frequency

Among the 34 sequences, an average of 29877.94 bases was observed per genome. The average number of A, C, G, T for 34 genomes were 8907.05, 5507.85, 5581.91 and 9581.23, respectively. This was calculated from observed frequencies (over 34 sequences) of 302840, 187267, 199985 and 325762 corresponding to A, C, G and T with propensities of 0.2981, 0.1843, 0.1969, and 0.3207, respectively. The average calculations excluded the unresolved bases ('N'), which can be due to base call errors, poor signal quality, etc. (Balaji et al., 2017). These ambiguous bases were given in Table 1.

<... Table 1 ...>

<... Figure 3 ...>

The codon usage pattern for the entire dataset is shown in Fig. 3. The GC% was calculated to be 38.1208% while the GT% turned out to be 51.7542%. This is due to a higher frequency of Thymine bases. It is also evident from the heat map (Fig. 4a) that T/U, followed by A is most frequent. The codon usage frequency indicated that the most frequent codon was UGU (343.6) code for cysteine, while the least frequent was GCG (23.6) code for alanine. The representation of this in Fig. 4b makes it clear that the UUU, UGU and UAA come under the high-frequency bracket, being marked with dark red. There are quite a few codons that come under low-frequency bracket, indicated by dark green. A relatively leucine-rich pattern is shown in Fig. 4b, followed by threonine and serine. A high-frequency tag of red colour confirms the presence of a large number of the ambiguous bases in the South 24 Parganas sequence (EPI_ISL_437536). Relative synonymous codon usage is shown in Fig. 4c. However, this is not related to amino acid usage (Xu et al., 2008). The amino acid frequency is showed in Fig. 4d. The most abundant amino acid was leucine (12.5%), and the least was histidine (2.45%).

<... Figure 4 ...>

In all the heat maps (Fig. 4), it is observed that there are three distinct clusters. The first cluster comprises West Bengal isolates of Kolkata, Darjeeling and Tehatta along with Kerala which consistently formed a cluster. The second cluster comprises of all the other sequences analysed except Karnataka. The Karnataka isolate is a single sequence that can be considered as a separate cluster if more Karnataka isolates are included. The reason for this is elaborated under the next section (3.3).

3.3. Proteome analysis

From the analysed dataset, it was found that few sequences differ in their reading frames for protein translation (open reading frames) in comparison with others. These sequences fall into 3 main groups, as discussed in section 3.2. The protein region 'ORF1a' falls into reading frame +1 for the group containing the West Bengal sequences and Kerala. The same region appears in reading frame +3 for the lone isolate of Karnataka, and it appears in reading frame +2 for all the remaining sequences (including Wuhan reference).

This concept of -1 frameshift has been discussed previously by **Maranon et al. (2020)** and **Rodnina et al. (2020)**.

It was observed that the majority of the variations occur in nsp6 region and S protein region, and a little lesser but a significant number in N protein. A leucine zipper motif exists in the Orf1ab, Orf3a and Orf7b proteins. A cysteine-rich (5329-5354) and glutamine-rich (929-977) regions were also found in the Orf1ab. The spike glycoprotein of Kerala isolate also contained such cysteine-rich region from residues 1234-1253. In contrast, it was from 1235 to 1254 for the rest of the isolates due to codon deletion present in the Kerala sequence. A serine-rich region was observed in the nucleocapsid protein from the residues 176-206. The Orf8 and 7a contain a protein of unknown function (Entry pfam_fs: DUF308 and DUF1443, respectively). Besides, N-myristoylation site, protein kinase C phosphorylation site, N-glycosylation site and casein kinase II phosphorylation site were found abundantly scattered throughout the proteome. There are a few variable regions in the proteome depicted in **Fig. 5a-f**, and the Orfs 6a and 10 are highly conserved (**Fig. 5g&h**).

<.... Figure 5 ...>

3.4. Phylogenomic analysis

The phylogenetic tree based on the genome sequences of Indian isolates is depicted in **Fig 6a**. Two significant clusters are branching out at the intermediary node (labelled with the bootstrap value of 99). The first cluster, along with the reference sequence, is labelled 'cluster A' and the subsequent cluster as 'B'. The cluster A contains geographically diverse strains of the North, South (Andhra Pradesh, and Tamil Nadu), East (e.g. Assam) and West (e.g. Rajasthan) along with Kerala and Wuhan reference among few others. The cluster B mainly contains strains from Gujarat (West) and West Bengal (East) with two sequences of the South (Telangana and Karnataka). This pattern repeats itself (i.e. Telangana and Karnataka clusters with Gujarat and West Bengal sequences, Tamil Nadu and Andhra Pradesh appeared together based on mutation analysis (section 3.5).

<.... Figure 6 ...>

This tree data is in agreement with **Saha et al. (2020)** for some of the common sequences viz., ISL- 430464, 65, 67, and 435056 were positioned in the same cluster as in cluster B (**Fig. 6**), and devoid of the cluster containing the reference sequence (cluster A). This also emphasises that ISL-430465 (Darjeeling WB) and 430464 (Kolkata WB) are closely related to each other and more distantly related to ISL-430467(E. Medinipur WB). These clusters are within the country; such instances were known since the first case of Wuhan in December 2019.

The phylogenetic tree of Indian isolates was also appended with global SARS-CoV-2 isolates from geographically distant locations and the clustering trend is depicted in **Fig 6b**. There are three distinct clusters A1, B1, and C1. The cluster A1 mainly contains the Indian isolates of North and North East India. However, it is interesting to know that the pairs of Karnataka-Telangana and Andhra-Tamil Nadu is clustered together. The cluster B1 consists sequences primarily from East (West Bengal) and the West (Gujarat) and can be referred to as the East-West cluster. This cluster also includes Spain and Peru isolates, and it matches with the results of variation analysis. The Gujarat cluster seems to be a subcluster of this East-West cluster. The cluster C1, is the international cluster (except the Kerala isolate). Since only the Kerala isolated is clustered with the international isolates, this could be be a reference point of divergence, from which Indian strains diverge and progressively become less similar to international strains. Based on the phylogenomic clustering (**Table S1**), it is evident that the Indian isolates are quite distinct from the international strains. This is in

agreement with the global variations described with three central strains A, B and C, where A being the primary ancestral strain. The B strain is confined to Eastern Asia, whereas A and C are found in Europe and the West. Hence, people outside East Asia are immune to the primary B strain. This selection pressure forced the formation of derived B strains (Forster et al., 2020).

3.5. Phylogeographical variations

The phylogeographical patterns of variations are mapped for all the representative states analysed (**Fig. 1**). These variations were specific only to the coding regions rather than the non-coding regions. However, these variations were observed neither in the Wuhan reference genome nor in Kerala. The Indian isolates are organised into regional groups (**Table 2**) and depicted in **Fig. 1**. These patterns are similar to the clusters A and B obtained from phylogenomic analysis (**Fig. 6**). Excluding the Wuhan and the Kerala sequence, the North-type I cluster is almost identical to the genomically derived cluster A (**Fig. 6**). Also, the pattern of majority Gujarat (West) and West Bengal (East) along with two specific sequences of the South, which forms the basis for the ‘East-West’ cluster that resembles the cluster B (**Fig. 6**). The percentage variations that are geographically specific to each region is showed in **Fig. 7**. Based on high sequence identity and variations, the following union territories such as Delhi and Ladakh were grouped into Uttar Pradesh (EPI_ISL_435099) and Jammu-Kashmir, respectively. These groupings are not to be considered as an act of disputing the current borders or divisions of states and union territories. They are merely clustered for easy comparison.

<... Table 2 ...>

<... Figure 7 ...>

Based on the type of mutations (**Table 3**), three predominant clusters are as follows.

- Gujarat cluster: Point mutation (Substitution) of either C or G with T in Orf3a, Orf6 and Orf10 restricted only to the genomes of Gujarat.
- North-type I: Point mutations in nsp3, nsp6, S & N proteins, predominantly in the Northern states.
- East-West cluster: Mutations in nsp3, S protein and RNA dependent RNA polymerase genes, which were specific to the Western and Eastern states.

<... Table 3 ...>

It was observed that the single point mutations were found to be geographically unique. There were five substitutions and three deletions found exclusively in the Kerala sequence (EPI_ISL_413522), which are not observed in the Wuhan reference or any other Indian states. However, these variations (especially the deletions) are not considered to be giving rise to a unique strain, as these variations are probably due to sequencing errors (**Balaji et al., 2017**). Since those deletions are not present in any of the analysed sequences, including the reference, they are dismissed as sequencing errors, in contrast to **Chatterjee (2020)**, who documented these as mutations. However, to confirm that these deletion patterns in Kerala isolate probably sequencing errors and are not due to a unique strain of distinct geographical origin, the Kerala isolate was compared thoroughly against Wuhan reference sequence. There were two direct identical repeats of 13 bases each (GGTGTATTACC) at base positions 21665-21677 and 21986-21998, respectively. 308 bases separated these direct repeats. This sequence pattern is exclusive to the Wuhan reference sequence, whereas,

in the Kerala isolate, only one repeat is found at base positions 21652-21664. The other repeat has three missing bases 'TTA' in 'GGTGTTTA - - - CC' between 21972 and 21982 bases and the resulting pattern 'GGTGTTTACC' is with 10 bases.

The positional differences are due to truncated residues (~13 bases) in the 5'-terminus of Kerala isolate. It is evident that the missing bases are probably due to errors in sequencing or contig assembly. The missing bases were misinterpreted as either mutation or error in alignment techniques as proposed by **Chatterjee (2020)**. This may further mislead researchers if the same misinterpretation is published as it is. Moreover, there is no mention of the direct repeats. The sequence comparison of direct repeats containing tandem 'TTA' repeat clarify that the Kerala isolate probably have sequence errors that might have originated during sequencing/assembly by the interference of tandem repeats (**Torresen et al., 2019**) rather than mutations. Some of the sequence variations may be laboratory-specific, which might have come from sample preparation, contamination, base call errors, recurrent sequencing errors, the sequencing technology used and consensus calling approaches (**Balaji et al., 2017; Maio et al., 2020**).

3.6. *Origin of new strains*

There are research gaps put forth by **Saha et al. (2020)**, to determine whether there are any mutations originated in India? It is of interest to identify those strains. To address this gap, we have classified the variations with the possibility of Indian origin in the following manner.

An international travel ban to and from China, Iran and EU, EU FTA, Turkey, UK was imposed on 17th March 2020 by the Government of India (Ministry of home affairs). This was followed by a complete international travel ban from 22nd March 2020 onwards (<https://boi.gov.in/>). Hence, it is to be noted that no passenger would have entered India from other countries after 22nd March 2020. Moreover, the median time for the appearance of symptoms is 5-6 days and a maximum of 14 days, according to WHO (<https://www.who.int>). Thus presuming the last person who was exposed to the virus abroad, enters on 22nd March 2020, the person may develop symptoms within 5 days or between 5 and 14 days. Accordingly, the sequences have been divided based on their collection dates (**Table S2**), i.e., (i) before 22nd March 2020, (ii) after 5 days (27th March 2020 onwards), and (iii) after 14 days (5th April 2020 onwards).

Based on these classes, mutations in sequences belonging to the first category would be of foreign origin as International travel was allowed and India was in its nascent stage of the pandemic, with very few states affected. In the second category, there is a probability of mutation being in Indian origin, after ruling out that the same mutation does not exist in the first category sequences. The third category of sequences has a very high probability of containing mutations of Indian origin as 14 days after 22nd March 2020 is the last day a virus from another country could show symptoms in India (**Table 4**). The original strain might have undergone multiple transmissions throughout India, which is the primary reason for such mutation, as suggested by **Shu (2020)**. One interesting point to note is that the Gujarat cluster, which is placed in the high probability of Indian origin contains variations that are not present in any of the international sequences (collected as of 3rd April).

<.... Table 4 ...>

4. Conclusions

The question of identifying mutations in SARS-CoV-2 isolates of Indian origin was addressed based on phylogenomic, proteomic, and phylogeographic analysis (reflecting inter-state differences). The Indian isolates indicated that the majority of sequence variations are associated with protein-coding regions than the non-coding regions. The genomes analysed had a high entropy due to length differences as well as due to the absence of poly-A tail. Although the polyadenylation signal hexamer AAUAAA is invariant in the analysed dataset, the absence of poly-A tail in the dataset hinders the understanding of viral regulation during infection. The comparison of direct repeats in Wuhan reference sequence with Kerala isolate clarified that the missing bases 'TTA' is due to sequence errors that might have originated while sequencing/assembly by the interference of tandem repeats. The ORF predictions indicated that the theoretical protein translation for 'ORF1a' in West Bengal and Kerala isolates had +1 reading frame, whereas for Karnataka, the reading frame is +3, and for the rest, it is +2. Further analysis of the ribosomal frameshifting and transcriptional slippage events are crucial in understanding the strain-specific variations. The phylogeographical patterns of variations based on mutations of Indian origin was partially solved. However, the fact that international isolates deposited till the early April did not contain the same variations as of the Gujarat clusters, this supports the theory of Indian origin. However, the study is limited only based on the identified variations with only one representative for each state and does not include all states and union territories of India. Hence, further analysis of the nature and patterns of mutations can be done with a larger dataset having fewer sequence errors. The reasons for certain mutations being restricted to particular regions should requires further probing, and more light needs to be shed on the geographic driving forces, primarily in those sequences with confirmed mutations of Indian origin. Intra-state variations can also be studied, keeping the Gujarat cluster in mind and looking for explanations for variations within the same geographical region. It is to be noted that specific mutations in SARS-CoV-2 strains may be responsible for higher risks that may cause fatalities and mortalities. Hence, misinterpretation of mutations along with sequence errors is quite challenging, it is difficult to ascertain such errors that arise due to handling of large sample volumes, high throughput processing and sequencing in the ongoing pandemic. However, a quality check for sequences is crucial not only for the understanding mutations on virulence, pathogenicity, antigenicity and resistance but also for the development of vaccines.

5. Accession numbers

NC_045512; NC_004718.3; YP_009725318.1; EPI_ISL_413522; EPI_ISL_428481; EPI_ISL_430464-67; EPI_ISL_431102; EPI_ISL_435049, 55, 56, 60-62; EPI_ISL_435075-77; EPI_ISL_435088-90; EPI_ISL_435098, 101, 107, 112; EPI_ISL_436420-21; EPI_ISL_436443; EPI_ISL_437438; EPI_ISL_444486; EPI_ISL_437534, 36, 37; EPI_ISL_447545, 46, 48

Conflict of Interest

There is no conflict of interest.

References

1. Virology: Coronaviruses. (1968). *Nature*, 220(5168), 650. <https://doi.org/10.1038/220650b0>
2. *Computational network analysis of HCV, GBV-C, and HIV-1 viral proteins and their interactions with human*. (n.d.). 576.
3. The epidemiological characteristics of an outbreak of 2019 novel coronavirus diseases (COVID-19) in China. (2020). *Zhonghua liu xing bing xue za zhi = Zhonghua liuxingbingxue zazhi*, 41(2), 145–151. <https://doi.org/10.3760/cma.j.issn.0254-6450.2020.02.003>

4. Aggarwal, A., Mehta, S., Gupta, D., Sheikh, S., Pallagatti, S., Singh, R., & Singla, I. (2012). Clinical & immunological erythematosus patients characteristics in systemic lupus Maryam. *Journal of Dental Education*, 76(11), 1532–1539. <https://doi.org/10.4103/ijmr.IJMR>
5. Anindita, B., Rakesh, S., Mitra, S., Mahadeb, L., Shanta, D., & Mamta, C.-S. (2020). *The novel Coronavirus enigma: Phylogeny and mutation analyses of SARS-CoV-2 viruses circulating in India during early 2020* (Vol. 21).
6. Balaji S, Ramachandran A, Nandy K, Shapshak P (2017). *Sequence Accuracy in Primary Databases: A Case Study on HIV-1B*. In *Global Virology II - HIV and NeuroAIDS*, pp 779-822
7. Ceraolo, C., & Giorgi, F. M. (2020). Genomic variance of the 2019-nCoV coronavirus. *Journal of Medical Virology*, 92(5), 522–528. <https://doi.org/10.1002/jmv.25700>
8. Chatterjee, S. (2020). *An Overview of Mutations Occurring within the Coronavirus 2 genome*. [available in bioRxiv SSRN 3632241].
9. Crooks, G., Hon, G., Chandonia, J., & Brenner, S. (2004). NCBI GenBank FTP SiteWebLogo: a sequence logo generator. *Genome Res*, 14, 1188–1190. <https://doi.org/10.1101/gr.849004.1>
10. Forster, P., Forster, L., Renfrew, C., & Forster, M. (2020). Phylogenetic network analysis of SARS-CoV-2 genomes. *Proceedings of the National Academy of Sciences*, 202004999. <https://doi.org/10.1073/pnas.2004999117>
11. Fuke, H., & Ohno, M. (2008). Role of poly (A) tail as an identity element for mRNA nuclear export. *Nucleic Acids Research*, 36(3), 1037–1049. <https://doi.org/10.1093/nar/gkm1120>
12. Giorgio, S. Di, Martignano, F., Torcia, M. G., Mattiuz, G., & Conticello, S. G. (2020). Evidence for RNA editing in the transcriptome of 2019 Novel Coronavirus. *BioRxiv*, 2020.03.02.973255. <https://doi.org/10.1101/2020.03.02.973255>
13. Hao, P., Chen, M., Zhang, G., He, W., & Li, Y. (2006). Bioinformatics Research on the SARS Coronavirus (SARS CoV) in China. *Current Pharmaceutical Design*, 12(35), 4565–4572. <https://doi.org/10.2174/138161206779010404>
14. Hassan, D. S. S., Choudhury, P. P., Roy, B., & Jana, S. S. (2020). Missense Mutations in SARS-CoV2 Genomes from Indian Patients. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.3609566>
15. He, J., Tao, H., Yan, Y., Huang, S.-Y., & Xiao, Y. (2020). Molecular mechanism of evolution and human infection with the novel coronavirus (2019-nCoV). *BioRxiv*, 2020.02.17.952903. <https://doi.org/10.1101/2020.02.17.952903>
16. Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., ... Higgins, D. G. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics (Oxford, England)*, 23(21), 2947–2948. <https://doi.org/10.1093/bioinformatics/btm404>
17. Ledergerber, C., & Dessimoz, C. (2011). Base-calling for next-generation sequencing platforms. *Briefings in Bioinformatics*, 12(5), 489–497. <https://doi.org/10.1093/bib/bbq077>
18. Liu, T., Hu, J., Kang, M., Lin, L., Zhong, H., Xiao, J., ... Ma, W. (2020). Time-varying transmission dynamics of Novel Coronavirus Pneumonia in China. *BioRxiv*, 2020.01.25.919787. <https://doi.org/10.1101/2020.01.25.919787>
19. Longxian Lv, Gaolei Li, Jinhui Chen, Xinle Liang, Y. L. (2020). *Comparative genomic analysis revealed specific mutation pattern between human coronavirus SARS-CoV-2 and Bat-SARSr-CoV RaTG13* *Longxian*. 21(1), 1–9.
20. MacLachlan, N. J., & Dubovi, E. J. B. T.-F. V. V. (Fifth E. (Eds.). (2017). *Chapter 24 - Coronaviridae*. <https://doi.org/https://doi.org/10.1016/B978-0-12-800946-8.00024-6>
21. Maier, H. J., Bickerton, E., & Britton, P. (2015). Coronaviruses: Methods and protocols. *Coronaviruses: Methods and Protocols*, 1–282. <https://doi.org/10.1007/978-1-4939-2438-7>
22. Maranon, D. G., Anderson, J. R., Maranon, A. G., & Wilusz, J. (2020). The interface between coronaviruses and host cell RNA biology: Novel potential insights for future therapeutic intervention. *Wiley Interdisciplinary Reviews. RNA*, (May), e1614. <https://doi.org/10.1002/wrna.1614>
23. Marra, M. A., Jones, S. J. M., Astell, C. R., Holt, R. A., Brooks-wilson, A., Butterfield, Y. S. N., ... Roper, R. L. (2003). *The Genome Sequence of the SARS-Associated Coronavirus*. 300(May), 1399–1405.
24. Metsalu, T., & Vilo, J. (2015). ClustVis: A web tool for visualising clustering of multivariate data using Principal Component Analysis and heatmap. *Nucleic Acids Research*, 43(W1), W566–W570. <https://doi.org/10.1093/nar/gkv468>
25. Paraskevis, D., Kostaki, E. G., Magiorkinis, G., Panayiotakopoulos, G., Sourvinos, G., & Tsiodras, S. (2020). Full-genome evolutionary analysis of the novel coronavirus (2019-nCoV) rejects the hypothesis of

- emergence as a result of a recent recombination event. *Infection, Genetics and Evolution*, 79. <https://doi.org/10.1016/j.meegid.2020.104212>
26. Peng, Y. H., Lin, C. H., Lin, C. N., Lo, C. Y., Tsai, T. L., & Wu, H. Y. (2016). Characterisation of the role of hexamer AGUAAA and poly(A) tail in coronavirus polyadenylation. *PLoS ONE*, 11(10), 7–13. <https://doi.org/10.1371/journal.pone.0165077>
 27. Preiss, T., Muckenthaler, M., & Hentze, M. W. (1998). Poly(A)-tail-promoted translation in yeast: implications for translational control. *RNA (New York, N.Y.)*, 4(11), 1321–1331. <https://doi.org/10.1017/s1355838298980669>
 28. Rodnina, M. V., Korniy, N., Klimova, M., Karki, P., Peng, B. Z., Senyushkina, T., ... Peske, F. (2020). Translational recoding: canonical translation mechanisms reinterpreted. *Nucleic Acids Research*, 48(3), 1056–1067. <https://doi.org/10.1093/nar/gkz783>
 29. Saha, I., Ghosh, N., Maity, D., Sharma, N., Sarkar, J. P., & Mitra, K. (2020). Genome-wide analysis of Indian SARS-CoV-2 genomes for the identification of genetic mutation and SNP. *Infection, Genetics and Evolution: Journal of Molecular Epidemiology and Evolutionary Genetics in Infectious Diseases*, 85(July), 104457. <https://doi.org/10.1016/j.meegid.2020.104457>
 30. Salemi, M., Fitch, W. M., Ciccozzi, M., Ruiz-Alvarez, M. J., Rezza, G., & Lewis, M. J. (2004). Severe acute respiratory syndrome coronavirus sequence characteristics and evolutionary rate estimate from maximum likelihood analysis. *Journal of Virology*, 78(3), 1602–1603.
 31. Sayers, E. W., Cavanaugh, M., Clark, K., Ostell, J., Pruitt, K. D., & Karsch-Mizrachi, I. (2020). GenBank. *Nucleic Acids Research*, 48(D1), D84–D86. <https://doi.org/10.1093/nar/gkz956>
 32. 2020). Genome-wide data inferring the evolution and population demography of the novel pneumonia coronavirus (SARS-CoV-2). *BioRxiv*, 2020.03.04.976662. <https://doi.org/10.1101/2020.03.04.976662>
 33. Simms, C. L., Yan, L. L., Qiu, J. K., & Zaher, H. S. (2019). Ribosome Collisions Result in +1 Frameshifting in the Absence of No-Go Decay. *Cell Reports*, 28(7), 1679–1689.e4. <https://doi.org/10.1016/j.celrep.2019.07.046>
 34. Sturman, L. S., & Holmes, K. V. (1983). *The Molecular Biology of Coronaviruses* (M. A. Lauffer & K. B. T.-A. in V. R. Maramorosch, eds.). [https://doi.org/https://doi.org/10.1016/S0065-3527\(08\)60721-6](https://doi.org/https://doi.org/10.1016/S0065-3527(08)60721-6)
 35. Su, Z., & Wu, Y. (2020). A Multiscale and Comparative Model for Receptor Binding of 2019 Novel Coronavirus and the Implication of its Life Cycle in Host Cells. *BioRxiv*, (718), 2020.02.20.958272. <https://doi.org/10.1101/2020.02.20.958272>
 36. Tørresen, O. K., Star, B., Mier, P., Andrade-Navarro, M. A., Bateman, A., Jarnot, P., ... Linke, D. (2019). Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases. *Nucleic Acids Research*, 47(21), 10994–11006. <https://doi.org/10.1093/nar/gkz841>
 37. Tørresen, O. K., Star, B., Mier, P., Andrade-Navarro, M. A., Bateman, A., Jarnot, P., ... Linke, D. (2019). Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases. *Nucleic Acids Research*, 47(21), 10994–11006. <https://doi.org/10.1093/nar/gkz841>
 38. Vijaykrishna, D., Smith, G. J. D., Zhang, J. X., Peiris, J. S. M., Chen, H., & Guan, Y. (2007). Evolutionary Insights into the Ecology of Coronaviruses. *Journal of Virology*, 81(8), 4012–4020. <https://doi.org/10.1128/jvi.02605-06>
 39. Weiss, E. A., Gilmartin, G. M., & Nevins, J. R. (1991). Poly(A) site efficiency reflects the stability of complex formation involving the downstream element. *EMBO Journal*, 10(1), 215–219. <https://doi.org/10.1002/j.1460-2075.1991.tb07938.x>
 40. Woo, P. C. Y., Huang, Y., Lau, S. K. P., & Yuen, K. Y. (2010). Coronavirus genomics and bioinformatics analysis. *Viruses*, 2(8), 1805–1820. <https://doi.org/10.3390/v2081803>
 41. Wu, F., Zhao, S., Yu, B., Chen, Y. M., Wang, W., Song, Z. G., ... Zhang, Y. Z. (2020). A new coronavirus associated with human respiratory disease in China. *Nature*, 579(7798), 265–269. <https://doi.org/10.1038/s41586-020-2008-3>
 42. Wu, F., Zhao, S., Yu, B., Chen, Y.-M., Wang, W., Hu, Y., ... Zhang, Y.-L. Y.-Z. (2020). Complete genome characterisation of a novel coronavirus associated with severe human respiratory disease in Wuhan, China. *BioRxiv*, 2020.01.24.919183. <https://doi.org/10.1101/2020.01.24.919183>
 43. Wu, H. Y., & Brian, D. A. (2010). Subgenomic messenger RNA amplification in coronaviruses. *Proceedings of the National Academy of Sciences of the United States of America*, 107(27), 12257–12262. <https://doi.org/10.1073/pnas.1000378107>
 44. Wu, H. Y., Ke, T. Y., Liao, W. Y., & Chang, N. Y. (2013). Regulation of Coronaviral Poly(A) Tail Length during Infection. *PLoS ONE*, 8(7), 1–12. <https://doi.org/10.1371/journal.pone.0070548>

45. Xia, X. (2018). DAMBE7: New and Improved Tools for Data Analysis in Molecular Biology and Evolution. *Molecular Biology and Evolution*, 35(6), 1550–1552. <https://doi.org/10.1093/molbev/msy073>
46. Xiaolu Tang, Changcheng Wu, Xiang Li, Yuhe Song, Xinmin Yao, Xinkai Wu, Yuange Duan, Hong Zhang, Yirong Wang, Zhaohui Qian, J. C. and J. L. (n.d.). *On the origin and continuing evolution of SARS-CoV-2*.
47. Xiong, C., Jiang, L., Chen, Y., & Jiang, Q. (2020). Evolution and variation of 2019-novel coronavirus. *BioRxiv*, 2020.01.30.926477. <https://doi.org/10.1101/2020.01.30.926477>
48. Xu, X. Z., Liu, Q. P., Fan, L. J., Cui, X. F., & Zhou, X. P. (2008). Analysis of synonymous codon usage and evolution of begomoviruses. *Journal of Zhejiang University: Science B*, 9(9), 667–674. <https://doi.org/10.1631/jzus.B0820005>
49. Yachdav, G., Hecht, M., Pasmanik-Chor, M., Yeheskel, A., & Rost, B. (2014). HeatMapView: Interactive display of 2D data in biology. *F1000Research*, 3, 1–6. <https://doi.org/10.12688/f1000research.3-48.v1>
50. Yvonne CF Su, Danielle E Anderson, Barnaby E Young, Feng Zhu, Martin Linster, Shirin Kalimuddin, Jenny GH Low, Zhuang Yan, Jayanthi Jayakumar, Louisa Sun, Gabriel Z Yan, Ian H Mendenhall, Yee-Sin Leo, David Chien Lye, Lin-Fa Wang, G. J. S. (2020). *Discovery of a 382-nt deletion during the early evolution of SARS-CoV-2*.
51. Zhao, Z., & Sokhansanj, B. A. (2020). *Characterising geographical and temporal dynamics of novel coronavirus SARS-CoV-2 using informative subtype markers*.
52. Zhou, H., Chen, X., Hu, T., Li, J., Song, H., Liu, Y., ... Shi, W. (2020). A novel bat coronavirus reveals natural insertions at the S1/S2 cleavage site of the Spike protein and a possible recombinant origin of HCoV-19. *BioRxiv*, 2020.03.02.974139. <https://doi.org/10.1101/2020.03.02.974139>
53. Zhou, P., Yang, X. Lou, Wang, X. G., Hu, B., Zhang, L., Zhang, W., ... Shi, Z. L. (2020). A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*, 579(7798), 270–273. <https://doi.org/10.1038/s41586-020-2012-7>

Figure legends

Fig 1. Indian map depicting the locations from which the samples were collected. The map illustration was created using mapchart.net

Fig 2. Entropy plot

Fig 3. Codon usage pattern for the analyzed dataset

Fig 4. Heatmaps of a) Nucleotide frequency b) Codon usage, c) RSCU and d) Amino acid frequency

Fig 5. SARS CoV-2 proteomes of Indian isolates (a) envelope protein, (b) nucleocapsid phosphoprotein, (c) Orf3a, (d) Orf7a, (e) spike protein, (f) Orf7b, (g) Orf6a and (h) Orf10

Fig 6. (a) Cladogram of Indian isolates, (b) Cladogram of Indian and international clusters. Both the trees are generated by using Fast-ME algorithm with bootstrap (100 trials), the values are displayed for above 95% confidence, and the SARS-CoV-1 is used as an outgroup.

Fig 7. Percentage variations that define the patterns of the different clusters of (a) North –type I cluster, (b) North- type II cluster, and (c) East-West cluster

Table 1: The analyzed dataset with accession numbers, sampled geographic location, sequence length, GC%, sequence ambiguity and the sequenced laboratory details are showed

Sl. No	Accession ID	Location	Sequence length (bp)	GC %	Ambiguous bases								Submitted lab
					R	Y	K	M	S	W	N	Total	
1	EPI_ISL_413522	Kerala	29854	38.0184	0	0	0	0	0	0	0	0	Indian Council of Medical Research - National Institute of Virology
2	EPI_ISL_428481	Karnataka	29840	38.0948	0	0	0	0	0	0	257	257	National Institute of Mental Health and Neuroscience (NIMHANS)
3	EPI_ISL_430464	Kolkata (West Bengal)	29858	38	0	0	0	0	0	0	0	0	National Institute of Biomedical Genomics
4	EPI_ISL_430465	Darjeeling (West Bengal)	29858	38	0	0	0	0	0	0	0	0	National Institute of Biomedical Genomics
5	EPI_ISL_430466	Tehatta (West Bengal)	29857	38.011	0	0	0	0	0	0	0	0	National Institute of Biomedical Genomics
6	EPI_ISL_430467	East Medinipur (West Bengal)	29856	38.0125	0	0	0	0	0	0	0	0	National Institute of Biomedical Genomics
7	EPI_ISL_431102	Hyderabad (Telangan a)	29903	37.966	0	0	0	0	0	0	0	0	Gandhi Medical College and Hospital, Secendrabad, Hyderabad
8	EPI_ISL_435049	Ahmedabad (Gujarat)	29903	38.0597	0	0	0	0	0	0	118	118	Gujarat Biotechnology Research Centre
9	EPI_ISL_435055	Gandhinagar (Gujarat)	29903	38.0346	0	0	0	0	0	0	103	103	Gujarat Biotechnology Research Centre
10	EPI_ISL_435056	Mansa (Gujarat)	29903	38.053	0	0	0	0	0	0	144	144	Gujarat Biotechnology Research Centre

11	EPI_ISL_435061	Delhi	29866	38.0064	0	0	1	0	0	0	7	8	NCDC/CSIR-IGIB
12	EPI_ISL_435062	Punjab	29867	37.9985	1	1	1	0	0	0	17	20	NCDC/CSIR-IGIB
13	EPI_ISL_435075	Tamil Nadu	29866	38.2023	0	2	1	0	2	1	506	512	NCDC/CSIR-IGIB
14	EPI_ISL_435076	Haryana	29866	38.1856	1	2	4	1	1	3	393	405	NCDC/CSIR-IGIB
15	EPI_ISL_435077	Maharashtra	29866	38.2023	1	4	2	1	0	1	453	462	NCDC/CSIR-IGIB
16	EPI_ISL_435088	Odisha	29866	38.1086	0	0	1	1	0	1	261	264	NCDC/CSIR-IGIB
17	EPI_ISL_435089	Andhra Pradesh	29866	38.517	2	1	2	0	0	0	1472	1477	NCDC/CSIR-IGIB
18	EPI_ISL_435090	Jammu	29867	38.0135	0	0	1	0	1	1	28	31	NCDC/CSIR-IGIB
19	EPI_ISL_435098	Nepal	29866	38.0081	1	2	2	2	1	0	14	22	NCDC/CSIR-IGIB
20	EPI_ISL_435101	Ladakh	29866	38.0064	0	0	0	0	0	0	4	4	NCDC/CSIR-IGIB
21	EPI_ISL_435107	Kargil	29866	38.137	1	2	1	2	0	3	428	437	NCDC/CSIR-IGIB
22	EPI_ISL_435112	Bihar	29866	38.0014	0	1	1	2	1	0	11	16	NCDC/CSIR-IGIB
23	EPI_ISL_436420	Rajasthan	29866	38.0014	0	0	1	1	1	2	5	10	NCDC/CSIR-IGIB
24	EPI_ISL_436421	Assam	29866	38.0131	0	3	2	3	2	0	34	44	NCDC/CSIR-IGIB
25	EPI_ISL_436443	Madhya Pradesh	29866	38.1822	2	4	19	20	1	0	618	664	NCDC/CSIR-IGIB
26	EPI_ISL_437438	Surat (Gujarat)	29903	38.033	0	0	0	0	0	0	110	110	Gujarat Biotechnology Research Centre
27	EPI_ISL_437536	South 24 Parganas (West Bengal)	29903	39.996	0	0	0	0	0	0	4822	4822	National Institute of Biomedical Genomics
28	EPI_ISL_437537	Howrah (West Bengal)	29903	38.1383	0	0	0	0	0	0	343	343	National Institute of Biomedical Genomics
29	EPI_ISL_444486	Dhansura (Gujarat)	29903	38.0163	0	0	0	0	0	0	76	76	Gujarat Biotechnology Research Centre
30	EPI_ISL_447534	Modasa (Gujarat)	29903	38.0296	0	0	0	0	0	0	78	78	Gujarat Biotechnology Research Centre

31	EPI_ISL_447545	Prantij (Gujarat)	29903	38.0263	0	0	0	0	0	0	60	60	Gujarat Biotechnology Research Centre
32	EPI_ISL_447546	Himatnagar (Gujarat)	29903	38.0196	0	0	0	0	0	0	78	78	Gujarat Biotechnology Research Centre
33	EPI_ISL_447548	Dahegam (Gujarat)	29903	38.0246	0	0	0	0	0	0	79	79	Gujarat Biotechnology Research Centre

Table 2: Regional grouping based on geographical directions

Regions	Colour code	States of India
North	Red	Nepal, Haryana, Jammu, Punjab, Delhi, Madhya Pradesh
South	Blue	Karnataka, Kerala, Tamil Nadu, Andhra Pradesh, Telangana
East	Green	Assam, West Bengal, Bihar, Odisha
West	Yellow	Gujarat, Rajasthan, Maharashtra

Table 3: Sequence data organized into regional groups based on the type of mutations

Description/Name	Mutation	AA change	Base number	Gene	States/Sequences involved
Gujarat Cluster	T>C	F>L	2836	nsp3	9, 29, 32
	A>G	I>V	16078	RDRP	9,30, 33
	T>C	L(silent)	18877	3'-to-5' exonuclease	9,29,30,32,33
	T>C	S>P	22444	S protein	9,29, 32,33
	T>G	H>Q	25563	Orf3a	9,29,30,32,33
	T>C	I>T	26735	Orf6	9,29,30,32,33
	T>C	L>S	28854	N Protein	9,29,30,32,33
North-type I	M>C A>C	R>S (M=A) or silent R>S	6310	Nsp3	22, 15, 25, 18, 26, 17

	A>C	K>T	6312	Nsp3	19, 22,24, 18, 23, 26, 15, 16, 25, 13, 17
	K>T	V>L (K=G) or silent.	8022	Nsp3	11,12, 19, 22,24, 23, 15, 18, 14,20, 16, 13 17
	G>T	V>L			
	T>G	L>V	11083	Nsp6	19, 22,24, 23, 26, 18, 14,20,21, 16, 13, 17 25, 14
	K>G	L>V (K=T) or silent			
	T>C	V>A	13730	RDRP	18, ,19, 22, 24, 16, 23, 15, 26 , 25, 13,17
	T>C	Stop codon>Q (Nonsense)	23929	S protein	19, 22, 23, 15, 26 ,24, 18, 16, 25, 13
East-West	T>C	S>P	28311	N Protein	19, 22, 23, 26, 18, 16, 17 24, 14 , 15, 25, 13,
	Y>C	S>P (Y=T) or silent			
	T>C	C>R	241	nsp1	2,7 8,9,10,29,30,31,32,33, 3,4,5,6, 11,12
	T>C	L (silent)	3037	nsp3	2,7 8,9,10,29,30,31,32,33, 3,4,5,6, 11,12
North-type II	T>C	L>P	14408	RDRP	2,7 8,9,10,29,30,31,32,33, 3,4,5,6, 11,12
	G>A	G=G (Silent)	23403	S protein	2,7 8,9,10,29,30,31,32,33, 3,4,5,6, 11,12
	M>A	R>S if (M=C,S=G) P>S if (M=C,S=C) Silent if (M=A,S=G) I>S if (M=A,S=C)	24389	S protein	19, 22,24, 23, 15, 25
Other	S>A		24390		
	T>C Y>C	L (silent)	19524	3'-to-5' exonuclease	22, 26 , 18, 17, 15
Other	A>G	K>R	28881	N Protein	2, 3,4

	A>G	(Part of the same codon)	28882		
	C>G	R>G	28883		
Kerala exclusive	C>T		2277	Nsp2	1
	T>C		6695	Nsp3	
	T>C		14657	RDRP	
	T>C		17373	helicase	
		Deletion	21993,94,95	S protein	
	T>G		22785	S protein	

Table 4: Probable origin of variations in India based on collection dates (Table S1)

Cluster name	Variation/Mutation position	Gene region	Probability of Indian origin
North-I	8022	Nsp3	Low
	11083	Nsp6	
East-West	241	Nsp1	
	3037	Nsp3	
	14408	RDRP	
	23403	S protein	
Other	28881	N protein	
	28882		
	28883		
North-I	6310	Nsp3	Medium (Quite probable)
	6312	Nsp3	
	13730	RDRP	
	23929	S protein	
	28311	N Protein	
North-II	24389	S protein	
	24390	S protein	
Gujarat	2836	nsp3	High
	16078	RDRP	
	18877	3'-to-5' exonuclease	
	22444	S protein	
	25563	Orf3a	
	26735	Orf6	
	28854	N Protein	

Figure 1

[Click here to access/download;Figure;Fig 1.tif](#)

Figure 2

Figure 4

Figure 5

Figure 7

Supplementary Table S1: The phylogenetic clusters of SARS-CoV-2 isolates depicting differentiating features based on probable mutations

ClusterStrain numberLocation of isolatesAccession numbersGenome size						Differentiating features									
						nsp1	nsp2	nsp3	nsp6	RDRP	Exonucleas e	S	Orf3a	Orf6	N
Gujarat cluster	a	GBRC-8,55,85	Gandhinagar, Dhansura, Himatnagar	EPI_ISL_444486, 447534, 447546	29903		T>C (F>L) 2841		A>G (I>V) 16090						
	b	GBRC-8,76,87	Gandhinagar, Modasa, Dahegam	EPI_ISL_444486, 447545, 447548											
	c	GBRC-8,55,85,87	Gandhinagar, Dhansura, Prantij, Himatnagar, Dahegam	EPI_ISL_444486, 447534, 447546, 447548			T>C (L silent) 18889								
	d	GBRC-8,55,76,85,87	Gandhinagar, Dhansura, Modasa, Prantij, Himatnagar, Dahegam	EPI_ISL_444486, 447534, 447545, 447546, 447548			T>C (S>P) 22472		T>G (H>Q) 25592						
North-type I	a	NCDC-02330, 02323, 02326, 02251, 3961, GBRC10	Maharashtra, Andhra Pradesh, Jammu, Bihar, Madhya Pradesh, Surat	EPI_ISL_435077, 435089, 435090, 435112, 436443, 437438,	29866, 29866, 29867, 29866, 29866, 29903,		M>C (R>S or silent) A>C (R>S) 6322								
	b	NCDC-02331, 2330, 2336, 2323, 2326, 2240, 2251, 2522, 2525, 3961,	Tamil Nadu, Maharashtra, Odisha, Andhra Pradesh, Jammu, Nepal, Bihar, Rajsathan, Assam, Madhya Pradesh	EPI_ISL_435075, 435077, 435088, 435089, 435090, 435098, 435112, 436420, 436421, 436443,	29866, 29866, 29866, 29866, 29867, 29866, 29866, 29866, 29866,		A>C (K>T) 6324								

c	NCDC-01537, 1538, 2331, 2329, 2330, 2336, 2323, 2326, 2240, 1441, 2251, 2522, 2525,	Delhi, Punjab, Tamil Nadu, Haryana, Maharashtra, Odisha, Andhra Pradesh, Jammu, Nepal, Ladakh, Bihar, Rajsathan, Assam	EPI_ISL_435061, 435062, 435075, 435076, 435077, 435088, 435089, 435090, 435098, 435101, 435112, 436420, 436421,	29866, 29867, 29866, 29866, 29866, 29866, 29866, 29867, 29866, 29866, 29866, 29866, 29866,
d	NCDC-02331, 2329, 2336, 2323, 2326, 2240, 1441, 2370, 2251, 2522, 2525, 3961; GBRC10, BRA/SP02	Tamil Nadu, Haryana, Odisha, Andhra Pradesh, Jammu, Nepal, Ladakh, Kargil, Bihar, Rajsathan, Assam, Madhya Pradesh, Surat , Brazil	EPI_ISL_435075, 435076, 435088, 435089, 435090, 435098, 435101, 435107, 435112, 436420, 436421, 436443, 437438; MT126808	29866, 29866, 29866, 29866, 29867, 29866, 29866, 29866, 29866, 29866, 29866, 29866, 29903, 29876
e	NCDC-02331, 2330, 2336, 2323, 2326, 2240, 2251, 2522, 2525, 3961, GBRC10,	Tamil Nadu, Maharashtra, Odisha, Andhra Pradesh, Jammu, Nepal, Bihar, Rajsathan, Assam, Madhya Pradesh, Surat	EPI_ISL_435075, 435077, 435088, 435089, 435090, 435098, 435112, 436420, 436421, 436443, 437438,	29866, 29866, 29866, 29866, 29867, 29866, 29866, 29866, 29866, 29866, 29903,
f	NCDC-02331, 2330, 2336, 2323, 2326, 2240,	Tamil Nadu, Maharashtra, Odisha, Andhra Pradesh, Jammu, Nepal,	EPI_ISL_435075, 435077, 435088, 435089, 435090, 435098,	29866, 29866, 29866, 29866, 29867, 29866,

K>T (V>L or
silent)
G>T(V>L)
8034

K>G
(L>V or
silent)
T>G
(L>V)
11095

T>V (V>A)
13742

T>V (Stop
codon>Q)
23958

[illegible]

North-type II	a	NCDC-02330, 02240, 02251, 2522, 2525, 3961	Maharashtra, Nepal, Bihar, Rajsathan, Assam, Madhya Pradesh	EPI_ISL_435077 , 435098, 435112 , 436420, 436421, 436443	29866						<u>M>A & S>A</u> 24418,2441 9R>S if (M=C,S=G) P>S if (M=C,S=C) Silent if (M=A,S=G) I>S if (M=A,S=C)				
	b	NCDC-02330, 02323, 02326, 02251; GBRC10	Maharashtra, Andhra Pradesh, Jammu, Bihar, Surat	EPI_ISL_435077, 435089, 435090, 435112, 437438	29866, 29866, 29867, 29866, 299903						<u>T>C & Y>C</u> (L silent) 19536				
Other	a	nimh-0130;S3,S5; Peru-10,Valencia7	Karnataka, Kolkata, Darjeeling, Peru, Spain	EPI_ISL_428481, 430464, 430465; MT263074, MT233522	29840, 29858, 29856, 29782										A>G (K>R) 28933

Supplementary Table S2: Sample collection dates, to aid origin prediction

<i>Category</i>	<i>Accession ID</i>	<i>Location</i>	<i>Collection date</i>
<i>Pre 22/3/2020</i>	EPI_ISL_413522	Asia / India / Kerala	2020-01-27
	EPI_ISL_430464	Asia / India / West Bengal / Kolkata	2020-03-21
	EPI_ISL_431102	Asia / India / Telangana / Hyderabad	2020-03-11
	EPI_ISL_435061	Asia / India / Delhi	2020-03-16
	EPI_ISL_435062	Asia / India / Punjab	2020-03-16
	EPI_ISL_437536	Asia / India / West Bengal / South 24 Parganas	2020-03-19
	EPI_ISL_435101	Asia / India / Ladakh	2020-03-15
<i>27/3/2020 onwards</i> <i>(5-day buffer)</i>	EPI_ISL_430465	Asia / India / West Bengal / Darjeeling	2020-03-28
	EPI_ISL_430466	Asia / India / West Bengal / Tehatta	2020-03-26*
	EPI_ISL_435075	Asia / India / Tamil Nadu	2020-03-29
	EPI_ISL_435076	Asia / India / Haryana	2020-03-29
	EPI_ISL_435077	Asia / India / Maharashtra	2020-03-29
	EPI_ISL_435088	Asia / India / Odisha	2020-03-29
	EPI_ISL_435089	Asia / India / Andhra Pradesh	2020-03-29
	EPI_ISL_435090	Asia / India / Jammu	2020-03-29
	EPI_ISL_435098	Asia / India / Nepal	2020-03-28
	EPI_ISL_435107	Asia / India / Kargil	2020-03-29
	EPI_ISL_435112	Asia / India / Bihar	2020-03-28
	EPI_ISL_436420	Asia / India / Rajasthan	2020-03-31
	EPI_ISL_436421	Asia / India / Assam	2020-03-31
	EPI_ISL_437537	Asia / India / West Bengal / Howrah	2020-03-30
	EPI_ISL_430467	Asia / India / West Bengal / East Medinipur	2020-04-03
<i>5/4/2020 onwards</i> <i>(14-day buffer)</i>	EPI_ISL_428481	Asia / India / Karnataka	2020-04-06
	EPI_ISL_435049	Asia / India / Gujarat / Ahmedabad	2020-04-13
	EPI_ISL_435055	Asia / India / Gujarat / Gandhinagar	2020-04-22
	EPI_ISL_435056	Asia / India / Gujarat / Mansa	2020-04-21

EPI_ISL_436443	Asia / India / Madhya Pradesh	2020-04-09
EPI_ISL_437438	Asia / India / Gujarat / Surat	2020-04-24
EPI_ISL_444486	Asia / India / Gujarat / Dhansura	2020-05-04
EPI_ISL_447534	Asia / India / Gujarat / Modasa	2020-05-05
EPI_ISL_447545	Asia / India / Gujarat / Prantij	2020-05-05
EPI_ISL_447546	Asia / India / Gujarat / Himatnagar	2020-05-05
EPI_ISL_447548	Asia / India / Gujarat / Dahegam	2020-04-28

Figures

Figure 1

Indian map depicting the locations from which the samples were collected. The map illustration was created using mapchart.net . Note: The designations employed and the presentation of the material on this map do not imply the expression of any opinion whatsoever on the part of Research Square

concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. This map has been provided by the authors.

Figure 2

Entropy plot

Figure 3

Codon usage pattern for the analyzed dataset

Figure 4

Heatmaps of a) Nucleotide frequency b) Codon usage, c) RSCU and d) Amino acid frequency

Figure 5

SARS CoV-2 proteomes of Indian isolates (a) envelope protein, (b) nucleocapsid phosphoprotein, (c) Orf3a, (d) Orf7a, (e) spike protein, (f) Orf7b, (g) Orf6a and (h) Orf10

Figure 6

(a) Cladogram of Indian isolates, (b) Cladogram of Indian and international clusters. Both the trees are generated by using Fast-ME algorithm with bootstrap (100 trials), the values are displayed for above 95% confidence, and the SARS-CoV-1 is used as an outgroup.

Figure 7

Percentage variations that define the patterns of the different clusters of (a) North type I cluster, (b) North-type II cluster, and (c) East-West cluster