Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Inferring the genetic variability in Indian SARS-CoV-2 genomes using consensus of multiple sequence alignment techniques

Indrajit Saha, Nimisha Ghosh, Debasree Maity, Nikhil Sharma, Kaushik Mitra

PII: S1567-1348(20)30353-1
DOI: https://doi.org/10.1016/j.meegid.2020.104522
Reference: MEEGID 104522
To appear in: Infection, Genetics and Evolution

Received date: 10 July 2020
Revised date: 31 July 2020
Accepted date: 27 August 2020

Please cite this article as: I. Saha, N. Ghosh, D. Maity, et al., Inferring the genetic variability in Indian SARS-CoV-2 genomes using consensus of multiple sequence alignment techniques, Infection, Genetics and Evolution (2019), https://doi.org/10.1016/j.meegid.2020.104522

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2019 Published by Elsevier.
Inferring the Genetic Variability in Indian SARS-CoV-2 Genomes using Consensus of Multiple Sequence Alignment Techniques

Indrajit Saha\textsuperscript{a,f,*}, Nimisha Ghosh\textsuperscript{b,f}, Debasree Maity\textsuperscript{c}, Nikhil Sharma\textsuperscript{d}, Kaushik Mitra\textsuperscript{e}

\textsuperscript{a}Department of Computer Science and Engineering, National Institute of Technical Teachers’ Training and Research, Kolkata, West Bengal, India
\textsuperscript{b}Department of Computer Science and Information Technology, Institute of Technical Education and Research, Siksha ‘O’ Anusandhan (Deemed to be University), Bhubaneswar, Odisha, India
\textsuperscript{c}Department of Electronics and Communication Engineering, MCKV Institute of Engineering, Howrah, West Bengal, India
\textsuperscript{d}Department of Electronics and Communication Engineering, Jaypee Institute of Information Technology, Noida, Uttar Pradesh, India
\textsuperscript{e}Department of Community Medicine, Burdwan Medical College, Burdwan, West Bengal, India
\textsuperscript{f}Equally contributed

Abstract
Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2) is a threat to the human population and has created a worldwide pandemic. Daily thousands of people are getting affected by the SARS-CoV-2 virus; India being no exception. In this situation, there is no doubt that vaccine is the primary prevention strategy to contain the wave of COVID-19 pandemic. In this regard, genome-wide analysis of SARS-CoV-2 is important to understand its genetical variability. This has motivated us to analyse the 566 Indian SARS-CoV-2 sequences using multiple sequence alignment techniques viz. ClustalW, MUSCLE, ClustalO and MAFFT to identify the lists of mutations as substitution, deletion, insertion and SNP. Thereafter, a consensus of these results, called as Consensus Multiple Sequence Alignment (CMSA), is considered to have the final list of mutations so that the advantages of all four alignment techniques can be preserved. The analysis shows 767, 2025 and 54 unique substitutions, deletions and SNPs in Indian SARS-CoV-2 genomes. More precisely, out of 54 SNPs, 4 SNPs are present close to the 60\% of the virus population. The results of this experiment can be useful for the virus classification, designing and

* Corresponding author: indrajit@nitttrkol.ac.in
defining the dose of vaccine for the Indian population.

*Keywords:* Multiple Sequence Alignment, Point mutation, SNP, SARS-CoV-2

1. **Introduction**

Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2) causes a disease COVID-19 which was originated in Wuhan, China (Zhu et al., 2020). This disease has brought a wave of pandemic throughout the world. In this worrisome situation, undoubtedly vaccine is the primary prevention strategy to contain this virus. However, the development process of vaccine is time consuming and requires the analysis of genetical variability of the virus population so that effective and safe vaccine can be developed for heterogeneous human population (Poland, 2020).

In this regard, Tung Phan has performed a genetical analysis to show the evolution of SARS-CoV-2 (Phan, 2020), subsequently, we have also analysed Indian SARS-CoV-2 genomes in (Saha et al., 2020). In continuation to these works, we have further conducted the genetical analysis of 566 SARS-CoV-2 virus population to identify the mutation as substitution, deletion, insertion and Single Nucleotide Polymorphism (SNP). Generally, 1% of the population which is being affected by substitution is termed as SNP. We have concerned ourselves with only non-synonymous mutations as they are responsible for changes in the amino acids. In order to find the genetic variability, the multiple sequence alignment (MSA) is very much essential in presence of reference sequence. On the other hand, it is also known that the multiple sequence alignment technique provides a near optimal result. Therefore, different alignment techniques can provide different results on the same pool of sequences. Hence, we have used four different well-known multiple sequence alignment techniques viz. ClustalW (Thompson et al., 1994), MUSCLE (Edgar, 2004), ClustalO (Sievers et al., 2011) and MAFFT (Katoh et al., 2019) in order to align the Indian 566 SARS-CoV-2 sequences and identified the lists of mutations as substitution, deletion, insertion and SNP. Thereafter, a consensus of these results, called as Consensus Multiple Sequence Alignment (CMSA), is considered to have the final list of mutations so that the advantages of all four alignment techniques can be preserved. It is important to mention that the identification of SNPs can be helpful for the classification of virus strain and accordingly designing of vaccine and defining the dose of the vaccine can be done effectively (Jeon et al., 2016).
Recently the metagenomic analysis using Next-Generation Sequencing (NGS) (Lu et al., 2020, Zhou et al., 2020) shows that the SARS-CoV-2 is a single-stranded enveloped RNA virus with a genome length ranges from 27 to 32 kilobases (Vellingiri et al., 2020). As reported in NCBI, SARS-CoV-2 has 11 coding regions that can encode ORF1ab polyproteins, spike (S) glycoprotein, envelope (E) protein, membrane (M) glycoprotein, nucleocapsid (N) protein and accessory proteins such as ORF3a, ORF6, ORF7a, ORF7b, ORF8 and ORF10. It has further been reported that Open Reading Frame (ORF) can encode several Non-Structural Proteins (NSP). The genomic orientation of SARS-CoV-2 virus is shown in Figure S1(a) and their coordinates in Table S1 in supplementary. This is important to mention that the virus has a novel strain. Furthermore, the understanding of its genetic variability in different countries is still limited. This is yet another motivation to conduct this research for Indian SARS-CoV-2 sequences.

2. Materials and Methods

The recent genomic sequences of Indian SARS-CoV-2 virus have been collected from Global Initiative on Sharing All Influenza Data (GISAID)\(^1\) in fasta format. It contains 566 complete and near complete genomes with sequence ID. The average length of the 566 genomes is 29831 bp. In addition to this, we have downloaded the Reference Sequence (NC_045512.2)\(^2\) from National Center for Biotechnology Information (NCBI) to conduct the experiment with 566 Indian SARS-CoV-2 genomes. For the data visualization and editing, BioEdit has been used.

We have applied four different Multiple Sequence Alignment (MSA) techniques: ClustalW (Thompson et al., 1994), MUSCLE (Edgar, 2004), ClustalO (Sievers et al., 2011) and MAFFT (Katoh et al., 2019) in order to align the Indian 566 SARS-CoV-2 sequences. ClustalW, ClustalO and MAFFT are progressive alignment techniques whereas MUSCLE is an iterative progressive alignment technique. Progressive alignment techniques use methods such as Needleman-Wunsch algorithm, Smith-Waterman algorithm etc. to complete the pairwise alignments and then the sequences are clustered together to show their relationships by using methods such as \(k\)-means algorithm. Iterative progressive alignment technique works similarly as the progressive alignment techniques but dynamic programming is repeatedly applied over here to

\(^1\) https://www.gisaid.org/

\(^2\) https://www.ncbi.nlm.nih.gov/nuccore/1798174254
realign the initial sequences. At the same time, it also appends new sequences to the growing MSA. The difference in the characteristics of the four methods adopted in this work is worth mentioning over here. ClustalW initially performs pairwise alignment of all sequences by using the \( k \)-tuple method. Next, MSA is created by progressively aligning the most closely related sequences based on Neighbour-Joining guide tree method. ClustalO uses the \( k \)-tuple method to produce pairwise alignment. Then mBed is used to cluster the sequences followed by \( k \)-means clustering algorithm. Next, the guide tree is built using Unweighted Pair Group Method with Arithmetic Mean (UPGMA) method. Finally, MSA is constructed using the HHalign package. MAFFT uses two different heuristic methods, progressive (FFT-NS-2) and iterative refinement (FFT-NS-i). The main aim of MAFFT is to merge local and global algorithms for MSA. Initially, FFT-NS-2 is used to calculate all-pairwise distances to create a provisional MSA from which refined distances are calculated. Then, FFT-NS-i is performed to get the final MSA. In MUSCLE technique, two distance measures are used: \( k \)-mer for unaligned pairs and Kimura method for aligned pairs of sequences. Initially, a draft MSA is produced in MUSCLE using the \( k \)-mer method. Next, a progressive alignment is constructed based on the guide tree as produced by the UPGMA method. This initial tree is then re-estimated using the Kimura distance method after which UPGMA method is once again used to produce a new guide tree, thereby creating a second MSA. New MSAs are finally created by realigning the two sequences created previously. Thus, we have used these four techniques to identify the mutation in Indian virus sequences. These methods are providing lists of mutations as substitution, deletion, insertion and SNP. Thereafter, a consensus of these results, called as Consensus Multiple Sequence Alignment (CMSA), is considered to have the final list of mutations so that the advantages of all four alignment techniques can be preserved. This is to be noted that the identification of list of mutations from the alignment result can be found in our previous article (Saha et al., 2020). The overall pipeline of the experiment is shown in Figure 1(A).

3. Results

The results of the experiment is reported in Table 1. It shows that the CMSA identified unique 767, 2025 and 54 substitutions, deletions and SNPs while the zero insertions are common among four methods. The detailed results are provided in Table S2 in supplementary. In Table S2, the results
of each method and CMSA are provided separately. It contains coordinate position of mutation, number of occurrence of mutation in virus genome (frequency of mutation), change in nucleotide, change in amino acid, entropy to measure change at genomic location and mapping with coding region so that mutation point can be identified crisply. The venn diagram of this result is also shown in Figure 1(C) while few examples of substitution, SNP and deletion are shown in Figure 1(B) using BioEdit software.

Further, the results of CMSA are also visualized using BioCircos plot in Figure 1(D). Here whole virus genome with the coding regions is shown in the outer track while substitution, deletion, insertion and SNPs are shown in subsequent tracks where frequency of mutation is shown through bar and dot plots. It gives a complete visual representation about the occurrence of the mutations. For example, in case of SNP, coordinate positions like 241, 3037, 14410, 23405 have changed their nucleic acid close to 60% of the Indian virus population. Out of these four major changes, two of them are belonging in ORF1ab and one in the Spike gene. Apart from that, ORF3a, Membrane, ORF8, Nucleocapsid genes are also having SNPs. In Figure 1(E), SNPs present in more than 10% of the Indian SARS-CoV-2 population is shown. Moreover, the aligned sequences, code and additional results are provided in supplementary website.3

4. Conclusion
This study shows the genetical variability of Indian SARS-CoV-2 genomes using consensus of multiple sequence alignment techniques. The analysis shows 767, 2025 and 54 unique substitutions, deletions and SNPs in Indian SARS-CoV-2 genomes. More precisely, out of 54 SNPs, 4 SNPs are present close to the 60% of the virus population. These mutations are non-synonymous in nature. The reason behind of these frequent mutations needs to study in future research. However, the motivation behind the finding of SNPs is mainly to recognise the genomic locations that can be targeted to categorise the virus strain in India. Not withstanding these, once the proper strain of the virus has been identified, these SNPs can be used to define the correct dose of vaccine. In future, these SNPs can be used for protein modelling so that drugs can be designed to target such proteins. This is a very promising area of research in which we are currently working on. We hope the outcomes of our work can help the other researchers as well.

3 http://www.nitttrkol.ac.in/indrajit/projects/COVID-ConsensusMutation-India/
Ethics approval and consent to participate

The ethical approval or individual consent was not applicable.

Availability of data and Supplementary materials

The ClustalW, MUSCLE, ClustalO and MAFFT aligned 566 Indian SARS-CoV-2 genomes with reference and consensus genomes, software to find mutation and supplementary are available at “http://www.nitttrkol.ac.in/indrajit/projects/COVID-ConsensusMutation-India/”. Moreover, Indian SARS-CoV-2 genomes used in this work are publicly available at GISAID database.

Funding

This work has been partially supported by CSIR short term research grant on COVID-19 (CVD/2020/000991) from Science and Engineering Research Board (SERB), Department of Science and Technology, Govt. of India.

Role of funding source

No funding source had a role in the writing of the manuscript or the decision to submit it for publication. No author was paid by a pharmaceutical company or other agency to write this article.

Conflict of interest

The authors declare that they have no conflict of interest.

Author contributions

Indrajit Saha: Conceptualization; Data curation; Supervision; Funding acquisition; Formal analysis; Investigation; Methodology; Project administration; Resources; Validation; Visualization; Writing - original draft; Writing - review & editing. Nimisha Ghosh: Formal analysis; Software; Validation; Visualization; Writing - review & editing. Debasree Maity:
Acknowledgement

We thank all those who have contributed sequences to GISAID database and reviewers for the valuable comments to improve the article.

References

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research, 32*, 17921-1797. doi:10.1093/nar/gkh340.

Jeon, J. S., Won, Y. H., Kim, I. K., Ahn, J. H., Shin, O. S., Kim, J. H., & Lee, C. H. (2016). Analysis of single nucleotide polymorphisms among Varicella-Zoster Virus and identification of vaccine-specific sites. *Virology, 496*, 277–286. doi:10.1016/j.virol.2016.06.017.

Katoh, K., Rozewicki, J., & Yamada, K. D. (2019). MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Briefings in Bioinformatics, 20*, 1160–1166. doi:10.1093/bib/bbx108.

Lu, I.-N., Muller, C. P., & He, F.-Q. (2020). Applying next-generation sequencing to unravel the mutational landscape in viral quasispecies. *Virus Research, 283*, 197963. doi:10.1016/j.virusres.2020.197963.

Phan, T. (2020). Genetic diversity and evolution of SARS-CoV-2. *Infection, Genetics and Evolution, 81*, 104260. doi:10.1016/j.meegid.2020.104260.

Poland, G. A. (2020). Tortoises, hares, and vaccines: A cautionary note for SARS-CoV-2 vaccine development. *Vaccine, 38*, 4219-4220. doi:10.1016/j.vaccine.2020.04.073.

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, (p. 104457).*

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Soding, J., Thompson, J. D., & Higgins, D. G. (2011). Fast, scalable
generation of high-quality protein multiple sequence alignments using clustal omega.

*Molecular System Biology*, 7. doi:10.1038/msb.2011.75.

Thompson, J. D., Higgins, D. G., & Gibson, T. J. (1994). Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research*, 22, 4673–4680. doi:10.1093/nar/22.22.4673.

Vellingiri, B., Jayaramayya, K., Iyer, M., Narayanasamy, A., Govindasamy, V., Giridharan, B., Ganesan, S., Venugopal, A., Venkatesan, D., Ganesana, H., Rajagopalan, K., Rahman, P. K., Cho, S.-G., Kumar, N. S., & Subramaniam, M. D. (2020). COVID-19: A promising cure for the global panic. *Science of the Total Environment*, 725, 138277. doi:10.1016/j.scitotenv.2020.138277.

Zhou, P., Yang, X. L., Wang, X. G., Hu, B., Zhang, L., Zhang, W., Si, H. R., Zhu, Y., Li, B., Huang, C., Chen, H., J.Chen, Y.Luo, Guo, H., R Jiang, Liu, M., Chen, Y., Shen, X., Wang, X., Zheng, X., Zhao, K., Chen, Q., F. Deng, J. L., Yan, B., Zhan, F., Wang, Y., Xiao, G., & Shi, Z. (2020). A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*, 579, 270–273. doi:10.1038/s41586-020-2012-7.

Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., Zhao, X., Huang, B., Shi, W., Lu, R., Niu, P., Zhan, F., Ma, X., Wang, D., Xu, W., Wu, G., Gao, G. F., & Tan, W. (2020). A novel coronavirus from patients with pneumonia in china, 2019. *The New England Journal of Medicine*, 382, 727–733. doi:10.1056/NEJMoA2001017.

| Method   | All Mutations | Substitution | Deletion | Insertion | SNP |
|----------|---------------|--------------|----------|-----------|-----|
| ClustalW | 3384          | 933          | 2449     | 2         | 64  |
| MUSCLE   | 3344          | 848          | 2494     | 2         | 65  |
| ClustalO | 3397          | 893          | 2502     | 2         | 66  |
| MAFFT    | 3396          | 888          | 2506     | 2         | 66  |
| CMSA     | 2792          | 767          | 2025     | 0         | 54  |

Table 1: Mutation results of different methods on Indian SARS-CoV-2 Genomes
| Coordinate of Mutation | Occurrence of Mutation in Genomes | Type of Mutation | Change in Nucleotide | Change in Amino Acid | Avg. Entropy | Mapped with Coding Region |
|------------------------|-----------------------------------|-----------------|---------------------|---------------------|-------------|--------------------------|
| 241                    | 342                               | Substitution    | C>T                 | S>L                 | 0.7012      | 5’-UTR                   |
| 3037                   | 340                               | Substitution    | C>T                 | S>F                 | 0.6944      | ORF1ab                   |
| 14410                  | 332                               | Substitution    | C>T                 | P>L                 | 0.7174      | ORF1ab                   |
| 18879                  | 117                               | Substitution    | C>T                 | S>F                 | 0.5216      | ORF1ab                   |
| 22446                  | 69                                | Substitution    | C>T                 | T>I                 | 0.3702      | Spike                    |
| 23405                  | 334                               | Substitution    | A>G                 | D>G                 | 0.7125      | Spike                    |
| 23931                  | 165                               | Substitution    | C>T                 | T>I                 | 0.6789      | Spike                    |
| 25565                  | 122                               | Substitution    | G>T                 | F>S                 | 0.5207      | ORF3a                    |
| 26737                  | 112                               | Substitution    | C>T                 | T>I                 | 0.4969      | Membrane                 |
| 28313                  | 174                               | Substitution    | C>T                 | P>L                 | 0.6883      | Nucleocapsid             |
| 28856                  | 71                                | Substitution    | C>T                 | S>L                 | 0.3899      | Nucleocapsid             |
| 28883                  | 64                                | Substitution    | G>A                 | R>K                 | 0.3825      | Nucleocapsid             |
| 28884                  | 64                                | Substitution    | G>A                 | G>N                 | 0.3848      | Nucleocapsid             |
| 28885                  | 64                                | Substitution    | G>C                 | G>T                 | 0.3848      | Nucleocapsid             |

Table 2: Mutation as SNP present in more than 10% of population of Indian SARS-CoV-2 genomes

Figure 1: (A) Pipeline of the workflow (B) Examples of mutations like substitution, deletion and SNP (C) Venn diagram to represent the consensus results of four alignment techniques (D) BioCircos plot to represent the whole virus genome with the frequency of mutations in different tracks (E) SNPs present in more than 10% of population of Indian SARS-CoV-2 genomes
Figure 1

A. Input of 566 Indian SARS-CoV-2 (COVID-2019) Genomes

- ClustalW
- MUSCLE
- ClustalO
- MAFFT

Preparation of Consensus Genome from Aligned set of Genomes for each method

Use of four Consensus Genomes to Identify the different sets of Mutations as Substitution, Deletion, Insertion and SNP for each method

Consensus of four Mutation results to get the final set of Substitution, Deletion, Insertion and SNP

B. Example of Substitutions and SNPs

C. Example of Deletions

D. Tracks with Coding Regions, Substitutions, Deletions, Insertions and SNPs

E. % of virus population vs SNP Coordinates

Figure 1