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.

Infection, Genetics and Evolution 85 (2020) 104522

Contents lists available at ScienceDirect

Infection, Genetics and Evolution

journal homepage: www.elsevier.com/locate/meegid

Short communication

Inferring the genetic variability in Indian SARS-CoV-2 genomes using consensus of multiple sequence alignment techniques

Indrajit Saha[a,*,1], Nimisha Ghosh[b,1], Debasree Maity[c], Nikhil Sharma[d], Kaushik Mitra[e]

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

ARTICLE INFO

Keywords:
Multiple sequence alignment
Point mutation
SNP
SARS-CoV-2

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 genetic variability. This has motivated us to analyse 566 Indian SARS-CoV-2 sequences using multiple sequence alignment techniques viz. ClustalW, MUSCLE, ClustalO and MAFFT to align and subsequently 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 prepared 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 virus classification, designing and defining the dose of vaccine for the Indian population.

1. Introduction

Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2) causes a disease COVID-19 which 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 genetic 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 genomic 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 genomic 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. Subsequently, these aligned results are used 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 prepared 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

* Corresponding author.
E-mail address: indrajit@nitttrkol.ac.in (I. Saha).
[1] Equally contributed.

https://doi.org/10.1016/j.meegid.2020.104522
Received 10 July 2020; Received in revised form 31 July 2020; Accepted 27 August 2020
Available online 01 September 2020

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 Fig. S1(A) and their coordinates in Table S1 in supplementary. This is important to mention that the virus has a novel strain and 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)[2] in fasta format. It contains 566 complete and near complete genomes with sequence ID. The average length of the 566 genomes is 29,831 bp. In addition to this, we have downloaded the Reference Sequence (NC_045512.2)[3] 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 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. Thereafter, 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. Thereafter, 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 align and using these aligned sequences, identify the mutation in Indian virus sequences. These methods and the subsequent identification technique provide lists of mutations as substitution, deletion, insertion and SNP. Thereafter, a consensus of these results, called as Consensus Multiple Sequence Alignment (CMSA), is prepared 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 technique to have the list of mutations from the alignment result can be found in our previous article (Saha et al., 2020) as well as mentioned in Fig. S1(B). The overall pipeline of the experiment is shown in Fig. 1(A).

3. Results

The results of the experiment are reported in Table 1. It shows that the CMSA identifies 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 reported 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. Moreover, the venn diagrams of these results are shown in Fig. 1(C) while few examples of substitution, SNP and deletion are shown in Fig. 1(B) using BioEdit software.

Further, the results of CMSA are also visualized using BioCircos plot in Fig. 1(D). Here whole virus genome with the coding regions is shown in the outer track while substitution, deletion, insertion and SNP are visualised in subsequent tracks where frequency of mutations is illustrated through bar and dot plots. It gives a complete visual representation about the frequency of the mutations. For example, in case of SNP, at coordinate positions, 241, 3037, 14,410, 23,405, the change in nucleic acid is 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 Fig. 1(E), SNPs present in more than 10% of the Indian SARS-CoV-2 population is shown and subsequently reported in Table 2. Moreover, the aligned sequences, code and additional results are provided in supplementary website.[4].

4. Conclusion

This study shows the genetic 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 are present 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 these frequent mutations needs to be studied in future research. However, the motivation to find the SNPs is mainly to recognise the genomic locations that can be used to categorise the virus strain in India. Not withstanding these, once the strain of the virus is identified, the proper vaccine can be used. In future, these SNPs can be used for protein modelling so that drugs can be designed to target such proteins.

Ethics approval and consent to participate

The ethical approval or individual consent was not applicable.

[2] https://www.gisaid.org/

[3] https://www.ncbi.nlm.nih.gov/nuccore/1798174254

[4] http://www.nitttrkol.ac.in/indrajit/projects/COVID-ConsensusMutation-India/

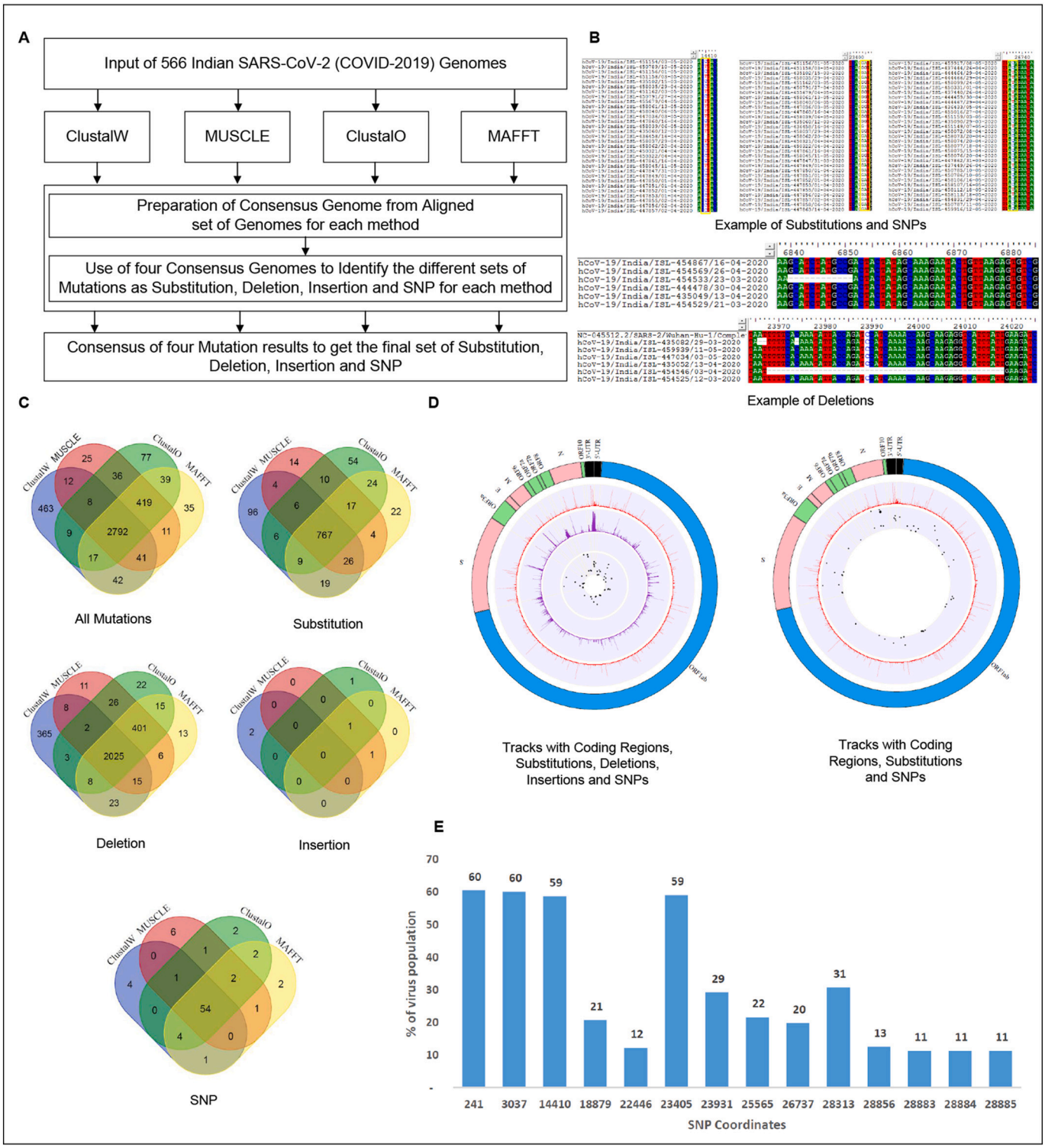

Fig. 1. (A) Pipeline of the workflow (B) Examples of mutations like substitution, deletion and SNP (C) Venn diagram to represent the consensus results of four alignment techniques (D) BioCircos plot to represent the whole virus genome with the frequency of mutations in different tracks (E) SNPs present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

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 CRG short term research grant on COVID-19 (CVD/2020/000991) from Science and Engineering Research Board (SERB), Department of Science and Technology, Govt.

Table 1
Mutation results of different methods on Indian SARS-CoV-2 Genomes.

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

Coordinate of Mutation	Frequency 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
14,410	332	Substitution	C > T	P > L	0.7174	ORF1ab
18,879	117	Substitution	C > T	S > F	0.5216	ORF1ab
22,446	69	Substitution	C > T	T > I	0.3702	Spike
23,405	334	Substitution	A > G	D > G	0.7125	Spike
23,931	165	Substitution	C > T	T > I	0.6789	Spike
25,565	122	Substitution	G > T	R > I	0.5207	ORF3a
26,737	112	Substitution	C > T	T > I	0.4969	Membrane
28,313	174	Substitution	C > T	P > L	0.6883	Nucleocapsid
28,856	71	Substitution	C > T	S > L	0.3899	Nucleocapsid
28,883	64	Substitution	G > A	R > K	0.3825	Nucleocapsid
28,884	64	Substitution	G > A	G > N	0.3848	Nucleocapsid
28,885	64	Substitution	G > C	G > T	0.3848	Nucleocapsid

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.

Declaration of Competing Interest

The authors declare that they have no conflict of interest.

Acknowledgement

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

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.meegid.2020.104522.

References

Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 17921797. https://doi.org/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 polymorphism among varicella-zoster virus and identification of vaccine-specific sites. Virology 496, 277–286. https://doi.org/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. Brief. Bioinform. 20, 11601166. https://doi.org/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 Res. 283, 197963. https://doi.org/10.1016/j.virusres.2020.197963.

Phan, T., 2020. Genetic diversity and evolution of SARS-CoV-2. Infect. Genet. Evol. 81, 104260. https://doi.org/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. https://doi.org/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. Infect. Genet. Evol. 85, 104457. https://doi.org/10.1016/j.meegid.2020.104457.

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. Mol. Syst. Biol. 7. https://doi.org/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 Res. 22, 4673–4680. https://doi.org/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. Sci. Total Environ. 725, 138277. https://doi.org/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., Chen, J., Luo, Y., Guo, H., Jiang, R., Liu, M., Chen, Y., Shen, X., Wang, X., Zheng, X., Zhao, K., Chen, Q., Deng, F., L, 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. https://doi.org/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. N. Engl. J. Med. 382, 727–733. https://doi.org/10.1056/NEJMoa2001017.