METHODOLOGY

Exploring Diversity of COVID-19 Based on Substitution Distance

Yi-Hau Chen [1]
Hsiuying Wang [2]

[1]Institute of Statistical Science, Academia Sinica, Nankang, Taipei, Taiwan; [2]Institute of Statistics, National Chiao Tung University, Hsinchu, Taiwan

Background: The number of COVID-19 infections worldwide has reached 10 million. COVID-19 caused by SARS-CoV-2 is more contagious than SARS-CoV-1. There is a dispute about the origin of COVID-19. Study results showed that all SARS-CoV-2 sequences around the world share a common ancestor towards the end of 2019.

Methods: Virus sequences from COVID-19 samples at the early time should be less diversifiable than those from samples at the later time because there might be more mutations when the virus evolutes over time. The diversity of virus nucleotide sequences can be measured by the nucleotide substitution distance. To explore the diversity of SARS-CoV-2, we use different nucleotide substitution models to calculate the distances of SARS-CoV-2 samples from 3 different areas, China, Europe, and the USA. Then, we use these distances to infer the origin of COVID-19.

Results: It is known that COVID-19 originated in Wuhan China and then spread to Europe and the USA. By using different substitution models, the distances of SARS-CoV-2 samples from these areas are significantly different. By ANOVA testing, the p-value is less than 2.2e-16. The analyzed results in most substitution models show that China has the lowest diversity, followed by Europe and lastly by the USA. This outcome coincides with the virus transmission time order that SARS-CoV-2 starts in China, then outbreaks in Europe and finally in the USA.

Conclusion: The magnitude of nucleotide substitution distance of SARS-CoV-2 is closely related to the transmission time order of SARS-CoV-2. This outcome reveals that the nucleotide substitution distance of SARS-CoV-2 may be used to infer the origin of COVID-19.

Keywords: coronavirus, diversity, nucleotide sequences, nucleotide substitution model, SARS-CoV-2

Introduction

The coronavirus disease 2019 (COVID-19) was caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). This coronavirus outbreak has spread rapidly to more than 183 countries and the number of COVID-19 infections worldwide has reached 10 million. It has caused many deaths, especially in China, the USA, and Europe. The first SARS-CoV-2 was identified in Wuhan in December 2019. Several patients in Wuhan were linked to a seafood market.[1] After that, this disease has rapidly spread from Wuhan to other cities in China and other countries. The symptoms of COVID-19 include fever, tiredness, dry cough, aches and pains, nasal congestion, runny nose, sore throat, or diarrhea. Respiratory droplet transmission is the main route of transmission, and it can also be transmitted by aerial droplets.[2]

Correspondence: Hsiuying Wang
Institute of Statistics, National Chiao Tung University, 1001 University Road, Hsinchu 30010, Taiwan
Tel +886-3-5712121 ext 56813
Email wang@stat.nctu.edu.tw

3887

A dispute over the origin of the SARS-CoV-2 has been raised. Study results showed that all SARS-CoV-2 sequences around the world sharing a common ancestor towards the end of 2019.[3] Nevertheless, it is hard to conclude the SARS-CoV-2 origin. Severe acute respiratory syndrome (SARS) was reported in the early 2000s caused by severe acute respiratory syndrome coronavirus (SARS-CoV or SARS-CoV-1). Another coronavirus causing Middle East respiratory syndrome (MERS) is called MERS-coronavirus (MERS-CoV). SARS-CoV-2 binds the same cell entry receptor—angiotensin-converting enzyme II (ACE2) as SARS-CoV.[4,5] But the spike glycoprotein (S) of MERS-CoV targets another cellular receptor, dipeptidyl peptidase 4 (DPP4).[6,7]

A study compared 6 receptor-binding domain (RBD) amino acids binding to ACE2 receptors and concluded that SARS-CoV-2 may bind human ACE2 with high affinity. This high-affinity binding of the SARS-CoV-2 spike protein to human ACE2 may be the evidence that SARS-CoV-2 is not the product of purposeful manipulation.[8] Key residues in the spike protein that make contact with the ACE2 receptor are marked with blue boxes in SARS-CoV-2 and related viruses in Figure 1.[8]

The diversity of the virus sequences in an area may be used to track back the origin of the virus. In the evolutionary process of the virus, the divergence degree of the descended sequences in an early stage may be smaller than those in a later stage. Therefore, the divergence degrees of sample sequences from different areas may be a key clue that can be used to infer the origin of the virus. A basic process in the evolution of nucleotide sequences is the substitution of one nucleotide for another during evolution. To detect the divergence degree of nucleotide sequences, we can compare the proportion of different sites of two sequences that have descended from a common ancestral sequence. When the proportion of different sites between these two sequences is

small, the chance for more than one substitution to have occurred at a site is negligible. However, if the proportion of the different sites is large, the observed number of differences is likely to be smaller than the actual number of substitutions due to multiple hits at the same site. Many nucleotide substitution models have been proposed to correct for multiple hits and the substitution distances are used to measure the divergence degree between two sequences.[9–17]

The first whole-genome sequence of SARS-CoV-2 was published on January 5, 2020.[18] Then large amounts of genomes from different countries had been sequenced and were deposited on the National Center for Biotechnology Information (NCBI) Genbank. In this study, we compare the divergence of SARS-CoV-2 sequences from the three areas, China, the USA, and Europe. We apply several different nucleotide substitution models to calculate the pairwise distances of sequences in each area and then analyze the characteristics of the distances of the three areas. The analyzed results are consistent in most of these substitution models, and the results show China has the lowest diversity, followed by Europe and lastly by the USA. Since the magnitude of the substitution model distance is consistent with the virus transmission time order, we conclude that the substitution model distance method can be used to infer the origin of COVID-19.

Coronavirus

Coronaviruses are members of the family Coronaviridae, with single-stranded RNA genomes ranging from 26 to 32 kilobases in length, the largest known viral RNA genome.[19] Coronavirus virions are spherical or pleomorphic enveloped particles containing single-stranded RNA associated with a nucleoprotein within a capsid comprised of matrix protein. Hosts of coronavirus include avian and various mammals, such as bats, camels, and dogs. They dwell in bats and wild birds and then spread

	455	486	493	494	501	505
Human-SARS-CoV-2	L	F	Q	S	N	Y
Bat-RaTG13	L	L	Y	R	D	N
Pangolin	L	F	Q	S	N	Y
Human-SARS-CoV	Y	L	N	D	T	Y
Bat-SARS-CoV-related	W	P	R	P	A	H
Bat-SARS-CoV-related	S	F	N	D	N	Y

Figure 1 Key residues in the spike protein that make contact to the ACE2 receptor are marked with blue boxes in SARS-CoV-2 and related viruses. More details are referred to Andersen, et al (2020).[8]

to other animals and humans. The virus enters the host cell, and the virus replicates in cells of the ciliated epithelium, causing cell damage and inflammation.

The coronavirus subfamily is divided into four genera, α, β, γ, and δ coronaviruses. Human coronavirus infections are caused by α and β coronavirus.[20] Avian infectious bronchitis was first described in newborn chicks in 1931. It was reported in 1968 that coronaviruses caused not only avian bronchitis but also murine hepatitis and upper respiratory tract diseases in humans.[21] The virus of avian infectious bronchitis is classified as a gammacoronavirus. These viruses are members of a group that should be called the coronaviruses.[21] These viruses are highly diverse because they are easy to mutation and recombination. The first coronaviruses that infected humans were called 229E and OC43 with a mild common cold symptom. The coronaviruses did not cause serious human infections until the outbreaks of SARS and then MERS.

Materials and Methods
Data Acquisition
The SARS-Cov-2 nucleotide sequences are sampled, aligned, and downloaded from the NCBI Virus sequence database. The sequences are aligned using the online multiple alignment tool provided by the NCBI Virus website. We also use the visual checking tool provided by the NCBI Virus website to check the quality of sequence alignments. In each of the geographic regions considered (China, the USA, Europe), the inclusion criteria for the sampling of the sequences are: (i) the sequence length >20,000 bp, and (ii) the collection date is before (and including) March 31, 2020. The sequences from China or the USA meeting the inclusion criteria are then randomly sampled to form the analysis samples of 50 sequences for the two regions; the sampling rates are around 72% (for China) and 5% (for the USA). Only 43 sequences from Europe in the NCBI database satisfy the inclusion criteria and all of them are included in our analysis sample for Europe. The reason we consider only sequences whose collection dates are before (including) March 31, 2020, is because after that date, China has few COVID-19 local cases, although there are still plenty of sequences for the other two regions (Europe and USA). Since it may be more meaningful to compare sequences from different regions within the same calendar-time frame, our analysis only includes sequences with collection date before (including) March 31, 2020, so that sufficient numbers of sequences can be obtained in the same calendar-time frame from each of the

three regions. The other criterion we set for the sequences under analysis is sequence length: we select sequences >20,000 bp, which is a length we found to result in high-quality alignments of sequences. Although it may be possible to consider only complete sequences (about 30,000 bp), this will lead to a substantial reduction of sequence numbers for the three regions. Focusing on sequences with length>20,000 bp does achieve a good balance between sequence numbers and reliable sequence alignments.

Substitution Models
The simplest and most frequently used nucleotide substitution model may be the Jukes and Cantor one-parameter model.[22] From Jukes and Cantor (1969), we have the number of substitutions per site

$$K = -\frac{3}{4}\ln\left(1 - \frac{4}{3}\hat{p}\right)$$

where $\hat{p}$ is the observed proportion of different nucleotides between the two sequences. An approximated estimator for the sampling variance is

$$V(k) = \frac{\hat{p} - \hat{p}^2}{L\left(1 - \frac{4}{3}\hat{p}\right)^2},$$

where L is the length of the sequences.

In addition to the Jukes and Cantor one-parameter model, another frequently used model is Kimura two-parameter model.[13] Let $\hat{P} = \frac{X_1}{L}$ and $\hat{Q} = \frac{X_2}{L}$ be the observed proportions of transitional and transversional differences between the two sequences, respectively, where X_1 and X_2 are the numbers of transitional and transversional differences between the two sequences. Then the number of nucleotide substitutions per site between the two sequences, K_2, is estimated by

$$K_2 = \frac{1}{2}\ln\left(\frac{1}{1 - 2\hat{P} - \hat{Q}}\right) + \frac{1}{4}\ln\left(\frac{1}{1 - 2\hat{Q}}\right)$$

The sampling variance is approximately given by

$$V(K_2) = \frac{1}{L}\left[\hat{P}\left(\frac{1}{1 - 2\hat{P} - \hat{Q}}\right)^2 \right.$$
$$+ \hat{Q}\left(\frac{1}{2 - 4\hat{P} - 2\hat{Q}} + \frac{1}{2 - 4\hat{Q}}\right)^2$$
$$\left. - \left(\frac{\hat{P}}{1 - 2\hat{P} - \hat{Q}} + \frac{\hat{Q}}{2 - 4\hat{P} - 2\hat{Q}} + \frac{\hat{Q}}{2 - 4\hat{Q}}\right)^2\right]$$

Besides the most frequently used one-parameter and two-parameter models, there are many other substitution models including the F81 model, F84 model, and TN93 model and

others.[15–17] Tajima and Nei (1982) proposed a substitution number estimator for the F81 model, Tateno, Takezaki and Nei (1994) proposed a substitution number estimator and a variance estimator for the F84 model, and Tamura and Nei (1993) proposed a substitution number estimator and a variance estimator for the TN93 model.[15–17] The substitution number estimator can be used as the distance between two nucleotide sequences. In addition, another distance, called Jaccard distance, has been successfully served as an effective tool for monitoring and tracking the epidemic of pathogenic SARS-CoV-2, and the Jaccard distance of SNP variants was adopted in the phylogenetic analysis of genomes.[23]

In this study, we use R codes to analyze the data. Different substitution models can be selected to calculate the distance between two sequences. The *BiocManager, Biostrings, msa, seqinr*, and *ape* packages are needed to be installed to run the R codes. The code *dist.dna* is used to calculate the pairwise distance between two sequences. 17 substitution models can be selected in the R code *dist.dna* to calculate the sequence distances, including raw, N, TS, TV, JC69, K80, F81, K81, F84, BH87, T92, TN93, GG95, logdet, paralin, indel, and indelblock. We include the distance formulas K and K_2 for the one-parameter model (JC69) and the two-parameter model (K80) in this paper. The details of the other 15 models are referred to R *ape* package.[24]

Methods

The three sequence groups for China, Europe, and the USA include 50 samples, 43 samples, and 50 samples, respectively. The reason that we select this sample size near 50 is that there are only 43 samples for Europe in the NCBI database. To balance the data, we randomly select 50 samples for China and the USA.

The procedure of the analyzed method is as follows.

Step 1. Select a substitution model to calculate the pairwise distance between any two sequences in each group. As a result, there are $C_2^{50} = \frac{50 \times 49}{1 \times 2} = 1225$ distances for China and the USA groups, and there are $C_2^{43} = \frac{43 \times 42}{1 \times 2} = 903$ distances for the Europe group.

Step 2. Calculate the mean and standard deviation of the distances for the three groups. Compare the means and standard deviations of these three groups.

Step 3. To test whether the means of the three groups are equal, we apply the analysis of variance (ANOVA) to test it. To test whether the means of any two groups are equal, we use a *t*-test with pooled standard deviation to test it.

In this study, we use the "dist.dna" function of the "ape" package in R software to calculate the distances among aligned sequences. There are other softwares such as MEGA that can be used to calculate the distances. Although the other software may not provide all of the 17 substitution models discussed in this study, the most frequently used models such as the one-parameter and two-parameter models can be calculated in most related software. We provided R code and the IDs of the COVID-19 sequences for our sequence distance analysis in the Supplementary materials and the publicly accessible website https://reurl.cc/GrA5dd.

Results

We use 17 models to calculate the pairwise sequence distances of China, the USA, and Europe. The means and standard deviations of the three groups are shown in Table 1. The ANOVA shows that the means of the three groups are not equal with a p-value of less than 2.2e-16. Also, we use a *t*-test with a pooled standard deviation to test the pairwise means of the three groups. The pairwise means of the three groups also are significantly different with a p-value less than 2.2e-16. Among the 17 models, 14 models show a similar result. China has the smallest mean, followed by Europe and lastly by the USA. Only the 3 models "TV", "logdet", and "indelblock" have different results. In "TV" model, China has the smallest mean, followed by the USA and lastly by Europe; in "logdet" model, the USA has the smallest mean, followed by China and lastly by Europe; in "indelblock" model, Europe has the smallest mean, followed by China and lastly by the USA. We conclude that China has the smallest mean, followed by Europe and lastly by the USA based on most of the models.

The distance between two virus sequences can be used to estimate the number of substitutions (or measure the divergence degree) between two sequences. As a result, we use the mean of these distances, but not the variance, to estimate the divergence degree of sequences. In addition, since it is reasonable to assume that the virus sequences in the early stage are less diversifiable than those in the later stage, the sequence group with the smallest mean indicates the early origin of the virus.

The boxplots of the 4 models, "JC69", "TV", "logdet", and "indelblock" are presented in (Figure 2A–D). The performances of the boxplots for the other 13 models are similar to "JC69" model. We present them in a supplementary file (Figures S1-S13). From (Figure 2A), China has the smallest mean, followed by Europe and lastly by the USA. Figure 2B

Table 1 The Means and Standard Deviations for China, Europe, and the USA

Models	Mean (Standard Deviation)			p-value** (ANOVA)
	China	Europe	USA	
raw	0.0001048666 (7.896388e-05)	0.0002572972 (1.860513e-04)	0.0003481381 (2.058586e-04)	<0.00001
N	3.119048(2.348623)	5.634551(4.074337)	6.955102 (4.112643)	<0.00001
TS	2.420068 (1.904639)	3.813953 (2.869819)	5.372245 (3.655761)	<0.00001
TV*	0.6989796 (0.9026857)	1.8205980 (1.7671983)	1.5828571 (1.4279653)	<0.00001
JC69	0.0001048781 (7.897694e-05)	0.0002573644 (1.861160e-04)	0.0003482471 (2.059534e-04)	<0.00001
K80	0.0001048810 (7.897975e-05)	0.0002573747 (1.861239e-04)	0.0003482756 (2.059788e-04)	<0.00001
F81	0.0001048783 (7.897721e-05)	0.0002573658 (1.861173e-04)	0.0003482493 (2.059553e-04)	<0.00001
K81	0.0001048812 (7.898006e-05)	0.0002573751 (1.861241e-04)	0.0003482766 (2.059792e-04)	<0.00001
F84	0.0001048816 (7.898048e-05)	0.0002573779 (1.861267e-04)	0.0003482824 (2.059848e-04)	<0.00001
BH87	0.0001079938 (8.461228e-05)	0.0002608000 (1.950313e-04)	0.0003598358 (2.233953e-04)	<0.00001
T92	0.0001048816 (7.898047e-05)	0.0002573778 (1.861266e-04)	0.0003482823 (2.059847e-04)	<0.00001
TN93	0.0001048886 (7.898629e-05)	0.0002573992 (1.861434e-04)	0.0003483246 (2.060194e-04)	<0.00001
GG95	7.778456e-05 (0.0001004559)	2.573644e-04 (0.0002498545)	3.663608e-04 (0.0003305383)	<0.00001
Logdet*	0.03027763 (8.818082e-05)	0.03196718 (1.899310e-04)	0.02781502 (2.252642e-04)	<0.00001
paralin	0.0001102436 (8.301758e-05)	0.0002670095 (1.924433e-04)	0.0003671794 (2.187228e-04)	<0.00001
indel	59.52721 (40.39569)	73.61240 (65.53877)	117.59265 (97.64568)	<0.00001
Indelblock*	2.709184 (0.8447791)	1.843854 (1.2770286)	3.005714 (0.6435705)	<0.00001

Notes: *The three models with different results are marked with gray color. **The ANOVA p-values are all less than 2.2e-16. We denote them as "<0.00001".

shows that China has the smallest mean, followed by the USA and lastly by Europe. In Figure 2C, the USA has the smallest mean, followed by China and lastly by Europe. Figure 2D shows that Europe has the smallest mean.

In this study, we apply 17 approaches to calculate pairwise distances between two sequences. Since these 17 approaches adopt different substitution models, they lead to different distances. Some substitution models infer that China has the smallest mean, and other models infer that the USA or EU has the smallest mean. Although different models lead to different results, 14 among the 17 models infer that China has the smallest mean. Thus, we based on most models to conclude that China sequences have the lowest diversity.

Discussions

SARS was reported in the early 2000s caused by severe acute respiratory syndrome coronavirus (SARS-CoV or SARS-CoV-1). Now COVID-19 caused by SARS-CoV-2 appears to be less severe than SARS but is more infectious than SARS. Generally, SARS-CoV-2 has high transmissibility and its clinical manifestations have more systematic symptoms and more severe radiological abnormalities than SARS.[25] The origin of SARS-CoV-1 was believed to have originated from China. The first SARS was reported in November 2002 in the Guangdong province of southern

China and has been linked to its wet markets. Bats have been identified as a natural reservoir for SARS-CoV-1.[26]

Compared to SARS-CoV-1, the origin of SARS-CoV-2 is more disputable. The virus was suspected to be made in a laboratory. However, analysis shows natural evolution evidence was supported by data on SARS-CoV-2's backbone.[8] Nevertheless, the virus is believed to have originated from Wuhan, China. Researchers believed that SARS-CoV-2 was most likely transmitted from bats to humans via an intermediary animal such as the pangolin. Recent research revealed that bats are the most possible host of the SARS-CoV-2 because this virus is 96% identical at the whole-genome level to a bat coronavirus.[27] Since China is the country where the coronavirus outbreak began, it is reasonable to suppose that the China SARS-CoV-2 sequence sample should have a different performance compared with other areas. From our analysis, among the 17 substitution models, 14 models show that China samples have the smallest distance mean, followed by Europe and lastly by the USA. This distance mean order is consistent with the virus transmission time order that SARS-CoV-2 starts in China, then outbreaks in Europe, and finally in the USA. This fact reveals that the SARS-CoV-2 sequence has more mutations in the later stage than in the earlier stage. The magnitude of nucleotide substitution distance of SARS-CoV-2 is closely related to the transmission time order of
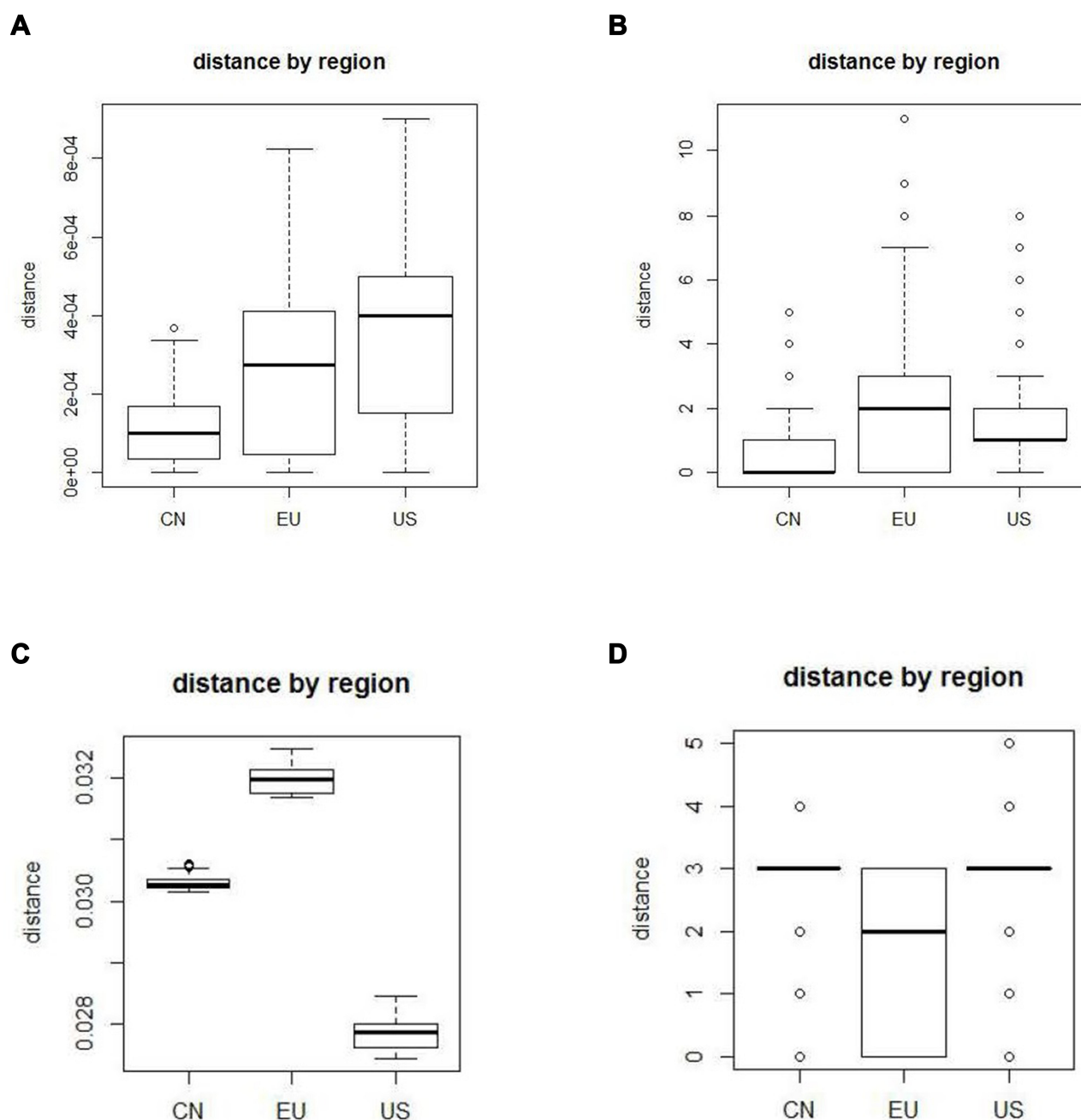
A

B

C

D

Figure 2 (**A**) The boxplot of "JC69" model, (**B**) the boxplot of " TV" model, (**C**) the boxplot of "logdet" model, (**D**) the boxplot of " indelblock" model.

SARS-CoV-2. This outcome reveals that the nucleotide substitution distance of SARS-CoV-2may be used to infer the origin of COVID-19.

Conclusions

The SARS-CoV-2 pandemic is not only a critical public health problem, but it also affects economic development that has led to serious unemployment. The global effect of SARS-CoV-2 has been far more than SARS-CoV-1. Fortunately, the plague has eased in many countries lately. Although many researchers

devoted to the SARS-CoV-2 study, so far no effective SARS-CoV-2 vaccines have been developed. One of the reasons is the diversity of SARS-CoV-2 and close to 200 recurrent genetic mutations have been identified in the virus.[3] In this study, we propose using substitution distance to infer the origin of COVID-19. Based on the samples from China, Europe, and the USA, our outcome shows that China samples have the smallest distance mean value, followed by Europe and lastly by the USA, which consists with the virus transmission time order that SARS-CoV-2 starts in China, then

outbreaks in Europe and finally in the USA. This outcome reveals that the nucleotide substitution distance of SARS-CoV-2 may be used to infer the origin of COVID-19.

Data Sharing Statement

Acknowledgments

Author Contributions

All authors made substantial contributions to conception and design, acquisition of data, or analysis and interpretation of data; took part in drafting the article or revising it critically for important intellectual content; agreed on the journal to which the article will be submitted; gave final approval of the version to be published; and agree to be accountable for all aspects of the work.

Disclosure

The authors declare no conflicts of interest for this work.

References

1. Zhu N, Zhang D, Wang W, et al. A Novel Coronavirus from Patients with Pneumonia in China, 2019. *N Engl J Med*. 2020;382(8):p. 727–-733. doi:10.1056/NEJMoa2001017
2. Jin YH, Cai L, Cheng Z-S, et al. A rapid advice guideline for the diagnosis and treatment of 2019 novel coronavirus (2019-nCoV) infected pneumonia (standard version. *Military Med Res*. 2020;7 (1):1. doi:10.1186/s40779-020-0233-6
3. van Dorp L, Acman M, Richard D, et al. Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. *J Infect Genetics Evolution*. 2020;83:104351. doi:10.1016/j.meegid.2020.104351
4. Zhou P, Yang X-L, Wang X-G, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*. 2020;579 (7798):p. 270–273. doi:10.1038/s41586-020-2012-7
5. Hoffmann M, Kleine-Weber H, Schroeder S, et al. SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor. *Cell*. 2020;181(2):271–280.e8. doi:10.1016/j.cell.2020.02.052
6. Wang N, Shi X, Jiang L, et al. Structure of MERS-CoV spike receptor-binding domain complexed with human receptor DPP4. *Cell Res*. 2013;23(8):p. 986–93. doi:10.1038/cr.2013.92

7. Lu L, Liu Q, Zhu Y, et al. Structure-based discovery of Middle East respiratory syndrome coronavirus fusion inhibitor. *Nat Commun*. 2014;5:p. 3067. doi:10.1038/ncomms4067
8. Andersen KG, et al. The proximal origin of SARS-CoV-2. *Nat Med*. 2020;26(4):p. 450–452.
9. Felsenstein J. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J Mol Evol*. 1981;17(6):p. 368–76. doi:10.1007/BF01734359
10. Felsenstein J. Distance Methods for Inferring Phylogenies: A Justification. *Evolution*. 1984;38(1):p. 16–24. doi:10.1111/j.1558-5646.1984.tb00255.x
11. Hasegawa M, Kishino H, Yano T. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J Mol Evol*. 1985;22(2):p. 160–74.
12. Jukes T, Cantor C. Evolution of protein molecules. In: Munro HN, editor. *Mammalian Protein Metabolism*. New York: Academic Press; 1969:21–132.
13. Kimura M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol*. 1980;16(2):p. 111–120. doi:10.1007/BF01731581
14. Nei M, Kumar S. *Molecular Evolution and Phylogenetics*. Oxford university press; 2000.
15. Tajima F, Nei M. Biases of the estimates of DNA divergence obtained by the restriction enzyme technique. *J Mol Evol*. 1982;18 (2):p. 115–120. doi:10.1007/BF01810830
16. Tamura K, Nei M. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol Biol Evol*. 1993;10(3):p. 512–526.
17. Tateno Y, Takezaki N, Nei M. Relative efficiencies of the maximum-likelihood, neighbor-joining, and maximum-parsimony methods when substitution rate varies with site. *Mol Biol Evol*. 1994;11(2):p. 261–277.
18. Wu F, et al. A new coronavirus associated with human respiratory disease in China. *Nature*. 2020;579(7798):p. 265–269.
19. Su S, Wong G, Shi W, et al. Epidemiology, genetic recombination, and pathogenesis of coronaviruses. *Trends Microbiol*. 2016;24(6):p. -490–502. doi:10.1016/j.tim.2016.03.003
20. Li G, Fan Y, Lai Y, et al. Coronavirus infections and immune responses. *J Med Virol*. 2020;92(4):p. 424–432. doi:10.1002/jmv.25685
21. Virology. Coronaviruses. *Nature*. 1968;220(5168):650. doi:10.1038/220650b0
22. Jukes TH, Cantor CR. *Evolution of Protein Molecules. (Munro HN Ed.) Mammalian Protein Metabolism, III*. New York: Academic Press; 1969:21–132.
23. Yin C. Genotyping coronavirus SARS-CoV-2: methods and implications. *Genomics*. 2020;112(5):p. 3588–3596. doi:10.1016/j.ygeno.2020.04.016
24. Paradis E, Schliep KJB. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *J sci biotech*. 2019;35 (3):526–528.
25. Yuen K-S, Ye ZW, Fung S-Y, et al. SARS-CoV-2 and COVID-19: the most important research questions. *Cell Biosci*. 2020;10(1):p. 1–5. doi:10.1186/s13578-020-00404-4
26. Wang L-F, Shi Z, Zhang S, et al. Review of bats and SARS. *Emerg Infect Dis*. 2006;12(12):p. 1834. doi:10.3201/eid1212.060401
27. Perlman S. Another Decade, Another Coronavirus. *N Eng J Med*. 2020;382(8):760–762. doi:10.1056/NEJMe2001126