 doi:10.20944/preprints202003.0316.v1

# Similarities and Evolutionary Relationships of COVID-19 and Related Viruses

Yanni Li[1*], Bing Liu[2*], Jiangtao Cui[1], Zhi Wang[1], Yulong Shen[1], Yueshen Xu[1], Kaicheng Yao[1], Yuanfang Guan[3]

[1] School of Computer Science and Technology, Xidian University

[2] Wangxuan Institute of Computer Technology, Peking University & University of Illinois at Chicago

[3] Department of Computational Medicine and Bioinformatics, Michigan Medicine, University of Michigan

*** Corresponding authors:** yannili@mail.xidian.edu.cn（Y. Li）, dcsliub@pku.edu.cn (B. Liu)

## Abstract

Today, we are all threatened by an unprecedented pandemic: COVID-19. How different is it from other coronaviruses? Will it be attenuated or become more virulent? Which animals may be its original host? In this study, we analyzed 377 publicly available complete genome sequences for the COVID-19 virus, the previously known flu-causing coronaviruses (HCov-229E, HCov-OC43, HCov-NL63 and HCov-HKU1) and the lethal, pathogenic P3/P4 viruses, SARS, MERS, Victoria, Lassa, Yamagata, Ebola, and Dengue. We found strong similarities between the current circulating COVID-19 and SARS and MERS, as well as COVID-19 in rhinolophines and pangolins. On the contrary, COVID-19 shares little similarity with the flu-causing coronaviruses and the other P3/P4 viruses. Strikingly, we observed divergence of COVID-19 strains isolated from human hosts has steadily increased from December 2019 to March 2020, suggesting COVID-19 is actively evolving in human hosts. From all existing human COVID-19 genome sequences, we calculated the first common model that represents the shared sequences of the human COVID-19 strains, which provides important information for vaccine and antibody development. Geographic and time-course analysis of the evolutionary trees of the human COVID-19 reveals possibly heterogeneous evolutional paths among strains from 21 countries. This finding has important implications to the management of COVID-19 and the development of vaccines.

## Introduction

Since its first report in December 2019, the severe infectious pneumonia caused by the new COVID-19 virus has spread widely from the Wuhan City, across China, and now to more than 100 countries. On March 11, WHO announced COVID-19 outbreak a pandemic, the first of its kind since the 2009 Swine Flu. Internationally, as of mid-March, 2020, COVID-19 has resulted in more than 150,000 cases and nearly 6000 deaths. COVID-19 is currently the biggest health, economical and survival threat to the entire human race. We are in urgent need to understand this virus, find treatment and develop vaccines to combat it.

One challenge in developing effective antibodies and vaccines for COVID-19 is that we do not yet understand this virus. How far away is it from other coronaviruses? Has it undergone any changes since its first discovery? These questions are critical for us to find cures and design effective vaccines, and critical for managing this virus. The study of COVID-19 began only recently[1-6]. So far, pioneering studies related to the virus have been limited to a few complete genome sequences and a few related viruses[7-8]. One study used six COVID-19 sequences from patients in Wuhan and compared them with those of SARS and MERS[9]. Another two studies used nine and five sequences respectively, and found that COVID-19 is similar to SARS[10-11]. These pioneering efforts paved the foundation for this work, which involves 377 complete genome sequences, covering 194 genomes isolated from COVID-19 in human hosts from 21 countries, 21 genomes from animals and the environment, 101 genomes from the previously known flu-

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 March 2020 doi:10.20944/preprints202003.0316.v1

causing coronavirus (HCov-229E (3 genomes), HCov-OC43 (78 genomes), HCov-NL63 (16 genomes) and HCov-HKU1 (4 genomes)), and 61 genomes from seven potentially lethal pathogenic P3/P4 viruses, SARS (11 genomes), MERS (11 genomes), Victoria (5 genomes), Lassa (6 genomes), Yamagata (5 genomes), Ebola (11 genomes), and Dengue (12 genomes). This collection allows us to analyze the evolution of COVID-19 in depth.

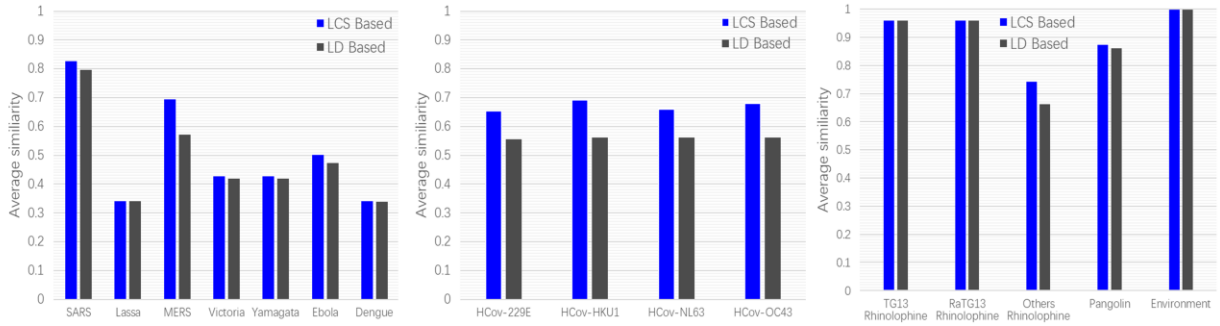
In this article, we report strong shared similarity between the currently circulating COVID-19 and the SARS virus, as well as strong shared similarities with COVID-19 in rhinolophines (especially with two strains) and in pangolins. On the contrary, COVID-19 shares a moderate sequence similarity to the flu-causing coronaviruses, despite reported similar symptoms. Strikingly, we observed the divergence of the COVID-19 strains isolated from human hosts has steadily increased from December 2019 to March 2020, suggesting COVID-19 is now actively evolving in human hosts. This may potentially explain the differences in the death rate in different areas, as the virus might have evolved into strains of different lethality. From all existing complete genome sequences of COVID-19 in humans, we derived the first common model for the COVID-19 sequences that represents the shared sequences of all 194 human COVID-19 strains, which will be critical to inform the future studies of vaccines and antibody design. Geographic and time-course analysis of the evolutionary trees of the human COVID-19 reveals heterogeneous evolutional relationships among strains from 21 countries and identified 13 virus strains that are very likely to be linked to or can potentially help the researchers find the first generation COVID-19 virus. Overall, the findings in this paper provide important information to the understanding and the management of COVID-19 and also to the development of vaccines for the virus in the near future.

## Results:

### Relatively strong similarities between COVID-19 and SARS, but relatively week similarity to several other P3/P4 viruses

There has been an active debate over whether COVID-19 is related to the SARS virus and other virulent viruses sporadically and temporarily appeared in populations. We compared the similarities between the complete genome sequences of COVID-19 and those of the seven pathogenic P3/P4 viruses (SARS, Lassa, MERS, Victoria, Yamagata, Ebola and Dengue). Two similarity metrics were employed, LCS (Eq. 1) and LD (Eq. 2) (described in Methods). The information about these viruses including the strain name, serial number, data source, sampling time, sampling location, and sequence length, is provided in Tables S8-S14 in supplementary materials. Fig. 1(A) shows the average similarities (LCS/LD) of the sequences of COVID-19 and the seven P3/P4 viruses, which are 0.825863/0.795447 for SARS, 0.341151/0.340103 for Lassa, 0.693463/0.570777 for MERS, 0.426403/0.418833 for Victoria, 0.426455/0.418873 for Yamagata, 0.500381/0.472106 for Ebola and 0.340755/0.339583 for Dengue. It can be observed that SARS have higher similarities with COVID-19, followed by MERS. We would like to highlight that COVID-19 and Ebola shares a fewer similarity. As we will see later, the similarity between COVID-19 and Ebola is even smaller than that of between COVID-19 and that of other flu-causing viruses. This inspires us to reflect about the roadmap of alternative medicines and therapies that we should develop for COVID-19.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 March 2020 doi:10.20944/preprints202003.0316.v1




(A) (B) (C)

Fig. 1. (A) average similarities between COVID-19 and seven deadly pathogenic P3/P4 viruses with human as host; (B) average similarities between the four flu coronaviruses and COVID-19 with human as host; (C) average similarities between the COVID-19 viruses in different hosts.

## Similarity between COVID-19 and other known flu-causing coronaviruses

It has been reported that many symptoms of the COVID-19 patients resemble those of the influenza patients infected with four known flu coronaviruses. Therefore, we computed the sequence similarities between COVID-19 and the four known flu coronaviruses, HCov-229E, HCov-OC43, HCov-NL63 and HCov-HKU1. The data source information of the four flu coronaviruses are provided in Tables S4-S7 in supplementary materials. The similarity matrices, computed with Eq. 1 (LCS) and Eq. 2 (LD), respectively, between the four flu viruses and COVID-19 with human as host are provided in supplementary materials (Similarity Matrices). Fig. 1(B) shows the average similarities, calculated using the LCS/LD method, between the sequences of the four known flu coronaviruses and COVID-19 that take humans as hosts, are 0.653182/0.555697 for HCov-229E, 0.680588/0.561878 for HCov-OC43, 0.659644/0.560697 for HCov-NL63 and 0.690928/0.561203 for HCov-HKU1. We observe that the difference in the similarity values between the two (LCS and LD) metrics is about 10%, but the trends of the two results are consistent.

Compared to the shared similarity between COVID-19 and the seven lethal P3/P4 strains, the similarity between COVID-19 and other known flu-causing coronaviruses is in general higher, other than SARS and MERS. This pinpoints the importance of revisiting the treatment of flus and whether drug repurposing could possibly alleviate the current COVID-19 crisis.

## COVID-19 viruses in human is not very different from COVID-19 in rhinolophine and pangolin

To help trace the original or the intermediate host of COVID-19 and to assist the finding of natural remedies, we analyzed the similarity between COVID-19 viruses in different hosts, including human, rhinolophine, pangolin, and environmentally collected strains. The detailed virus data information is provided in Tables S15-S17 in supplementary materials. Fig. 1(C) shows the similarities of the COVID-19 sequences in different hosts.

We found that the COVID-19 virus living in the environment is highly similar to that living in the human body. The average similarity can reach 0.997209/0.997148 (LCS/LD). This is expected, as this is likely to reflect what is being transmitted right now among the human population. We also found relatively strong similarities between TG13 and RaTG13 (rhinolophine host) and COVID-19 (human host), reaching 0.959925/0.958431 (LCS/LD) and 0.959992/0.958498 (LCS/LD), respectively. But the average similarities between COVID-19 with human host and the other COVID-19 virus strains with rhinolophine host are not very high, 0.741621/0.663063, lower than the similarity with the COVID-19 virus strains (pangolin host),

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 March 2020 doi:10.20944/preprints202003.0316.v1

0.874151/0.860406, by 13% and 20% (LCS/LD).

## Clustering and homologous/evolutionary relationships among viruses

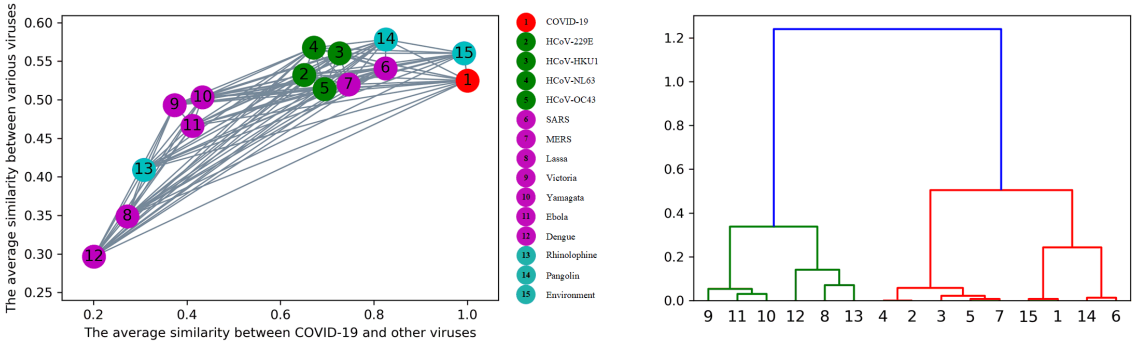
As a high similarity between sequences often implies a close relationship, to further investigate the inherent relationships among various viruses, we first compute the MLCS (Multiple Longest Common Subsequences) of all sequences from each type of virus using our I-MLCS tool (discussed in Methods) as the common models or shared representations of the 15 types of viruses. We then calculate the similarity matrix using the LCS similarity metric between the 15 types of viruses, shown in Table 1. Using the similarity matrix, we construct a fully connected weighted graph shown in Fig. 2(A) for the 15 types of viruses, where a vertex represents a type of virus, and the weight of each edge is the similarity between the two connected vertices. We then cluster the graph using the hierarchical clustering algorithm AGNES[12]. The clustering result is given in Fig. 2(B).

Fig. 2 shows that the similarities among viruses of No. 9 (Victoria), No. 10 (Yamagata) and No.11 (Ebola) are high, and the three types of viruses are in the same cluster. The similarities among No. 1 (COVID-19), No. 2 (HCoV-229E), No. 3 (HCoV-HKU1), No. 4 (HCoV-NL63), No. 5 (HCoV-229E), No. 6 (SARS), No. 7 (MERS), No. 14 (Pangolin host), and No. 15 (Environment) viruses are also high, and are in another cluster. Notice that No. 1 (COVID-19) and No. 15 (Environment) viruses have the highest similarity. No. 6 (SARS) and No. 14 (Pangolin host) viruses also have a high similarity.

Table 1. The similarity matrix of the 15 types of viruses

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 100.0 | 65.04 | 67.59 | 65.11 | 68.93 | 82.45 | 69.54 | 22.24 | 42.36 | 42.34 | 41.15 | 20.09 | 39.76 | 82.54 | 99.26 |
| 2 | 65.04 | 100.0 | 67.06 | 75.30 | 65.02 | 65.08 | 64.85 | 24.60 | 44.99 | 44.96 | 43.85 | 22.22 | 42.39 | 66.94 | 64.79 |
| 3 | 67.59 | 67.06 | 100.0 | 68.09 | 75.58 | 67.05 | 67.29 | 23.24 | 42.76 | 42.77 | 41.37 | 21.05 | 40.16 | 65.61 | 67.30 |
| 4 | 65.11 | 75.30 | 68.09 | 100.0 | 65.40 | 64.84 | 64.97 | 24.58 | 44.50 | 44.53 | 43.26 | 22.26 | 42.53 | 67.09 | 64.84 |
| 5 | 68.93 | 65.02 | 75.58 | 65.40 | 100.0 | 68.22 | 68.94 | 22.15 | 41.91 | 41.90 | 40.58 | 20.00 | 38.53 | 63.96 | 68.96 |
| 6 | 82.45 | 65.08 | 67.05 | 64.84 | 68.22 | 100.0 | 69.01 | 22.42 | 42.68 | 42.64 | 41.67 | 20.25 | 40.08 | 77.07 | 82.14 |
| 7 | 69.54 | 64.85 | 67.29 | 64.97 | 68.94 | 69.01 | 100.0 | 22.25 | 42.29 | 42.26 | 41.36 | 20.09 | 38.66 | 64.57 | 69.23 |
| 8 | 22.24 | 24.60 | 23.24 | 24.58 | 22.15 | 22.42 | 22.25 | 100.0 | 40.77 | 40.87 | 42.36 | 59.15 | 44.38 | 25.59 | 22.08 |
| 9 | 42.36 | 44.99 | 42.76 | 44.50 | 41.91 | 42.68 | 42.29 | 40.77 | 100.0 | 93.24 | 63.97 | 38.50 | 56.88 | 46.45 | 42.13 |
| 10 | 42.34 | 44.96 | 42.77 | 44.53 | 41.90 | 42.64 | 42.26 | 40.87 | 93.24 | 100.0 | 63.88 | 38.54 | 56.88 | 46.45 | 42.11 |
| 11 | 41.15 | 43.85 | 41.37 | 43.26 | 40.58 | 41.67 | 41.36 | 42.36 | 63.97 | 63.88 | 100.0 | 39.28 | 57.99 | 45.26 | 40.94 |
| 12 | 20.09 | 22.22 | 21.05 | 22.26 | 20.00 | 20.25 | 20.09 | 59.15 | 38.50 | 38.54 | 39.28 | 100.0 | 41.04 | 23.13 | 19.94 |
| 13 | 39.76 | 42.39 | 40.16 | 42.53 | 38.53 | 40.08 | 38.66 | 44.38 | 56.88 | 56.88 | 57.99 | 41.04 | 100.0 | 44.83 | 39.52 |
| 14 | 82.54 | 66.94 | 65.61 | 67.09 | 63.96 | 77.07 | 64.57 | 25.59 | 46.45 | 46.45 | 45.26 | 23.13 | 44.83 | 100.0 | 82.23 |
| 15 | 99.26 | 64.79 | 67.30 | 64.84 | 68.96 | 82.14 | 69.23 | 22.08 | 42.13 | 42.11 | 40.94 | 19.94 | 39.52 | 82.23 | 100.0 |

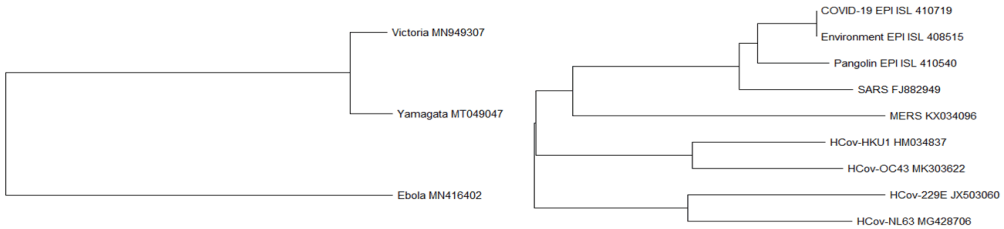
*Preprints* (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 March 2020 doi:10.20944/preprints202003.0316.v1




(A) (B)

Fig. 2. (A) the fully connected weighted graph and (B) the clustering result of the 15 types of viruses.

We may infer that different viruses that reside in the same cluster are closely related. To verify this inference, we build their evolutionary trees for these viruses' complete genome sequences in the clusters using the MEGA 6.0 tool, which are shown in Fig. 3. Fig. 3(A) shows an evolutionary tree for the cluster with the cluster-members, Nos. 9, 10 and 11, which are very similar to each other. Fig. 3(B) shows another evolutionary tree for the cluster with cluster members, Nos. 1, 2, 3, 4, 5, 6, 7, 14, and 15, which are also very similar to each other. For each cluster, we build five random trees. In building each tree, we randomly select one virus sequence from each virus type (e.g., one sequence from each of the Nos. 1, 2, 3, 4, 5, 6, 7, 14, and 15 viruses) and feed them into MEGA 6.0. As the resulting 5 trees for each cluster all have the same structure, only one tree is shown in Fig. 3(A) or Fig. 3(B). The rest are given in supplementary materials (Evolutionary trees of COVID-19 strains and viruses' cluster/Clusters' evolutionary trees). Note that since Nos. 8, 12, and 13 are far from each other (which is difficult to see in Fig. 2(A) due to the projection to the 2-dimensional space), we did not compute their evolutionary trees.




(A) (B)

Fig. 3. (A) an evolutionary tree of the cluster {Nos. 9, 10 and 11}, and (B) an evolutionary tree of the cluster {Nos.1, 2, 3, 4, 5, 6, 7, 14, and 15}

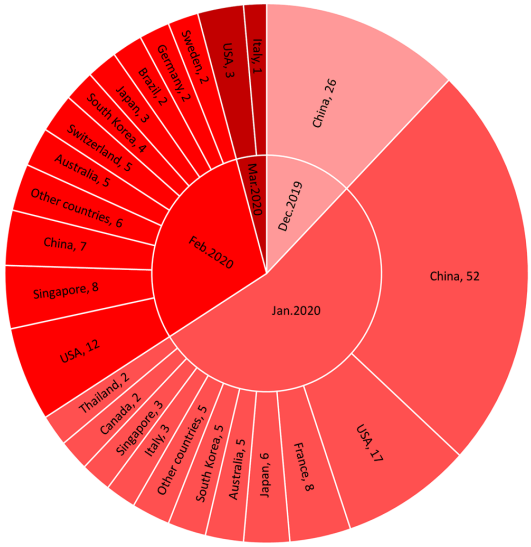
Comparing Fig. 2(B) with Fig. 3, we can clearly see that their results are consistent and Fig. 3 further reveals the homologous and evolutionary relationships of different virus sequences in the same cluster. Studying the viruses in the same cluster and their homologous and evolutionary relationships may help more accurately understand these different viruses.

## Time-course analysis of COVID-19 in human host reveals active divergence in humans from December 2019 to March 2020

We analyzed 194 publicly released complete genome sequences of the COVID-19 virus strains that take humans as hosts, which were sampled between December 2019 and March 2020 in 21 countries. The sample sources are illustrated in Fig. 4, and the details about these sequences are given in Table S3 in supplementary materials. We have extracted the common model (MLCS) (see Methods) of all sequences of COVID-19

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 March 2020 doi:10.20944/preprints202003.0316.v1

and provided them in supplementary materials (The MLCSs of COVID-19, SARS, MERS and 4 flu coronaviruses/COVID-19—MLCS.txt). These common models represent the shared information (common subsequences) across all strains for each country of the virus species. This knowledge is critical for designing effective vaccines and antibodies in the near future.




| No. | Country | Totality | No. | Country | Totality |
|---|---|---|---|---|---|
| 1 | China | 85 | 12 | Thailand | 2 |
| 2 | USA | 32 | 13 | England | 2 |
| 3 | Singapore | 11 | 14 | Brazil | 2 |
| 4 | Australia | 10 | 15 | Canada | 2 |
| 5 | Japan | 9 | 16 | Nepal | 1 |
| 6 | South Korea | 9 | 17 | Cambodia | 1 |
| 7 | France | 9 | 18 | Vietnam | 1 |
| 8 | Italy | 5 | 19 | Belgium | 1 |
| 9 | Switzerland | 5 | 20 | Mexico | 1 |
| 10 | Germany | 3 | 21 | New Zealand | 1 |
| 11 | Sweden | 2 | | | |

Fig. 4. Sources of data and sampling/sequencing times.

We calculated the average similarities between the sequences of COVID-19 strains across samples collected from different months. Although the overall similarity of these human strains is high, we observed a reduction of similarity in later months, indicating mutations within the human population is already occurring (Table 2). In Dec. 2019, the similarity was on average 0.999054, corresponding to an average of 29.75 nucleotide differences. In Mar. 2020, this number has dropped to 0.988468, corresponding to an average of 348.33 nucleotide differences. Such changes imply evolutional changes of this virus, which might result in attenuation or more virulent strains. This difference is statistically significant ($p$<0.0026). The detailed similarity analysis results (*i.e.*, similarity matrices) of the 194 COVID-19 sequences are also given in supplementary materials (Similarity Matrices). These results lead to an important conclusion: the COVID-19 has already begun its divergence in the human population, which potentially explains the underlying differences in virulence and alerts us to consider this divergence in designing antibodies and vaccines.

Table 2. Average similarities between the sequences of COVID-19 strains of different months

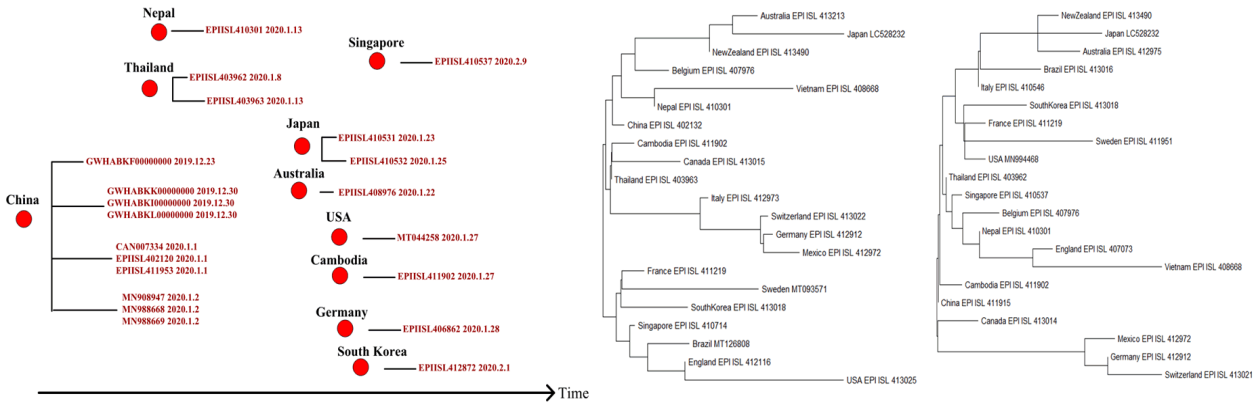
| The Average Similarity (LCS/LD similarity metric) | | | | |
|---|---|---|---|---|
| Time Period | Dec. 2019 | Jan. 2020 | Feb. 2020 | Mar. 2020 |
| Dec. 2019 | 0.999054/0.999015 | 0.998707/0.998638 | 0.997936/0.997870 | 0.992879/0.992504 |
| Jan. 2020 | 0.998707/0.998638 | 0.998462/0.998350 | 0.997784/0.997677 | 0.992709/0.992173 |
| Feb. 2020 | 0.997936/0.997870 | 0.997784/0.997677 | 0.997195/0.997085 | 0.992415/0.991771 |
| Mar. 2020 | 0.992879/0.992504 | 0.992709/0.992173 | 0.992415/0.991771 | 0.988468/0.988334 |
| The average similarity between any two sequences in the 194 COVID-19 strains is 0.995561/0.995342. | | | | |

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 March 2020 doi:10.20944/preprints202003.0316.v1

## Geographic evolution of COVID-19 in human hosts reveals possibly independent multi-route spreading of the virus

We next analyzed the evolutionary relationships of the COVID-19 strains from 21 countries, which are China, Japan, South Korea, USA, Sweden, France, Singapore, Australia, Thailand, Italy, Germany, Nepal, Cambodia, Vietnam, England, Switzerland, Mexico, Canada, Brazil, Belgium and New Zealand. Our main goal is to discover the evolutionary relationships of the virus strains from these countries, to study the spreading of the virus, and to identify the possible first generation strains.

As China is widely regarded as the origin of the virus, we first conducted 85 independent experiments, using each of the 85 sequences collected from China. For the rest 20 countries, we randomly selected one strain from each country. We also ensured that every sequence from every country has appeared in the 85 experiments at least once. The 21 resulting sequences for each experiment are then fed into MEGA 6.0 to compute the evolutionary tree of the sequences. A virus strain from any country that has appeared as a first generation strain in any of the 85 evolutionary trees is given in Fig. 5(A), with its country, serial number and sequencing date attached. We found 21 such strains from 10 countries (all in Fig. 5(A)). Fig. 5(A) also uses a time line to order and to mark these 21 strains' sequencing dates. We would like to stress that due to our experiment setting above, any sequence that can possibly be in the first generation of the COVID-19 virus has at least one chance to appear in an evolutionary tree, which means that no possible first generation strain is missed. As examples, 2 (out of 85) generated evolutionary trees are shown in Fig. 5(B) and Fig. 5(C). The rest 83 evolutionary trees are given in the supplementary material (Evolutionary trees of COVID-19 strains from 21 countries).




(A) (B) (C)

Fig. 5. (A) countries, serial numbers, and sequencing dates of the potential first-generation COVID-19 virus strains. (B) and (C) two evolutionary trees of 21 COVID-19 strains from 21 countries (one strain per country).

This above set of experiments has narrowed down the search for the first generation COVID-19 virus strains to the 21 strains (or sequences) in Fig. 5(A). However, we still do not know the relationships of these 21 strains or sequences. We thus perform another computational experiment, i.e., using only these 21 strains to build an evolutionary tree. Fig. 6 shows the output evolutionary tree. We can observe that three strains from China and the strains from Germany, Australia, Cambodia, Singapore and Nepal are not likely to be first generation strains. These exclusions are also cross-verified with the 85 evolutionary trees discussed above. The rest 13 strains (China (7), Thailand (2), Japan (2), USA (1), and South Korea (1)) are very likely to belong to and/or can lead us to find the first generation COVID-19 virus strains and the origin of the virus, for which further investigation is needed. It is also interesting to see that those strains from Japan, South Korea, and the USA appear among the first generation strains although they were sequenced much

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 March 2020 doi:10.20944/preprints202003.0316.v1

later than the others.

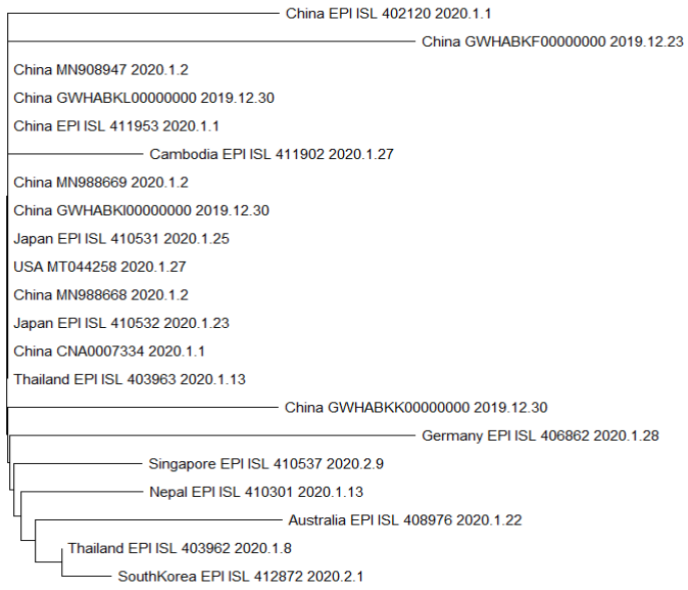



Fig. 6. The evolutionary tree of the 21 COVID-19 strains in Fig. 5(A).

## Discussion and Conclusion

Pathogenic mechanism, virus detection, and vaccine and drug developments all heavily depend on the analysis of the complete genome sequences of COVID-19. This study provides important information to support the decision making of medical and healthcare professionals in tracking COVID-19's mutation paths, developing virus detection tools, vaccines and drugs, and controlling the epidemic. We would like to reiterate several key findings.

First, the genome sequences of COVID-19 viruses in humans have already gone through mutations over the past four months. This has important implication for developing COVID-19 test kits, vaccines and antibody treatments. Recently, efforts to isolate antibodies for COVID-treatment have been announced by several pharmaceutical companies, and vaccines are being actively developed by many research institutions around the world. The breadth of the coverage of the antibodies and vaccines will be critical in determining its efficacy.

Second, COVID-19 shares little similarity with Ebola, but more with the four previously known flu-causing coronaviruses (HCov-229E, HCov-OC43, HCov-NL63 and HCov-HKU1), and even more with SARS. The sequence analysis suggests treatment to SARS and other flu-inducing coronaviruses might be another roadmap that we should explore. We recommend considering this fact during medication and treatment development.

Third, the COVID-19 virus strains from most countries may have multiple evolutional paths. Extensive analyses of the COVID-19 strains from different countries show that about 13 strains from China (7/85), Thailand (2/2), Japan (2/9), USA (1/32), and South Korea (1/9) are most likely to be linked to or can potentially lead us to find the first generation COVID-19 virus and its origin. Note that by no means do we imply that the host patients of the 13 COVID-19 virus strains are from those countries or contracted the virus in those countries due to international travels. As the data shown here, at the national scale, COVID-19 could have already been with us through multiple origins. This also highlights the need to develop more aggressive isolation and quarantine procedures for anyone demonstrating suspicious symptoms, even without direct or known contact of a patient.

*Preprints* (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 March 2020 doi:10.20944/preprints202003.0316.v1

In sum, we carried out a comprehensive sequence analysis of the complete genome sequences of COVID-19 virus as well as comparison against those of other viruses, in the hope that some of the above information will guide us to a clear roadmap for preventing further spread, treating the patients of this virus in the near future.

## Methods:

## Data collection

In this work, we collected 215 publicly available complete genome sequences of the COVID-19 virus (194 with human as host, 13 with the rhinolophine as host, 6 with the pangolin as host, and 2 in the environment), of the previously known flu-causing coronaviruses, HCov-229E (3 sequences), HCov-OC43 (78 sequences), HCov-NL63 (16 sequences) and HCov-HKU1 (4 sequences), and of seven deadly pathogenic P3/P4 viruses, SARS (11 sequences), MERS (11 sequences), Victoria (5 sequences), Lassa (6 sequences), Yamagata (5 sequences), Ebola (11 sequences), and Dengue (12 sequences). The total number of the sequences is 377. These sequences are downloaded from the following databases: GenBank or NCBI[13] (National Center for Biotechnology Information), GISAID[14] (Global Initiative on Sharing All Influenza Data), and CDC[15] (Center for Disease Control and Prevention). The average sequence length is approximately 30,000.

## Similarity metrics

As mining the big sequence data is still a difficult problem[16-19]. The key challenge is the computational complexity. We have been designing novel algorithms[17,18] to tackle this challenge. Our recent work has proposed several efficient analysis techniques for big sequences, and also developed an automated tool (I-MLCS), which is mainly used in the similarity analysis.

Based on the similarity metric design criteria and a common method for extracting subsequences among sequences in bioinformatics and computational biology[19,20], we give the following definitions and equations for computing the similarity of big sequences.

**Definition 1 (MLCS)**: The task of mining MLCS[17-20] (**M**ultiple **L**ongest **C**ommon **S**ubsequence) is to discover all *longest common subsequences* from multiple given sequences of equal length or unequal lengths. We use $d$ ($d \geq 3$) to represent the number of sequences. We call MLCS the *common model* of the $d$ sequences.

**Definition 2 (LCS)**: Mining of LCS[21] (**L**ongest **C**ommon **S**ubsequence) is to discover all LCSs of two given sequences ($d = 2$). We call LCS the *common model* of a pair of sequence $s_i$ and $s_j$.

We define an *LCS-based similarity* of a pair of sequences $s_i$ and $s_j$ as

$$sim(s_i, s_j)=|\mathrm{LCS}|/\max(|s_i|,|s_j|), \quad (1)$$

where |LCS| represents the length of the LCS's mined from the pair of sequences $s_i$ and $s_j$. $|s_i|$ and $|s_j|$ represent the lengths of sequences $s_i$ and $s_j$, respectively.

**Definition 3 (LD):** Lowenstein/edit distance LD[22,23,24] is the minimum number of operands required to convert a character sequence $s_i$ to another sequence $s_j$ using the operations of inserting, deleting or changing a character.

LD is the most commonly used measure of similarity between two sequences. The edit distance-based similarity between a pair of sequences $s_i$ and $s_j$ is defined as[20]

$$sim(s_i, s_j)=1-\mathrm{LD}/\min(|s_i|, |s_j|) \quad (2)$$

We use two similarity metrics/measures for each analysis experiment, one based on LCS (Eq. 1) and the other based on Lowenstein/edit distance (LD, Eq. 2). The two alternative results enable medical professionals and biological researchers to cross-verify or cross-compare, and possibly decide which method makes more biological sense.

We used a similarity matrix[20] to represent the similarities between a set of sequences, which can reveal some potential biological evolutionary or genetic relationships of different species quantitatively. In this study, we report all similarity matrices of the complete genome sequences of COVID-19 among themselves and also between the complete genome sequences of COVID-19 with those of the other related viruses. Notice that the similarity matrix of homogeneous sequences is a symmetric matrix, which represents pairwise comparisons between sequences of the same virus type; otherwise an asymmetric matrix, which represents pairwise comparisons between sequences of two different virus types. The average similarity between sequences in the same virus type is computed using all the elements of the upper/lower half of the symmetric similarity matrix except the diagonal elements, while the average similarity between sequences of two different virus classes is calculated using all the elements of the asymmetric similarity matrix.

## Computing platforms and tools

This paper's investigation is carried out using two main computational tools, our big sequence data analysis tool I-MLCS (for similarity analysis) and the existing MEGA 6.0 system[24] (for evolutionary relationship analysis). I-MLCS (the Integrated Multiple Longest Common Subsequence mining system) was developed based on our latest research in designing effective and efficient algorithms for analyzing big sequence data (i.e., sequences with length over $10^4$). All the calculations were done on a computing cluster of 18 nodes (Intel(R) Xeon(R) Gold 5115 CPU, 2 chip, 10 cores/chip, 2 threads/core, @2.4 GHz and 96GB RAM).

## Acknowledgments

Yanni Li's work was supported by the National Natural Science Foundation of China (No. 61472296).

## References

[1] WORLD HEALTH ORGANIZATION, et al. Laboratory testing for coronavirus disease 2019 (COVID-19) in suspected human cases: interim guidance, 2 March 2020. World Health Organization, 2020.

[2] Sahin, A. R., Erdogan, A., Agaoglu, P. M., Dineri, Y., Cakirci, A. Y., Senel, M. E., et al. 2019 Novel Coronavirus (COVID-19) Outbreak: A Review of the Current Literature. EJMO. 2020, 4(1):1-7.

[3] Peeri, N. C., Shrestha, N., Rahman, M. S., Zaki, R., Tan, Z., Bibi, S., et al. The SARS, MERS and novel coronavirus (COVID-19) epidemics, the newest and biggest global health threats: what lessons have we learned?. International Journal of Epidemiology. 2020.

[4] Kim, J. M., Chung, Y. S., Jo, H. J., Lee, N. J., Kim, M. S., Woo, S. H., et al. Identification of Coronavirus Isolated from a Patient in Korea with COVID-19. Osong Public Health and Research Perspectives. 2020, 11(1):3.

[5] Zu, Z. Y., Jiang, M. D., Xu, P. P., Chen, W., Ni, Q. Q., Lu, G. M., & Zhang, L. J. Coronavirus Disease 2019 (COVID-19): A Perspective from China. Radiology, 200490.

[6] Heymann, D. L., & Shindo, N. COVID-19: what is next for public health?. The Lancet. 2020.

[7] Bhanu, D., & Alluri, A. Analysis of Whole Genome. 2020.

[8] Cai, X. An Insight of comparison between COVID-19 (2019-nCoV) and SARS-CoV in pathology and pathogenesis. 2020.

[9] Xu, X., Chen, P., Wang, J. et al. Evolution of the novel coronavirus from the ongoing Wuhan outbreak and modeling of its spike protein for risk of human transmission. Sci. China Life Sci. 63, 457–460 (2020). https://doi.org/10.1007/s11427-020-1637-5.

[10] Lu, R., Zhao, X., Li, J., Niu, P., Yang, B., Wu, H., et al. Genomic characterization and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. The Lancet, 2020, 395(10224):542-545.

[11] Zhou P., et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature 2020 Feb 3; [e-pub]. (https://doi.org/10.1038/s41586-020-2012-7).

[12] Sobczak, G., Pikuła, M., & Sydow, M. Agnes: a novel algorithm for visualizing diversified graphical entity summarisations on knowledge graphs. In International Symposium on Methodologies for Intelligent Systems. 2012: 182-191.

[13] NCBI: https://www.ncbi.nlm.nih.gov/

[14] GISAID: https://www.gisaid.org/

[15] CDC: https://www.cdc.gov/

[16] Maier, D. The complexity of some problems on subsequences and supersequences[J]. Journal of the ACM (JACM), 1978, 25(2): 322-336.

[17] Li, Y., Wang, Y., Zhang, Z., Wang, Y., Ma, D., & Huang, J. A novel fast and memory efficient parallel MLCS algorithm for long and large-scale sequences alignments[C]//2016 IEEE 32nd International Conference on Data Engineering (ICDE). IEEE, 2016: 1170-1181.

[18] Li, Y., Li, H., Duan, T., Wang, S., Wang, Z., & Cheng, Y. A real linear and parallel multiple longest common subsequences (mlcs) algorithm[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016: 1725-1734.

[19] Sayeed, S. D., Rahman, M. S., & Rahman, A. On Multiple Longest Common Subsequence and Common Motifs with Gaps[C]//International Workshop on Algorithms and Computation. Springer, Cham, 2018: 207-215.

[20] Jin, X., Jiang, Q., Chen, Y., Lee, S. J., Nie, R., Yao, S., et al. Similarity/dissimilarity calculation methods of DNA sequences: a survey[J]. Journal of Molecular Graphics and Modelling, 2017, 76: 342-355.

[21] Hsu, W. J., & Du, M. W. Computing a longest common subsequence for a set of strings[J]. BIT Numerical Mathematics, 1984, 24(1): 45-59.

[22] Lowenstein, J. H. Differential vertex operations in Lagrangian field theory[J]. Communications in Mathematical Physics, 1971, 24(1): 1-21.

[23] Ristad, E. S., & Yianilos, P. N. Learning string-edit distance[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1998, 20(5): 522-532.

[24] MEGA: Molecular Evolutionary Genetics Analysis. https://www.megasoftware.net/, 2020

[25] Hofacker, I. L., Huynen, M. A., Stadler, P. F., & Stolorz, P. E. Knowledge Discovery in RNA Sequence Families of HIV Using Scalable Computers[C]//KDD. 1996: 20-25.

## Supplementary Materials:

Files include:

85 evolutionary trees of 21 countries

Clusters' evolutionary trees (5+5)

Similarity Matrices（132 excel files）

Tables S3~S17

The MLCSs of COVID-19, SARS, MERS and 4 flu coronaviruses (11 text files）

The most likely 1st-generation strains