Research on the Evolutionary Path of COVID-19

Yunyuan Dong*, Wei Guo and Qiong Ren
School of Mathematics and Computer Science, Jianghan University, Wuhan 430056, P.R.China
Email: yunyuan_dong@jhun.edu.cn
*Corresponding author

Abstract. World Health Organization reported that a novel coronavirus (COVID-19) was identified as the causative virus of pneumonia of unknown etiology first reported by Chinese authorities on 7 January, 2020. Our objectives were to discover the evolutionary path and evolutionary characteristics of COVID-19. As time goes on, the evolution of COVID-19 has shown a variety of trends. In this paper, we proposed a method of constructing the evolutionary network of COVID-19 in chronological order, and analysed the evolutionary tree and haplotype of COVID-19.

1. Introduction
In December of 2019, an outbreak of pneumonia caused by an unknown etiology occurred in Wuhan. The pathogen was soon identified to be a novel coronavirus, which was named 2019-nCoV by World Health Organization (WHO) [1]. And it was denoted as COVID-19 (the 2019 novel coronavirus disease) also by WHO at February 2020. With the passage of time, the COVID-19 led to 3084048 persons infected and 216780 deaths cross 212 countries or regions (till 17:00 on April 29, 2020, Beijing time). Fighting the epidemic caused by COVID-19 is a global responsibility.

Professor Yong-Zhen Zhang's team of Fudan University first identified a new type of virus and determined the genome of the RNA virus by sequencing technology, and published the genome sequence of the virus for the first time on January 10, 2020. The sequence obtained from this study can be found in the SRA database of NCBI (BioProject registration number PRJNA603194). The complete genome of the virus has been deposited in GenBank with the registration number of MN908947. This genome is now available on GenBank and an updated version has been posted as MN908947.3. And at the first time, two papers about the new coronavirus were continuously published in Nature and they confirmed the path of virus into cell [2-3].

China National Center for Bioinformation (CNCB)/National Genomics Data Center (NGDC) construct the 2019 Novel Coronavirus Resource (2019nCoVR, https://bigd.big.ac.cn/ncov), in order to promote the data sharing and make all relevant information of 2019-nCoV publicly available. 2019nCoVR integrated of genomic and proteomic sequences as well as their metadata information from the Global Initiative on Sharing All Influenza Data, National Center for Biotechnology Information, China National GeneBank, National Microbiology Data Center and CNCB/NGDC. All the data we used are all from the website 2019nCoVR [4]. As of 29 April, 2020, Beijing time, there are 13199 sequences of COVID-19 and 12946 papers on the website of 2019nCoVR.

The outbreak of COVID-19 has been over five months. It confirmed that the seafood market in Southern China is not the only source [5-7].But there are many things we haven't figured out about this virus [8]. It is of great significance to identify the source of the virus and identify the intermediate host, control the epidemic situation and prevent the outbreak again. In this paper, we proposed a method of constructing the evolutionary network of COVID-19 in chronological order, and analysed the...
evolutionary tree and haplotype of COVID-19. Hope to provide help for clinical treatment and vaccine development.

2. Data and Methods

2.1. Data
The first genome sequence of COVID-19 (MN908947.3) was published on January 10, 2020. All the results are based on the comparative analysis and identification with MN908947.3.

We choose 22 genome sequences of COVID-19, which are isolated from humans; high quality and the length are same as MN908947.3. Due to the reason of permission, we just get the 22 genome sequences from GenBank and Genome Warehouse. The 22 genome sequences came from China, Japan, Greece, U.S.A, Brazil and some other countries. Table 1 is the detailed information of all 22 genome sequences.

| NO. | Location          | Virus Strain Name                  | Collection Date |
|-----|------------------|-----------------------------------|-----------------|
| 1   | China / Hubei / Wuhan | Wuhan-Hu-1                        | 2019-12-31      |
| 2   | South Korea      | SNU01                             | 2020-01         |
| 3   | China / Yunnan   | SARS-CoV-2/Yunnan-01/human/2020/CHN | 2020-01-17     |
| 4   | United States / Washington | SARS-CoV-2/WA3-UW1/human/2020/USA | 2020-02-27     |
| 5   | China / Beijing  | SARS-CoV-2/105/human/2020/CHN     | 2020-01-26     |
| 6   | China / Beijing  | SARS-CoV-2/231/human/2020/CHN     | 2020-01-28     |
| 7   | China / Beijing  | SARS-CoV-2/233/human/2020/CHN     | 2020-01-28     |
| 8   | China / Beijing  | SARS-CoV-2/235/human/2020/CHN     | 2020-01-28     |
| 9   | United States / Washington | SARS-CoV-2/WA6-UW3/human/2020/USA | 2020-02-29     |
| 10  | United States / Washington | SARS-CoV-2/WA7-UW4/human/2020/USA | 2020-03-01     |
| 11  | Japan            | TKYE6182_2020                      | 2020-01         |
| 12  | United States / Washington | hCoV-19/USA/WA-UW205/2020        | 2020-03-13      |
| 13  | United States / WA | SARS-CoV-2/WA-UW315/human/2020/USA | 2020-03-22     |
| 14  | United States / Virginia | hCoV-19/USA/VA-DCLS-0046/2020 | 2020-03-28     |
| 15  | China / Anhui / Fuyang | SARS-CoV-2/Fuyang_FY002/human/2020/CHN | 2020-03-10 |
| 16  | United States / Washington | hCoV-19/USA/USA-UW-1595/2020 | 2020-03-19     |
| 17  | South Africa / KwaZulu-Natal | SARS-CoV-2/human/ZAF/R03006/2020 | 2020-03-07     |
| 18  | Greece           | SARS-CoV-2/human/GRC/10/2020      | 2020-03-18     |
| 19  | Greece           | SARS-CoV-2/human/GRC/13/2020      | 2020-03-18     |
| 20  | United States / Idaho | SARS-CoV-2/human/USA/ID-UW-4355/2020 | 2020-03-27     |
| 21  | United States / Idaho | SARS-CoV-2/human/USA/ID-UW-4380/2020 | 2020-03-27     |
| 22  | Brazil           | SARS-CoV-2/human/BRA/SP02cc/2020  | 2020-03-18     |

2.2. Method

2.2.1. Constructing the evolutionary network of COVID-19. We construct the matrix from the first gene sequence in chronological order. With the development of time, if the gene sites in the same position changes, it is considered that the mutation occurs; the corresponding position of the matrix is set to 1. The more sites change in one gene sequence, the larger the corresponding matrix element value.

2.2.2. Evolutionary tree. Evolutionary tree [9], also named phylogenetic tree, is a tree that shows the evolutionary relationship between species that are considered to have common ancestors. It is used to represent the results of phylogenetic research and is a method to describe the correlation between different organisms in bioinformatics. Phylogenetic analysis can help people understand the evolution
of all organisms. There is a lot of software to draw evolutionary trees, e.g. MEGA (Molecular Evolutionary Genetics Analysis) [10], PhyML (http://www.atgc-montpellier.fr/phyml/), FastTree (http://www.microbesonline.org/fasttree/), BEAST (Bayesian evolutionary analysis by sampling trees) [11] and so on. In this paper, the evolutionary tree of COVID-19 was drawn by MEGA and BEAST.

2.2.3. Haplotype analysis. Haplotype is the abbreviation of haploid genotype. Haplotype is the combination of a series of genetic mutations that coexist on a single chromosome, each of which has its own unique haplotypes [12]. During meiosis, recombination between homologous non sister chromatids can produce new haplotypes. Haplotypes contain a complete set of genetic information, which is the basis of describing individual genome and an essential aspect of genome research.

The haplotype network map can reveal the genetic distance and evolution relationship between different virus haplotypes, and the root node and leaf node indicate the variation and evolution direction of virus.

When the amount of data is large enough and the sampling randomness is good enough, the haplotypes with a large number of viruses indicate a large population, suggesting that this type of virus spreads quickly in the population.

3. Results and Discussions

3.1. The Analysis of Evolutionary Path of COVID-19

3.1.1. The changes of each site. First, we counted the changes of each site. There are 87 sites changes. For a single site, the number of changed genomes ranges from 1 to 9. There are 62 sites only have one change. In order not to affect the readability of Figure 1, we only reserve the sites where two or more genome sequence changes, as shown in Figure 1. That means some sites are conservative and not easy to change, while others are not.

![Figure 1. The genome sites changed.](image)

3.1.2. The matrix of evolutionary path. Second, we construct the matrix of evolutionary path of COVID-19. The more sites change in one gene sequence, the larger the corresponding matrix element value, as shown in Figure 2. The numbers in Figure 2 represent the corresponding genome sequence in Table 1. The first genome sequence of COVID-19 (MN908947.3) is the centre of the whole evolutionary network of COVID-19. We build networks in chronological order, so the first genome sequence is at the centre of the whole network. If more sites of a genome sequence change, it means
that it is far away from the first gene sequence. As shown in Figure 2, with the change of time, there are more and more genomes with site changes.

COVID-19 is RNA virus. It is more likely to mutate. There are many reasons for the genetic material of the virus to change. Many chemical and physical factors can be used to induce mutation, such as nitrous acid, hydroxylamine, and high temperature and so on. But at present, the information we get is limited. We only analyse it from the point of view of topology according to the order of time.

![Figure 2. The evolutionary network of COVID-19.](image)

3.2. Evolutionary Tree
Based on the available high-quality sequence variation of COVID-19, evolutionary trees were constructed by UPGMA and Bayesian evolutionary analysis.

In this paper, the evolutionary tree of COVID-19 was drawn by MEGA (Figure 3) and BEAST (Figure 4). In Figure 3, the tree scale is 0.0001, which represents the number of differences between sequences (e.g. 0.1 means 10% differences between two sequences). In Figure 4, the tree scale is 0.05, which represents days before the time of lastly sampled genomes by scale*365.

3.3. Haplotype Analysis
Haplotype network of COVID-19 genomes were used to reveal the genetic distance and evolution relationship between different COVID-19 haplotypes.

With the development of time, there are more and more virus mutations and the haplotype network of virus genome will become more and more complex, which will lead to the unrecognizability of human eyes. Therefore, we only selected the data of China up to 2020-1-27 to display, corresponding to 1 and 3, genome sequences in Table 1.

Circle (node) represents a haplotype. Nodes of different colours represent different provinces in Figure 5. The number in the node indicates the number of COVID-19 belonging to this haplotype. The line represents the distance between the two haplotypes, the longer the line, the farther the distance (the greater the difference).
As of 2020-01-17, 57 strains of virus have been sampled, with a total of 30 haplotypes. The first virus genome is one of 18 genomes in the middle of Figure 5, and the third genome is the green one in Figure 5.

**Figure 3.** Evolutionary tree by MEGA

**Figure 4.** Evolutionary tree by BEAST

**Figure 5.** Haplotypes of COVID-19 as of 2020-01-17

### 4. Conclusions

For humans, COVID-19 is an unknown virus. Its understanding is based on the hard work of countless researchers. Just as the understanding of the disease caused by the virus is constantly refreshing, so is the understanding of the virus. As we study the virus, it is mutating and evolving. Hope to provide help for clinical treatment and vaccine development, in this paper, we proposed a method of
constructing the evolutionary network of COVID-19 in chronological order, and analysed the evolutionary tree and haplotype of COVID-19.

5. Acknowledgments
This work is supported by the National Science Foundation of China (Grant No: 31600981)

6. References
[1] Hongzhou Lu, Charles W. Stratton, Yi-Wei Tang, Outbreak of pneumonia of unknown etiology in Wuhan, China: The mystery and the miracle, J Med Virol, Special Issue: 2019 Novel Coronavirus Origin, Evolution, Disease, Biology and Epidemiology: Part-I, Vol 92, Issue4, pp 401-402, 2020.
[2] Wu, F., Zhao, S., Yu, B. et al. A new coronavirus associated with human respiratory disease in China. Nature, 2020, https://doi.org/10.1038/s41586-020-2008-3
[3] Zhou, P., Yang, X., Wang, X. et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature, 2020, https://doi.org/10.1038/s41586-020-2012-7
[4] Zhao WM, Song SH, Chen ML, et al, The 2019 novel coronavirus resource, Yi Chuan, 2020, 42(2): 212–221. doi:10.16288/j.yczz.20-030 [PMID: 32102777]
[5] Liangsheng Zhang, Jian-Rong Yang, Zhenguo Zhang, Zhenguo Lin, Genomic variations of SARS-CoV-2 suggest multiple outbreak sources of transmission, 2020, doi: https://doi.org/10.1101/2020.02.25.20027953
[6] Aiping Wu, Peihua Niu, et al, Mutations, Recombination and Insertion in the Evolution of 2019-nCoV, 2020, doi: https://doi.org/10.1101/2020.02.29.971101
[7] Yu Wen-Bin, Tang Guang-Da, Zhang Li, Corlett Richard T, Decoding evolution and transmissions of novel pneumonia coronavirus using the whole genomic data, 2020. [ChinaXiv:202002.00033]
[8] David Cyranoski, Mystery deepens over animal source of coronavirus, Nature, 2020, 579, 18-19, doi: 10.1038/d41586-020-00548-w
[9] Redei, George P, Evolutionary Tree, In: Encyclopedia of Genetics, Genomics, Proteomics and Informatics. Springer, Dordrecht, 2008.
[10] Kumar S, Dudley J, Nei M & Tamura K, MEGA: A biologist-centric software for evolutionary analysis of DNA and protein sequences, Briefings in Bioinformatics, 2008, 9: 299-306.
[11] Drummond, A.J., Rambaut, A, BEAST: Bayesian evolutionary analysis by sampling trees, BMC Evol Biol, 2007, 7, 214, https://doi.org/10.1186/1471-2148-7-214.
[12] Li SS, Zhang YX, Fan CM, et al. Advances in haplotype analysis technique. Chinese Journal of Biotechnology, 2018, 34(6): 852-861.