Current Trends in OMICS

Volume 1 Issue 1, Spring 2021
ISSN(P): 2221-6510 ISSN(E): 2409-109X
Journal DOI: https://doi.org/10/32350/cto
Issue DOI: https://doi.org/10/32350/cto.11
Homepage: https://journals.umt.edu.pk/index.php/CTO/index

Article: **Evolutionary Frequency of Initially Sequenced Human Coronavirus Genomes**

Author(s): [1]Sidra Majaz, [2]Hamid ur Rahman, [1]Aamir Saeed, [1]Ashfaq Ahmad

Affiliation: [1]Department of Bioinformatics, Hazara University, Mansehra, Khyber Pakhtunkhwa, Pakistan
[2]Department of Zoology, Hazara University, Mansehra, Khyber Pakhtunkhwa, Pakistan

Article History: Received: June 12, 2021
Revised: July 15, 2021
Accepted: July 15, 2021
Available Online: August 2, 2021

Citation: Majaz S, Rahman HU, Saeed A, Ahmad A. Evolutionary frequency of initially sequenced human coronavirus genomes. *Curr Trend OMICS.* 2021;1(1):08–17. https://doi.org/10/32350/cto.11/01

Journal QR

Article QR

Sidra Majaz

A publication of the
Department of Knowledge and Research Support Services University of Management and Technology, Lahore, Pakistan

Evolutionary Frequency of Initially Sequenced Human Coronavirus Genomes

Sidra Majaz[1], Hamid ur Rahman[2], Aamir Saeed[1], Ashfaq Ahmad[1*]

[1]Department of Bioinformatics, Hazara University, Mansehra, Khyber Pakhtunkhwa, Pakistan
[2]Department of Zoology, Hazara University, Mansehra, Khyber Pakhtunkhwa, Pakistan

Abstract

A novel, human-infecting coronavirus causing COVID-19 was first identified in Wuhan, China in December 2019. Within a short span of time the virus recorded more than 1 million deaths, worldwide. This study addresses the overall evolutionary process from complete genomes of COVID-19. Addressing the complexity of the task, network-based approaches were used in mapping samples to their reported locations. A total of 473 complete human-coronavirus genomes from 20 different countries were studied, including samples from 17 states of the United States and samples from the Cruise-Diamond Princess. The phylodynamic network of a global scale was classified into five clusters containing two clusters of the samples from the USA. Cluster B was a shared cluster of samples from China and the USA, while clusters A and C were of a diverse nature. Chinese samples aggregated in clusters A and B which aided in retaining the homogeneous viral genomic pool. In contrast, samples from the USA and Spain were split into distinct clusters which indicated multiple port entries and a possibility of implying a delay in quarantine measures. In the intra-USA samples, we found that sequences reported from Washington and Virginia were scattered indicating evolutionary diversity. This report provides an insight into the transmission pattern of CoV2, which is complicated to evaluate exclusively through the conventional surveillance means.

Keywords: human-coronavirus, phylodynamic network, evolutionary diversity, genomic pool

Introduction

A novel, human-infecting coronavirus called SARS-CoV2 causing COVID-19 was first identified with the use of next-generation sequencing in Wuhan, China in late December 2019 [1]. Contagion in medical workers and family clusters was also reported confirming human-to-human transmission [2]. Patients infected with COVID-19 exhibit a high fever, sore throat, dyspnea, with invasive lesions

*Corresponding Author: ashfaqahmad82@hotmail.com, ashfaq.binfo@hu.edu.pk

present in both lungs as revealed by chest radiography [2, 3]. Within a period of 4 months the virus spread to more than 210 countries becoming an international emergency where European Region, Region of the Americas, Western Pacific and Eastern Mediterranean Region were the worst affected. As of April 13, 2020, more than 1773084 confirmed cases were reported around the world, with 111640 fatalities (www.cdc.gov). SARS-CoV2 is an RNA virus due to which it has high mutation rate which alternatively allows for estimating the underlying genealogy connecting sampled viruses [4]. SARS-CoV2 shares 96.3% of genetic similarity with the bat coronavirus RaTG13, which was obtained from bats in Yunnan in 2013 and is used as an out group in recent studies [5]. Identifying the origin and transmission pattern of such a pathogen is imperative to block the means of further spread [6].

Several approaches are being employed to combat the pandemic. Treatment with antiviral drugs, chloroquine, corticosteroids, and convalescent plasma transfusion are being tested with limited success [7-12]. Development of a potential vaccine is a time-consuming process and till then conventional public health procedures, such as isolation, quarantine, community distancing and social containment, can be used to stop the spread of this viral disease [13]. In order to successfully test these tactic phylogenetic methods they can be employed in clinical studies to investigate the pathogen spread in an individual and within communities. Moreover, understanding the global transmission and phylodynamic pattern of COVID-19 can assist in tracking undocumented infection sources and trace the route of infection transmission. New cases are being reported every day and with that sequencing data is also readily accessible. In our study we included sequence entries from 20 different countries, analyzed and mapped 473 complete CoV2 genomes and connected those through network-based distances retrieved from whole-genome sequencing.

Materials & Methods

All the sequences used in this study are retrieved from the NIH NCBI Virus database (http://www.ncbi.nlm.nih.gov/labs/virus). Entries with incomplete genome were removed and we were left with a final dataset of 473 sequences, containing 355-USA, 71-China, 18-Spain, 04-Korea, 03-Taiwan, 02 sequences each from India, Pakistan, Vietnam, Nepal, Israel and Iran and 01 sequence each from Australia, Finland, France, Peru, Brail, Japan, Sweden and Colombia. Prior to perform an alignment with MAFT online server [14], the non-genomic alphabets were removed. Including the Bat corona sequence, alignment was performed with a strategy mode keeping 1PAM/k=2 substitution matrix and 1.53

gap-penalty score. Alignment file was manually adjusted by removing the 5'-prime 30-40 nucleotides and 3'-prime poly-A sequences. The aligned dataset was transported to MEGA for generation to time-tree where neighbor-joining approach was utilized [15]. The DNASp6 packages were utilized for the data format conversion purposes [16]. The PopART version 1.7 was used to convert all the time trees into median joining network using the epsilon value of "0" and the final networks were drawn with iteration value of 5000 [17, 18]. For graphical manipulations, the Microsoft package paint.net was considered. To reproduce this data, the alignment file of all 473 genomes can be accessed from supplementary section.

Results

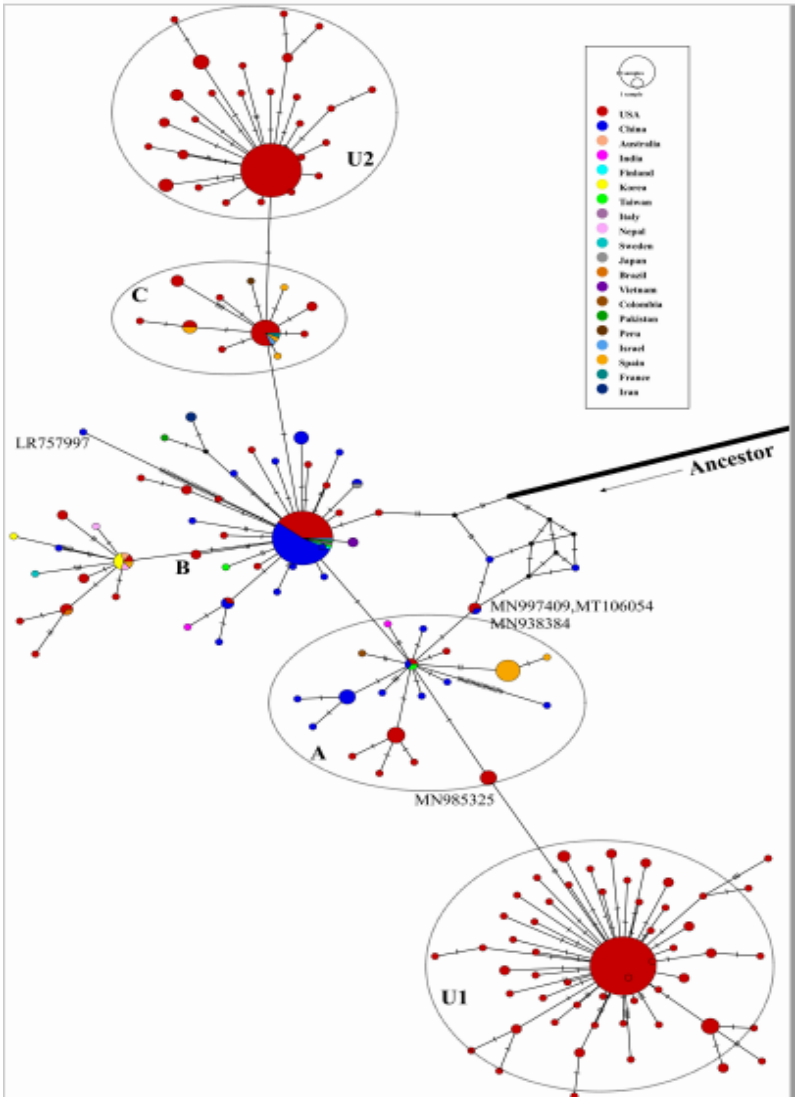
To understand the spread and evolving dynamics of CoV2, all the genomes available were mapped on NCBI virus database (www.ncbi.nlm.nih.gov/labs/virus). Total of 473 complete CoV2 genomes comprising of sequence entries from 20 different countries were selected for analyses. Based on available reports, Bat-CoV genome was used as an out group source [5]. Our analyses remained consistent with other reports which show that samples from Wuhan (MT291831) and Shenzhen/Hongkong (MN975262) are closest to the source. The former sample spread out into two clusters A and B engaging three samples (MN997409-Arizona, MT106054-Texas and MN938384-Hongkong/Shenzhen) to connect with cluster B and one sample, MT304489-Taxas for cluster A, sharing one and four mutations each (Figure 1).

For better understanding, we have classified the whole network into five clusters, where the distant U1 and U2 are rich in samples of the USA. Cluster B is mainly a shared cluster of China and USA while A and C are diverse.

The center of cluster A is shared by samples from USA, China and Taiwan while the Chinese source shares ancestry (two mutations each) to Colombian (MT256924) and Indian (MT050493) sample respectively. The sample from Taiwan provides a sole out group (MN985325) to cluster U1 which densely contains the sequences from Washington DC, USA. Cluster B is heavily centered to USA and China and provides direct descendants to Vietnam, Israel, India, Pakistan, Italy, Nepal, Australia, Sweden and Korea sharing one to four mutations. Interestingly, the Swedish sample uses Australian node rather than Chinese. Second cluster of the USA, U2 is connected to cluster B by a rather small cluster C that contained European and South American samples from Spain, France and Peru. The French sample of cluster C provides an out group to the U2 cluster that contained sequences from different states of the USA.

Collectively, our global scale CoV2 spreading dynamics indicate countries with multiple or different source entries that are assisting viral evolution at a rapid phase.

Figure 1. Global Scale Phylodynamics Mapping of the Cov2 Samples

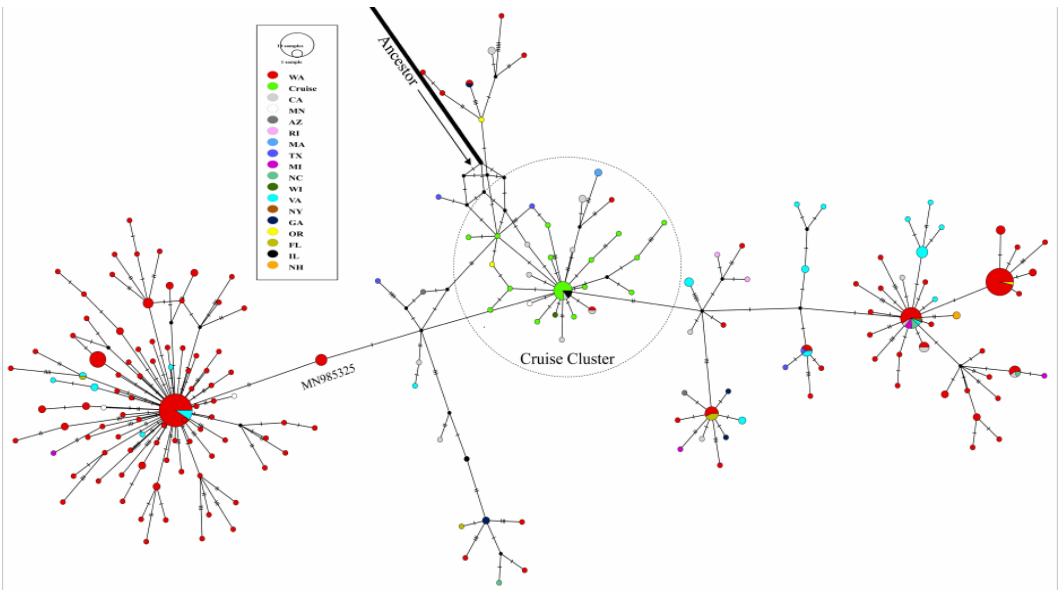


All the representative entries are colored to geographical locations. Smaller black nodes are arbitrary links while the number of cross-hatches on individual branches indicates the number of mutations. Bold representation of branch length denotes the ancestral Bat connection.

Phylodynamics of the USA

Until April 13, 2020 there were more than 400 sequences from the USA. Here, we have analyzed the 355 complete genome samples of the USA reported from seventeen different states including 24 samples from the Cruise Ship Diamond Princess that had 3771 passengers on board out of which more than 700 confirmed cases of CoV2 [19]. Since the cruise was carrying CoV2 positive patients from Hongkong, we used Bat-CoV genome as an out group. To our interest, the cruise samples grouped next to the ancestor, here we call it Cruise-cluster, along with the Cruise-cluster one sample each from Oregon (OR, MT304487) and Texas (TX, MT276331) stayed closer to the ancestor (Figure 2).

Figure 2. Phylodynamics Updates of the Cov2 Population Reported from the USA



The data includes the representation of 17 different states and samples are collected from the Cruise-diamond princess and reported from the USA. States include Washington, California, Minnesota, Arizona, Rhodes Island, Massachusetts, Texas, Michigan, North Carolina, Wisconsin, Virginia, New York, Georgia, Oregon, Florida, Illinois and New Hampshire and are represented by short names WA, CA, MN, AZ, RI, MA, TX, MI, NC, WI, VA, NY, GA, OR, FL, IL and NH and different colors.

The OR sample provides a base for one sample each for California (CA), Georgia (GA) and five for Washington (WA). The central base of the Cruise-

cluster is shared with the Arizonian sample directly infected from China (discussed above). Overall, the C-cluster shares similarity with majority of the samples from CA and further bifurcated. The left side group of WA samples is in the same group we previously mentioned as U1 and is connected by an arbitrary ancestor to the C-group suggesting that cruise samples are not the direct source for U1. Ultimately the only valid source left is from Taiwan. Similar case can be observed in the right cluster where the Cruise-cluster is not providing an actual ancestral link.

Discussion

Previously, phylodynamic is used to describe immunodynamics, epidemiology, and evolutionary biology' to understand how infectious diseases are transmitted and evolved [20]. A variety of evolutionary models assumes a tree to facilitate the testing and discussion of hypotheses. However, the increase in population size as a complex evolutionary scenario is poorly described by such models [21]. Such limitations have led to the development of a number of different types of phylogenetic networks. To estimate the evolutionary frequency of the available human CoV2 genomes and map them on to the geographical locations the study presents the analysis through median-joining network.

Analyzing the global scale evolution and spread of human CoV2, we have noticed the presence of Chinese samples only in cluster A and B highlighting the efficacy of tight quarantine practices of Chinese citizens that proved to be efficient in retaining the homogeneous viral genomic pool. On the other hand, samples from the USA were split into distinct clusters indicating multiple port entries of the virus and implying a delay in quarantine measures. Although USA had restrictions in place on all the traffic coming from China but such measures were not applied to the traffic coming from rest of the world, hence the virus was not contained as efficiently as it was contained in China. A similar phenomenon was observed in Spanish samples located in three different clusters (A, B and C) and shares ancestors from Taiwan, China, USA and Israel separately. Contrary, genomes reported from the USA population indicate that the passengers from the Cruise Diamond Princess were efficiently quarantined and treated and are not the major source for the spread of infection in the USA. The clustering of the cruise samples near the ancestral node are justified by two main reasons. Firstly, passengers were carrying the virus from the epicenter, China and secondly, they remained isolated inside the cruise which restricted viral evolution. Specifically, sequences of WA and VA have shown diversity and are scattered almost in every cluster. Overall, our data emphasizes that the CoV2 spread is higher in the USA

due to heterogeneity in viral pool when compared with the rest of the affected countries. Besides the US government needs to take some strict measures to keep the viral spread limited to the source by restricting the free movements of the citizens.

Conclusions

This study design mainly emphasizes on two factors which include, quarantine measures and viral spread across the borders. In order to present a fair idea, whole genome sequences used in this study were selected from the initial three months only. Based on the resultant clusters, our findings suggest the importance of quarantine measures. Samples from the countries which took timely measures stayed converged as compared to others. This study has used samples from the Cruise Diamond Princess as a control with other samples sequences from different states of the USA that further strengthen the notion that stringent and timely quarantine measures could have prevented the calamitous spread of virus and have saved several lives lost to this pandemic. Overall, this study design will be helpful for the policy makers and scientists to counter other pandemics in future in the developed world. This can also be considered a beacon of hope for underdeveloped countries to fight this pandemic where health facilities are limited and vaccine is currently unavailable.

Acknowledgements

The authors acknowledge all the researchers who are working hard in collecting, sequencing and data deposition of CoV2-19 samples and thank all the medical professionals who are working on the front-line fighting COVID-19.

References

[1] Zhu N, Zhang D, Wang W, et al. A Novel Coronavirus from Patients with Pneumonia in China, 2019. *New Engl J Med*. 2020;382(8):727-733.

[2] Chan JF, Yuan S, Kok KH, et al. A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. *Lancet.* 2020;395(10223):514-523. https://doi.org/10.1016/S0140-6736(20)30154-9

[3] Huang C, Wang Y, Li X, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *Lancet*. 2020;395(10223):497-506. https://doi.org/10.1016/S0140-6736(20)30183-5

[4] Drake JW. Rates of spontaneous mutation among RNA viruses. *Proc Natl Acad Sci*.1993;90(9):4171-4175. https://doi.org/10.1073/pnas.90.9.4171

[5] Forster P, Forster L, Renfrew C, Forster M. Phylogenetic network analysis of SARS-CoV-2 genomes. *Proc Natl Acad Sci*. 2020;117(17):9241-9243. https://doi.org/10.1073/pnas.2004999117

[6] Xiao C, Li X, Liu S, Sang Y, Gao SJ, Gao F. HIV-1 did not contribute to the 2019-nCoV genome. *Emerg Microbes Infect.* 2020;9(1):378-381. https://doi.org/10.1080/22221751.2020.1727299

[7] Chan KS, Lai ST, Chu CM, et al. Treatment of severe acute respiratory syndrome with lopinavir/ritonavir: a multicentre retrospective matched cohort study. *Hong Kong Med J*. 2003;9(6):399-406.

[8] Holshue ML, DeBolt C, Lindquist S, et al. First Case of 2019 Novel Coronavirus in the United States. *New Engl J Med.* 2020;382(10):929-936.

[9] Wang Z, Chen X, Lu Y, Chen F, Zhang W. Clinical characteristics and therapeutic procedure for four cases with 2019 novel coronavirus pneumonia receiving combined Chinese and Western medicine treatment. *Biosci Trends*. 2020;14(1):64-68. https://doi.org/10.5582/bst.2020.01030

[10] Savarino A, Di Trani L, Donatelli I, Cauda R, Cassone A. New insights into the antiviral effects of chloroquine. *Lancet Infect Dis.* 2006;6(2):67-69. https://doi.org/10.1016/S1473-3099(06)70361-9

[11] Jie Z, He H, Xi H, Zhi Z. Expert consensus on chloroquine phosphate for the treatment of novel coronavirus pneumonia. *Zhonghua Jie He He Hu Xi Za Zhi*. 2020;43(3):185-188. 10.3760/cma.j.issn.1001-0939.2020.03.009

[12] Gao J, Tian Z, Yang X. Breakthrough: Chloroquine phosphate has shown apparent efficacy in treatment of COVID-19 associated pneumonia in clinical studies. *Biosci Trends*. 2020;14(1):72-73. https://doi.org/10.5582/bst.2020.01047

[13] Wilder-Smith A, Freedman DO. Isolation, quarantine, social distancing and community containment: pivotal role for old-style public health measures in the novel coronavirus (2019-nCoV) outbreak. *J Travel Med*. 2020;27(2):1-4.

[14] Katoh K, Rozewicki J, Yamada KD. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Briefings Bioinform*. 2019;20(4):1160-1166. https://doi.org/10.1093/bib/ bbx108

[15] Huson DH, Bryant D. Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol.* 2006;23(2):254-267. https://doi.org/10.1093/molbev/msj030

[16] Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Mol Biol Evol.* 2018;35(6):1547-1549. https://doi.org/10.1093/molbev/msy096

[17] Rozas J, Ferrer-Mata A, Sanchez-DelBarrio JC, et al. DnaSP 6: DNA Sequence Polymorphism Analysis of Large Data Sets. *Mol Biol Evol.* 2017;34(12):3299-3302. https://doi.org/10.1093/molbev/msx248

[18] Bandelt HJ, Forster P, Rohl A. Median-joining networks for inferring intraspecific phylogenies. *Mol Biol Evol*. 1999;16(1):37-48. https://doi.org/10.1093/oxfordjournals.molbev.a026036

[19] Mizumoto K, Kagaya K, Zarebski A, Chowell G. Estimating the asymptomatic proportion of coronavirus disease 2019 (COVID-19) cases on board the Diamond Princess Cruise Ship, Yokohama, Japan, 2020. *Euro Surveill.* 2020;25(10):2000180.

[20] Frost SD, Pybus OG, Gog JR, Viboud C, Bonhoeffer S, Bedford T. Eight challenges in phylodynamic inference. *Epidemics*. 2015;10:88-92. https://doi.org/10.1016/j.epidem.2014.09.001

[21] Leigh JW, Bryant D. Popart: full-feature software for haplotype network construction. *Method in Ecology and Evolution*. 2015;6(9):1110-1116. https://doi.org/10.1111/2041-210X.12410