Scientific Research Publishing

Advances in Infectious Diseases, 2020, 10, 29-39
https://www.scirp.org/journal/aid
ISSN Online: 2164-2656
ISSN Print: 2164-2648

Comparison of Whole-Genome Sequences of COVID-19 Strains in Wuhan-China, USA and Spain to Determine Source of Epidemic

Zaidan Khlaif Imran*, Kawther Mohammed Ali Hasan

Department of Biology, College of Science for Women, University of Babylon, Babylon, Iraq
Email: *zaidan_omran@yahoo.com

How to cite this paper: Imran, Z.K. and Hasan, K.M.A. (2020) Comparison of Whole-Genome Sequences of COVID-19 Strains in Wuhan-China, USA and Spain to Determine Source of Epidemic. *Advances in Infectious Diseases*, **10**, 29-39.
https://doi.org/10.4236/aid.2020.103004

Received: April 25, 2020
Accepted: May 24, 2020
Published: May 27, 2020

Abstract

Background: The World Health Organization declared the SARS-CoV-2 outbreak a global public health emergency from Wuhan/China to others countries. **Methodology:** Genetic analyses of sixty eight complete genomes of SARS-CoV-2 (38 reference strain from China, 24 strain from USA and 7 strains from Spain and others from Japan and Korea) were performed. By the Bio Edit software a multiple alignment of the COVID-19 sequences was performed for each of America (24 strains) and China (38 strains) separately, also, multiple alignments were made with 68 sequences of strains of the virus from each of America, China and Spain. The mutation in complete genome of virus was detected. Phylogeny tree representative strains from China, USA, Spain and Japan with two MERS strains one from KSA and one from Korea as comparable. **Results:** The result shown that 31.6% had point mutations in Chinese strains, 73.9% in USA and 71.4 mutant strains in Spain. Most of the mutation occurred ORF1ab, OFR7a, S gene and ORF6a respectively in China, USA and Spain strains. **Conclusion:** The conclusion shown that there is a high genetic identity between the selected strains of virus in China-Wuhan and those that spread in America and Spain, which indicates that the epidemic center is coming from China. These observations provided evidence of the genetic diversity and rapid evolution of this novel coronavirus.

Keywords

COVID-19, USA, China, Epidemic, Multiple Alignment, Mutation, Phylogeny Tree

1. Introduction

In December 2019, a new coronavirus strain was discovered in Wuhan-China,

which is officially named COVID-19. Within two months of the discovery of the first patient, it has now out brake across China and in all areas globally [1]. More than two millions people were infection till time, the highest infection in USA, Europe and Asia especially in China [2] [3].

Due to unlock down of the city of Wuhan as the epidemic center, and many people (Students, workers, officials and visitors) traveled for business or spend a nice time in new year without caution and inadvertency from all countries in America, Europe and Asia, many events contributed speeded the outbreak of SARS Epidemic. Some of these factors not activation the quarantine clauses, the biology of virus was unknown causes high contamination with virus [4] [5]. Regularly updated information on severe acute respiratory syndrome coronavirus COVID-19 outbreak is available on ECDC's website [6], the European Commission website [7] and the World Health Organization's website [3].

Nucleotide substitution has been proposed to be one of the most significant components of viral advancement in nature [8]. The fast spread of SARS-CoV-2 brings up captivating issues, for example, regardless of whether its development is driven by changes. Since December 2019, the SARS-CoV-2 has infected more than 2 million people and caused deaths among them (with high case-fatality rate) and has spread to most world countries according to WHO reports 90 - 91 [1] [2].

Therefore, it is critical to track and characterize the COVID-19 genomic variants in different geographic locations of epidemic. In this study, aim to determination the genetic variation in coding and non-coding regions of viral genome and made comparison between China, USA and Spanish strains to give explanation of sources of epidemic.

2. Methodology

2.1. Analysis Data Methods

In this study, the sequencing data of 68 SARS-CoV-2 strains were retrieved for from genbank. A length of 3000 base of 38 SARS-CoV-2 strains from China were read separately from 24 SARS-CoV-2 virus from USA and 7 SARS-CoV-2 strains from Spain. Multiple alignment was performed for SARS-CoV-2 strains group based on Bio Edit software.

The multiple alignment nucleotide gathering course of action was performed by Bio Edit software and the succession of the strain China/Wuhan. Mu-1/2020/ was utilized as leading sequence alignment for China strains. while, the MT258379.1 USA/CZB-RR057-007/2020 was utilized for alignment USA strains.

2.2. Mutation Detection

We aligned the clean data to SARS-CoV-2 complete genomes (MN908947.3 as RefSeq as Leading sequence). Based on compartment triple codon of ORF1ab, S, ORF6a and other genes, the mutation types were determined the Missense mutation, Silent mutations and Insertion/Deletion.

2.3. Phylogenetic Analysis

Phylogenetic tree construction by the UPGMA method was performed using MEGA X software [9].

3. Results and Discussion

3.1. Epidemic Outbreak

The recent epidemic map of SARS COV2 through the world was shown that the COVID-19 outbreak globally, the SARS-CoV-2 has infected more than 5,000,000 people and caused more than 300,000 deaths, this epidemic was spread to Western Pacific Region 19 countries; European Region 60 countries; South-East Asia Region 10 countries; Eastern Mediterranean Region 22; Region of the Americas 54 countries and African Region 47 countries according to the Data as received by WHO from national authorities, 14 May 2020 (WHO Update on the Coronavirus disease 2019 (COVID-19) Outbreak, 2020) USA, Span, Italia, China representative the high infection, while other countries shown low incidence.

The highly outbreak of epidemic in USA, Russia, Spain, Italy and others countries in Europe, USA, Asia, Africa and Australia (Figure 1). The outbreak of epidemic may correlated with many factors: The epidemic outbreak is associated with the average age of the communities, the nature of the habits in those societies, congestion in mass transportation such as metro transportation, and present of high percentage of old age groups of 75 - 100 years. Another factor that helped spread the epidemic is the silence of the Chinese authorities on Disclosure of the

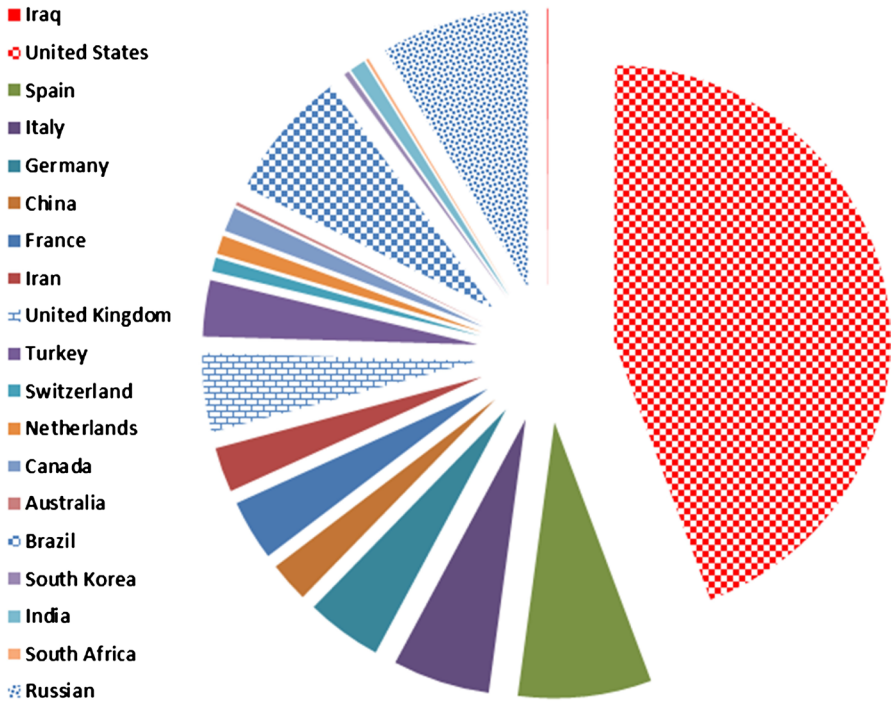

Figure 1. Schematic diagram of the outbreak of COVID-19 epidemic percentage through the world shown high percent infection in USA (44%) followed Russian and Spain (8% for each).

spread of the virus in Wuhan for many days, and this silence was associated with the Lunar New Year, continued flow of workers, students and citizens from China to various countries of the world, especially America, Italy and Spain.

The low beware response in the initial stages of the pandemic is the same aspect in the world. The response from some USA, Spain and others worlds' cities was grim, while others appeared to be holding steady in these early days of the outbreak. Without taking these precautionary measures. For example, Till March 11, 2020, New York City's subway system, the nation's largest, was down 18.65 percent, compared to the same day last year, whereas China announced a large outbreak of the epidemic in January and February and declared Quarantine in most cities.

The extension of COVID-19 epidemic through the four months shown different infection values, USA representative the highest area followed by European countries while others countries shown low values Figure 2.

3.2. Molecular Characteristics of Coronaviruses

COVID-19 are enveloped with a non-segmented, positive-sense, single-strand RNA, with size about 3000 bases, and consider as largest known genome among RNA viruses [10].

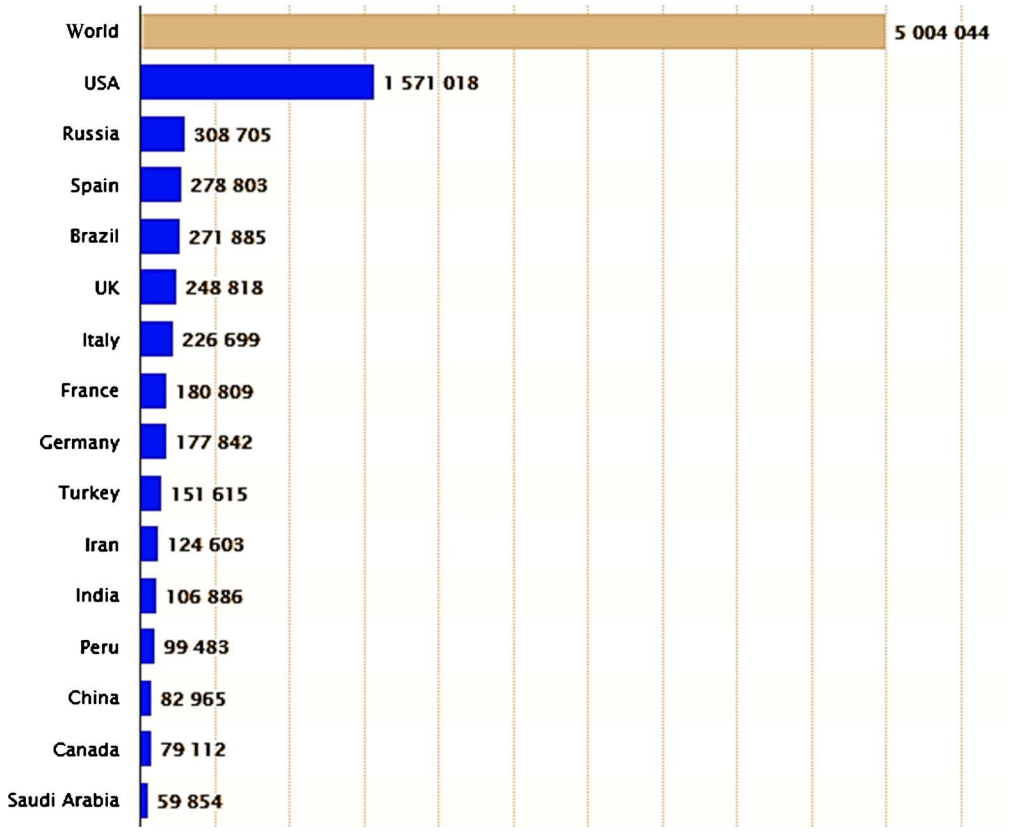

Figure 2. Number of coronavirus (COVID-19) cases worldwide as of May 20, 2020, by country. Figure cited from:
(https://www.statista.com/statistics/1043366/novel-coronavirus-2019ncov-cases-worldwide-by-country/).

The genomic structure of COVID-19 is compose of ORF1ab and S (Spike), E (Envelope), M (Membrane), N (Nucleocapsid). There are accessory genes interspersed within the structural genes at the 3' end of genome [11]. Some of point mutation some of which have been shown to may be an important roles in viral pathogenesis [12]. The S protein is responsible for receptor-binding and subsequent viral entry into host cells, the M and E proteins play important roles in viral assembly, and the N protein is necessary for RNA synthesis [13].

3.3. Distribution of Variants in SARS-CoV-2 Genome

After aligning the sequencing data to the SARS-CoV-2 genome, we detected the mutations of 38 samples from China. 24 from USA and 7 from Spain. Based on the mutations of SARS-CoV-2 in 36 samples, Most of the mutations occupied a large part of the gene S, N and M: S, N, and M. Since most of the mutations were located more than one sample, we deduced that the functions of the discovered mutation need to be further verified according to lab experiments. Surprisingly, we also found that 7 among 11 samples were heterogeneous at loci 8782. This may imply that the hosts have two different virus strains. Also, we find creditable Del. (Deletions), during the mutation analysis.

3.3.1. Multiple Sequence Alignment of 38 Chinese COVD-19 References Strains

The results of multiple sequence alignment of 38 Chinese COVID-19 shown 22 out of 38 of Chinese strain had 100% identity with reference strain MN908947.3, while 12 out of 38 strains shown 97% - 99% identity [14]. The multiple alignment shown consistent variation among COVID-19 strains. Interestingly, the variant (8785: C > T) is identified in 9 of samples. The variant (17,376: C > T) occur in 3 samples, and variant (28,147: T > C) shows consistent heterozygosity in 9 samples, while others variants occur in one or two samples. Most of variants occurred in ORF1ab gene, S gene, ORF3a gene and ORF6 **Figure 3**.

3.3.2. Multiple Sequence Alignment of 24 USA COVD-19 RefSeq Strains

The results of multiple sequence alignment of 24 USA COVID-19 strains shown 4 out of 24 of USA strain had 100% identity with reference strain MT258379.1, while 20 out of 24 strains shown 93% - 99% identity. The multiple alignment shown consistent variation among COVID-19 strains. Interestingly, the variant in non-coding region (241: T < C) in 11 samples, (313: C < T < Y) occurred in 3 samples; The variant in ORF1ab coding regions was (2446: T < C) occurred in two samples, (3037: T > C) is identified in 11 of samples, The variant (11,083: G > T) occur in 4 samples, and variant (14,407: T > C) occurred in 11 samples. In ORF6 gene the variant (25,563: G < T) occurred in 4 samples, In N gene the variant (28,144: T < C) occurred in 4 samples. All the previous variants shows consistent heterozygosity in all samples. The isolates from States: Rhode Island (RI); Arizona (AZ); Georgia (GA:MT304479.1; MT276328.1) shown 100% identity with reference leading sequence strain, while the strains of States: WA, MA,

TX, IL, GA and OR shows consistent heterozygosity in all samples (T < C) Figure 4.

3.3.3. Determined the Mutations Type Along COVID-19 Genome

Most mutation types which observed in this genetic analysis study were transition mutation (pyrimidine to Pyrimidine (C > T), transverion mutation (Purine

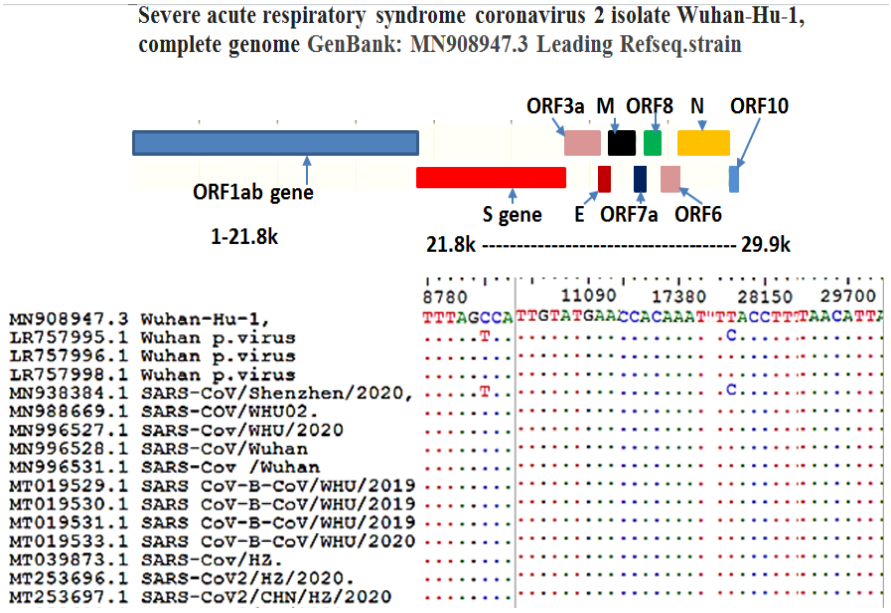

Figure 3. Multiple sequence alignment of Chinese SARS-CoV2 virus strains, shown 98% - 100% identity with reference strain MN908947.3. representative genes locations OFR1ab, S gene, ORF3a, E, M, ORFa, ORF8, ORF6, N, and ORF10.

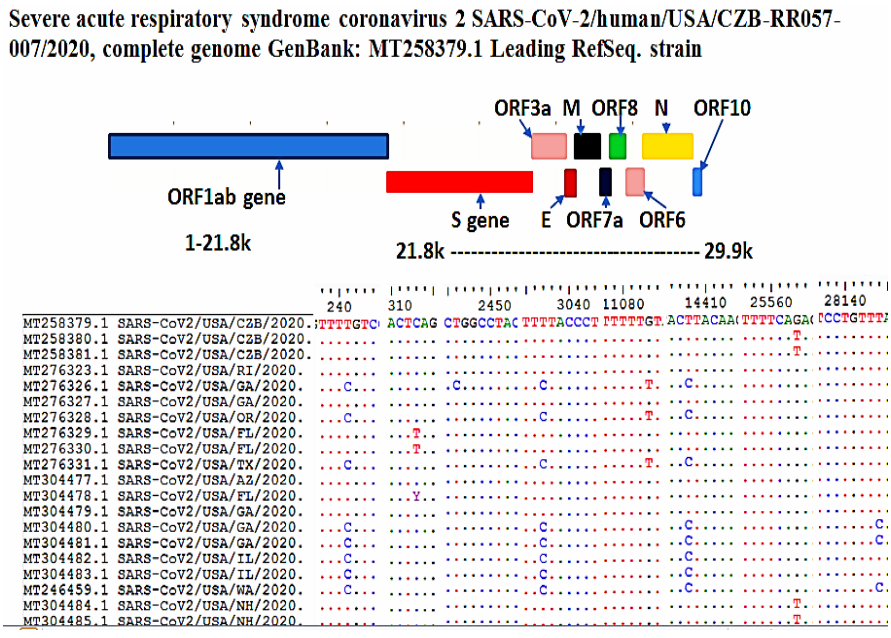

Figure 4. Multiple sequence alignment of USA SARS-CoV2 virus strains, shown 93% - 99% identity with reference strain MT258379.1. representative genes locations: OFR1ab, S gene, ORF3a, E, M, ORFa, ORF8, ORF6, N, and ORF10.

to Pyrimidine: G > T) and deletion. Most frequently mutations observed one is C > T as shown in **Figure 5(a)**. Our results consistent with the results of Variant analysis of COVID-19 genomes [15] [16].

3.3.4. Evaluation the Genetic Variations of Different Genes of 23 COVID-19 Strain from Chinese and USA

The result of multiple alignment of 23 strains from Chinese and USA was verified two aims: the first performed compartment between Chinese strains & five samples from Wuhan the center of epidemic and one from Fuyang, and USA strains most of them from Washington State (13 samples from WA), two samples from MN and CZB for each.

The results shown that all the six Chinese samples have 100% identity with reference sequence leading strand (MT259226.1), this highly identity percentage was emphasis the unity form genetic pool of COVID-19 in China. At the same time the present of two type of mutation (Transition and transverion) and all the previous variants (C > T; T > C) and A > G; G > T) shows consistent heterozygosity in all samples. This may explain the USA COVID-19 strains undergo high rate of mutation, and this development in genetic asses the aggressiveness of virus and encouraged it pathogenesis in USA and Europe, this results agree with Lyons and Lauring [17].

Whole genome sequencing analysis of 24 USA strains of the COVID-19 virus retrieve from gene bank representative different Countries between the end of December 2019 and mid-march 2020 showed 93% - 99.9% homology, with significant mutation, the validity of mutation converted many amino acid in the proteins of ORF1ab, S gene, ORF3a and ORF8. The most amino acid changed were Praline to Lucien, Lucien to Phenyl alanine in ORF1ab protein, Aspartic acid to Glycine in S protein, Glutamate to Histidine ORF3a, Serine to Lysine and lysine to Serine in ORF8 protein (**Figure 5(b)**). our result differ from result of Holshue *et al.* [18], when the found only 3 nucleotides and 1 amino acid that differed at open reading frame 8 between this patient's virus and the 2019-nCoV reference sequence (NC_045512.2).

3.3.5. Phylogeny Tree

The UPGMA phylogenetic tree (**Figure 6**) using MEGA-X was performed on OFR1ab gene sequences of selected 25 reference strains of COVID-19 deposit in genbank in 2020, is represented in SARS clade. In the tree, MERS virus sequences KT029139.1 MERS-CoV/KOR/KNIH/002_05_2015, KT026453.1 MERS Hu/Riyadh_KSA_2959_2015 [19], and KC776174.1 Human betacoronavirus 2c Jordan-N3/2012 formed a distinct clade (clade MERS), and SARS clade representative all other SARS virus strains of 2019-nCoV from China, USA, Spain and Japan, clustering together one clade. The 2019-nCoV strains from China, USA, Japan and Spain were significantly closely related with each other in on clade SARS (**Figure 6**).

The result of phylogeny tree (**Figure 6**) shown close convergence between the

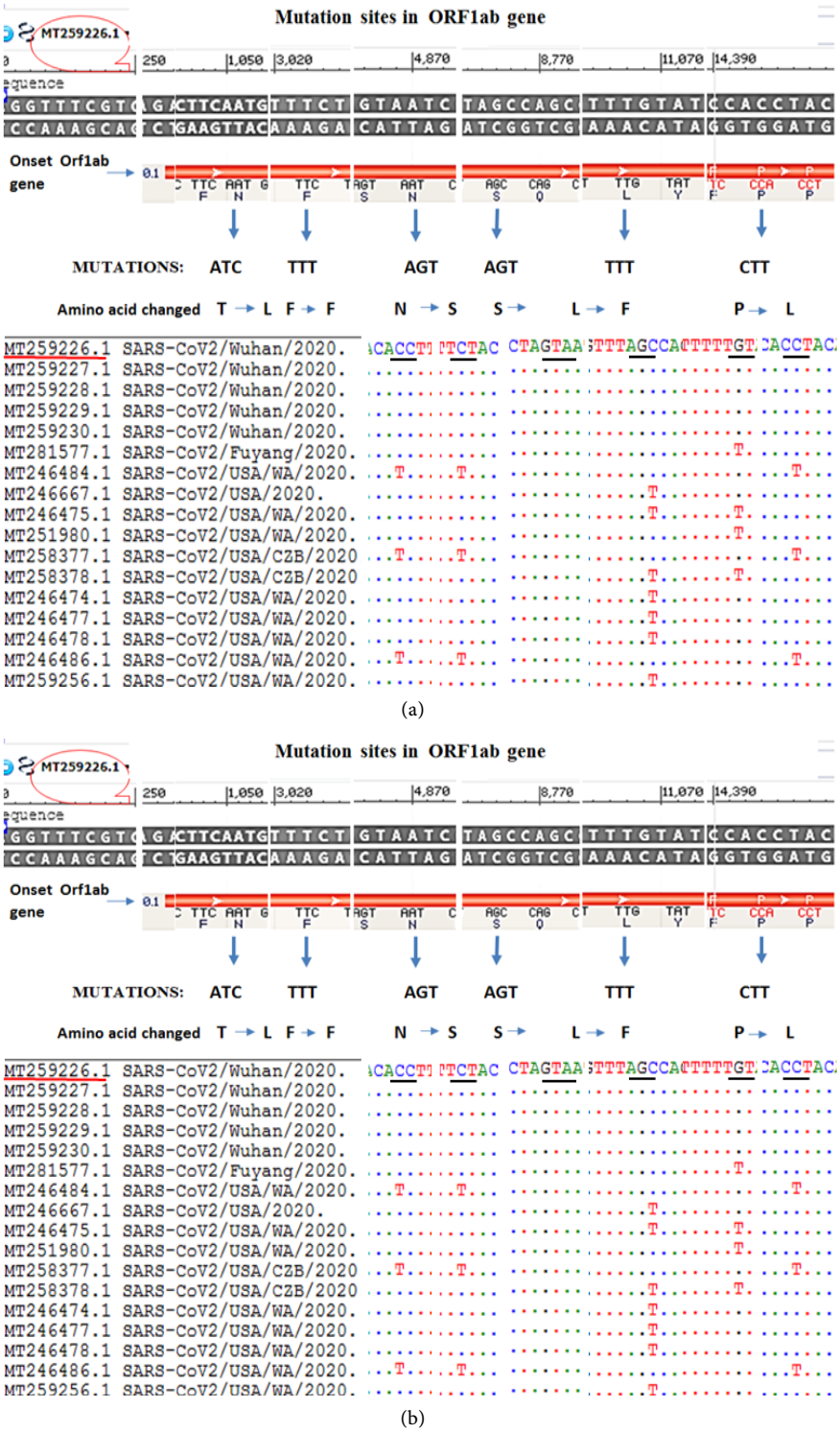

Figure 5. (a) Multiple sequence alignment ORF1ab gene region (250 - 1500 nt) of SARS-CoV2 virus strains, 4 Chinese strains::MT259227.1-MT259229.1 shown 100% identity with reference strain MT258379.1, while other Chinese and USA strains shown 93% - 99% identity. (b) Multiple sequence alignment ORF1ab gene region (250 - 1500 nt) of SARS-CoV2 virus strains, 4 Chinese strains::MT259227.1-MT259229.1 shown 100% identity with reference strain MT258379.1, while other Chinese and USA strains shown 93% - 99% identity.

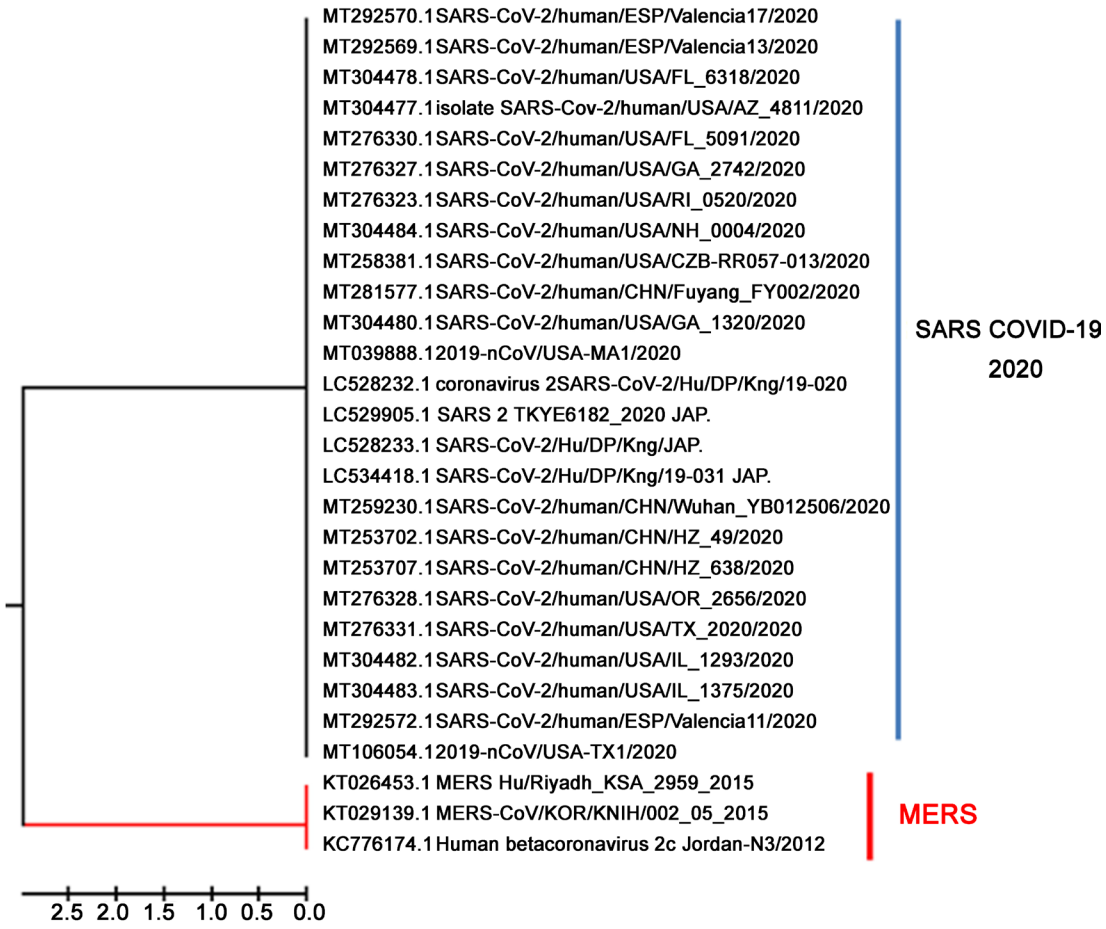

Figure 6. Phylogenetic analysis of the 25 COVID-19 virus strains from China, USA, Spain and Japan, clustering together one clade (SARS COVID-19, 2020) with two MERS virus strain KT029139.1 CoV/KOR/KNIH/002_05_2015, KT026453.1 Hu/Riyadh_KSA_2959_2015, and one strain KC776174.1 betacoronavirus 2c Human Jordan-N3/2012.

strains of the COVID-19 in the evolutionary tree, which indicates, even without doubt, that all strains of the COVID-19 belong to one evolutionary origin [20], and the source of the virus and its spread based on tracking epidemic genetically of the virus strains, this results consistent with Rambaut [21]. The phylogeny tree (**Figure 6**) of the COVID-19 was confirm that the source of epidemic is the Chinese city of Wuhan, and that the differences that were established within the multiple alignment of different sequences from their spread sites in China, America, Japan and Spain may return to a level of readiness to occur counting mutations among the strains of interest.

4. Conclusion

The conclusion shown that there is a high genetic identity between the selected strains of virus in China-Wuhan and those that spread in America and Spain, which indicates that the epidemic center is coming from China. These observations provided evidence of the genetic diversity and rapid evolution of this novel coronavirus.

Author's Contribution

This work was done in collaboration between both authors. The first author (ZKI) designed the study, developed the questionnaire, collected data, wrote the draft manuscript and corresponded with the journal. The second authors (KAH) participated in the development of the data collection tool. Both authors contributed to the literature searches and approved the final manuscript.

Ethical Approval

Both authors hereby declare that all actions have been examined and approved by the appropriate ethics committee and have therefore been performed in accordance with the ethical standards laid down in the 1964 Declaration of Helsinki.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Wang, C., Horby, P.W., Hayden, F.G. and Gao, G.F. (2020) A Novel Coronavirus Outbreak of Global Health Concern. *The Lancet*, **395**, 470-473. https://doi.org/10.1016/S0140-6736(20)30185-9

[2] WHO (2020) Coronavirus Disease 2019 (COVID-19) Situation Report 91. https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200420-sitrep-91-covid-19

[3] WHO (2020) Coronavirus Disease (COVID-19) Outbreak. https://www.who.int/emergencies/diseases/novel-coronavirus-2019

[4] Velavan, T.P. and Meyer, C.G. (2020) The COVID-19 Epidemic. *Tropical Medicine & International Health*, **25**, 278-280. https://doi.org/10.1111/tmi.13383

[5] Nishiura, H., *et al.* (2020) Estimation of the Asymptomatic Ratio of Novel Coronavirus Infections (COVID-19). *International Journal of Infectious Diseases*, **94**, 154-155. https://doi.org/10.1101/2020.02.03.20020248

[6] European Centre for Disease Prevention and Control (ECDC) (2020) COVID-19. https://www.ecdc.europa.eu/en/novel-coronavirus-china

[7] European Commission (EC) COVID-19. https://ec.europa.eu/health/coronavirus_en

[8] Lauring, A.S. and Andino, R. (2010) Quasispecies Theory and the Behavior of RNA Viruses. *PLOS Pathogens*, **6**, e1001005. https://doi.org/10.1371/journal.ppat.1001005

[9] Kumar, S., Stecher, G., Li, M., *et al.* (2018) MEGAx: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Molecular Biology and Evolution*, **35**, 1547-1549. https://doi.org/10.1093/molbev/msy096

[10] Weiss, S.R. and Navas-Martin, S. (2005) Coronavirus Pathogenesis and the Emerging Pathogen Severe Acute Respiratory Syndrome Coronavirus. *Microbiology and Molecular Biology Reviews*, **69**, 635-664. https://doi.org/10.1128/MMBR.69.4.635-664.2005

[11] Fehr, A.R. and Perlman, S. (2015) Coronaviruses: An Overview of Their Replication and Pathogenesis. *Methods in Molecular Biology*, **1282**, 1-23. https://doi.org/10.1007/978-1-4939-2438-7_1

[12] Zhao, L., Jha, B.K., Wu, A., Elliott, R., Ziebuhr, J., *et al.* (2012) Antagonism of the Interferon-Induced OAS-RNase L Pathway by Murine Coronavirus ns2 Protein Is Required for Virus Replication and Liver Pathology. *Cell Host Microbe*, **11**, 607-616. https://doi.org/10.1016/j.chom.2012.04.011

[13] Song, Z., Xu, Y., Bao, L., Zhang, L., *et al.* (2019) From SARS to MERS, Thrusting Coronaviruses into the Spotlight. *Viruses*, **11**, 59. https://doi.org/10.3390/v11010059

[14] Wu, F., Zhao, S., Yu, B., Chen, Y.-M., Wang, W., Hu, Y., *et al.* (2020) Complete Genome Characterisation of a Novel Coronavirus Associated with Severe Human Respiratory Disease in Wuhan, China. https://doi.org/10.1101/2020.01.24.919183

[15] Koyama, T., Platt, D. and Parida, L. (2020) Variant Analysis of COVID-19 Genomes. *Bulletin of the World Health Organization*. https://doi.org/10.2471/BLT.20.253591

[16] Lu, R., Zhao, X., Li, J., Niu, P., Yang, B., Wu, H., *et al.* (2020) Genomic Characterisation and Epidemiology of 2019 Novel Coronavirus: Implications for Virus Origins and Receptor Binding. *The Lancet*, **395**, 565-574.

[17] Lyons, D.M. and Lauring, A.S. (2017) Evidence for the Selective Basis of Transition-to-Transverion Substitution Bias in Two RNA Viruses. *Molecular Biology and Evolution*, **34**, 3205-3215. https://doi.org/10.1093/molbev/msx251

[18] Holshue, M.L., DeBolt, C., Lindquist, S., Lofy, K.H., Wiesman, J., Bruce, H., *et al.* (2020) First Case of 2019 Novel Coronavirus in the United States. *New England Journal of Medicine*, **382**, 929-936. https://doi.org/10.1056/NEJMoa2001191

[19] Chan, J.F.-W., Yuan, S., Ko, K., To, K., Chu, H., Yang, J., *et al.* (2020) A Familial Cluster of Pneumonia Associated with the 2019 Novel Coronavirus Indicating Person-to-Person Transmission: A Study of a Family Cluster. *The Lancet*, **395**, 514-523. https://doi.org/10.1016/S0140-6736(20)30154-9

[20] Cotton, M., Watson, S.J., Kellam, P., Al-Rabeeah, A.A., Makhdoom, H.Q., Assiri, A., *et al.* (2013) Transmission and Evolution of the Middle East Respiratory Syndrome Coronavirus in Saudi Arabia: A Descriptive Genomic Study. *The Lancet* (*London*, *England*), **382**, 1993-2002. https://doi.org/10.1016/S0140-6736(13)61887-5

[21] Rambaut, A. (2020) Phylogenetic Analysis of 23 nCoV-2019 Genomes. http://virological.org/t/phylogenetic-analysis-of-23-ncov-2019-genomes-2020-01-23/335