DOI: 10.7860/JCDR/2020/45029.13891

Review Article

Microbiology Section

Tracing the Origin and Transmission of Severe Acute Respiratory Syndrome-CoV-2: Does Genome Research Hold the Key?

THANGAM MENON

ABSTRACT

The Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) was first identified in December 2019 in Wuhan, China. Since then, the virus has been rapidly spreading across countries, resulting in the present Corona Virus Disease (COVID-2019) pandemic. Thousands of genomes of different strains of the virus have been sequenced and made available in the public domain which has helped to detect viral mutations and track movement of the virus across the globe. The whole genome sequence of SARS-CoV-2 isolated from a human strain has 96.2% similarity to a bat coronavirus. The strains of SARS-CoV-2 found in different countries are genetically diverse and infections have been shown to be caused by multiple introductions in the same country. Though there are several theories regarding the origin of this virus, genetic studies indicate that it may have appeared by natural selection in humans following zoonotic transfer. This review highlights the recent knowledge of the origin of the coronavirus, its diversity in different geographical regions and mutations, which have aided it to infect the human host.

Keywords: Coronavirus 2, Diversity, Mutation, Pandemic

INTRODUCTION

The current pandemic caused by SARS-CoV-2 has had an unprecedented impact on the human population across the globe. Though there have been other coronaviruses such as SARS-CoV and MERS-CoV which has caused serious disease in human beings, the SARS-CoV-2 has shown a much more efficient human-to-human transmission, resulting in a rapid spread of the disease across countries. The first official case was reported from China on 31st December. 2019 [1]. However, according to scientists the first patient reported symptoms on December 1, 2019 [2]. The first whole genome sequence of SARS-CoV-2 was published in January 2020 (GenBank accession No. MN908947) and the virus was isolated from a sample of broncho-alveolar lavage fluid from a patient who worked in the seafood market and was admitted to the Central Hospital of Wuhan on 26th December 2019 with severe respiratory symptoms [3]. The strain of virus was called Wuhan-Hu-1. Before the end of January 2020, more than 2000 cases were confirmed in Wuhan. Next-generation sequencing of samples from nine patients, from the first outbreak in Wuhan showed that the genome sequences were extremely similar, exhibiting more than 99.98% sequence identity [4]. Since then, the genome sequence of SARS-CoV-2 from clinical samples has been obtained by several laboratories and researchers from different laboratories, have made the sequence data available in the public domain [5-7]. Analysis of genomic data gives insights into the origin of the virus as well as its manner of transmissibility in the human host besides providing information for drug design and vaccine development.

According to the live world statistics reported on Worldometer (https://www.worldometers.info/coronavirus/), as of 11 June 2020, this emerging infection has been reported in 215 countries, causing over **74,58,646** infections with 4,19,020 deaths. In India, there are **287,155** reported cases and over 8107 deaths.

ORIGIN OF COVID-19

The evolutionary history of RNA viruses has always been difficult to resolve and SARS-CoV-2 is no different. The origin of the disease COVID-19 still remains elusive. There are media reports that Wei Guixian, a 57-year-old shrimp seller in Huanan wholesale seafood market, who reported flu like symptoms on December 10th 2019 was the first patient to contract the disease. However, scientists do not believe that she was the first case. The earliest cases reported were linked to the Huanan wet market in Wuhan and thus an animal source was suspected. On the basis of current sequence databases, all human coronaviruses have animal origins: SARS-CoV, Middle East Respiratory Syndrome Coronavirus (MERS-CoV), Human Coronavirus NL63 (HCoV-NL63) and Human Coronavirus-229E (HCoV-229E) are considered to have originated in bats; HCoV-OC43 and HKU1 are likely to have originated from rodents [8]. Several researchers have confirmed the genetic similarity between SARS-CoV-2 and a bat coronavirus of the sub-genus Sarbecovirus. The whole-genome sequence identity of SARS-CoV-2 isolated from a human strain has 96.2% similarity to a coronavirus (RaTG13) isolated from the bat *Rhinolophus affinis* from Yunnan province of China [9]. Phylogenetic analysis of the full-length genome and the gene sequences of *RdRp* and spike protein(S) has shown that RaTG13 is closely related to SARS-CoV-2, suggesting that SARS-CoV-2 may have originated in bats [10].

ANIMAL RESERVOIR

In the first cohort of patients studied in Wuhan, 34% of them reported no exposure at the seafood market [2] and hence it was presumed that the virus had an intermediate host as is the case with SARS-CoV and MERS-CoV which usually pass into intermediate hosts, such as civets or camels, before infecting humans [11]. However, the specific route of transmission of SARS-CoV-2 from natural reservoirs to humans remains unclear.

Malayan pangolins (*Manis javanica*) contain coronaviruses similar to SARS-CoV-2. Autopsy findings in two dead pangolins showed pulmonary fibrosis of the lungs and it was thus thought that the dead Malayan pangolins may carry a new CoV closely related to SARS-CoV-2. The genome of Pangolin-CoV was 91.02% identical to SARS-CoV-2 [12]. A high amino acid similarity was found in the receptor binding domain of the S (spike) protein, between a strain isolated from a pangolin (GD410721), and

SARS CoV-2. However, the pangolin coronaviruses are unlikely to be directly linked to the outbreak and the direct progenitor of SARS-CoV-2 is yet to be identified.

GENOME DIVERSITY

The genetic diversity of populations of viruses circulating in different countries when analysed showed that there was a significant haplotype diversity in samples from countries outside China indicating accelerated mutation rates in these strains of SARS CoV-2 virus [13]. Strains isolated from the countries most affected had the greatest genetic diversity and it was difficult to identify the "Patient Zero" in these countries, where there would have been extensive transmission in the early stages of the epidemic [14]. Generally, diversity of strains has been seen in every country which has sequenced a reasonably large number of its isolates, with representatives of all major clades being present in each of the affected countries. This indicates that each of these local epidemics could have been started by several independent introductions of the virus which may have taken place extremely early on in the pandemic. The pattern however was not seen in China, the country where the outbreak had started [15].

Scientists who have analysed full genomes of strains available (https://www.gisaid.org/) have identified three variants which differed based on amino acid changes and named them as A, B and C types. Type A was the ancestral type and closest to the bat outgroup coronavirus. B was the most common type in East Asia, whereas the A and C types were found in Europe and America. Americans who had lived in Wuhan had mutated versions of the A type. The major European type was the C type. The earliest introductions of the virus into Italy were from Germany and Singapore. Wuhan's major virus type, 'B', was prevalent in patients from across East Asia and appeared to be environmentally adapted to the East Asian population. The B type was not found in regions beyond East Asia without further mutations and it was only the mutated derived B type which could cause infections outside Asia suggesting that type B was immunologically or environmentally adapted to the East Asian population [16]. Others have disagreed with this, stating that random emergence of new mutations is not uncommon when a viral strain is introduced in a new population, such random mutations can be propagated without them being selected or advantageous, and the dataset of 160 genomes analysed in the study was not truly representative of the entire population [17]. A study which analysed 176 full length genomes showed that recently isolated strains were showing more divergence, which is characteristic of fast evolving RNA viruses; however there was limited genetic variation indicative of a relatively recent common ancestor for all these viruses [18]. The Most Recent Common Ancestor (MRCA) represents the point where the ancestral virus of all the sampled cases was in the same host and estimation of the date of MRCA indicates a period from the beginning of December to mid December 2019 which matches the earliest reported date of symptom onset for the initial cluster of pneumonia cases in Wuhan [2].

This timing of the MRCA of the sampled genomes is being supported as more genomes are being generated and analysed. According to the genome sequence data there has been no animal reservoir responsible for generating new cases and all the cases after December 2019 have occurred by human to human transmission. The virus may still exist in one or more non-human animal species, but that will be of no consequence to the current pandemic [18].

MUTATIONS

Mutations have emerged independently multiple times suggesting that the virus is continuously adapting to its new human host and 80% of such mutations are non-synonomous giving rise to changes at the protein level. A dataset of 7666 public genome assemblies found independent recurrent mutations at 198 sites in the genome which included a mutation at nucleotide position 21,575 in the spike protein, and at position 11,083 in ORF1a. Mutations were also found in the ORF1ab region in two sites coding Nsp11 and Nsp13. The mutation at site 11,083 in the SARS-CoV-2 genome is in a region of Orf1a encoding Nsp6 which overlaps a putative immunogenic peptide predicted to result in both CD4+ and CD8+ T-cell reactivity. The mutation in nucleotide position 21,575 corresponds to the SARS-CoV-2 spike protein, but this homoplasy falls outside of the N-terminal and receptor binding domains of the spike protein [18]. These appeared to be the strongest putative regions under selection in this dataset.

From a global perspective, the SARS-CoV-2 appears to have accumulated only moderate genetic diversity so far with an average difference between any two genomes of 9.6 single nucleotide polymorphisms (SNPs), indicating that it has evolved from a relatively recent common ancestor [13]. A recent study analysed the complete genome sequences of 95 strains of SARS -CoV-2 which were published between December 2019 and April 2020 and found 156 total variants and 116 unique variants. Of the 95 genomes analysed, 24 samples did not exhibit any variants. There were several variants of which the most common ones were 28144T>C (ORF8) seen in 14 samples and 8782C>T (ORF1ab) which was seen in 13 samples. Both of these occurrences were found to be linked and the variant substrains were frequently seen outside of Wuhan. Among the non-structural proteins Nsp1 to Nsp16, Nsp3 had the largest number of variants. The most frequently observed base change was C>T [18,19].

The SNPs at location 8,782 and 28,144 define the two major types (L and S) of SARS-CoV-2 viruses. L type (~70%) is more prevalent than the S type (~30%). The L type was more common in the early stages of the outbreak in Wuhan, but decreased thereafter due to selective pressure caused by human intervention. The S type is the ancestral version is less aggressive and causes milder infections [20].

INDIAN SEQUENCES

Analysis of the complete genome sequences (29,851 nucleotides) of SARS-CoV-2 isolated from the first two positive cases who arrived in India in January 2020 from Wuhan revealed that the sequences of Indian strains of SARS-CoV-2 were very similar to the Wuhan strain (accession number: NC 045512) with a homology of 99.98%. Phylogenetic analysis showed that they belonged to different clusters indicating that they were two different introductions into the country. Both strains clustered with strains of the Sarbecovirus subgenus and had the closest homology (96.09%) with the RaTG13 strain of the bat coronavirus. Two mutations (408 Arg→Ile and 930 Ala→Val) in the spike protein sequences which were observed in SARS-CoV-2 strains isolated from India were localised over the S1 and S2 domains but not in the region of the ACE2-binding interface [21].

ORIGIN OF SARS-COV-2

A progenitor of SARS-CoV-2, may have been circulating in the community causing asymptomatic infections in human. During this process it may have, through adaptation, acquired certain genomic features, which would have enabled it to rapidly spread in the community by human-to-human transmission. The sequence data available currently suggests that virus would have emerged in late November or early December 2019. Prior to this, a period of unrecognised transmission in humans would have occurred between the initial zoonotic event and the acquisition of the genetic adaptation increasing its transmissibility in humans. Many prior zoonotic events that produced short chains of human-to-human transmission over an extended period can give rise to widespread disease. On the other hand containment measures produce short transmission chains which eventually resolve in the absence of

adaptation to sustained transmission, as was seen in the case of MERS-CoV [22].

It appears that SARS-CoV-2 appeared by natural selection in humans following zoonotic transfer. The receptor-binding domain in the spike protein of SARS-CoV-2 is significantly different from other closely related betacoronaviruses and has a higher binding affinity with human Angiotensin-converting enzyme 2 receptor than any of the other viruses. When sequences are very dissimilar, they are more likely to be due to natural selection rather than genetic manipulation. The spike protein of SARS-CoV-2 shows the addition of O-linked glycans and an inserted proline which flank a polybasic cleavage site which is responsible for the high level of infectivity and host range of the virus [23]. The addition of such glycans typically occurs under immune selection. Such major changes in the molecular structure of SARS-CoV-2 compared with other coronaviruses provide evidence that the changes have occurred by natural selection. If the virus was engineered, it would not have differed substantially from those of already known pathogenic coronaviruses.

An alternate theory of the origin of this strain is that it may have been selected during passage in cell cultures since research involving passage of bat SARS-CoV-like coronaviruses in cell cultures has been going on for many years in laboratories across the world and an inadvertent release of the virus outside the laboratory containment zone could have happened by accident. However, this appears to be less likely and studies of the genetic sequences of the virus and their acquired mutations indicate that a genetically similar progenitor virus has not been described so far [24].

CONCLUSION(S)

Understanding the origins of a pandemic are important in the prevention of future zoonotic events. Scientific and technological advancements have helped us to understand this epidemic better than previous epidemics. Genome sequencing, genetic engineering and an understanding of host-pathogen interactions at a molecular level have given us insights of the transmissibility and pathogenesis of the virus. However, more scientific data is needed, particularly with viral sequencing from animal sources to identify the zoonotic source, which will be crucial in preventing future epidemics. Extensive sequence studies of genomes of the virus isolated from infected patients would be the key to developing an optimal drug and an effective vaccine. Identifying invariant regions of the genome is important because these regions are ideal targets for drugs and vaccines. It is important to continuously monitor genomic changes in the virus for drug and vaccine design and to avoid drug resistance and vaccine evasion. In addition, it may also be important to analyse archival specimens to provide insights into the past demography of SARS-CoV-2.

REFERENCES

[1] Singhal T. A review of coronavirus disease-2019 (COVID-19). Indian J Pediatr. 2020;87(4):281-86. doi:10.1007/s12098-020-03263-6.

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

[3] Wu F, Zhao S, Yu B, Chen Y, Wang W, Song Z, et al. A new coronavirus associated with human respiratory disease in China. Nature. 2020;579(7798):265-69. doi:10.1038/s41586-020-2008-3.

[4] Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: Implications for virus origins and receptor binding. Lancet. 2020;395(10224):565-74. doi:10.1016/S0140-6736(20)30251-8.

[5] Chan JF, Kok KH, Zhu Z, Chu H, To KK, Yuan S, et al. Genomic characterization of the 2019 novel human-pathogenic coronavirus isolated from a patient with atypical pneumonia after visiting Wuhan. Emerg Microbes Infect. 2020;9:221-36. https://doi.org/10.1080/22221751.2020.1719902.

[6] Chen L, Liu W, Zhang Q, Xu K, Ye G, Wu W, et al. RNA based mNGS approach identifies a novel human coronavirus from two individual pneumonia cases in 2019 Wuhan outbreak. Emerg Microbes Infect. 2020;9(1):313-19. https://doi.org/10.1080/22221751.2020.1725399.

[7] Chan JFW, Yuan S, Kok KH, To KK, Chu H, Yang J, 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;395:514-23. https://doi.org/10.1016/S0140-6736(20)30154-9.

[8] Cui J, Li F, Shi Z. Origin and evolution of pathogenic coronaviruses. Nat Rev Microbiol. 2019;17:181-92. https://doi.org/10.1038/s41579-018-0118-9.

[9] Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W, et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature. 2020;579:270-73. doi.org/10.1038/s41586-020-2012-7.

[10] Zheng J. SARS-CoV-2: An Emerging coronavirus that causes a global threat. Int J Biol Sci. 2020;16(10):1678-85. doi:10.7150/ijbs.45053.

[11] Omrani AS, Al-Tawfiq JA, Memish ZA. Middle East Respiratory Syndrome Coronavirus (MERS-CoV): Animal to human interaction. Pathog Glob Health. 2015;109(8):354-62. doi:10.1080/20477724.2015.1122852.

[12] Zhang T, Wu Q, Zhang Z. Probable pangolin origin of SARS-CoV-2 associated with Covid-19 outbreak. Current Biology. 2020;30:1346-51.

[13] Yu W, Tang G, Zhang L, Corlett RT. Decoding the evolution and transmissions of the novel pneumonia coronavirus (SARS-CoV-2 / HCoV-19) using whole genomic data. Zool Res. 2020;41(3):247-57. doi: 10.24272/j.issn.2095-8137.2020.022 https://www.sciencedaily.com/releases/2020/05/200505190550.htm.

[14] https://www.sciencedaily.com/releases/2020/05/200505190550.htm.

[15] Dorp L.van, Acman M, Richard D, Shaw LP, Ford CE, Ormond L, et al. Emergence of genomic diversity and recurrent mutations in SARS-CoV-2, Infection, Genetics and Evolution.2020;83:1043517. https://doi.org/10.1016/j.meegid.2020.104351.

[16] Forster P, Forster L, Renfrew C, Forster M. Phylogenetic network analysis of SARS-CoV-2 genomes. Proceedings of the National Academy of Sciences. 2020;117(17):9241-43; DOI: 10.1073/pnas.2004999117.

[17] Mavian C, Pond SK, Marini S, Magalis BR, Vandamme AM, Dellicour S, et al. Sampling bias and incorrect rooting make phylogenetic network tracing of SARS-COV-2 infections unreliable, Proceedings of the National Academy of Sciences. 2020;117(23):12522-23; DOI: 10.1073/pnas.2007295117.

[18] Rambaut A. Virological.org http://virological.org/t/356 (2020).

[19] Khailany RA, Safdar M, Ozaslan M. Genomic characterization of a novel SARS-CoV-2 [published online ahead of print, 2020 Apr 16]. Gene Rep. 2020;19:100682. doi:10.1016/j.genrep.2020.100682.

[20] Tang X, Wu C, Li X, Song Y, Yao X, Wu X, et al. On the origin and continuing evolution of SARS-CoV-2. National Science Review, 2020, nwaa036. https://doi.org/10.1093/nsr/nwaa036.

[21] Yadav PD, Potdar VA, Choudhary ML, Nyayanit DA, Agrawal M, Jadhav SM, et al., Full-genome sequences of the first two SARS-CoV-2 viruses from India. Indian J Med Res. 2020;151:200-09. DOI: 10.4103/ijmr.IJMR_663_20.

[22] Dudas G, Carvalho LM, Rambaut A, Bedford T. MERS-CoV spillover at the camel-human interface. Elife. 2018;7:e31257. DOI: 10.7554/eLife.31257.

[23] Walls AC, Walls, Park YJ, Tortorici MA, Wall A, McGuire AT, et al. Structure, function, and antigenicity of the SARSCoV-2 spike glycoprotein. Cell. 2020;180:281-92. https://doi.org/10.1016/j.cell.2020.02.058.

[24] Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF. The proximal origin of SARS-CoV-2. Nat Med. 2020;26:450-52. https://doi.org/10.1038/s41591-020-0820-9.

PARTICULARS OF CONTRIBUTORS:
1. UGC BSR Faculty, Department of Microbiology, University of Madras, Chennai, Tamil Nadu, India.

NAME, ADDRESS, E-MAIL ID OF THE CORRESPONDING AUTHOR:
Dr. Thangam Menon,
Department of Microbiology, University of Madras, Taramani Campus,
Chennai, Tamil Nadu, India.
E-mail: thangam56@gmail.com

PLAGIARISM CHECKING METHODS: [Jain H et al.]
- Plagiarism X-checker: May 13, 2020
- Manual Googling: Jun 15, 2020
- iThenticate Software: Jul 13, 2020 (21%)

ETYMOLOGY: Author Origin

AUTHOR DECLARATION:
- Financial or Other Competing Interests: None
- Was informed consent obtained from the subjects involved in the study? No
- For any images presented appropriate consent has been obtained from the subjects. NA

Date of Submission: **May 12, 2020**
Date of Peer Review: **Jun 10, 2020**
Date of Acceptance: **Jun 15, 2020**
Date of Publishing: **Aug 01, 2020**