ORIGINAL ARTICLE

COMPARISON OF SARS-COV2 GENOME SEQUENCES FROM PAKISTAN WITH GENOME SEQUENCES FROM OTHER COUNTRIES

Muhammad Tariq[1]✉ , Fatima Jehan[1], Nooria Ali Liaqat[1], Mehnaz[1], Umar Farooq[2]

ABSTRACT

OBJECTIVE: To analyze SARS-CoV-2 whole genome reported of Pakistan and compare them with other viral strains collected from other world region to better understand the origin and genetic characterization of the virus.

METHODS: All the available genomic information of SARS-CoV-2 including Pakistani strains were collected from various online sources. Phylogenetic analysis of 131 sequences from 11 countries (Brazil, China, India, Italy, Nepal, Pakistan, Spain, Sweden, Taiwan, USA and Viet-Nam) were performed and compared with other related coronaviruses to find the evolution of virus and its origin. Individual SARS-CoV-2 gene, spike (S) glycoprotein and the receptor binding domain (RBD) with closely related coronaviruses were compared to further explore genetic variations and the likely RBD properties of the virus.

RESULTS: The analysis shows that genome of all analyzed 131 SARS-CoV-2 strains collected from different geographical area were extremely similar, exhibiting >99% sequence identity. Notably, genome of the SARS-CoV-2 has high similarity (89.1% sequence identity) with the two bat-derived severe acute respiratory syndrome (SARS) like betacoronaviruses, bat-SL-CoVZC45 and bat-SL-CoVZXC21 but was sufficiently divergent from SARS-CoV (82.3% similarity) and MERS-CoV (50% similarity). Phylogenetic analysis shows that the SARS-CoV-2 has relatively similar spike glycoprotein with bat-SL-CoVZC45, however, the RBD was more like that of SARS-CoVGZ02.

CONCLUSION: Using different bioinformatics tools, we determined that SARS-CoV-2 has high similarities to bat-derived SARS like betacoronaviruses than SARS-CoV at the whole genome level, however, the RBD was more like that of SARS-CoVGZ02, which shows that they use similar ACE2 as a cell receptor.

KEYWORDS: SARS-CoV-2 (MesH); SARS-CoV-2 variants (MesH); Whole Genome Sequencing (MesH); Phylogenetics (MeSH); spike protein, SARS-CoV-2 (MeSH); Receptor Binding Domain (Non-MeSH).

1: Centre for Omic Sciences, Islamia College University, Peshawar, Pakistan

2: College of Life Science, Northeast Forestry University, Harbin 150040, China

Contact #: +92-346-9421336

Email✉: mtariq@icp.edu.pk

INTRODUCTION

The ongoing pandemic of coronavirus disease 2019 (COVID-19) cause by the new coronavirus, called the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), appeared in Wuhan city of China in December 2019 [1] was declared a global pandemic by March 2020. In past two decades, two other coronaviruses namely SARS-CoV (2002-2003), that emerged in Guangdong, southern China,[2] and the Middle East Respiratory Syndrome Coronavirus (MERS-CoV) [2012-present], which appeared first in Saudi Arabia,[3,4] have also caused world-wide outbreaks. Most of the viruses from the coronavirus family infect mammals, whereas others infect a variety of birds.[5,6]

All coronaviruses including SARSCoV-2 have a single positive stranded RNA genome ranging from 26 to 32 kilobases in length.[7] Since the emergence of SARS-CoV-2, genome analysis has played an important role in public health response by informing the design of appropriate molecular diagnostics and documenting epidemiological efforts to trace contacts.

Further, genome of SARS-CoV-2 is composed of a long sequence (ORF1ab), which encode proteins used for RNA replication and genes encoding the non-structural proteins and the structural proteins. The virus four structural proteins are: Spike protein (S), envelope gene (E), membrane protein (M), and the nucleoprotein (N). All these genes are also found in other coronaviruses.[8,9] These genes are working in different viral processes including formation of the virus particles, the spike protein (S) is a glycoprotein, playing essential role by attaching the virion to host cell receptor thereby changing its tropism and pathogenicity,[10] which is the major target of virus to neutralize the antibodies.[11] Mutations in this region induce structural changes, which leading to antigenicity changing.

Data regarding genomic characterization of SARS-CoV-2 is limited from our region. This study was planned to describe genomic characterization of SARS-CoV-2 from 11 countries including Pakistan, Brazil, China, India, Italy, Nepal, Spain, Sweden, Taiwan, USA and Viet-Nam and closely related coronaviruses strains. Furthermore, phylogenetic analysis of 131 whole SARS-CoV-2 genomes and other closely related coronaviruses were also done, to determine evolutionary distance between SARS-CoV-2, other related member of the family of SARS-CoV and the bat SARS-like coronaviruses. Furthermore, individual SARS-CoV-2 gene, spike (S) glycoprotein and the Receptor Binding Domain (RBD) with closely related coronaviruses were compared to further explore genetic variations and the likely RBD properties of the virus as well as to determine the possible functional implications.

TABLE I: COMPARISON OF SARS-COV-2 SEQUENCES WITH THE BAT SARS-LIKE CORONAVIRUSES, BAT-SL-COVZC45 (ACCESSION NO. Mg772933), BETACOV/YN2018C (ACCESSION NO. Mk211377), BAT-SL-COVZXC21 (ACCESSION NO.MG772934) AND THE SARS-COV GZ02 (ACCESSION NO. Ay390556).

	Strain	Complete Genome (%)	Gene Region (%)												
			1ab	1a	1b	S	3	E	M	7	8	10b	N	13	14
Nucleotide Sequences	Bat-SL-CoVZC45	89.1	89.1	90.7	86.0	83.7	87.8	98.7	93.9	95.2	88.8	88.8	91.3	89.1	97.2
	BetaCoV/YN2018C	82.3	82.3	84.2	80.8	80.3	87.7	93.9	85.4	95.1	87.6	88.7	88.4	88.9	97.2
	Bat-SL-CoVZXC21	88.4	88.6	90.3	86.2	83.3	88.9	98.7	93.4	95.2	89.1	88.5	91.2	89.5	96.7
	SARS-CoVGZ02	82.3	79.8	75.5	86.6	78.5	76.2	93.9	85.7	74.9	82.3	--	88.5	--	--
Amino Acid Sequences	Bat-SL-CoVZC45	--	95.6	95.6	95.8	81.0	90.6	100.0	98.7	93.4	87.6	94.2	94.3	73.2	92.9
	BetaCoV/YN2018C	--	86.4	86.7	84.3	75.3	88.3	94.7	90.5	89.5	86.8	90.1	90.3	71.4	90.2
	Bat-SL-CoVZXC21	--	95.2	95.1	95.8	81.0	90.6	100	98.7	93.4	88.4	94.2	94.3	73.2	92.9
	SARS-CoVGZ02	--	86.1	80.5	95.6	76.2	72.7	94.7	90.5	68.9	85.2	--	90.5	--	--

METHODS

DATA COLLECTION

The entire genome of Wuhan-HU-1 or the reference genome (with accession number NC_045512) and other 130 other viral strains sequences data from 11 countries were retrieved from the data base known as GISAID (Global Initiative on Sharing Avian Influenza Data) (https://www.gisaid.org/). The sequences included from Brazil (1), China (n=27), India (n=2), Italy (n=1), Nepal (n=1), Pakistan (n=1), Spain (n=4), Sweden (n=1), Taiwan (n=2), USA (n=89) and Viet Nam (n=2). Additionally, thirty-one closely related coronavirus complete genome sequences data were collected from the NCBI GeneBank (http://www.ncbi.nlm.nih.gov/genbank/), which is always accessible publicly. Genome analyses were performed on these sequences, and pairwise intragenomic identity, pairwise variability and homology sequence modeling with alignment have been executed as well.

PHYLOGENETIC ANALYSIS

For phylogenetic analysis, we performed multiple sequence alignment of SARS-CoV-2 and reference sequences using
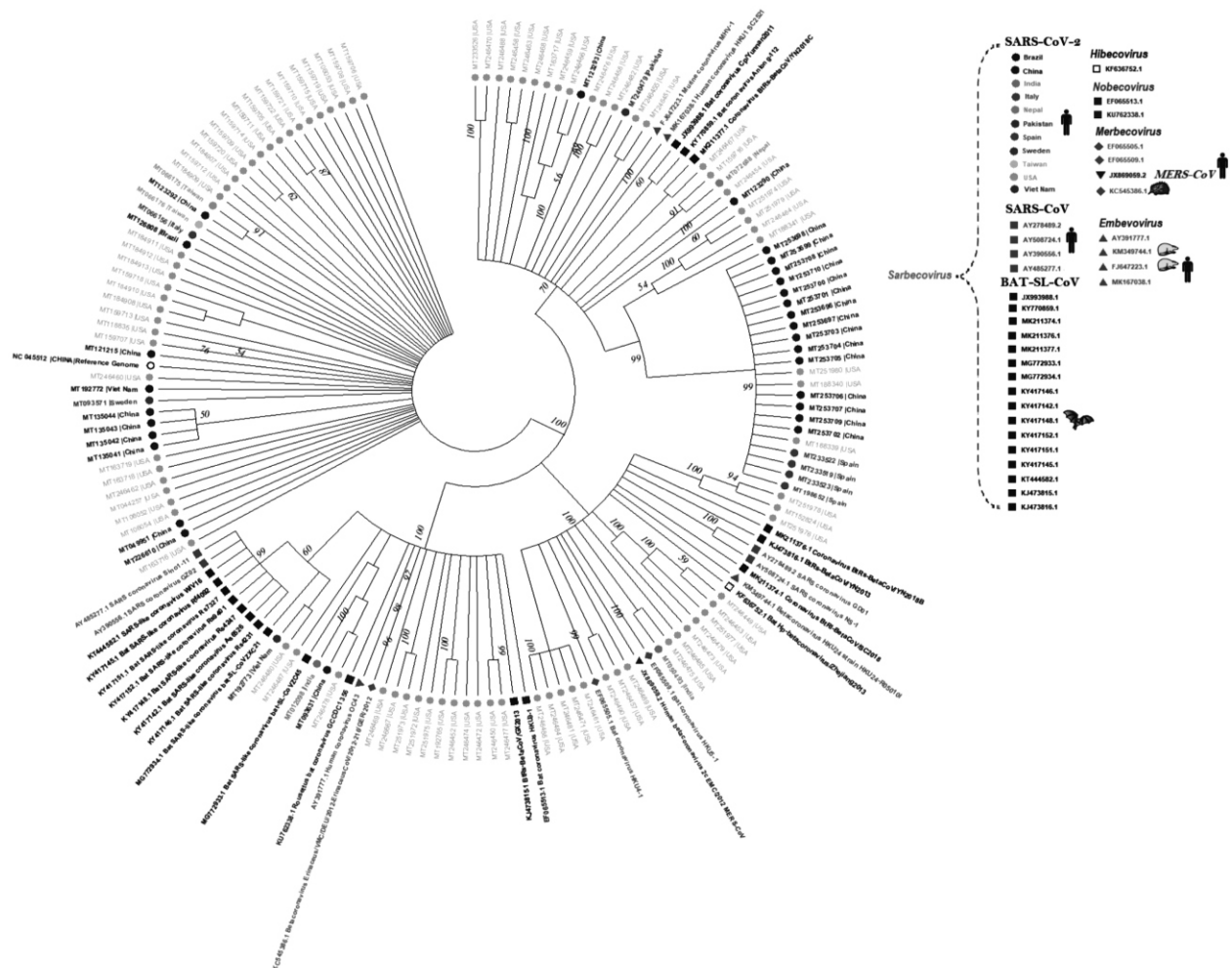

Figure 1: Phylogenetic analysis of whole genomes of SARS-CoV-2 from eleven countries and representative of betacoronaviruses
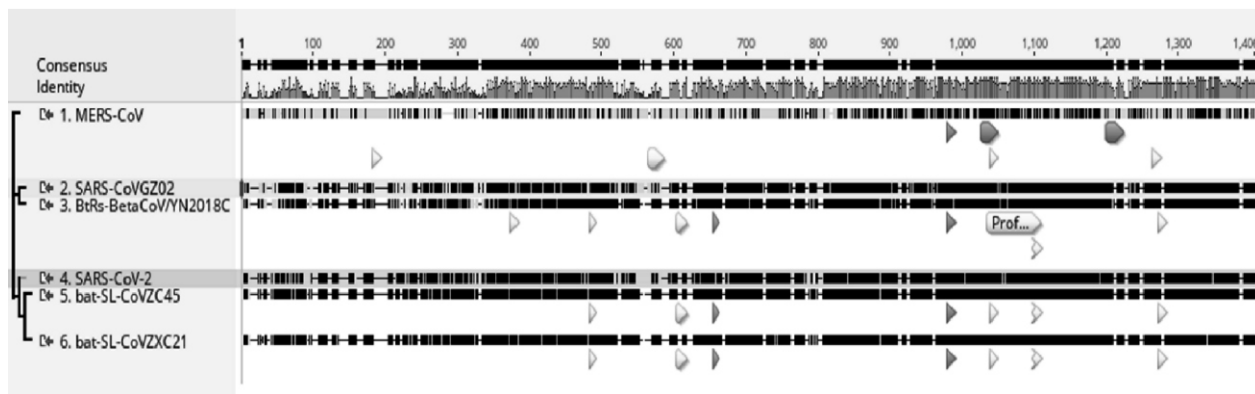
Figure 2: Comparison of SARS-CoV-2 spike protein (S) amino acid sequence with other Coronaviruses

the ClustalW program with MEGA-X software.[12] Phylogenetic analysis of whole genome and major coding regions were done with the same MEGA-X software[12], using the Neighbor-Joining (NJ) method.[13] The consensus bootstrap tree implied from 1000 replicates were used for representing the evolutionary history of the analyzed taxa[14] and the evolutionary distance were calculated using the maximum likelihood (ML) methods.[15] The analysis involved 131 nucleotide sequences. The ambiguous nucleotide positions were removed for each pair.

Additionally, Comparison of individual SARS-CoV-2 gene, spike protein (S) and the RBD with closely related corona-viruses was done to further explore genetic variations and the likely PBD properties of the virus and the results were visualized in heatmap and phylogenetic trees using the same Geneious software.

RESULTS

GENOME VARIATION ANALYSIS OF SARS-CoV-2

To understand genome variations that occurs with the geographical area, we first compared whole genomes of 131 SARS-CoV-2 from Brazil, China, India, Italy, Nepal, Pakistan, Spain, Sweden, Taiwan, USA and Viet-Nam using the viral strain of Wuhan, China (accession number NC_045512) as a reference. We found that at the whole genome level sequences similarities between all the viral strains were greater > 99%, whereas at the amino acid level the similarity was 100%, indicating that the virus emerges recently into human

population, however some recent sequences were more diverse.

Blastn search of the whole SARS-CoV-2 genomes revealed that the most closely related viruses available were bat-SL-CoVZC45 (accession number MG772933; 89.1%; query coverage 95%) and another bat SARS-like coronavirus, bat-SL-CoVZXC21 (accession number Mg772934; 88.4 %; query coverage 94%). When compared at the individual gene level the sequences identities of these two viruses in five genes (E, M, 7, N and 14) were greater than 90%. The highest similarities between the two viruses was observed in the E gene, with sequence identity of 98.7% (Table I) whereas the lowest sequence identity was observed in the S gene (83%). Furthermore, sequence identity in the 1b gene was 86%, which was lower than that of 1a gene (about 90%; Table I). We also found more than 90% sequence identity in three genes (E,7, and 14) between SARS-CoV0-2 and another bat SARS-Like coronavirus, BetaCoV/YN2018C (accession no. MK211377). Similarly, most of the encoded proteins have high sequence identity between SARS-CoV-2 and related bat SARS-like coronaviruses. The notable exception was the spike (S) protein, which has only 81% sequence identity with bat-SL-CoVZC45, bat-SL-CoVZXC21 and 75% sequence identity with BetaCoV/YN2018C, but were more significantly diverse from SARS-CoV (about 82.3%) and MERS-CoV (about 50%) (Table I)

Phylogenetic analysis of all 131 SARS-CoV-2 and 31 closely related corona-viruses genomes revealed, that most of the SARS-CoV-2 strains were close related to some betacoronaviruses

within the subgenus sarbecovirus detected in bats (Figure 1) but were distinct from SARS-CoV and MERS-CoV. Of the 131 SARS-CoV-2 genomes analyzed, the strain collected from USA (MT246478), China (MT093631) and from India (MT012098) were closely clustered with bat-SARS-like corona-virus: bat-SL-CoVZC45). Another strain from USA (accession number MT246480) and from Viet-Nam (MT192773) were closely clustered with bat-SL-CoVZXC21, whereas the third viral strain collected from USA (accession number MT246467) was closely related with bat-SL-CoV YN2018C (Figure I).

Next, we compared the envelope spike (S) proteins of SARS-CoV-2 with its closest homologs (bat-SL-CoVZC45, bat-SL-CoVZXC21, bat-SL-CoV YN2018C, SARS-CoVGZ02 and MERS-CoV). The results show that the spike protein (S) of SARS-CoV-2 has high sequence homology with bat-SL-CoVZC45, indicating potential recombination events in the envelope spike (S) (Figure 2) which was also closer at the whole genome level.

We also analyze the RBD of SARS-CoV-2 and other closely related corona-viruses. The RBD directly interact with the peptidase domain of human receptor, angiotensin-converting enzyme 2 (ACE2) as an SARS-CoV,[16] MERS-CoV[17,18] and an BatCoV HKU4.[19] Through our phylogenetic analysis and heatmap of the RBD of six closely related coronaviruses (Figure 3A and B), we found that, although SARS-CoV-2 was closer to bat-SL-CoVZC45, bat-SL-CoVZXC21 and bat-SL-CoV YN2018C at the whole genome level, the RBD of SARS-CoV-2 was much closer to that of
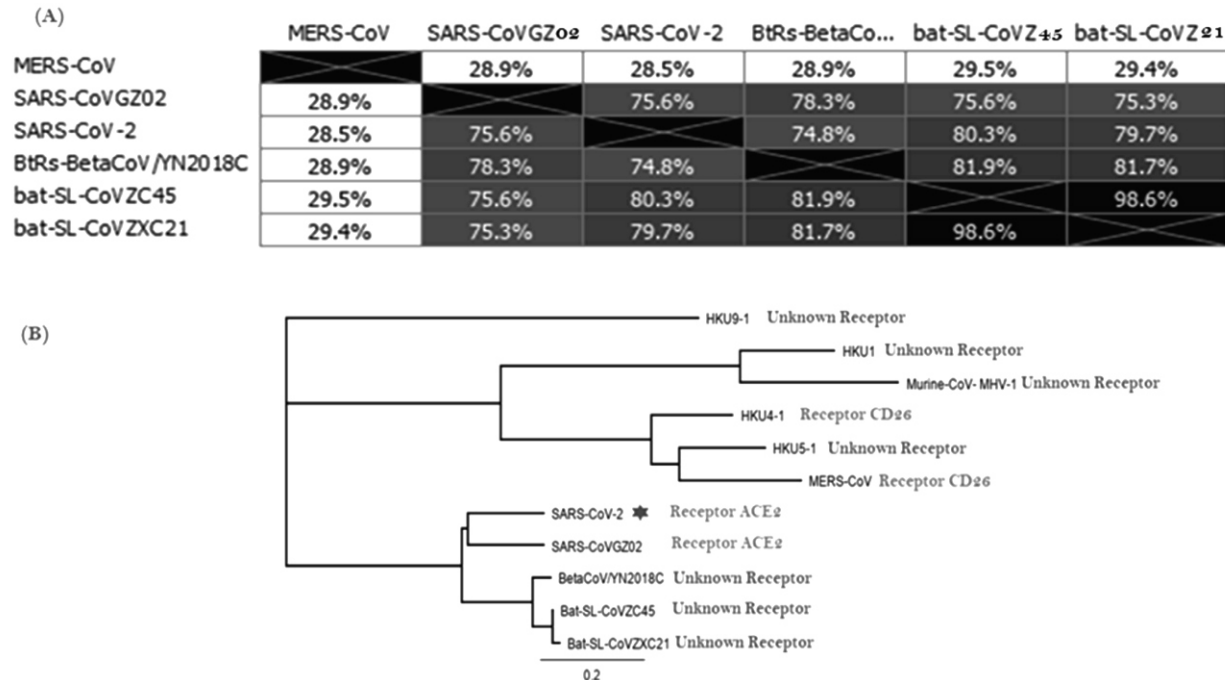
(A)

	MERS-CoV	SARS-CoVGZ02	SARS-CoV-2	BtRs-BetaCo...	bat-SL-CoVZ45	bat-SL-CoVZ21
MERS-CoV		28.9%	28.5%	28.9%	29.5%	29.4%
SARS-CoVGZ02	28.9%		75.6%	78.3%	75.6%	75.3%
SARS-CoV-2	28.5%	75.6%		74.8%	80.3%	79.7%
BtRs-BetaCoV/YN2018C	28.9%	78.3%	74.8%		81.9%	81.7%
bat-SL-CoVZC45	29.5%	75.6%	80.3%	81.9%		98.6%
bat-SL-CoVZXC21	29.4%	75.3%	79.7%	81.7%	98.6%	

(B)

Figure 3: (A) Heatmap of Receptor Binding Domain (RBD) identity between SARS-CoV-2 and other coronaviruses
(B) Phylogenetic tree based on comparison of RBD sequence, SARS-CoV-2 is highlighted with star.

SARS-CoVGZ02 (Figure 3A, B).

DISCUSSION

In this study we have investigated genome variations of 131 SARS-CoV-2 strains collected from 11 countries including Brazil, China, India, Italy, Nepal, Pakistan, Spain, Sweden, Taiwan, USA and Viet-Nam and compared its whole genome, individual gene and part of the coding regions with 31 closely related coronaviruses strains to better understand evolutionary history, and the likely PBD properties of the virus.

The results show that all SARS-CoV-2 sequences collected from different countries were found nearly identical at the whole genome level with nucleotide similarity above 99% and 100% identity at the amino acid level. These findings suggest that the SARS-CoV-2 originated from a single source in a short time and was quickly detected, however the recently reported sequences to gene bank was found more diverse.

Phylogenetic analysis of all 131 SARS-CoV-2 genomes revealed, that most of the SARS-CoV-2 strains closely related to three bat-derived coronaviruses; bat-SL-CoVZC45, bat-SL-CoVZXC21[20] and bat-SL-CoV YN2018C,[21] collected previously in bats (Figure 1). These results show that bat is the main source for SARS-CoV-2, however there many facts which suggest that there is another animal that acted as intermediate host between humans and bats. First, as the coronavirus case was reported first at the month of December, when almost all bat species were hibernating. Second, no bat species were sold in the Huanan seafood market, when the infection was in high peak. Third, sequences identity between SARS-CoV-2 and its close relative bat-SL-CoVZC45, bat-SL-CoVZXC21 and bat-SL-CoV YN2018C was found less than 90% which is also prominent in the phylogenetic tree as long branch between bat-SL-CoVZC45 and bat-SL-CoVZXC21 (Figure 1) and therefore bat-SL-CoVZC45 and bat-SL-CoVZXC21 are no the direct ancestors of SARS-CoV-2. Fourth, bats also acted as natural reservoir in both SARS-CoV and MERS-CoV, however another animal known as the masked palm civet for SARS-CoV[22] and dromedary camel for MERS-CoV,[23] acting as intermediate host and from them it transferred to humans. These facts suggest that SARS-CoV-2 might also be initially hosted by bats and then might have been transferred to human through unknown animal.

Furthermore, comparison of SARS-CoV-2 spike proteins (S) with its closest homologs (bat-SL-CoVZC45, bat-SL-CoVZXC21, bat-SL-CoVYN2018C, SARS-CoVGZ02 and MERS-CoV), indicates that SARS-CoV-2 has relatively high amino acid sequence homology compared to other coronaviruses (Figure 2), although the sequence varies at some position (mostly from 1 to ~ 400). This finding suggests that there are potential recombination events in the envelope spike (S) of bat-SL-CoVZC45 and SARS-CoV-2.

In the past several researchers have identified that different coronaviruses can bind to the human ACE2 for SARS-CoV[16] and the CD26 used by MERS-CoV.[17] The results of this study shows that, although the SARS-COV-2 whole genome is closely related to bat- SARS-like coronaviruses (Figure 1) and has also relatively similar spike protein (S) (Figure 2), the RBD of virus is much similar to that of SARS-CoVGZ02 (Figure 3A, B), suggesting that they use the same ACE2 as a cell receptor.[24]

CONCLUSION

We have explored the genetic variation among the various viral strains of SARS-

CoV-2 and found that it has close affinities with three bat-SARS-like corona viruses at the whole genome level: bat-SL-CoVZC45, bat-SL-CoVZXC21, bat-SL-CoVYN2018C, SARS-CoVGZ02. The spike proteins (S) of virus is more like bat-SARS-like coronavirus compared to SARS-CoV, however the RBD is much like that of SARS-CoV The virus has unique genetic features which needs further investigation to ascertain their role in viral replication cycle and pathogenesis. More animal samples to determine the natural animal reservoir and intermediate animal host is important.

ACKNOWLEDGMENT

REFERENCES

1. 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(10223):497-506. https://doi.org/10.1016/S0140-6736(20)30183-5

2. Peiris JS, Guan Y, Yuen KY. Severe acute respiratory syndrome. Nat Med 2004;10(12 Suppl):S88-97. https://doi.org/10.1038/nm1143.

3. Zaki AM, van Boheemen S, Bestebroer TM, Osterhaus AD, Fouchier RA. Isolation of a novel coronavirus from a man with pneumonia in Saudi Arabia. N Engl J Med 2012;367(19):1814-20. https://doi.org/10.1056/NEJMoa1211721.

4. Hajjar SA, Memish ZA, McIntosh K. Middle East Respiratory Syndrome Coronavirus (MERS-CoV): a perpetual challenge. Ann Saudi Med 2013;33(5):427-36. https://doi.org/10.5144/0256-4947.2013.427.

5. Tang Q, Song Y, Shi M, Cheng Y, Zhang W, Xia XQ. Inferring the hosts of coronavirus using dual statistical models based on nucleotide composition. Sci Rep 2015;5:17155. https://doi.org/10.1038/srep17155.

6. Fehr AR, Perlman S. Coronaviruses: an overview of their replication and pathogenesis. Methods Mol Biol 2015;1282:1-23. https://doi.org/10.1007/978-1-4939-2438-7_1.

7. Su S, Wong G, Shi W, Liu J, Lai ACK, Zhou J, et al. Epidemiology, Genetic Recombination, and Pathogenesis of Coronaviruses. Trends Microbiol 2016;24(6):490-502. https://doi.org/10.1016/j.tim.2016.03.003.

8. Marra MA, Jones SJ, Astell CR, Holt RA, Brooks-Wilson A, Butterfield YS, et al. The Genome sequence of the SARS-associated coronavirus. Science 2003;300(5624):1399-404. https://doi.org/10.1126/science.1085953.

9. Ruan YJ, Wei CL, Ee AL, Vega VB, Thoreau H, Su ST, et al. Comparative full-length genome sequence analysis of 14 SARS coronavirus isolates and common mutations associated with putative origins of infection. Lancet 2003;361(9371):1779-85. https://doi.org/10.1016/s0140-6736(03)13414-9.

10. Fung TS, Liu DX. Human Coronavirus: Host-Pathogen Interaction. Annu Rev Microbiol 2019;73:529-57. https://doi.org/10.1146/annurev-micro-020518-115759.

11. Yu F, Du L, Ojcius DM, Pan C, Jiang S. Measures for diagnosing and treating infections by a novel coronavirus responsible for a pneumonia outbreak originating in Wuhan, China. Microbes Infect 2020;22(2):74-9. https://doi.org/10.1016/j.micinf.2020.01.003.

12. 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):1547-9. https://doi.org/10.1093/molbev/msy096.

13. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 1987;4(4):406-25. https://doi.org/10.1093/oxfordjournals.molbev.a040454.

14. Felsenstein J. Confidence limits on phylogenies: an approach using the bootstrap. Evolution 1985;39(4): 783-91. https://doi.org/10.1111/j.1558-5646.1985.tb00420.x.

15. Tamura K, Nei M, Kumar S. Prospects for inferring very large phylogenies by using the neighbor-joining method. Proc Natl Acad Sci USA 2004;101(30):11030-5. https://doi.org/10.1073/pnas.0404206101.

16. Li F, Li W, Farzan M, Harrison SC. Structure of SARS coronavirus spike receptor-binding domain complexed with receptor. Science 2005;309(5742):1864-8. https://doi.org/10.1126/science.1116480.

17. Lu G, Hu Y, Wang Q, Qi J, Gao F, Li Y, et al. Molecular basis of binding between novel human coronavirus MERS-CoV and its receptor CD26. Nature 2013;500(7461):227-31. https://doi.org/10.1038/nature12328.

18. Wang N, Shi X, Jiang L, Zhang S, Wang D, Tong P, et al. Structure of MERS-CoV spike receptor-binding domain complexed with human receptor DPP4. Cell Res 2013;23(8):986-93. https://doi.org/10.1038/cr.2013.92.

19. Wang Q, Qi J, Yuan Y, Xuan Y, Han P, Wan Y, et al. Bat origins of MERS-CoV supported by bat coronavirus HKU4 usage of human receptor CD26. Cell Host Microbe 2014;16(3):328-37. https://doi.org/10.1016/j.chom.2014.08.009.

20. Hu D, Zhu C, Ai L, He T, Wang Y, Ye F, et al. Genomic characterization and infectivity of a novel SARS-like coronavirus in Chinese bats. Emerg Microbes Infect 2018;7(1):154. https://doi.org/10.1038/s41426-018-0155-5.

21. Han Y, Du J, Su H, Zhang J, Zhu G, Zhang S, et al. Identification of Diverse Bat Alphacoronaviruses and Betacoronaviruses in China Provides New Insights Into the Evolution and Origin of Coronavirus-Related Diseases. Front Microbiol 2019;10:1900. https://doi.org/10.3389/fmicb.2019.01900.

22. Guan Y, Zheng BJ, He YQ, Liu XL, Zhuang ZX, Cheung CL, et al. Isolation and characterization of viruses related to the SARS coronavirus from animals in

southern China. Science 2003;302 (5643):276-8. https://doi.org/10.1126/science.1087139.

23. Alagaili AN, Briese T, Mishra N, Kapoor V, Sameroff SC, Burbelo PD, et al. Middle East respiratory syndrome coronavirus infection in dromedary camels in Saudi Arabia. mBio 2014;5(2):e00884-14. https://doi.org/10.1128/mBio.00884-14.

24. Zhou P, Yang X-L, Wang X-G, Hu B, Zhang L, Zhang W, et al. Discovery of a novel coronavirus associated with the recent pneumonia outbreak in humans and its potential bat origin. bioRxiv 2020:2020.01. 22.914952. https://doi.org/10.1101/2020.01.22.914952.

AUTHOR'S CONTRIBUTION

Following authors have made substantial contributions to the manuscript as under:

MT: Conception and study design, acquisition of data, drafting the manuscript, critical review, approval of final version to be published

FJ: Conception and study design, analysis and interpretation of data, critical review, approval of final version to be published

NAL: Analysis and interpretation of data, critical review, approval of final version to be published

Me and UF: Drafting the manuscript, critical review, approval of final version to be published

Authors agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

DATA SHARING STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.