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

Muhammad Tariq1, Fatima Jehan1, Nooria Ali Liaqat1, Mehnaz1, Umar Farooq2

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) 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 were more like that of SARS-CoV-2GZ02.

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-CoV-2GZ02, 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).

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 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, and the Middle East Respiratory Syndrome Coronavirus (MERS-CoV) [2012-present], which appeared first in Saudi Arabia, have also caused worldwide outbreaks. Most of the viruses from the coronavirus family infect mammals, whereas others infect a variety of birds.

All coronaviruses including SARS-CoV-2 have a single positive stranded RNA genome ranging from 26 to 32 kilobases in length. 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. 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, which is the major target of virus to neutralize the antibodies. 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 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.
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
the ClustalW program with MEGA-X software. Phyllogenetic analysis of whole genome and major coding regions were done with the same MEGA-X software, using the Neighbor-Joining (NJ) method. The consensus bootstrap tree implied from 1000 replicates were used for representing the evolutionary history of the analyzed taxa and the evolutionary distance were calculated using the maximum likelihood (ML) methods. 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 coronaviruses 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-CoV-0-2 and another bat SARS-Like coronavirus, BetaCoVYN2018C (accession no. MK211377). Similarly, most of the encoded proteins have high sequence identity between SARS-CoV-2 and related bat SARS-like coronaviruses.

Phylogenetic analysis of all 131 SARS-CoV-2 and 31 closely related coronavirus 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 coronavirus: 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-CoVYN2018C (Figure 1).

Next, we compared the envelope spike (S) proteins of SARS-CoV-2 with its closest homologs (bat-SL-CoVZC45, bat-SL-CoVZXC21, bat-SL-CoVYN2018C, 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 coronaviruses. The RBD directly interact with the peptidase domain of human receptor, angiotensin-converting enzyme 2 (ACE2) as an SARS-CoV, MERS-CoV and an BatCoV HKU4. 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-CoVYN2018C at the whole genome level, the RBD of SARS-CoV-2 was much closer to that of...
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\(^{13}\) and bat-SL-CoV YN2018C\(^{14}\) 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 SARS-CoV-2 (Figure 1) and therefore bat-SL-CoVZC45 and bat-SL-CoVZXC21 are not 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\(^{15}\) and dromedary camel for MERS-CoV\(^{16}\) 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-CoV YN2018C, SARS-CoVZG02 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\(^{17}\) and the CD26 used by MERS-CoV\(^{18}\). 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-CoVZG02 (Figure 3A, B), suggesting that they use the same ACE2 as a cell receptor.\(^{19}\)

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 coronavirus viruses at the whole genome level: bat-SL-CoVZC45, bat-SL-CoVZXC21, bat-SL-CoVYMN2018C, SARS-CoVZG02. 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

We thankfull to all those who contributed sequences to NCBI (http://www.ncbi.nlm.nih.gov/genbank/) and the GISAID database (https://www.gisaid.org/). No significant financial support were acquired for this study.

REFERENCES


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.

CONFLICT OF INTEREST

Authors declared no conflict of interest

GRANT SUPPORT AND FINANCIAL DISCLOSURE

Authors have declared no specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors

DATA SHARING STATEMENT

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

This is an Open Access article distributed under the terms of the Creative Commons Attribution-Non Commercial 2.0 Generic License.