Amino acid mutations in the protein sequences of human SARS CoV-2 Indian isolates compared to Wuhan-Hu-1 reference isolate from China

Dr. Kunchur Guruprasad, Ph.D

ABREAST™, Plot Nos.14/A & 15, Sitaramnagar, Safilguda, Hyderabad-500056, India

E.mail: abreastkgp@gmail.com, kunchur.guruprasad@gmail.com

Mobile: +91 7337554324

Website: www.abreast.in

Abstract:

The human SARS CoV-2 protein sequences from 22 Indian isolates were compared with the isolate from Wuhan Hu-1, China, epicentre of the current COVID-19 pandemic disease. The Indian isolates from Kerala, Telangana, Karnataka and Gujarat were analysed. Forty four distinct mutations associated with 11 SARS CoV-2 proteins were present in the Indian SARS CoV-2 isolates. The mutations and their associated proteins are; E381D, I671T, I476V, G662R, T265I in nsp2 protein, T2016K, P2376L, V2629A, P2144S, P2079L, S1534L, S1189T, G2035E in nsp3, A3143V in nsp4, L3606F in nsp6, P4715L, A4489V, A4798V, L4721I, V5272I, T5036M, A4577V, I4593L in RNA dependent RNA polymerase, S5490A, T5538I, P5828L in helicase protein, D6724G, D6719Y, V6600A in endoRNAse, Y28H, D614G, C1250F, Q271R, Y145-del, R408I, A930V in spike glycoprotein, P344S, P13L, R203K, T393I, S194L, S33I in nucleocapsid phosphoprotein N, Q57H in orf3a and L84S in orf8 protein. The D614G in the spike protein and P4715L mutations in RNA dependent RNA polymerase are present in relatively higher numbers among these SARS CoV-2 isolates. The mutations reported in this work add to the growing knowledge of human SARS CoV-2 mutations in the global context.

Introduction:

The novel severe acute respiratory syndrome coronavirus-2 (SARS CoV-2) responsible for the current COVID-19 pandemic disease has resulted in 4,429,810 coronavirus cases and 298,174 deaths worldwide (https://www.worldometers.info/coronavirus/). The confirmed coronavirus cases and deaths in India are 78,121 and 2551, respectively (https://www.worldometers.info/coronavirus/country/india/) as on 14th May 2020. The disease caused by the virus is known to spread via aerosol droplets released by the infected person upon coughing or sneezing. Efforts are underway by several groups in the world for developing a suitable vaccine/drug to treat the disease. Management of the spread of virus infection is currently being controlled by implementing countrywide lockdowns, promoting social distancing, incorporating sanitisation methods, advocating maintenance of regular personal hygiene and educating the masses via the media and other communication channels on the practise of proper care and prevention of the disease.

SARS CoV-2 is a positive sense single stranded RNA genome. The complete genome corresponding to the novel SARS CoV-2 was isolated from an infected individual in Wuhan, using metagenomic RNA sequencing and assembly (Wu et al., 2020) and the virus is now designated as 'Wuhan-Hu-1' coronavirus. This genome shares 79.6% sequence identity to SARS-CoV and is 96% identical at the whole-genome level to a bat coronavirus (SARS-CoV RaTG13) (Zhou et al., 2020). The first report of complete human SARS CoV-2 genome sequences from two infected individuals with a travel history from Wuhan, China was reported from the state of Kerala, India (Yadav et al., 2020). The genome sequences can be accessed publicly from the NCBI databank available at https://www.ncbi.nlm.nih.gov/. Subsequently, the complete genomes of human SARS CoV-2 collected from infected individuals from different states in India have been sequenced and deposited in NCBI databank. At the time of this communication, a total of 22 genomes of the human SARS CoV-2 Indian isolates are available in the databank. These include 2 from Kerala, 4 from the state of Telangana, 7 from Karnataka and 9 from Gujarat.

Among the human SARS CoV-2 proteins, the surface glycoprotein or spike protein has been the focus of intense study by several researchers worldwide as it recognizes the angiotensin converting enzyme 2 (ACE-2) receptor for entry into the human cells (Zhou et al., 2020) to cause infection resulting in the present COVID-19 disease and is therefore a prime target for therapeutic development.

Here, we compare all protein sequences coded by the gene products in the human SARS CoV-2 genomes from the 22 Indian isolates with the human SARS CoV-2 Wuhan-Hu-1 isolate from China used as the reference sequence for the comparisons (Wu et al., 2020). We report all mutations present in the proteins among the 22 human SARS CoV-2 isolates from India.

Materials and Methods:

The human SARS CoV-2 protein sequences were obtained from the publicly accessible NCBI databank. The multiple sequence alignments of individual family of proteins were generated using the online program CLUSTAL-OMEGA (Madeira et al., 2019) available at the EBI website (https://www.ebi.ac.uk/). The alignments were carefully examined to identify the mutations in the Indian isolates relative to the Wuhan-Hu-1 isolate.

Results and Discussion:

Proteins in human SARS CoV-2

The SARS CoV-2 isolate from Wuhan Hu-1 complete genome sequence locus QHD43415 (GenBank code:MN908947.3) contains 29,903 base pairs single stranded RNA and is identical to the reference sequence (NCBI Accession code: NC_045512.2). The gene "orf1ab" corresponding to the protein product "orf1ab polyprotein" comprises 7096 amino acid residues. The polyprotein codes for 15 protein products and additionally, there are 10 other proteins towards the 3' in the SARS CoV-2 genome. According to the annotation available in the NCBI databank, the large polyprotein "orf1ab" comprises the following protein products; "leader protein" (1-180 amino acids), non-structural proteins "nsp2" (181-818), "nsp3" (819-2763), "nsp4" (2764-3263), "3C-like proteinase" (3264-3569), "nsp6" (3570-3859), "nsp7" (3860-3942), "nsp8" (3943-4140), "nsp9" (4141-4253), "nsp10" (4254-4392), "RNA dependent RNA polymerase" (4393-5324), "helicase" (5325-5925), "3' -to-5' exonuclease" (5926-6452), "endoRNAse" (6453-6798),"2' -O-ribose methyltransferase" (6799-7096). The other proteins include the surface glycoprotein or spike protein comprising 1273 amino acid residues (GenBank id: QHD43416.1), the envelope protein 'E' comprising 75 amino acid residues (QHD43418.1), the membrane glycoprotein 'M' comprising 222 amino acid residues (QHD43419.1), the nucleocapsid phosphoprotein 'N' 419 amino acids (QHD43423.2). The other genes in the SARs-CoV-2 genome correspond to the open reading frames; orf3a protein comprising 275 amino acid residues (QHD43417.1), orf6 comprising 61 amino acid residues (QHD43420.1), orf7a comprising 121 amino acid residues (QHD43421.1), orf8 comprising 121 amino acid residues and the orf10 comprising 38 amino acid residues (QHI42199.1).

The NCBI genome accession codes, city/state/sequencing centres for the human SARS CoV-2 isolates from states of Kerala, Karnataka, Telangana and Gujarat in India along with genome references are shown in Table 1. The reference genome and protein codes for the human SARS CoV-2 Wuhan-Hu-1, China isolate is included.

Mutations in proteins of human SARS CoV-2 Indian isolates relative to the Wuhan-Hu-1 isolate:

A total of forty-four distinct mutations were identified among the 22 Indian human SARS CoV-2 isolates from the states of Kerala, Karnataka, Telangana and Gujarat. The mutations identified in the individual protein sequences corresponding to the different genomes is shown in Table 2. The table includes mutations observed in the "orf1ab" polyprotein comprising 15 proteins and the 10 other proteins with reference to the Wuhan Hu-1 isolate. Table 2 provides mutations associated with individual proteins across the genomes.

Mutations in "orf1ab" polyprotein sequence comprising 7096 amino acid residues:

The distinct mutations present in the Indian isolates relative to the Wuhan-Hu-1 human SARS CoV-2 isolate are mentioned below. The "nsp2" protein comprises five mutations; E381D, I671T, I476V, G662R, T265I. The "nsp3" protein comprises eight mutations; T2016K, P2376L, V2629A, P2144S, S1534L, S1189T, G2035E, P2079L. The "nsp4" is associated with a single mutation; A3143V. The "nsp6" also has a single mutation; L3606F. The RNA dependent RNA polymerase is associated with eight mutations; P4715L, A4489V, A4798V, L4721I, V5272I, T5036M, A4577V, I4593L. The helicase protein comprises three mutations; S5490A, T5538I, P5828L. The endoRNAse protein comprises three mutations; V6600A, D6724G and D6719Y. The spike glycoprotein comprises seven mutations; Y28H, D614G, C1250F, Q271R, Y145 deletion mutant, R408I, A930V. Six distinct mutations were observed in the nucleocapsid phosphoprotein 'N'; P344S, P13L, R203K, T393I, S194L, S33I. The single mutation observed in orf3a protein is Q57H and in the orf8 protein is L84S mutation.

The emergence of a more transmissible form of SARS CoV-2 associated with the D614G mutation reported by the Los Alamos National Laboratory, NM, USA (Korber et al., 2020) is present in the Telangana, Karnataka and Gujarat SARS CoV-2 isolates. In addition, we observe that the P4715L mutation is present in 14

isolates and the A4489V mutation present in 5 isolates in the RNA dependent RNA polymerase protein based on their distribution among the states of Telangana, Karnataka and Gujarat in India. Three mutations in the spike glycoprotein; Y145-del, R408I, A930V present in two SARS CoV-2 isolates from Kerala (Yadav et al., 2020) are not present in the isolates from Telangana, Karnataka or Gujarat.

The type of amino acid residue mutation at a particular position along the sequence was observed to be the same in all the genomes analysed so far. For instance, the P2376L mutation in the "nsp3" protein, proline at position 2376 is always observed to be mutated to leucine in all the genomes (MT396242.1, MT396243.1, MT396244.1, MT396245.1, MT396246.1, MT396247.1, MT396248.1). Q57H is present in three of the isolates from Gujarat in the orf3a protein. The P13L and R203K mutations in the nucleocapsid phosphoprotein 'N' is present in some of the isolates from Telangana and Karnataka. The P4715L mutation in RNA dependent RNA polymerase and the D614G spike protein mutations are present in some of the isolates from Telangana, Karnataka and Gujarat. The mutations in the human SARS CoV-2 first isolated and sequenced from two infected individuals in Kerala are distinct from the mutations present in the isolates from Karnataka, Telangana and Gujarat.

Conclusions:

Forty four distinct mutations are present in the proteins of twenty two human SARS CoV-2 isolates from the infected individuals from the States of Kerala, Telangana, Karnataka and Gujarat in India compared to the reference human SARS CoV-2 Wuhan-Hu-1 isolate from China. The mutations are associated with 11 distinct proteins in the human SARS CoV-2; “nsp2” protein, “nsp3” protein, “nsp4” protein, “nsp6” protein, “RNA dependent RNA polymerase”, “helicase”, “endoRNAse”, spike glycoprotein, nucleocapsid phosphoprotein ‘N’, and the orf3a, orf8 proteins. Our results point towards the emerging human SARS CoV-2 mutations in India. A pool of mutations gathered from large datasets would provide insights into how the virus is capable of adapting to the human host and strategies for targeted therapeutic development to combat the COVID-19 pandemic.

Acknowledgements:

The author sincerely acknowledges the several researchers and genome sequencing centres for making available the complete genomes in the NCBI public repository for access.

Conflict of interest:

The author declares no conflict of interest

Funding:

None

References:

Korber, B., Fischer, W., Gnanakaran, S. G., Yoon, H., Theiler, J., Abfalterer, W., ... & Partridge, D. G. (2020). Spike mutation pipeline reveals the emergence of a more transmissible form of SARS-CoV-2. *bioRxiv*.

Madeira, F., Park, Y. M., Lee, J., Buso, N., Gur, T., Madhusoodanan, N., ... & Lopez, R. (2019). The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic acids research*, *47*(W1), W636-W641.

Schäffer, A. A., Aravind, L., Madden, T. L., Shavirin, S., Spouge, J. L., Wolf, Y. I., ... & Altschul, S. F. (2001). Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic acids research*, *29*(14), 2994-3005.

Wu, F., Zhao, S., Yu, B., Chen, Y. M., Wang, W., Song, Z. G., ... & Yuan, M. L. (2020). A new coronavirus associated with human respiratory disease in China. *Nature*, *579*(7798), 265-269.

Yadav, P. D., Potdar, V. A., Choudhary, M. L., Nyayanit, D. A., Agrawal, M., Jadhav, S. M., ... & Cherian, S. S. (2020). Full-genome sequences of the first two SARS-CoV-2 viruses from India. *Indian Journal of Medical Research*, *151*(2), 200.

Zhou, P., Yang, X. L., Wang, X. G., Hu, B., Zhang, L., Zhang, W., ... & Chen, H. D. (2020). A pneumonia outbreak associated with a new coronavirus of probable bat origin. *nature*, *579*(7798), 270-273.

Table legends:

Table 1. Dataset of the genomes of human SARS CoV-2 Indian isolates analysed with reference to the Wuhan Hu-1 isolate from China

Table 2. Mutations in human SARS CoV-2 associated with different proteins in the Indian isolates with reference to the Wuhan Hu-1 isolate NC_045512.2 from China