AIJR Preprints
Section: Coronavirus
Article Id: 39, Version: 1, 2020
URL: https://preprints.aijr.org/index.php/ap/preprint/view/39
{Click on above link to see the latest available version of this article}

NOT PEER-REVIEWED

A Genetic Perspective of 2019-nCoV in Relation to Cross Species Transmission

Rimjhim Dasgupta

4NBIO, 2502, Glen Classic, Hiranandini Gardens, Powai, Mumbai

Version 1: Received: 15 April 2020 / Approved: 16 April 2020 / Online: 16 April 2020

ABSTRACT

Coronaviruses have caused two large scale pandemics severe acute respiratory syndrome (SARS) and Middle East respiratory syndrome (MERS) in last two decades. It was thought that SARS-related coronaviruses (SARSr-CoV) is mainly found in bats. Previous studies have shown that some bat SARSr-CoVs have the potential to infect humans. The current outbreak of viral pneumonia in the city of Wuhan, China, was caused by a novel coronavirus designated 2019-nCoV by the World Health Organization, as determined by sequencing the viral RNA genome. Many initial patients were exposed to wildlife animals at the Huanan seafood wholesale market, where poultry, snake, bats, and other farm animals were also sold. Here we have taken an attempt to understand the genetic structure of 2019-nCoV and subsequent sequence analysis of multiple regions of its genome to identify unique motifs, receptor binding domain, hypervariable region which may direct some insight to future research for developing effective treatment against this novel coronavirus. We have identified unique motif in spike protein, multiple hypervariable regions, amino acids polymorphism in ORF8 and N protein. These may affect the conformation of the peptide and shed some light to cross species transmission, and subsequent host adaptation.

Keywords: Genomics, Sequence Analysis, 2019-nCoV

1 Introduction

The Corona Virus Disease 2019 (COVID-19) caused by a novel coronavirus (CoV) named “2019 novel coronavirus” or “2019-nCoV” by the World Health Organization (WHO) is responsible for the recent pneumonia outbreak that started in early December, 2019 in Wuhan City, Hubei Province, China. This outbreak is associated with a large seafood and animal market, and investigations are ongoing to determine the origins of the infection. Many initial patients were exposed to wildlife animals at the Huanan seafood wholesale market, where poultry, snake, bats, and other farm animals were also sold.

Coronaviruses mainly cause respiratory and gastrointestinal tract infections and are genetically classified into four major genera: Alphacoronavirus, Beta coronavirus, Gamma coronavirus, and Delta coronavirus. The former two genera primarily infect mammals, whereas the latter two predominantly infect birds. Human CoVs include HCoV-NL63 and HCoV-229E, which belong to the Alpha coronavirus genus; and HCoV-OC43, HCoVHKU1, severe acute respiratory syndrome coronavirus (SARS-CoV), and Middle East respiratory syndrome coronavirus (MERS-CoV), which belong to the Beta coronavirus Genus. SARS-CoV and MERS-CoV are considered highly pathogenic, and it is very likely that both SARS-CoV and MERS-CoV were transmitted from bats to palm civets or dromedary camels, and finally to humans. There are still controversies about the source of the 2019-nCoV and its intermediate host. Many studies have proved the pathogen of COVID-19 is a novel coronavirus, which belongs to the Coronavirus family, Beta coronavirus genus and Sarbecovirus subgenus, with a linear single-stranded positive-strand RNA genome of about 30 kb (Ceraolo and Giorgi, 2020). An attempt has been taken in this article to analyse genetic perspective of 2019-nCoV by taking advantage of currently available genomic sequences from patients and literature information.

2 Results and Discussion

The genome includes a variable number (from 6 to 11) of open reading frames (ORFs) (Song et al 2019). The first ORF representing approximately 67% of the entire genome encodes 16 non-structural proteins (nsps), while the remaining ORFs encode accessory proteins and structural proteins (Figure 1). The four major structural proteins are the spike surface glycoprotein (S), small envelope protein (E), matrix protein (M), and nucleocapsid protein (N).

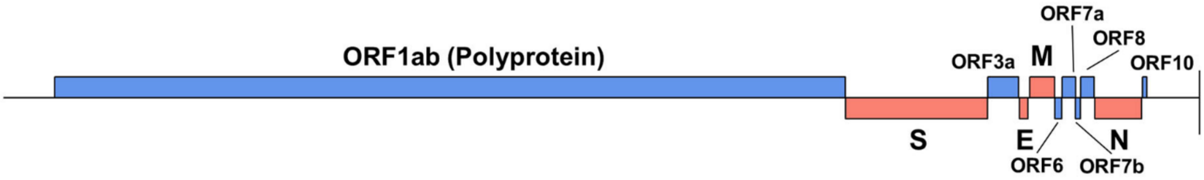
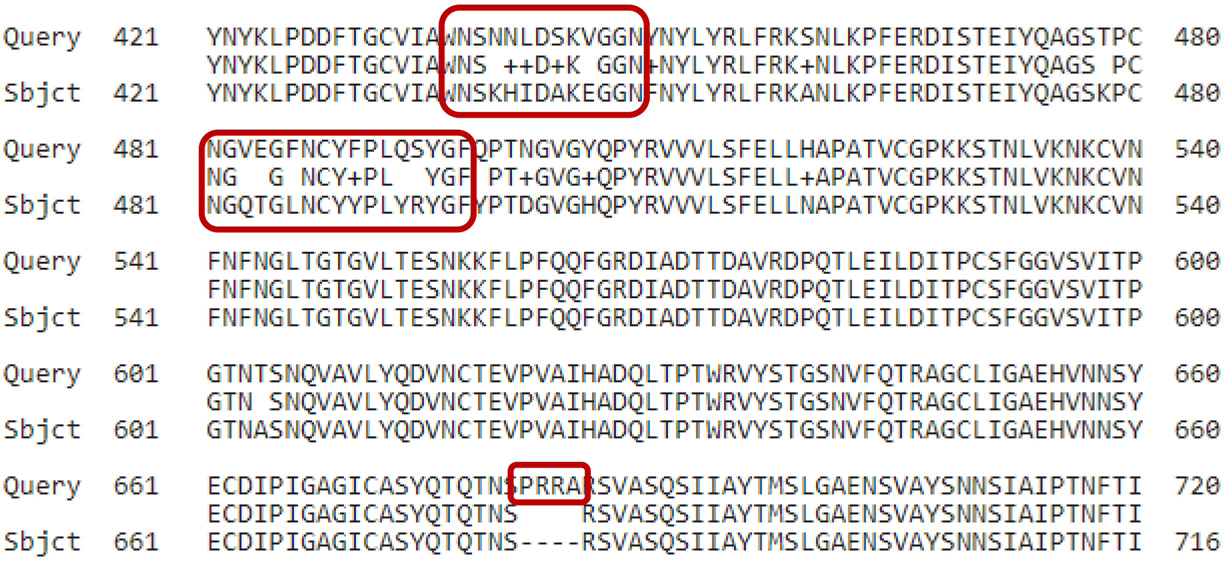


Figure 1: 2019-nCoV genomic sequence

A large gene encoding for a polyprotein (ORF1ab) at the 5' end of the genome is followed by four major structural protein-coding genes: S = Spike protein, E = Envelope protein, M= Membrane protein, and N =Nucleocapsid protein. There are also at least six other accessory open reading frames (ORFs).

Our sequence alignment result shows that 2019-nCoV and bat coronavirus RaTG13 (GenBank No.: MN996532) have the highest homology in the whole genome, ORF1ab, nucleocapsid protein (N), and spike protein (S). Furthermore, the amino acid homologies of ORF1ab, N, S proteins of the two viruses are 98.55, 99.05 and 97.41% respectively. These suggest the two viruses have a high genetic relationship (Wu et al. 2020)

Our study shows that S protein of the two strains (YP_009724390.1, QHR63300.2) has 33 different amino acids with major differences are located at 439–449 and 482–505 (Figure: 2, marked in box). Apart from that, the 2019-nCoV virus has a unique peptide (PRRA) insertion which is consistent with previous study (Li et al. 2020). It may be involved in the proteolytic cleavage of the S protein by cellular proteases, and impact host range and transmissibility. The PRRA motif is located at the 681 of the 2019-nCoV S protein, but not in the S protein of the bat coronavirus RaTG13. This may be involved in the proteolytic cleavage of the spike protein by cellular proteases, and thus could impact host range and transmissibility.



```
Query  421   YNYKLPDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPC  480
             YNYKLPDDFTGCVIAWNS ++D+K GGN+NYLYRLFRK+NLKPFERDISTEIYQAGS PC
Sbjct  421   YNYKLPDDFTGCVIAWNSKHIDAKEGGNFNYLYRLFRKANLKPFERDISTEIYQAGSKPC  480

Query  481   NGVEGFNCYFPLQSYGFQPTNGVGYQPYRVVVLSFELLHAPATVCGPKKSTNLVKNKCVN  540
             NG  G NCY+PL  YGF PT+GVG+QPYRVVVLSFELL+APATVCGPKKSTNLVKNKCVN
Sbjct  481   NGQTGLNCYYPLYRYGFYPTDGVGHQPYRVVVLSFELLNAPATVCGPKKSTNLVKNKCVN  540

Query  541   FNFNGLTGTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITP  600
             FNFNGLTGTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITP
Sbjct  541   FNFNGLTGTGVLTESNKKFLPFQQFGRDIADTTDAVRDPQTLEILDITPCSFGGVSVITP  600

Query  601   GTNTSNQVAVLYQDVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSY  660
             GTN SNQVAVLYQDVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSY
Sbjct  601   GTNASNQVAVLYQDVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSY  660

Query  661   ECDIPIGAGICASYQTQTNSPRRARSVASQSIIAYTMSLGAENSVAYSNNSIAIPTNFTI  720
             ECDIPIGAGICASYQTQTNS    RSVASQSIIAYTMSLGAENSVAYSNNSIAIPTNFTI
Sbjct  661   ECDIPIGAGICASYQTQTNS----RSVASQSIIAYTMSLGAENSVAYSNNSIAIPTNFTI  716
```

Figure 2: Sequence alignment Query: 2019-nCoV S protein; Subject: RaTG13 S protein

Moreover, our sequence alignment results show that the S genes of 2019-nCoV and RaTG13 are longer than other SARSr-CoVs. The major differences in the sequence of the S gene of 2019-nCoV are the three short insertions in the N-terminal domain as well as changes in four out of five of the key residues in the receptor-binding motif compared with the sequence of 2019-nCoV. Whether the insertions in the N-terminal domain of the S protein of 2019-nCoV confer sialic-acid-binding activity as it does in MERS-CoV needs investigation.

Spike protein (S), a structural protein located on the outer envelope of the virion, binds to the host-receptor angiotensin-converting enzyme 2 (ACE2). In the S1 subunit, the receptor-binding domain (RBD), spanning about 200 residues, consists of two subdomains: the core and external subdomains (Figure 3). The external subdomain contains two exposed loops on the surface, which bind with ACE2 (Liu et al. 2020). Investigating the evolutionary relationship of the RBD sequence in spike protein is helpful for understanding the virus origin trends.

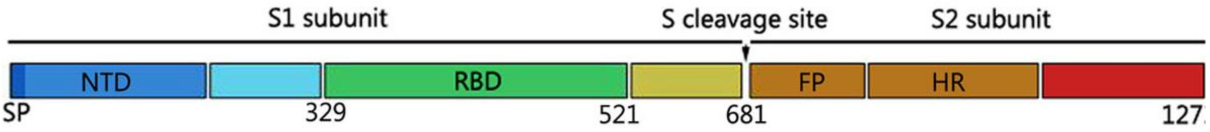


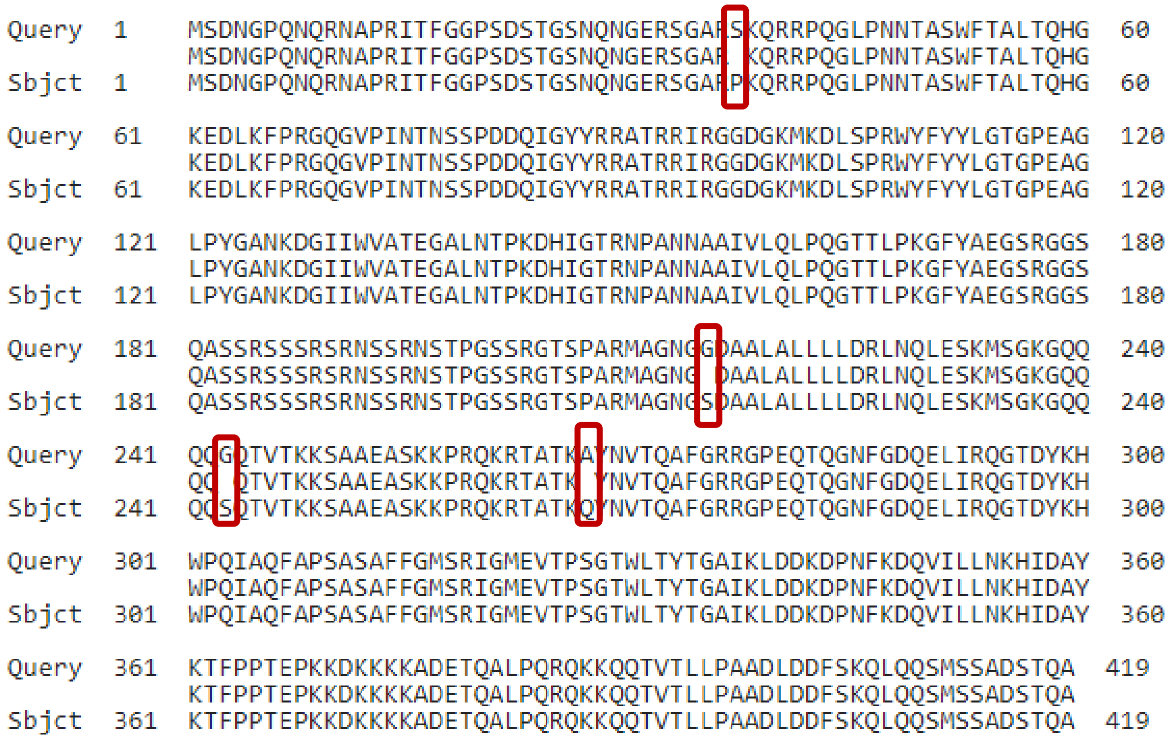
Figure 3: Spike protein

2019-nCoV RBD sequence possesses higher identity with pangolin SARS-like CoV SRR10168377 than bat RaTG13. Probably it indicates that if focusing on only the spike RBD, pangolin SARS-like CoV SRR10168377 has a higher probability to cross host barriers and infect humans.

It was reported that 2019-nCoV uses ACE2 as a cellular entry receptor. Virus infectivity studies using HeLa cells was conducted that expressed or did not express ACE2 proteins from humans. It showed that 2019-nCoV is able to use all ACE2 proteins as an entry receptor to enter ACE2-expressing cells, but not cells that did not express ACE2, indicating that ACE2 is probably the cell receptor through which 2019-nCoV enters cells. It has also been reported that 2019-nCoV does not use other coronavirus receptors, such as aminopeptidase N (APN) and dipeptidyl peptidase 4 (DPP4) (Zhou et al. 2020).

Moreover, we have performed sequence alignment (NCBI) of N protein of 2019-nCoV and bat coronavirus RaTG13 (YP_009724397.2, QHR63308.1)_had 4 different amino acids, which were 37S/P, 215G/S, 243G/S, and 267A/Q, respectively (Figure: 4). First three may be responsible for changing the conformation of the peptide.

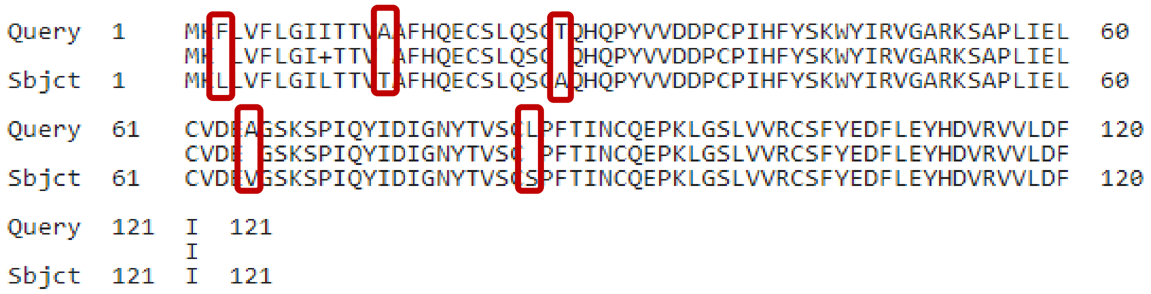
We found >95% identity in ORF8 between 2019-nCoV ORF8 and RaTG13 ORF8 (YP_009724396.1 QHR63307.1 respectively) with 5 changes in positions 3(F/L), 14(A/T), 26(T/A), 65(A/V) and 84 (L/S). Threonine (14 and 26) and Serine (84) are polar whereas Alanine (14) and Leucine (84) are nonpolar amino acids, these may affect the conformation of the peptide.



```
Query  1    MSDNGPQNQRNAPRITFGGPSDSTGSNQNGERSGARSKQRRPQGLPNNTASWFTALTQHG  60
            MSDNGPQNQRNAPRITFGGPSDSTGSNQNGERSGAR KQRRPQGLPNNTASWFTALTQHG
Sbjct  1    MSDNGPQNQRNAPRITFGGPSDSTGSNQNGERSGARPKQRRPQGLPNNTASWFTALTQHG  60

Query  61   KEDLKFPRGQGVPINTNSSPDDQIGYYRRATRRIRGGDGKMKDLSPRWYFYYLGTGPEAG  120
            KEDLKFPRGQGVPINTNSSPDDQIGYYRRATRRIRGGDGKMKDLSPRWYFYYLGTGPEAG
Sbjct  61   KEDLKFPRGQGVPINTNSSPDDQIGYYRRATRRIRGGDGKMKDLSPRWYFYYLGTGPEAG  120

Query  121  LPYGANKDGIIWVATEGALNTPKDHIGTRNPANNAAIVLQLPQGTTLPKGFYAEGSRGGS  180
            LPYGANKDGIIWVATEGALNTPKDHIGTRNPANNAAIVLQLPQGTTLPKGFYAEGSRGGS
Sbjct  121  LPYGANKDGIIWVATEGALNTPKDHIGTRNPANNAAIVLQLPQGTTLPKGFYAEGSRGGS  180

Query  181  QASSRSSSRSRNSSRNSTPGSSRGTSPARMAGNGGDAALALLLLDRLNQLESKMSGKGQQ  240
            QASSRSSSRSRNSSRNSTPGSSRGTSPARMAGNG DAALALLLLDRLNQLESKMSGKGQQ
Sbjct  181  QASSRSSSRSRNSSRNSTPGSSRGTSPARMAGNGSDAALALLLLDRLNQLESKMSGKGQQ  240

Query  241  QQGQTVTKKSAAEASKKPRQKRTATKAYNVTQAFGRRGPEQTQGNFGDQELIRQGTDYKH  300
            QQ QTVTKKSAAEASKKPRQKRTATK YNVTQAFGRRGPEQTQGNFGDQELIRQGTDYKH
Sbjct  241  QQSQTVTKKSAAEASKKPRQKRTATKQYNVTQAFGRRGPEQTQGNFGDQELIRQGTDYKH  300

Query  301  WPQIAQFAPSASAFFGMSRIGMEVTPSGTWLTYTGAIKLDDKDPNFKDQVILLNKHIDAY  360
            WPQIAQFAPSASAFFGMSRIGMEVTPSGTWLTYTGAIKLDDKDPNFKDQVILLNKHIDAY
Sbjct  301  WPQIAQFAPSASAFFGMSRIGMEVTPSGTWLTYTGAIKLDDKDPNFKDQVILLNKHIDAY  360

Query  361  KTFPPTEPKKDKKKKADETQALPQRQKKQQTVTLLPAADLDDFSKQLQQSMSSADSTQA  419
            KTFPPTEPKKDKKKKADETQALPQRQKKQQTVTLLPAADLDDFSKQLQQSMSSADSTQA
Sbjct  361  KTFPPTEPKKDKKKKADETQALPQRQKKQQTVTLLPAADLDDFSKQLQQSMSSADSTQA  419
```

Figure 4: Sequence alignment of N protein of 2019-nCoV and bat coronavirus RaTG13

We also checked the sequence identity between 2019-nCoV ORF8 and bat coronavirus RaTG13 ORF8. We noticed 5 differences as in figure 5. Some of these (A to T, T to A and L to S) could be associated with changing the conformation of the peptide.

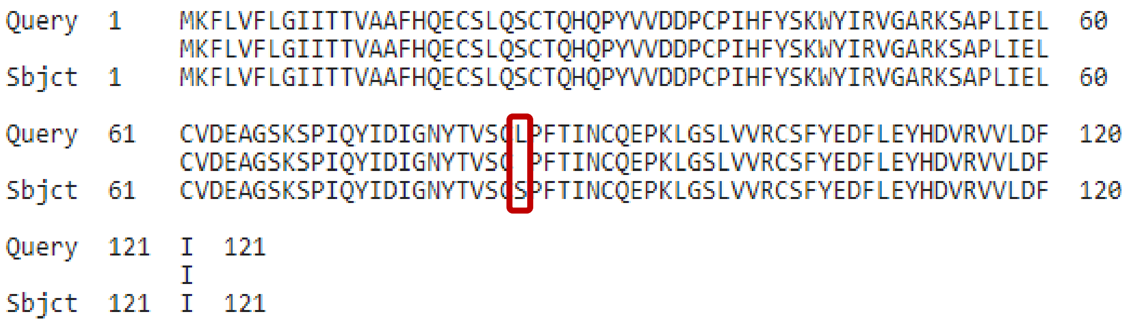


```
Query  1    MKFLVFLGIITTVAAFHQECSLQSCTQHQPYVVDDPCPIHFYSKWYIRVGARKSAPLIEL  60
            MK LVFLGI+TTV AFHQECSLQSC QHQPYVVDDPCPIHFYSKWYIRVGARKSAPLIEL
Sbjct  1    MKLLVFLGILTTVTAFHQECSLQSCAQHQPYVVDDPCPIHFYSKWYIRVGARKSAPLIEL  60

Query  61   CVDEAGSKSPIQYIDIGNYTVSCLPFTINCQEPKLGSLVVRCSFYEDFLEYHDVRVVLDF  120
            CVDE GSKSPIQYIDIGNYTVSC PFTINCQEPKLGSLVVRCSFYEDFLEYHDVRVVLDF
Sbjct  61   CVDEVGSKSPIQYIDIGNYTVSCSPFTINCQEPKLGSLVVRCSFYEDFLEYHDVRVVLDF  120

Query  121  I  121
            I
Sbjct  121  I  121
```

Figure 5: Sequence alignment Query: 2019-nCoV ORF8; Subject: RaTG13 ORF8

The high sequence similarity (>99%) was observed between the available genome sequences of 2019-nCoVs (from patient samples) with low variability. However, we found few hypervariable positions while aligning N protein from different patient samples (amino acid polymorphism 4N/D, 202S/N, 203R/K, 232S/T). First one may affect conformation of the peptide as Asn (N) is a neutral amino acid (with polar side chain) whereas Asp (D) is acidic amino acid.

However, there are at least two hotspots of hypervariability positions within protein-coding regions. Position aa24 and aa84 fall within ORF8 (Figure 6, provided alignment data for aa84 only) and these cause Ser to Leu and Leu to Ser, which can affect the conformation of the peptide, given that Serine is a polar amino acid, and Leucine is nonpolar. Aa24 and aa84 appear to be non-conserved also across other coronaviruses.



```
Query  1    MKFLVFLGIITTVAAFHQECSLQSCTQHQPYVVDDPCPIHFYSKWYIRVGARKSAPLIEL  60
            MKFLVFLGIITTVAAFHQECSLQSCTQHQPYVVDDPCPIHFYSKWYIRVGARKSAPLIEL
Sbjct  1    MKFLVFLGIITTVAAFHQECSLQSCTQHQPYVVDDPCPIHFYSKWYIRVGARKSAPLIEL  60

Query  61   CVDEAGSKSPIQYIDIGNYTVSCLPFTINCQEPKLGSLVVRCSFYEDFLEYHDVRVVLDF  120
            CVDEAGSKSPIQYIDIGNYTVSC PFTINCQEPKLGSLVVRCSFYEDFLEYHDVRVVLDF
Sbjct  61   CVDEAGSKSPIQYIDIGNYTVSCSPFTINCQEPKLGSLVVRCSFYEDFLEYHDVRVVLDF  120

Query  121  I  121
            I
Sbjct  121  I  121
```

Figure 6: Sequence alignment between patients 2019-nCoV ORF8

The levels of genetic similarity between the 2019-nCoV and RaTG13 suggests that the latter does not provide the exact variant that caused the outbreak in humans, but the hypothesis that 2019-nCoV has originated from bats is very likely. It was shown that the novel coronavirus (2019-nCov) is not-mosaic consisting in almost half of its genome of a distinct lineage within the beta coronavirus (Paraskevis et al 2020). However, these genomic features and their potential association with virus characteristics and virulence in humans need further attention. The comprehensive sequence analysis and comparison in conjunction with relative synonymous codon usage (RSCU) bias among different animal species based on the 2019-nCoV sequence suggests that the 2019-nCoV appears to be a recombinant virus between the bat coronavirus and an origin-unknown coronavirus. The recombination occurred within the viral spike glycoprotein, which recognizes cell surface receptor. It was suggested that snake could be the probable wildlife animal reservoir for the 2019-nCoV based on its RSCU bias which is closed to snake compared to other animals (Ji et al. 2020)

3 Conclusion

Taken together this analysis provides some insight about the genomic structure, sequence similarity with other viruses, unique motif in spike protein, receptor binding domain, core positions of high variability and amino acid polymorphism in ORF8 and N protein. The mutation in ORF8 resulting in one of its two variants, ORF8-L and ORF8-S, is predicted to be affecting the structural disorder of the protein. A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission. All together these findings may shed some cautiously light on the possibility of finding effective treatment for this novel coronavirus, starting from already existing anti-betacoronaviridae compounds, which will be dealing with a relatively homogenous viral population. Finally, considering the wide spread of 2019-nCoV in their natural reservoirs, future research should be focused on active surveillance of these viruses for broader geographical regions. Probably in the long term, broad-spectrum antiviral drugs and vaccines may be useful for emerging infectious diseases that are caused by this cluster of viruses in the future.

4 Competing Interests

The author declared that no conflict of interest exists related to the submitted work.

How to Cite:

Rimjhim Dasgupta. "A Genetic Perspective of 2019-nCoV in Relation to Cross Species Transmission". *AIJR Preprints*, 39, version 1, 2020. (URL: https://preprints.aijr.org/index.php/ap/preprint/view/39).

References

Ceraolo C, Giorgi FM. Genomic variance of the 2019-nCoV coronavirus. J Med Virol. 2020;92: 522–528. https://doi.org/10.1002/jmv.25700

Ji et al. Homologous recombination within the spike glycoprotein of the newly identified coronavirus 2019-nCoV may boost cross-species transmission from snake to human. doi: 10.1002/fut.22099

Li X, Zai J, Zhao Q, et al. Evolutionary history, potential intermediate animal host, and cross-species analyses of SARS-CoV-2. J Med Virol. 2020; 1–10. https://doi.org/10.1002/jmv.25731

Li et al. Genetic evolution analysis of 2019 novel coronavirus and coronavirus from other species. Infection, Genetics and Evolution. 2020, Volume 82, https://doi.org/10.1016/j.meegid.2020.104285

Liu Z, Xiao X, Wei X, et al. Composition and divergence of coronavirus spike proteins and host ACE2 receptors predict potential intermediate hosts of SARS-CoV-2. J Med Virol. 2020; 1–7. https://doi.org/10.1002/jmv.25726

Paraskevis et al. Full-genome evolutionary analysis of the novel corona virus (2019-nCoV) rejects the hypothesis of emergence as a result of a recent recombination event. Infection, Genetics and Evolution, 2020, 79, 104212. https://doi.org/10.1016/j.meegid.2020.104212

Song et al. From SARS to MERS, Thrusting Coronaviruses into the Spotlight**.** Viruses 2019, 11(1), 59; https://doi.org/10.3390/v11010059

Wu et al. Genome Composition and Divergence of the Novel Coronavirus (2019-nCoV) Originating in China. Cell Host & Microbe 27, March 11, 2020; https://doi.org/10.1016/j.chom.2020.02.001

Zhou et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature, 2020; Vol 579 https://doi.org/10.1038/s41586-020-2012-7