ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 18(10): 951-961 (2022)

BIOINFORMATION
Discovery at the interface of physical and biological sciences
BIOMEDICAL INFORMATICS
www.bioinformation.net
Volume 18(10)
Research Article

Received September 2, 2022; Revised October 3, 2022; Accepted October 6, 2022, Published October 31, 2022
DOI: 10.6026/97320630018951

Declaration on Publication Ethics:
The author's state that they adhere with COPE guidelines on publishing ethics as described elsewhere at https://publicationethics.org/. The authors also undertake that they are not associated with any other third party (governmental or non-governmental agencies) linking with any form of unethical issues connecting to this publication. The authors also declare that they are not withholding any information that is misleading to the publisher in regard to this article.

Declaration on official E-mail:
The corresponding author declares that lifetime official e-mail from their institution is not available for all authors

Comments from readers:
Articles published in BIOINFORMATION are open for relevant post publication comments and criticisms, which will be published immediately linking to the original article without open access charges. Comments should be concise, coherent and critical in less than 1000 words.

The authors did not declare that an unedited version of this work is already published at https://www.biorxiv.org/content/10.1101/2020.05.12.091199v2

Edited by P Kangueane
Citation: Som *et al.* Bioinformation 18(10): 951-961 (2022)

Recombination in sarbecovirus lineage and mutations/insertions in spike protein are linked to the emergence and adaptation of SARS-CoV-2

Anup Som*, Amresh Kumar Sharma & Priyanka Kumari

Centre of Bioinformatics, Institute of Interdisciplinary Studies, University of Allahabad, Prayagraj – 211002, India; *Corresponding author

Author contacts:
Anup Som - E-mail: som.anup@gmail.com
Amresh Kumar Sharma - E-mail: amresharma1@gmail.com
Priyanka Kumari - E-mail: priyanka.iids@gmail.com

Abstract:
The outbreak of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) in Wuhan city, China in December 2019 and thereafter its spillover across the world has created a global pandemic and public health crisis. Right after, there has been intense interest in understanding how the SARS-CoV-2 originated and evolved. This paper also aims to shed light on the origin and evolution of SARS-CoV-

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 18(10): 951-961 (2022)

2. A consensus result based on whole genome phylogeny, gene tree analysis, and genetic similarity study revealed that SARS-CoV-2 evolved from Bat-CoV-RaTG13. Furthermore, recombination analysis indicated that probable origin of SARS-CoV-2 is the results of ancestral intra-species recombination events between bat coronaviruses belonging to Sarbecovirus sub-genus. Multiple sequence alignment (MSA) revealed the insertion of four amino acid residues "PRRA" (Proline-Arginine-Arginine-Alanine) to the S1/S2 site in the spike protein of SARS-CoV-2, and structural modeling of spike protein of bat-CoV-RaTG13 also shows a high number of mutations at one of the receptor binding domains (RBD). Acquisition of the furin cleavage sites ("PRRA") along with high number of mutations at one of its RBD is probably responsible for the adaptation of SARS-CoV-2 into human systems. Furthermore, the codon adaptation index (CAI) was used to quantify the magnitude of adaptive efficacy of SARS-CoV-2 in human host in comparison with SARS-CoV. The CAI result showed a relatively less adaptive efficacy of the newly emerged SARS-CoV-2 to the human systems, which might be an indication of its mild clinical severity and progression compared to SARS-CoVs.

Keywords: Coronavirus; SARS-CoV-2; molecular phylogeny; recombination; codon adaptation index; spike protein; structural modeling

Background:

Coronaviruses are single-stranded RNA viruses of 26 to 32 kilo bases (Kb) nucleotide chain and consist of both structural and non-structural proteins. They have been known to cause lower and upper respiratory diseases, central nervous system infection and gastroenteritis in a number of avian and mammalian hosts including humans **[1-2]**. The recent outbreak of novel coronavirus (SARS-CoV-2) associated with acute respiratory disease called coronavirus disease 19 (commonly known as COVID-19) has caused a global pandemic. As of 30th June 2022, more than 551 million laboratories confirmed COVID-19 cases and approximately 6.34 million people have died and further COVID-19 appears as a global threat to public health as well as to the human civilization as economic and social disruption caused by the pandemic is devastating (WHO, COVID-19 situation reports). Coronaviruses are placed within the family *Coronaviridae*, which has two subfamilies namely *Ortho coronavirinae* and *Torovirinae*. *Ortho coronavirinae* has four genera: *Alpha coronavirus* (average genome size 28kb), *Betac oronavirus* (average genome size 30kb), *Gamma coronavirus* (average genome size 28kb), and *Delta coronavirus* (average genome size 26kb) **[3]**. Coronaviruses are typically harbored in mammals and birds. Particularly *Alpha coronavirus* and *Beta coronavirus* infect mammals, and *Gamma coronavirus* and *Delta coronavirus* infect avian species **[4-6]**. SARS-CoV-2 is a member of the genus *Beta coronavirus* and subgenus Sarbecovirus. Figure 1 depicts the taxonomical classification of SARS-CoV-2. The previous important outbreaks of coronaviruses are severe acute respiratory syndrome coronavirus (SARS-CoV or SARS-CoV-1) outbreak in China in 2002/03, Middle East respiratory syndrome coronavirus (MERS-CoV) outbreak in 2012 that resulted severe epidemics in the respective geographical regions **[7-9]**. The present outbreak of SARS-CoV-2 is the third documented spillover of an animal coronavirus to humans in only two decades that has resulted in a major pandemic **[10-12]**.

Since COVID-19 started, there has been intense research on the origin and evolution of the SARS-CoV-2, which resulted a very large number of publications on the origin and evolution of SARS-CoV-2. The key reported findings, out of the large number of research outcomes, are: bat and/or pangolin are the natural reservoir of SARS-CoV-2, Bat-CoV-RaTG13 is the closest relative of SARS-CoV-2, transmission of SARS-CoV-2 to human population took place via intermediate hosts, and mutations in the furin cleavage site in spike protein probably linked with the adaptation to the human systems etc. However significant progress has been made towards understanding the origin, transmission and adaptation mechanism of SARS-CoV-2 but the exact origin, cause of emergence and infection mechanism of SARS-CoV-2 are yet to be fully known. Therefore, as more-and-more datasets are generating, there are needs for further in-depth studies on the emergence of SARS-CoV-2.

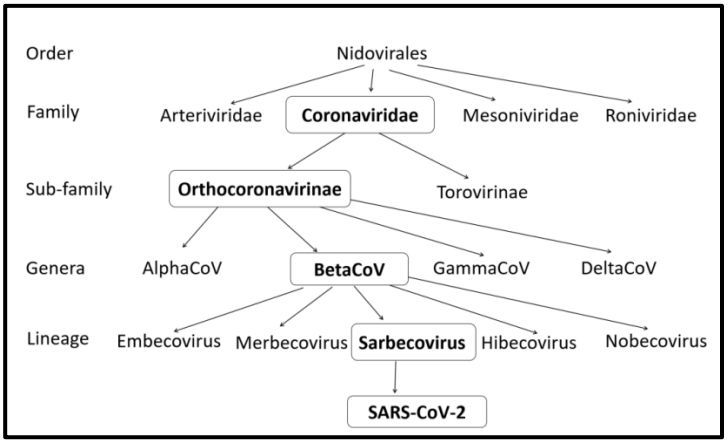

Figure 1: Taxonomical origin/classification of SARS-CoV-2.

Materials and Methods:

Data selection:

162 Orthocoronavirinae genomes were retrieved from NCBI (https://www.ncbi.nlm.nih.gov/), Virus Pathogen Database and Analysis Resource (https://www.viprbrc.org/). We only considered complete genome sequences having no unidentified nucleotide characters. Our datasets included 23 *Alpha coronavirus*, 92 *Beta coronavirus*, 32 *Delta coronavirus* and 15 *Gamma coronavirus* genomes belonging to different subgenus, diverse host species and from wide geographical location. Further for rooting the tree, we used two genome sequences from *Torovirus* and two from *Bafinivirus* belonging to domestic cow and fish respectively. The genera *Torovirus* and *Bafinivirus* belong to the sub-family *Torovirinave* of the family *Coronaviridae*. Overall, the phylogenetic analysis consists of 166 complete viral genomes (162 *Ortho coronavirinae* and four *Torovirinave genomes*).

Phylogenetic reconstruction

The genome sequences were aligned using the MAFFT alignment

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 18(10): 951-961 (2022)

tool **[13]**. Genome tree of the *Ortho coronavirinae* and *Beta coronaviruses* were reconstructed using maximum likelihood (ML) method and GTR+G+I model of nucleotide substitution as revealed by the model test with1000 boots trap support. The model test was performed for accurate phylogenetic estimation by using Model Finder, which is implemented in IQ-TREE version 1.5.4 **[14]**. Phylogenetic trees were reconstructed using IQ-TREE software **[15]**. The trees were visualized with iTOL tool **[16]**. Five gene trees of the *Beta coronaviruses* were reconstructed using Orf1ab, Spike (S), Envelope (E) Membrane (M), and Nucleocapsid (N) amino acid sequences. The ML method of tree reconstruction and protein-specific amino acids substitution model as revealed by Model Finder was used for gene tree reconstruction. Bootstrap test with 1000 bootstrap replicates was carried out to check the reliability of the gene trees.

Genome and gene recombination analysis:
Potential recombination events in the history of the *Beta coronaviruses* were assessed using the RDP5 package **[17]**. The RDP5 analysis was conducted based on the complete genome sequence using RDP, GENECONV, BootScan, MaxChi, Chimera, SiScan, and 3Scan methods. Putative recombination events were identified with a Bonferroni corrected P-value cut-off of 0.05 supported by more than four methods.

Sequence and structural analysis:
The homology and genetic variations analysis of sequences in different genomic regions of SARS-CoV-2 strain Wuhan Hu-01 (MN908947) is compared to bat-CoV-RaTG13 (MN996532) and pangolin-CoV-GX-P5E (MT040336) using CLUSTAL W (https://www.genome.jp/tools-bin/clustalw) and multiple sequence alignment (MSA) analysis of spike proteins were performed using CLUSTAL OMEGA (https://www.ebi.ac.uk/Tools/msa/clustalo/). The structures of the spike protein of SARS-CoV-2 Wuhan Hu-1 (PDB: 6XLU), bat-CoV-RaTG13 (PDB: 6ZGF) were retrieved from PDB database **[18]**. The spike protein for pangolin coronavirus was not available so it was modeled using SWISS-MODEL SERVER (https://swissmodel.expasy.org) with 6XR8 as template. These structures were compared using the structure superimposition/structure alignment tool of Chimera software **[19]**.

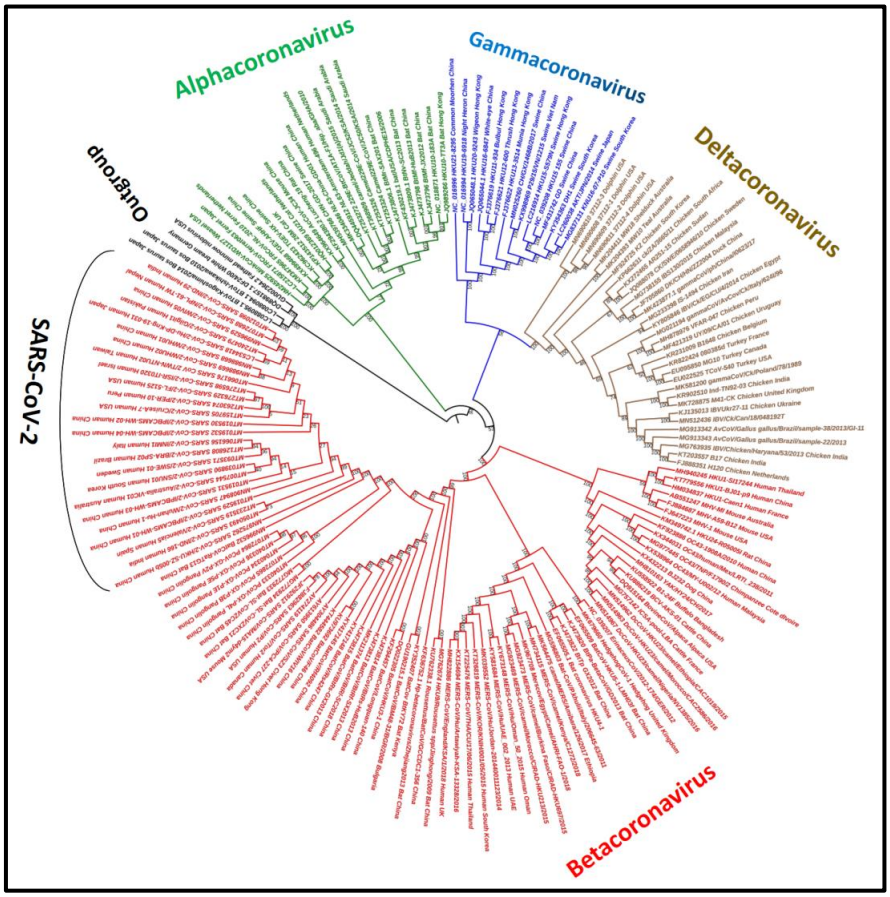

Figure 2: *Orthocoronavirinae* genome phylogeny. The genome tree consists of 162 complete *Orthocoronavirinae* genomes and four outgroups. Alignment consists of 58,538bp aligned nucleotide characters (9,384bp are completely aligned characters). Tree was reconstructed using ML method with GTR+G+I model of nucleotide evolution along with 1000 bootstrap replicates. Tree was rooted with the four Torovirinae genomes (outgroup). SARS-CoV-2 genomes are depicted in *Betacoronavirus*.

Results and Discussion:

This study aims to understand the origin and evolutionary trajectory of SARS-CoV-2 using molecular phylogenetic, genomic diversity, recombination, and structural analyses. Particularly, we carried out the phylogenetic study of human SARS-CoV-2 from their deep ancestral roots (i.e., from the sub-family (*Ortho coronavirinae*) to the lineage (SARS-CoV-2); **Figure** 1). Accordingly, the molecular phylogenetic analysis was based on two-stage; whole genome phylogeny followed by gene trees analyses. Firstly, reconstruction of genome phylogeny of the *Ortho coronavirinae* genomes and study the cladistic/evolutionary relationships of its four genera. Secondly, reconstruction of *Beta coronavirus* genome and gene phylogeny that included its five sub-genera namely *Embecovirus, Hibecovirus, Merbecovirus, Nobecovirus* and *Sarbecovirus,* and study the evolutionary relations of these five subgenera and find linage containing SARS-CoV-2.

***Ortho coronavirinae* phylogeny:**

The genome phylogeny of *Ortho coronavirinae* depicts that Alpha, Beta, Delta and Gamma coronaviruses clustered according to their cladistic relationships where *Alpha coronavirus* clade appeared as a basal radiation of the *Ortho coronavirinae* phylogeny (**Figure** 2). This result is consistent with the other studies **[20, 21]**. Furthermore, analysis of the clades found that *Gamma coronavirus* and *Delta coronavirus* clades are monophyletic (originated from a single common ancestor). This result is supported by their hosts' nature; as both types mostly infect avian species **[22]**. Further, a deeper analysis of the *Ortho coronavirinae* genome tree revealed that irrespective of their geographical locations, the host-specific strains are clustered together. This is probably due to the host-specific mutations, which is an important characteristic of viral genomes for their survival and replication **[23-25]**. For example, *Alphacoronavirus* strains from ferret_Japan and ferret_Netherland are monophyletic. Similarly cat_UK is monophyletic with cat_Netherland, and human_China is monophyletic with human_Netherland. Further analysis revealed all *Alpha coronavirus* camel strains of Saudi Arabia appeared in a distinct sub-clade where bat_Ghana strain appeared as outgroup which indicates interspecies transmission took place from bat_Ghana to camel. A number of scientific evidences, based on the independent datasets, also reported that coronavirus transmission took place to humans through intermediate hosts **[26-29]**.

Delta coronavirus and *Gamma coronavirus* clades exhibit a similar evolutionary pattern. In case of *Delta coronaviruses*, swine_Vietnam and swine_Hong Kong shared a single common ancestor. Similarly, swine_China and swine_South Korea are monophyletic clade and swine_Japan is monophyletic with swine_South Korea. In case of *Gamma coronaviruses* (whose natural hosts are avian species), chicken_Peru and chicken_Uruguay are monophyletic. Similarly, chicken_Iraq is monophyletic with chicken_Egypt strain. These observations reconfirm that coronaviruses are present in a large number of hosts those are widespread in different geographical location and coronaviruses undergo host-specific mutation/adaptation. This is not surprising as sequences of corona viruses isolated from different geographical locations and found that genetic changes through recurrent mutations of the virus are continuously arising, which ultimately promote host adaptation **[30,31]**.

***Beta coronavirus* phylogeny:**

Phylogenetic analysis of *Beta coronavirus* genomes revealed that the five sub genera clustered separately (**Figure** 3). Furthermore, like other three genera, the *Beta coronavirus* genome tree depicts that the host-specific strains from distance geographical locations formed monophyletic clades. For example, in Embecovirus clade, strain BJ01_P9_human_China is monophyletic with Caen1_human_France strain. Similarly, Embecovirus B1_24F_buffalo_Bangladesh is monophyletic with BCV_AKS_01_cattle_China.

SARS-CoV-2 (isolated from human) belongs to Sarbecovirus sub-genus. Sarbecoviruses formed three distinct clades (**Figure** 3), where Clade 1 consists of only bat as host species. InClade2, host species are bat, civet and human. Similarly, in Clade3 the host species are bat, pangolin and human and it depicts bat-CoV-RaTG13 (marked as MRCA in **Figure 3**) is closest to the human SARS-CoV-2 as all human SARS-CoV-2s clustered in a clade, and formed a monophyletic clade with bat-CoV-RaTG13 strain. Clade 3 also shown that pangolin (PCoV-GX-P5E) is the second closest relative of human SARS-CoV-2. Further, deep node analysis of Clade 3, shows that human SARS-CoV-2s, pangolin CoVs (strains PCoV-GX-P4L/P3B/P1E/P5E/P2V) and bat-CoVs (strains bat-SL-CoVZXC21 and bat-SL-CoVZC45) shared a single common ancestor (**Figure 3**). This observation suggests bat and pangolin is the natural host of SARS-CoV. The same inference had also been reported by a number of studies **[24, 26-28, 32]**.

Furthermore, phylogenetic analysis reveals that the MERS-CoVs, SARS-CoVs, SARS-CoV-2s are conserved in their respective hosts (e.g. all bat hosts clustered in Clade 2 and human hosts are in Clade 3). This observation led to the conclusion that host-specific mutations of MERS-CoVs, SARS-CoVs and SARS-CoV-2s occurred, which is probably to facilitate colonization and invade to the host immune system **[26, 33, 34]**.

Codon adaptation index analysis:

The codon adaptation index (CAI) was used to quantify the magnitude of adaptive efficacy of SARS-CoV-2 in human host in comparison with SARS-CoV. The CAI analysis was performed using the CAIcal server **[35]**. It was found that the average CAI value for SARS-CoV-2 with respect to human host is 0.692, which was considerably lower than for SARS-CoV (0.721). This result indicates a relatively less adaptive efficacy of the newly emerged SARS-CoV-2 to the human systems, which might be an indicative of its mild clinical severity and progression compared to SARS-CoVs. Further, in supports of this observation, we thoroughly review the existing scientific evidences on the host adaptation of SARS-Cov-2. A number of studies based on the CAI and relative synonymous codon usage (RSCU) reported that the host adaptation of SARS-CoV-2 occurred and probably this adaptation took place after SARS-CoV-2 diverge from RaTG13 because RaTG13 is less perfectly correlated with human cellular systems **[27,31,36]**.

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 18(10): 951-961 (2022)

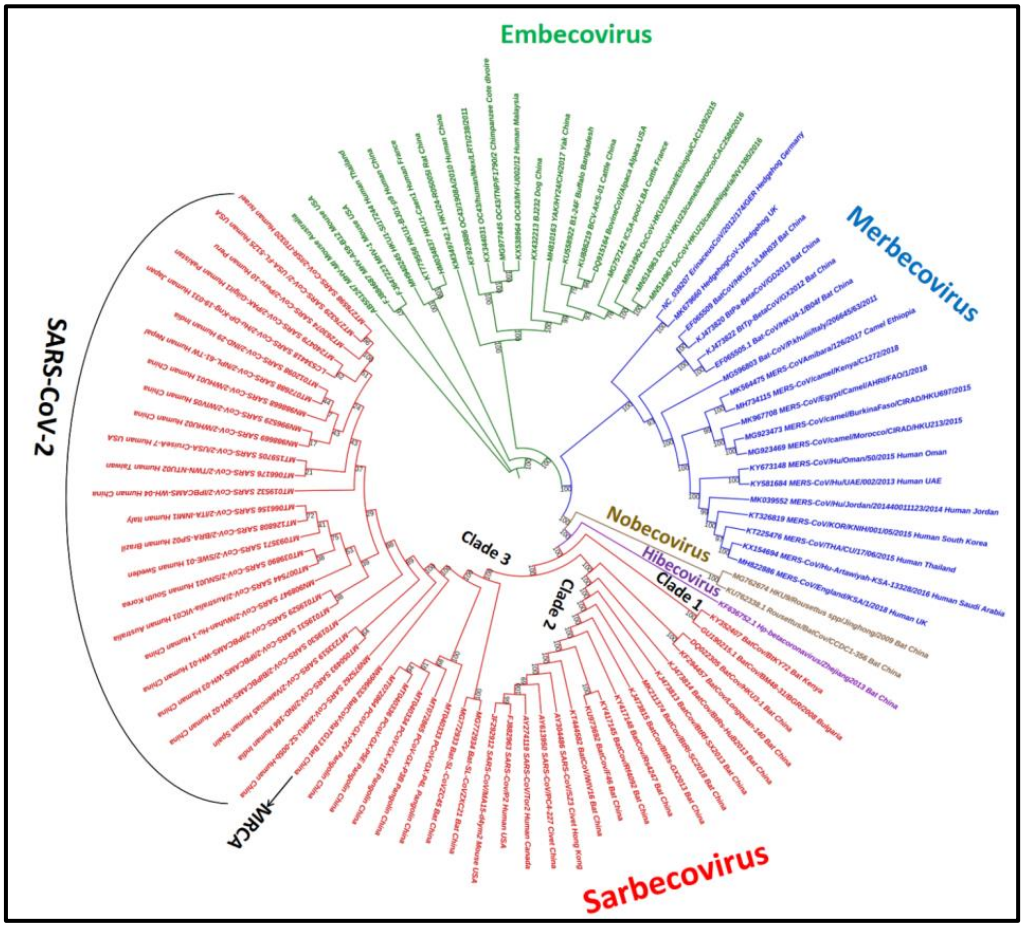

Figure 3: *Betacoronavirus* genome phylogeny. The genome tree consists of 92 complete *Betacoronavirus* genomes. Alignment consists of 41,054bp aligned nucleotide characters (23,064bp are completely aligned characters). Tree was reconstructed using ML method with GTR+G+I model of nucleotide evolution along with 1000 bootstrap replicates. Most recent common ancestor (MRCA) of SARS-CoV-2 is highlighted. Three distinct clades of Sarbecovirus are also depicted.

Average nucleotide identity analysis

To provide additional strength to our observation, we did average nucleotide identity (ANI) analysis, which routinely used to find the closeness between the virus/bacterial species including the identification of the species boundary. ANI between all the Sarbecovirus strains was calculated by using OrthoANIu tool **[37]**. The ANI results show, among all bat and pangolin CoVs, bat RaTG13 share highest ANI (~96%) with the human SARS-CoV-2s (**Figure** 4). Thus, this observation along with phylogeny based evidences indicates that SARS-CoV-2 is thought to have emerged from Bat coronavirus RaTG13.

Gene tree analysis

In addition to the whole genome phylogeny, gene tree analysis was also conducted as it provides a more reliable basis for studying species evolution. Five gene trees namely Orf1ab, Spike, Envelope, Membrane, and Nucleocapsid of the *Betacoronaviruses* were reconstructed for gene tree analysis (**Figure** 5 and Figures. S2-S5). Except Nucleocapsid genetree (**Figure** S5), other four gene trees have shown that the five subgenera clustered according to their cladistic relations where Embecovirus clade appeared as a basal radiation of the *Betacoronavirus* gene trees. Further, these gene trees were in concordance with the genome tree. The topological difference of Nucleocapsid gene tree with the *Betacoronavirus* genome/species tree might be possible as gene tree differs from species tree for various analytical and/or biological reasons **[38-40]**. Further, analysis on the gene trees found, except Envelope gene tree, other four gene trees exhibited bat-CoV-RaTG13 is the closest relative of SARS-CoV-2 followed by pangolin-CoV as found in the genome tree analysis (**Figure** 3 and **Figure** 5). Different evolutionary pattern of Envelope gene tree is probably due to stochastic error as its length is very small (average length 75 amino acids) **[40]**. Further analysis of the gene trees found though subgenera-wise four gene trees are similar, but within subgenera there are widespread phylogenetic incongruence **[41]**. This result led us to hypothesize that recombination events had occurred among *Beta coronaviruses* in the past that are caused to evolve new strains including the emergence of pathogenic lineage like SARS-CoV-2.

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 18(10): 951-961 (2022)

Recombination analysis:

Accordingly, we conducted both genome and gene recombination analysis of the *Beta coronaviruses* using RDP5 package **[17]**. The genome recombination analysis detected 21 putative recombination signals (Table 1). A recombination event was reported when five out of seven methods detected it. Recombination results show that major recombination events took place between bat coronaviruses belonging to the subgenus Sarbecoviruses. A recent study by Boni *et al.* (2020) also reported the Serbicoviruses lineage undergoes frequent recombination **[42]**. For further insights, we compared SARS-CoV-2 Hong Kong (HKU_SZ_005b) genome sequence with four closely related SARS-CoVs namely Bat-CoV-RaGT13, Bat-SL-CoVZC45, Bat-SL-CoVZXC21, and Pangolin-CoV-GX-P5E using simplot analysis (**Figure** 6). Simplot exhibits that bat-CoV-RaTG13 shows the highest similarity with SARS-CoV-2 genome including exchange of genetic materials at the different regions as shown in Figure 6. We classified the whole genomes into four regions (Regions1-4). In region 1 (which mostly covers ORF1a gene), we observed highest genetic divergence between pangolin and SARS-CoV-2 strains, and bat to bat recombination events were frequent. In region2 (ORF1b gene), recombination events mostly took place between bat and pangolin strains. In region3 (Spike gene), bat-CoV-RaTG13 genome shows divergence with SARS-CoV-2 genome and there is a good number of genetic recombination among the bat and pangolin strains. In region4 (E, M, N and ORF3/6-8/10 genes), all strains show high similarity and a few number of recombination events with the SARS-CoV-2 strain. Further, gene recombination analysis found that there are highest recombination events in spike protein (spotted nine events) followed by Orf1ab protein (six events). Membrane and Nucleocaspid proteins reported few recombination events and envelope protein did not show any recombination event. Overall, recombination results support our phylogenetic inference and suggest that the origin of SARS-CoV-2 is the results of ancestral intra-species recombination events between bat SARS-CoVs **[43-44]**. Details of recombination analysis are given in Table 1.

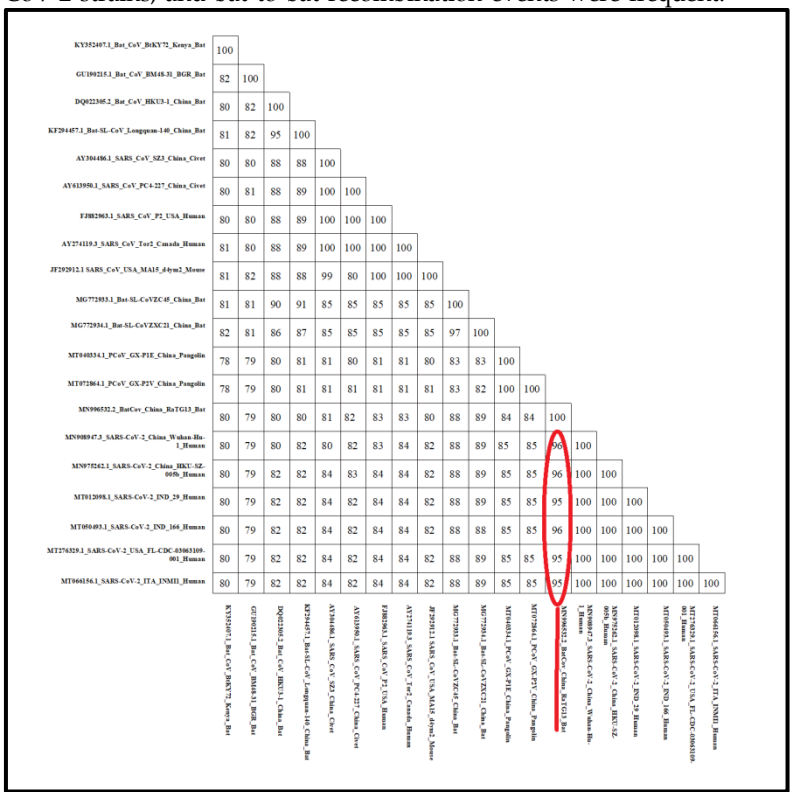

Figure 4: ANI values of the representative *Betacoronavuruses*. ANI values between SARS-CoV-2s and Bat-CoV-RaTG13 have been highlighted.

ISSN 0973-2063 (online) 0973-8894 (print)

Bioinformation 18(10): 951-961 (2022)

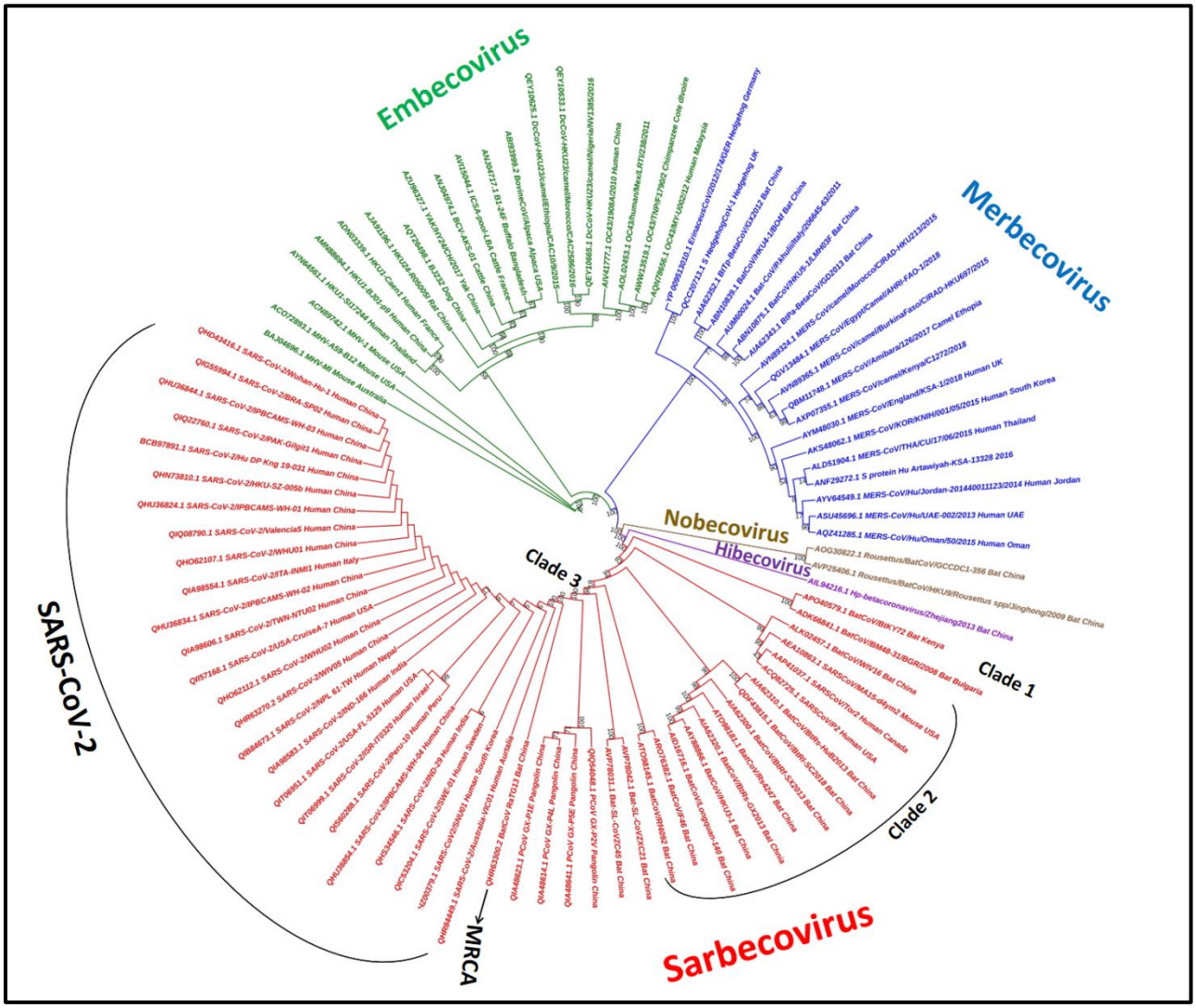

Figure 5: Spike (S) gene phylogeny. Alignment consists of 1,621 aligned amino acid characters (1,071bp are completely aligned characters). Tree was reconstructed using ML method and WAG+I+G4 model of protein evolution along with 1000 bootstrap replicates. Three distinct clades of sarbecovirus and most recent common ancestor (MRCA) of SARS-CoV-2 are depicted.

Table 1: Detected recombination events in the *Betacoronavirus* genomes with position of break and endpoints, and major and minor parents. Details of genome recombination analysis are given in the text.

	Alignment		Recombinant		Sequences			Detection Methods						
S.No	Begin	End	Begin	End	Recombinant	Major Parent	Minor Parent	RDP	GENECONV	Bootscan	Maxchi	Chimaera	SiSscan	3Seq
1	14696	23754	11713	20578	Bat_SL_CoV_ZC45 (MG772933)	Bat-CoV_RaTG13 (MN996532)	Bat-CoV_Longquan_140 (KF294457)	1.69E-296	1.24E-288	8.24E-305	1.02E-59	2.87E-68	3.61E-69	1.65E-274
2	24242	37168	21008	28069	Bat_SL_CoV_Rf4092 (KY417145)	Bat_SL_CoV_WIV16 (KT444582)	Bat_SL_CoV_F46 (KU973692)	4.37E-191	4.01E-180	9.96E-106	5.63E-03	2.11E-63	4.79E-88	2.85E-243
3	2239	3800	1672	2999	Bat-CoV_Longquan_140 (KF294457)	Bat-CoV_HKU3_1 (DQ022305)	Bat_SL_CoV_ZXC21 (MG772934)	1.97E-167	1.03E-184	8.95E-220	8.91E-44	1.17E-42	2.08E-35	2.78E-151
4	5254	24175	3019	20696	Bat_BtRs_Beta-CoV/GX2013 (KJ473815)	Bat-CoV_Longquan_140 (KF294457)	Bat_SL_CoV_WIV16 (KT444582)	2.56E-69	NS	3.44E-53	1.23E-46	1.97E-36	NS	1.95E-181
5	11942	21974	8988	18830	Bat_BtRl_SC2018 (MK211374)	Bat-CoV_Longquan_140 (KF294457)	Civet -CoV_SZ3 (AY304486)	4.91E-66	7.06E-90	2.19E-79	1.11E-12	1.71E-35	4.43E-46	1.11E-03
6	29816	33755	23176	25661	Bat_SL_CoV_Rs4247 (KY417148)	Bat_CoV_HKU3_1 (DQ022305)	Bat_SL_CoV_Rf4092 (KY417145)	NS	6.36E-22	5.76E-29	1.49E-20	6.95E-27	1.43E-12	3.37E-60
7	36762	37168	27497	27847	Bat_BtRs_Beta-CoV/GX2013 (KJ473815)	Bat-CoV_HKU3_1 (DQ022305)	Civet -CoV_SZ3 (AY304486)	4.16E-51	2.48E-54	1.71E-53	1.79E-15	4.65E-08	6.99E-39	NS
8	28696	33708	22540	25620	Bat_SL_CoV_F46 (KU973692)	Human_SARS-CoV_P2 (FJ882963)	Bat_BtRs_Beta-CoV/HuB2013 (KJ4738154)	1.14E-08	NS	4.80E-12	6.52E-19	8.32E-08	NS	2.63E-29
9	33666	35283	25557	26755	Bat_BtRl_BetaCoV/SC2018 (MK211374)	Bat_BtRs_Beta-CoV/HuB2013 (KJ4738154)	Human_SARS-CoV_P2 (FJ882963)	3.96E-23	1.04E-19	7.53E-26	2.62E-11	8.87E-10	1.15E-17	1.90E-11

10	38018	38494	28847	29235	SARS-CoV-2_SNU01 (MT039890)	Bat_SL_CoV_ZXC21 (MG772934)	Mouse-CoV_MA15-d4ym2 (JF292912)	1.03E-22	4.47E-20	6.09E-24	1.45 E-05	2.33E-05	3.06E-06	NS
11	31861	38021	27484	30152	Camel-CoV_HKU23-CAC1019 (MN514962)	Camel-CoV_HKU23-CAC2586 (MN514963)	Dog-CoV_BJ232-(KX432213)	2.36E-17	5.20E-14	1.00E -13	4.95 E-17	8.93E-17	7.99E -19	6.07E-15
12	7778	8147	5139	5469	Bat_BtRs_Beta-CoV/HuB2013 (KJ4738154)	Bat_SL_CoV_Rf4092 (KY417145)	Bat-CoV-HKU3-1 (DQ022305)	9.57E-15	1.18E-08	2.36E -10	6.81 E-03	1.09E-03	4.31E -03	3.92E-09
13	8662	10188	6304	7516	Mouse-MHV-1 (FJ647223)	Mouse-MHV-A59-B12 (FJ884687)	Mouse-MHV-MI (AB551247)	1.43E-12	7.77E-08	5.12E -10	7.33 E-09	1.70E-10	3.69E -09	NS
14	33682	34346	25354	25716	Bat_BtRf-BetaCoV/SX2013 (KJ473813)	Bat-CoV_BtRs_HuB2013 (KJ473814)	Mouse-CoV_MA15-d4ym2 (JF292912)	1.52E-11	1.16E-05	1.02E -09	1.90 E-08	4.94E-08	1.78E -08	2.49E-04
15	35580	38288	27048	29061	Bat_BtRl-BetaCoV-SC2018 (MK211374)	Bat-CoV_BtKY72 (KY352407)	Bat-CoV-RaTG13 (MN996532)	1.96E-26	1.94E-18	3.72E -14	3.25 E-10	7.50E-07	8.73E -30	NS
16	17604	18263	14579	15238	Bat-CoV-RaTG13 (MN996532)	PCoV_GX_P1E (MT040334)	Bat-SL-CoV_Rf4092 (KY417145)	6.32E-11	5.73E-03	1.14E -11	NS	0.0226 56	2.99E -06	3.81E-07
17	30785	31196	23930	24341	Bat-CoV_Longquan-140 (KF294457)	Bat_BtRs-BetaCoV/HuB2013 (KJ473814)	Bat_SL_CoVZXC21 (MG772934)	8.01E-11	4.97E-03	1.95E -09	2.02 E-06	8.07E-06	NS	1.25E-07
18	9439	9968	6687	7114	Civet-CoV_SZ3 (AY304486)	Bat_SL_CoV_Rs4247 (KY417148)	Bat_SL_CoV_F46 (KU973692)	3.45E-09	NS	1.65E -08	2.00 E-02	3.74E-04	1.55E -03	1.79E-06
19	19003	23780	15926	20547	Bat_SL_CoV_Rs4247 (KY417148)	Civet-CoV_SZ3 (AY304486)	Bat-CoV_BtRs_GX2013_Bat	NS	1.37E-02	1.43E -02	9.00 E-06	9.51E-09	NS	1.20E-08
20	30374	30696	23383	23635	Bat-CoV_BtRs_GX2013 (KJ473815)	Civet-CoV_PC4-227 (AY613950)	Bat-CoV_RaTG13 (MN996532)	2.76E-05	NS	3.77E -03	3.06 E-02	2.78E-02	NS	2.40E-05
21	33666	35283	25557	26755	Bat_BtRl_BetaCoV/SC2018 (MK211374)	Bat_BtRs_Beta-CoV/HuB2013 (KJ4738154)	Human_SARS-CoV_P2 (FJ882963)	3.96E-23	1.04E-19	7.53E -26	2.62 E-11	8.87E-10	1.15E -17	1.90E-11

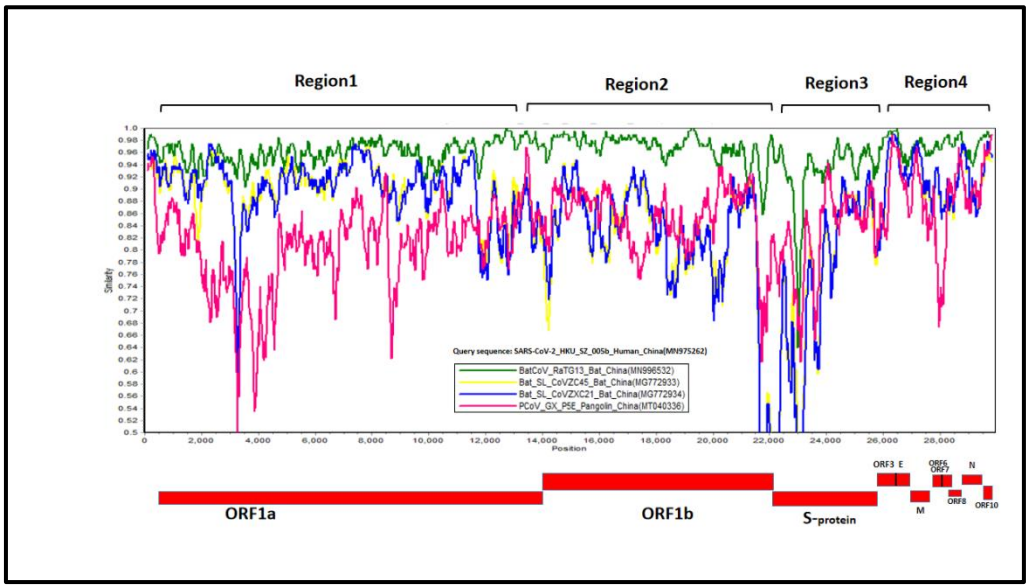

Figure 6: Similarity plot (Simplot) of SARS-CoV-2 HKU-China and its comparison with other Coronaviruses (Green, Bat-CoV-RaGT13; Pink, Pangolin-CoV-GX-P5E; Yellow Bat-SL-CoVZC45: and Blue, Bat-SL-CoVZXC21). Simplot depicts the Bat and Pangolin CoVs that show recombination. Four different regions (Regions 1-4) from the genomes showing recombination were highlighted.

Table 2: Homology and genetic variations in different genomic regions of SARS-CoV-2 Wuhan (MN908947) with respect to Bat-CoV-RaTG13 (MN996532) and Pangolin-CoV-GX-P5E (MT040336)

Strain	Envelop protein		Membrane protein		Spike protein		Nucleocaspid protein	
	Homology	**Genetic variation**	**Homology**	**Genetic variation**	**Homology**	**Genetic variation**	**Homology**	**Genetic variation**
Bat_RaTG13	100%	0%	98%	02%	97%	03%	99%	01%
PCoV_GX-P5E	100%	0%	98%	02%	92%	08%	93%	07%

Genetic variation analysis

Further we measured the genetic variation of bat-CoV-RaTG13 and pangolin-CoV-GX-P5E sequences with respect to SARS-CoV-2 Wuhan-Hu-1 strain, and found that spike protein has highest genetic variation 3% and 7 % respectively (Table 2). We further did MSA of the spike protein sequences and observed that the insertion of the novel amino acids "PRRA" in the spike protein of human SARS-CoV-2s (**Figure** 7). The "PRRA" insertion at the S1/S2 junction site which induces a furin cleavage motif needs to be investigated. Therefore, further detailed study on these residues would be required to shed light on molecular mechanism of interaction between SARS-CoV-2 and the host cells.

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 18(10): 951-961 (2022)


```
PCoV_GX-P5E_(QIA48641.1)            GTNTSNQVAVLYQDVNCTEVPMAIHAEQLTPAWRVYSAGANVFQTRAGCLVGAEHVNNSY  658
PCoV_GX-P2V_(QIQ54048.1)            GTNTSNQVAVLYQDVNCTEVPMAIHAEQLTPAWRVYSAGANVFQTRAGCLVGAEHVNNSY  660
Bat-CoV_RaTG13_(QHR63300.2)         GTNASNQVAVLYQDVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSY  660
SARS-CoV-2_WHU01_(QHO62107.1)       GTNTSNQVAVLYQDVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSY  660
SARS-CoV-2_HKU-SZ_(QHN73810.1)      GTNTSNQVAVLYQDVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSY  660
SARS-CoV-2_IND-166_(QIA98583.1)     GTNTSNQVAVLYQDVNCTEVPVAIHADQLTPTWRVYSTGSNVFQTRAGCLIGAEHVNNSY  660
                                    ***:****************:****:****:*****:*:**********:*********

PCoV_GX-P5E_(QIA48641.1)            ECDIPVGAGICASYHSMSS----FRSVNQRSIIAYTMSLGAENSVAYSNNSIAIPTNFTI  714
PCoV_GX-P2V_(QIQ54048.1)            ECDIPVGAGICASYHSMSS----FRSVNQRSIIAYTMSLGAENSVAYSNNSIAIPTNFTI  716
Bat-CoV_RaTG13_(QHR63300.2)         ECDIPIGAGICASYQTQTNS----RSVASQSIIAYTMSLGAENSVAYSNNSIAIPTNFTI  716
SARS-CoV-2_WHU01_(QHO62107.1)       ECDIPIGAGICASYQTQTNSPRRARSVASQSIIAYTMSLGAENSVAYSNNSIAIPTNFTI  720
SARS-CoV-2_HKU-SZ_(QHN73810.1)      ECDIPIGAGICASYQTQTNSPRRARSVASQSIIAYTMSLGAENSVAYSNNSIAIPTNFTI  720
SARS-CoV-2_IND-166_(QIA98583.1)     ECDIPIGAGICASYQTQTNSPRRARSVASQSIIAYTMSLGAENSVAYSNNSIAIPTNFTI  720
                                    *****:********:: :.     *** .:*****************************

PCoV_GX-P5E_(QIA48641.1)            SVTTEILPVSMTKTSVDCTMYICGDSIECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQE  774
PCoV_GX-P2V_(QIQ54048.1)            SVTTEILPVSMTKTSVDCTMYICGDSIECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQE  776
Bat-CoV_RaTG13_(QHR63300.2)         SVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQE  776
SARS-CoV-2_WHU01_(QHO62107.1)       SVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQE  780
SARS-CoV-2_HKU-SZ_(QHN73810.1)      SVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQE  780
SARS-CoV-2_IND-166_(QIA98583.1)     SVTTEILPVSMTKTSVDCTMYICGDSTECSNLLLQYGSFCTQLNRALTGIAVEQDKNTQE  780
                                    ************************ ********************************
```

Figure 7: Multiple sequence alignment of spike (S) protein consisting of six strains (three SARS-CoV-2s and three closest CoV strains from bat and pangolin).

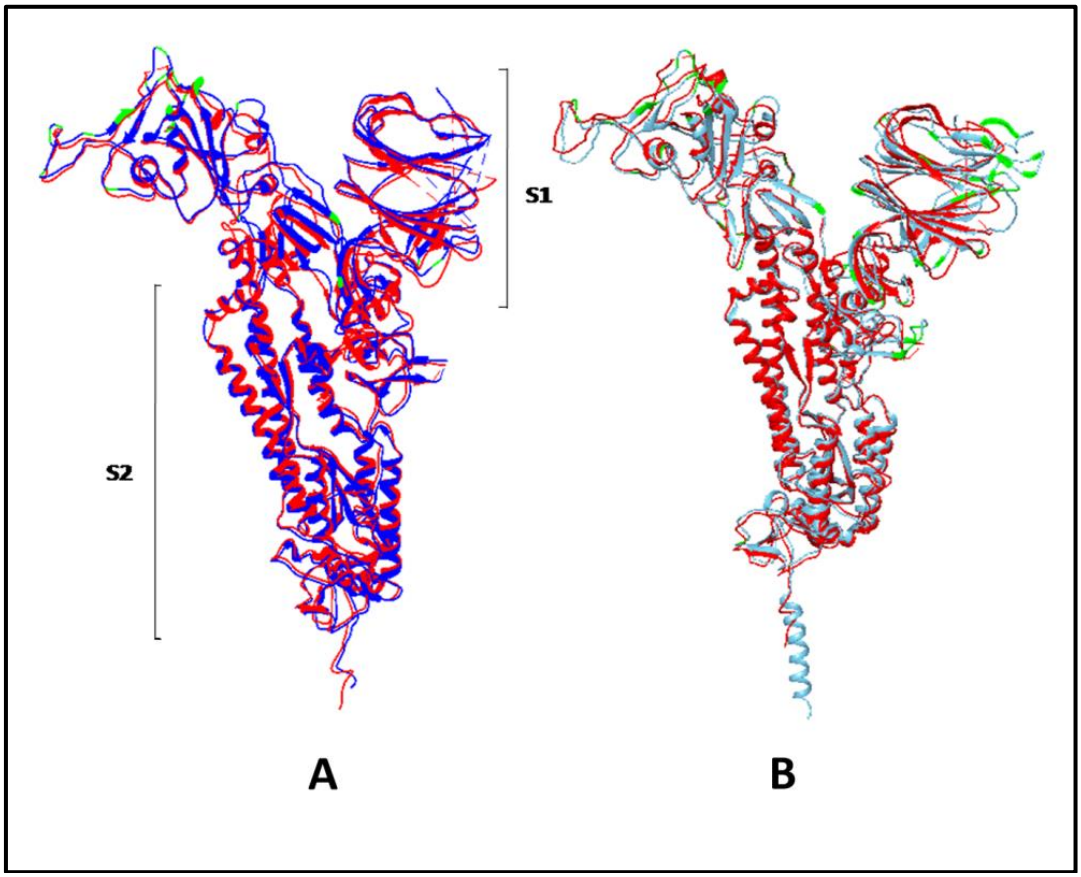

Figure 8: Structural representation of spike glycoprotein (S) and their comparison. Spike protein monomer superimposed structure of (**A**) SARS-CoV-2 Wuhan-Hu-1 (red) and bat-CoV-RaTG13 (blue), (**B**) SARS-CoV-2 Wuhan-Hu-1 & Pangolin-CoV-GX-P5E (sky blue). The green coloured highlighted regions represent the mutated amino acid residues.

Structural analysis of spike protein:
On the basis of MSA result, we compared the structure of spike protein of SARS-CoV-2 (PDB: 6XLU) with bat-CoV-RaTG13 (PDB: 6ZGF) and pangolin-CoV-GX-P5E (modeled protein) (**Figure** 8). The spike protein is a complex trimeric protein and monomer was used for structure comparison. It has two main units S1 and S2. The S1 subunit recognizes and binds to the host receptor enzyme via receptor-binding domains (RBDs) while the S2 subunit helps in fusion of viral cell membrane to host cell **[45-47]**. We found that structurally the spike protein of pangolin-CoV-GX-P5E is more diverse compared to SARS-CoV-2 (RMSD value 2.766 Å) while the bat-CoV-RaTG13 spike protein shows similarity to SARS-CoV-2 with RMSD 2.059 Å (**Figure** 8). It was observed that the bat-CoV-RaTG13 shows high number of mutations at one of the RBD (spotted 27 mutations: shown in green colour in Figure 8) while the pangolin-CoV-GX-P5E shows mutations at both the RBDs of S1 subunit (a total of 85 mutations). The changes in spike proteins have impact on the interaction of pathogen and host **[48, 49]**. Thus these mutations were probably responsible for the adaptation of SARS-CoV-2 into human systems. A number of studies reported that the mutations in spike protein of SARS-CoV-2 facilitate its adaptation into humans **[50-52]**. The insertion of the four amino acids "PRRA" found in the MSA represents an extended loop between the two parallel β-sheets (S1/S2 cleavage site). This cleavage point between the receptor binding domain (S1) and fusion peptide (S2) mediate cell-cell fusion and entry into human cell **[25,47]**. Thus structural analysis supports MSA results and suggests that SARS-Cov-2 is adapted to infect human systems.

Conclusions:
Outbreak of SARS-CoV-2 is the third documented spillover of an animal coronavirus to humans in only two decades that has resulted in a major pandemic. In quest of the emergence of SARS-CoV-2, this study finds that human SARS-CoV-2 emerged from Bat-CoV-RaTG13 through ancestral intra-species recombination events between bat corona viruses belonging to Sarbecovirus subgenus. Furthermore, acquisition/insertion of a furin cleavage motif("PRRA") to the S1/S2 site in the spike protein of SARS-CoV-2 along with high number of mutations at one of its RBD are probably responsible for the adaptation of SARS-CoV-2 into humans systems. Therefore, further detailed study on these residues would be required to shed light on molecular mechanism of interaction between SARS-CoV-2 and the host cells.

Conflicts of interest:
The authors declare that they have no conflict of interest.

Acknowledgments:
AKS and PK gratefully acknowledge the University Grants Commission (India) for financial assistance to carry out the research work. AS thanks Weintian Li for useful comments in improving the manuscript.

References:

[1] Zhu N *et al*. *N Engl J Med*. 2020 **382**:727–733 [PMID: 31978945]
[2] Gorbalenya AE *et al*. *Nature Microbiology* 2020 **5**:536-544 [PMID: 32123347]
[3] King AMQ *et al*. *Virus Taxonomy* 2012 785–795 [https://doi.org/10.1016/B978-0-12-384684-6.00066-5]
[4] Woo PCY *et al*. *Exp Biol Med (Maywood)* 2009 **234**:1117–1127 [PMID: 19546349]
[5] Woo PCY *et al*. *Viruses* 2010 **2**:1804–1820 [PMID: 21994708]
[6] Fan Y *et al*.*Viruse*s 2019 **11**:210 [PMID: 30832341]
[7] Eickmann M. *Science* 2003 **302**:1504b–11505 [PMID: 14645828]
[8] Vijaykrishna D *et al*. *JVI* 2007 **81**:4012–4020 [PMID: 17267506]
[9] Zumla A *et al*. *The Lancet*. 2015 **386**:995–1007 [PMID: 26049252]
[10] Velavan TP & Meyer CG, *Trop Med Int Health* 2020 **25**:278–280. [https://doi.org/10.1111/tmi.13383]
[11] Lai C-C *et al*. *Int J Antimicrob Agents* 2020 **55**:105924 [PMID: 32081636]
[12] Srivastava S *et al*. *J Biosci* 2021 **46**: 22 [PMID: 33737495]
[13] Katoh K, *Nucleic Acids Research* 2002 **30**:3059–3066 [PMID: 12136088]
[14] Kalyaanamoorthy S *et al*. *Nat Methods*2017 **14**:587–589 [PMID: 28481363]
[15] Nguyen L-T *et al*. *Molecular Biology and Evolution* 2015 **32**:268–274 [PMID: 25371430]
[16] Letunic I & Bork P, *Nucleic Acids Research* 2019 **47**:256–259 [PMID: 30931475]
[17] Martin DP *et al*. *Virus Evolution* 2015 **1**:vev003 [PMID: 27774277]
[18] Rose PW *et al*. *Nucleic Acids Res* 2017 **45**:D271–D281 [PMID: 27794042]
[19] Pettersen EF *et al*. *J Comput Chem* 2004 **25**:1605–1612 [PMID: 15264254]
[20] Luk HKH *et al*. *Infection Genetics and Evolution* 2019**71**:21–30. [PMID: 30844511]
[21] Wu F *et al*. *Nature* 2020 **579**:265–269. [PMID: 32015508]
[22] Wertheim JO *et al*. *J Virol* 2013 **87**:7039-7045 [PMID: 23596293]
[23] Songa H-D *et al*. *Proc Natl Acad Sc*i 2005 **102**:2430–2435 [PMID: 15695582]
[24] Fung TS & Liu DX *Annu Rev Microbiol* 2019 **73**:529-557 [PMID: 31226023]
[25] Andersen KG *et al*. *Nat Med* 2020 **26**:450–452 [PMID: 32284615]
[26] Montoya V *et al*. *J Evol Biol*. 2021 **34**:924-936 [PMID: 33751699]
[27] Roy A *et al*. *Front Microbiol* 2021 **12**: 548275[PMID: 33889134]
[28] York A *Nat Rev Microbiol* 2020 **18**:191–191 [PMID: 32051570]
[29] Zhou P *et al*. *Nature* 2020 **579**:270–273 [PMID: 32015507]
[30] Nakagawa S &Miyazawa T *Inflamm Regener* 2020 **40**:17 [PMID: 32834891]
[31] Tosta E *MemInstOswaldo Cruz* 2021 **116**: e210127 [PMID: 35019068]

[32] Cui J *et al.Nat Rev Microbiol* 2019 **17**:181–192 [PMID: 30531947]
[33] Ribet D & Cossart P *Microbes Infect.* 2015 **17**:173-183 [PMID: 25637951]
[34] Sheppard SK *et al. Nature Reviews Genetics* 2018 **19**:549–565 [PMID: 29973680]
[35] Puigbò P *et al. Biol. Direct*. 2008 **3**:38 [PMID: 18796141]
[36] Lu R *et al.The Lancet* 2020**395**:565–574 [PMID: 32007145]
[37] Yoon SH *et al. Antonie van Leeuwenhoek* 2017 **110**:1281–1286[PMID: 28204908]
[38] Degnan JH &Rosenberg NA *Trends Ecol Evol.* 2009 **24**:332–340 [PMID: 19307040]
[39] Som A *J Phylogen Evolution Biol.* 2013 **1**:116 [doi:10.4172/2329- 9002.1000116]
[40] Som A *Briefings in Bioinformatics* 2015 **16**:536–548 [PMID: 24872401]
[41] Jeffroy O *et al.Trends in Genetics* 2006 **22**:225–231 [PMID: 16490279]
[42] Boni MF *et al. Nat Microbiol* 2020 **5**:1408-1417 [PMID: 32724171]
[43] Flores-Alanis A *et al.BMC Res Notes* 2020 **13**:398 [PMID: 32854762]
[44] Li X *et al. SciAdv* 2020 **6**:eabb9153 [PMID: 32937441]
[45] Jaimes JA *et al. J Mol Biol.* 2020 **432**:3309–3325 [PMID: 32320687]
[46] Rehman SU *et al. Pathogens* 2020 **9**:240 [PMID: 32210130]
[47] Maitra A *et al. J Biosci* 2020 **45**:76. [PMID: 32515358]
[48] Li F *Annu Rev Virol.* 2016 **3**:237-261. [PMID: 27578435]
[49] Huang Y *et al. Acta Pharmacol Sin.*2020 **41**: 1141-1149 [PMID: 32747721]
[50] Choe H & Farzan M *Science* 2021 **372**: 466-467[PMID: 33926942]
[51] Isabel S *et al. Sci Rep* 2020 **10:**14031 [PMID: 32820179]
[52] Zhang Q *et al. Signal Transduct Target Ther.* 2021 **6**:233 [PMID: 34117216]