Vantage: Journal of Thematic Analysis **ISSN: 2582-7391**
A Multidisciplinary Publication of Centre for Research,
Maitreyi College, University of Delhi
October 2020, Volume 1, Issue 2

Original Research Article

Comparative Genomic and Phylogenetic Analysis of Spike and Nucleocapsid Proteins of Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) and 9 Other Taxonomically Related Coronaviruses using in-Silico Tools

Jaspreet Kaur*, Srishti and Shubhangi Sharma

Department of Zoology, Maitreyi College, University of Delhi, Chanakyapuri, New-Delhi-110021

*Correspondence: jkaur@maitreyi.du.ac.in

ABSTRACT

The world is witnessing a global pandemic due to COVID-19 disease, which is caused by Severe Acute Respiratory Syndrome Coronavirus-2. It is an enveloped ss-RNA virus, in which spike and nucleoprotein genes play an important role in the pathogenesis of Covid-19. Spike protein is required for attachment of virus to the host cell receptor, while nucleoprotein is important for replication of viral genome. Keeping this perspective in mind, we investigated nucleoprotein (N) and spike (S) genes in SARS-CoV-2 and 9 other taxonomically related coronaviruses using in-silico tools. The results obtained from our comparative genomics and phylogenetic analysis provided important evidences about how these organisms are evolutionarily related to each other. We found that N and S genes of these organisms were more adapted to the host (*Homo Sapiens*) and also found evidences for negative pervasive selection at different sites in the compared protein sequences of these genes. Thus, this study will help in understanding the epidemiology of SARS-CoV-2 in fine details.

Keywords: Severe Acute Respiratory Syndrome Coronavirus (SARS-CoV), spike (S), nucleoprotein or nucleocapsid (N), comparative genomics, phylogenetic analysis

1. INTRODUCTION

Coronavirus disease (COVID-19) is a fatal disease which is caused by Severe Acute Respiratory Syndrome Coronavirus-2 (SARS-CoV-2). It is an enveloped virus with positive-sense, non-segmented, single-stranded RNA genome and is composed of structural and non-structural components. The structural proteins include spike (S), envelope (E), membrane (M), and nucleocapsid (N) proteins. SARS-CoV-2, like many other human coronaviruses (HCoVs), including SARS-CoV and MERS-CoV, is a zoonotic pathogen that originated in wild animals. (Forni et al., 2017). Since the inception of the outbreak of COVID-19, there has been an exponential increase in the number of sequences of SARS CoV-2 isolates from across the globe. At present (on 5th October, 2020), there were 17, 223 complete and 8, 176 partial nucleotide sequences of SARS-CoV-2, making a total of 25, 399 sequences in the NCBI database. Out of these, 570 nucleotide sequences are from Indian geographic region (https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/).

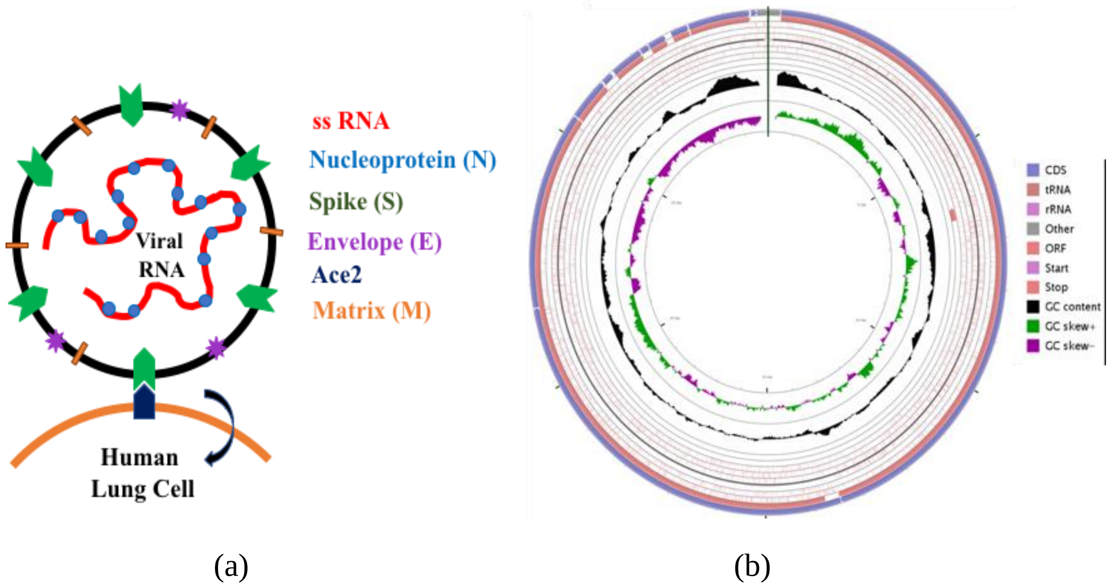


Figure 1: (a) Structure of SARS-CoV-2, showing spike (S) and nucleoprotein or nucleocapsid (N), Matrix (M), Envelope (E) proteins and angiotensin-converting enzyme 2 (ACE2) (b) Circular map of SARS-CoV-2 genome (constructed by CGView tool).

During the past few months, several studies related to sequence analysis of SARS-CoV-2 have been published. These studies have provided valuable insight into the probable origin of pandemic crisis (Zhou et al., 2020). In addition, several reports have highlighted the comparative genomic and phylogenetic analysis of different strains of SARS-CoV isolate sequenced so far (Kumar et al., 2020).

In the present study, we have used in-silico tools in order to understand the genomic features of SARS-CoV-2 and its relationship with other taxonomically related coronaviruses at the genetic level. There are two isolates of SARS-CoV with accession numbers: NC_004718.3 and MN908947.3 (Wuhan isolate or SARS-CoV-2) which have been included in this study. Thus, *we used these two isolates and 9 other related coronaviruses, making a total of 11 organisms which have been used in this study* (Table 1). In addition to exploring the genomic relatedness including detection of rearrangement and recombination events in genomes of different coronaviruses, we have focused our analysis on two important genes, namely, Spike (S) and Nucleocapsid or Nucleoprotein (N) in order to understand the pathogenesis and evolutionary constraints of these organisms. The spike protein is a glycoprotein, which forms a crown like appearance on the outer surface of the coronavirus and is responsible for the entry of virus into the host cells (Li, 2016). This spike protein binds to a molecule on the surface of human lung cells called the angiotensin-converting enzyme 2 (ACE2) (Figure 1a). The nucleoprotein, on the other hand is a phosphoprotein, which binds to the RNA molecule. These two genes undergo post-translational modifications (PTMs) *viz.* nucleoprotein is phosphorylated and spike protein is glycosylated. These modifications are used as powerful strategies by these viruses for increased affinity and stability of protein-protein interactions in order to evade the immune response of the host and hence survive successfully in it. But, at the same time, such PTMs can be used by researchers as targets for developing therapeutics like drugs and attenuated vaccines against these viruses (Fung & Liu, 2018). Thus, we have investigated the physico-chemical properties of these two genes in detail. We have also focussed on coiled coil regions of spike protein because coiled-coil domains are known for their characteristic heptad repeat and stability, thus making them excellent choices for vaccine development (Villard et al., 2007, McFarlane et al., 2009, Apostolovic et al., 2010). In addition, both the N and S genes in these coronaviruses are under negative selection, but there are reports of limited signals of positive selection in three viral ORFs (N protein, ORF8, and nsp1) of SARS-CoV-2 (Cagliani et al., 2020).

2. MATERIAL AND METHODS

Data Collection (Material): Complete genome sequences as well as nucleotide and protein sequences of Nucleoprotein and Spike genes of 11 CoVs were retrieved from NCBI (https://www.ncbi.nlm.nih.gov/). The selected CoVs taken in this study are given in Table 1 below:

Table 1: List of the selected CoVs employed in this study

S.No.	Name of CoVs (Type of Isolate)	Accession no.
1.	SARS CoV	NC_004718.3
2.	SARS CoV Wuhan isolate (Wu)/SARS-CoV-2	MN908947.3
3.	Bat CoV RaTG13	MN996532
4.	Pangolin CoV	MT084071
5.	Camel CoV	MK967708
6.	MERS-CoV	NC_019843.3
7.	Dromedarius CoV	MH259486
8.	H-Enteric CoV	FJ415324
9.	Canine CoV	KX432213
10.	Bovine CoV	NC_003045
11.	Avian CoV	NC_001451

The first epidemic had origins in Guangdong Province, China, during late 2002 and lasted until 2004. The outbreak was caused by SARS-CoV. The second epidemic was first characterized in a man with pneumonia in Saudi Arabia in 2012 and the causative organism was found to be MERS-CoV. The third pandemic was first reported in Hubei Province, China, in late 2019 and is still ongoing and is caused by SARS-CoV-2 or 2019-nCoV (Wong et al., 2020).

2.1. Comparative Genomics Analysis

Nucleotide diversity was assessed using an online calculator (https://www.science buddies.org/science-fair-projects/references/genomics-g-c-content-calculator).

Sequence similarity was calculated and dot plots were constructed using BLASTn and BLASTp tools (Altschul et al., 1990). Detection of gene order was done using Mauve software (Darling et al., 2004). Detection of potential recombination events and estimation of breakpoint locations was done by GARD (A Genetic Algorithm for Recombination Detection, Pond et al., 2001), implemented in Classic Datamonkey server (http://classic.datamonkey.org/). Detection of CpG islands was done using

MethPrimer 2.0 (Li et al., 2002) and CpG Island detection tool (https://www.bioinformatics.org/sms2/cpg_islands.html). Codon Usage Analysis was done by CodonW (https://galaxy.pasteur.fr/). The genetic code of the host (*Homo Sapiens*) was used in the analysis.

2.2. Physico-chemical analysis of S and N protein

Protein Sequence Analysis was done using Expasy ProtParam tool (Artimo et al., 2012). Glycosylation site prediction was done by NetNglyc tool (http://www.cbs.dtu.dk/services/NetNGlyc/) (Gupta & Brunak, 2002) at default value of 0.5. Phosphorylation site predictions: was done by DISPHOS 1.3 software (http://phospho.elm.eu.org/links.html) (Iakoucheva et al., 2004). Prediction of hydrophobic residues was done using web based Helixator (http://www.tcdb.org/progs/helical_wheel.php) and WHAT 2.0 tools (http://www.tcdb.org/progs/?tool=hydro) (Saier et al., 2006).

2.3. Phylogenetics and Detection of positive/negative selection

Multiple Sequence Alignment & Phylogenetic analysis was done using Clustal Omega (https://www.ebi.ac.uk/Tools/msa/clustalo/) (Sievers et al., 2011) and MegaX (Kumar et al., 2018). The terminal nucleotides not common to all sequences were trimmed. Detection of sites under positive or negative selection was done using HyPhy tool (integrated in MegaX) and by Selecton server (http://selecton.tau.ac.il/). All the computational tools were used at default parameters unless specified.

3. RESULTS AND DISCUSSION

The complete genome sequences and sequences of N and S genes from the selected CoVs were subjected to different computational analysis, the results of which are given in following sections.

3.1. Comparative Genome Analysis

We compared the genomes of selected coronavirus to find out the evolutionary relationship between them. In order to achieve this target, we first analysed the nucleotide composition in genome sequences of these CoVs and found that the frequency of uracil is significantly higher in all the genomes (Figure 2).

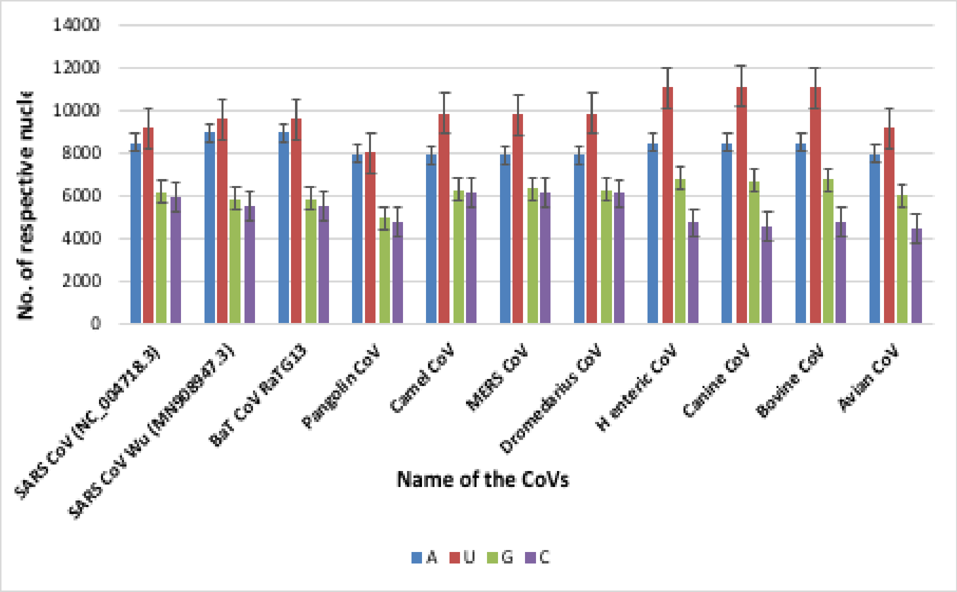


Figure 2: Nucleotide diversity in selected coronaviruses. Error bars represent standard deviations.

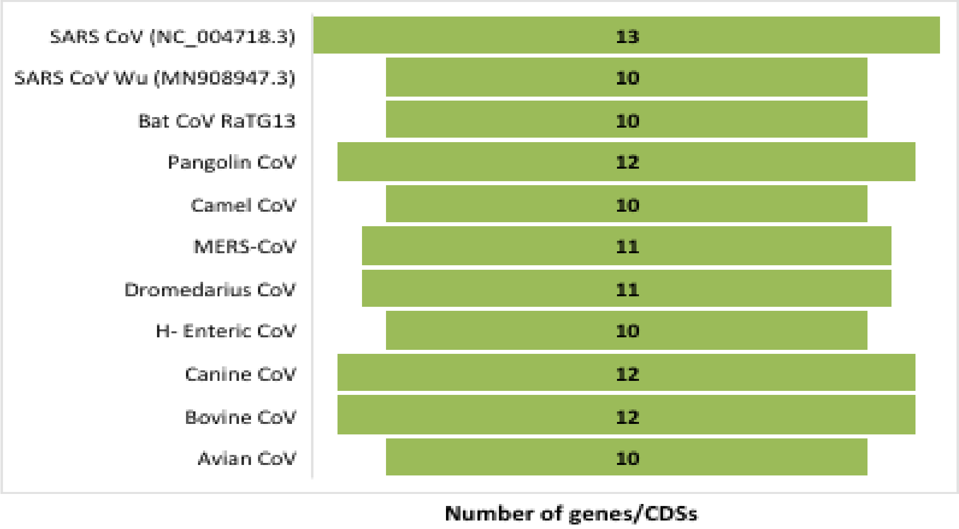


Figure 3: Comparison of number of coding sequences in different CoVs.

Also, in case of purines, adenine (A) is preferred over guanine (G). These results are in agreement with previous studies (Gupta et al., 2020). Overall, the AU content was higher (with percentage mean of 28.09 and 33.09 respectively) than the GC content (with percentage mean of 20.83 and 17.97 respectively). Our results are in agreement with earlier studies wherein GC% was found to be smaller than the AU or AT% in SARS-CoV-2 genome (Gupta et al., 2020). In addition, the genome size of the selected 11 coronaviruses ranged from 27 to 31 kb, with 11 as the average number of genes or coding sequences (CDSs) present in them (Figure 3) and SARS-CoV (NC_004718.3) has maximum no. of CDS, i.e. 13. It was interesting to note that with just a handful of genes, these viruses are able to control a complex system like the human cell!

Comparison of Nucleoprotein and Spike Genes: The average length of nucleoprotein was found to be ~1.28 kbp and average length of spike gene was ~ 3.88 kbp. (Figure 4).

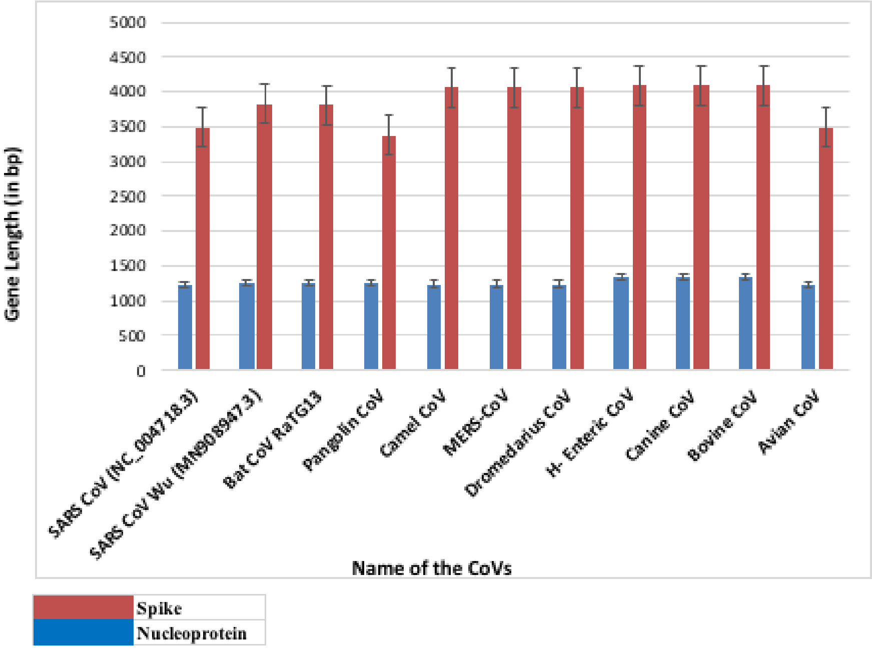
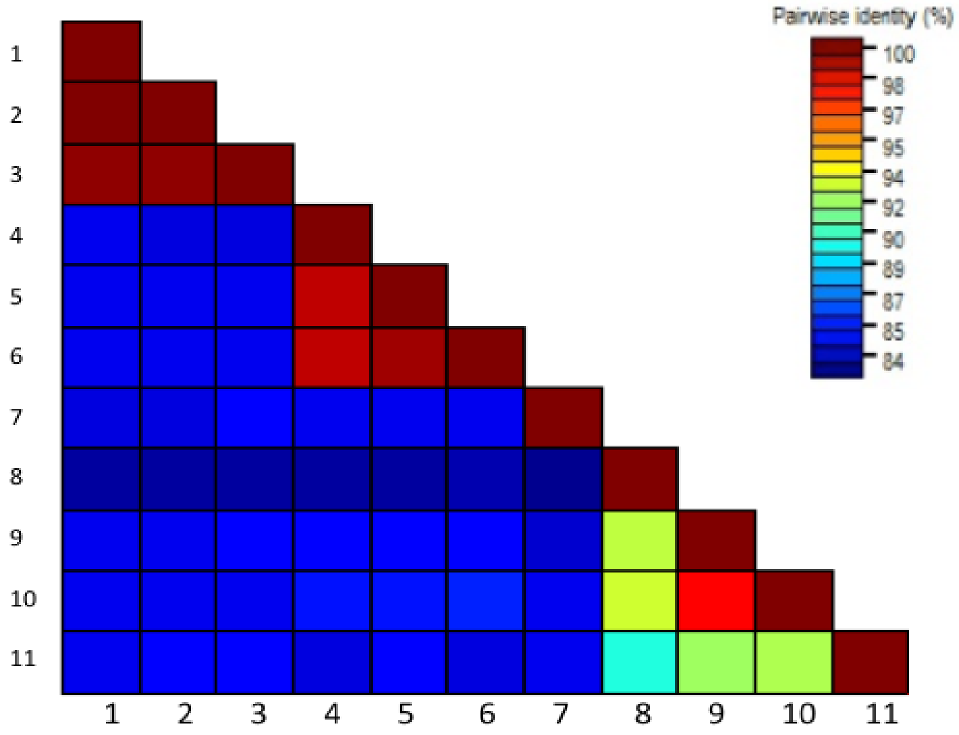


Figure 4: Comparison of gene length of Nucleoprotein and Spike genes. Error bars represent standard deviations.

BLASTn results: BLASTn results: SARS-CoV-2 (MN908947.3) showed maximum pairwise genomic sequence similarity with Bat CoV RaTG13 followed by Pangolin CoV and SARS-CoV (NC_004718.3), indicating the close genetic relatedness between these 4 organisms (Figure 5). The homology between these organisms was further confirmed by their respective dot plots (Figure 6), wherein continuous diagonal lines can be seen clearly, except in the dot plot of SARS CoV & Pangolin CoV, where insertions or deletions were observed as breaks or discontinuities in the diagonal line.

Our findings confirm the previous reports about the probable bat origin of SARS-CoV-2 (Li et al., 2020). Also, the distant relatives of SARS-CoV-2 (MN908947.3) included Camel, MERS, Dromedarius, H-Enteric, Canine, Bovine and Avian CoVs, with few local regions of sequence similarity. Thus, our findings are in agreement to earlier reports wherein a high divergence between canine CoVs and SARS-CoV-2 has been demonstrated (Sharun et al., 2020).



1. NC_019843.3: MERS-CoV; 2. MH259486.1: Dromedarius CoV; 3. MK967708: Camel CoV; 4. KX432213.1: Canine CoV; 5. NC_003045.1: Bovine CoV; 6. FJ415324.1: H-enteric CoV; 7. NC_001451.1: Avian CoV; 8. MT084071: Pangolin CoV; 9. MN996532.1: Bat CoV RaTG13; 10. MN908947.3: SARS Wu; 11. NC_004718.3: SARS CoV

Figure 5: Pairwise identity matrix between complete genome sequences of selected CoVs, constructed by SDTv1.2 tool.

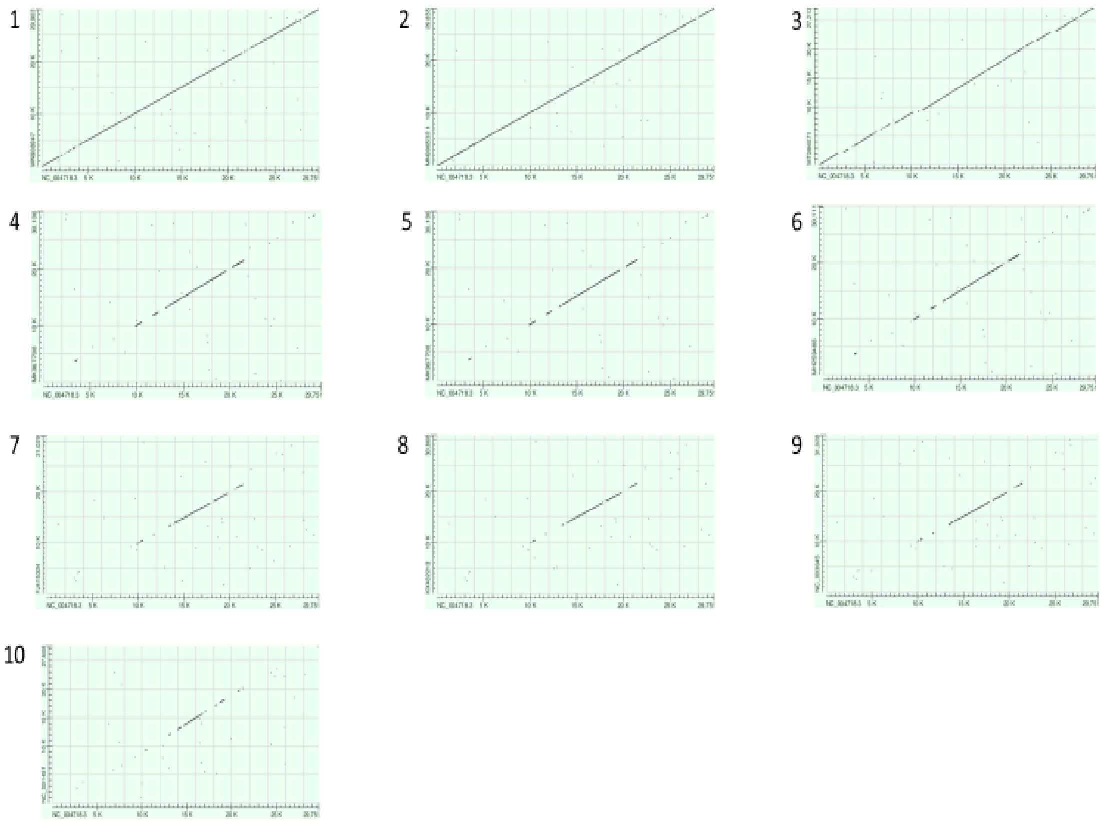


Figure 6: Dot plots. Key to the figure: 1. SARS CoV & SARS CoV Wu (MN908947.3), 2. SARS CoV & BaT CoV RaTG13, 3. SARS CoV & Pangolin CoV, 4. SARS CoV & Camel CoV, 5. SARS CoV & MERS CoV, 6. SARS CoV & Dromedarius CoV, 7. SARS CoV & H enteric CoV, 8. SARS CoV & Canine CoV, 9. SARS CoV & Bovine CoV, 10. SARS CoV & Avian CoV.

Genome Alignment: The detection of homologous regions between CoV genomes was done by multiple genome alignment using Mauve software. Such regions were depicted by locally collinear blocks (LCBs) with similar coloured blocks. The sequence elements conserved among all the genomes under study were also shown with connecting lines (of same colour as the LCBs). We also found some genomic rearrangements with respect to the reference genome, wherein orthologous region (dark green coloured LCB) in first 4 genomes was reordered and was actually present at the beginning of the genome sequences of Canine (KX432213.1), Dromedarius (MH259846.1) and Bovine (NC_00345.1) CoVs. Also, a light green coloured LCB, which started from ~30 kb was present only in 3 genomes: Pangolin (MT08407), H-Enteric (FJ1415324.1) and Avian (NC_001451) CoVs. Overall, the close genetic relatedness of genomes of SARS, Bat-RaTG13 and Pangolin CoVs (first 4 genomes in Figure 7) was clearly evident from these results also.

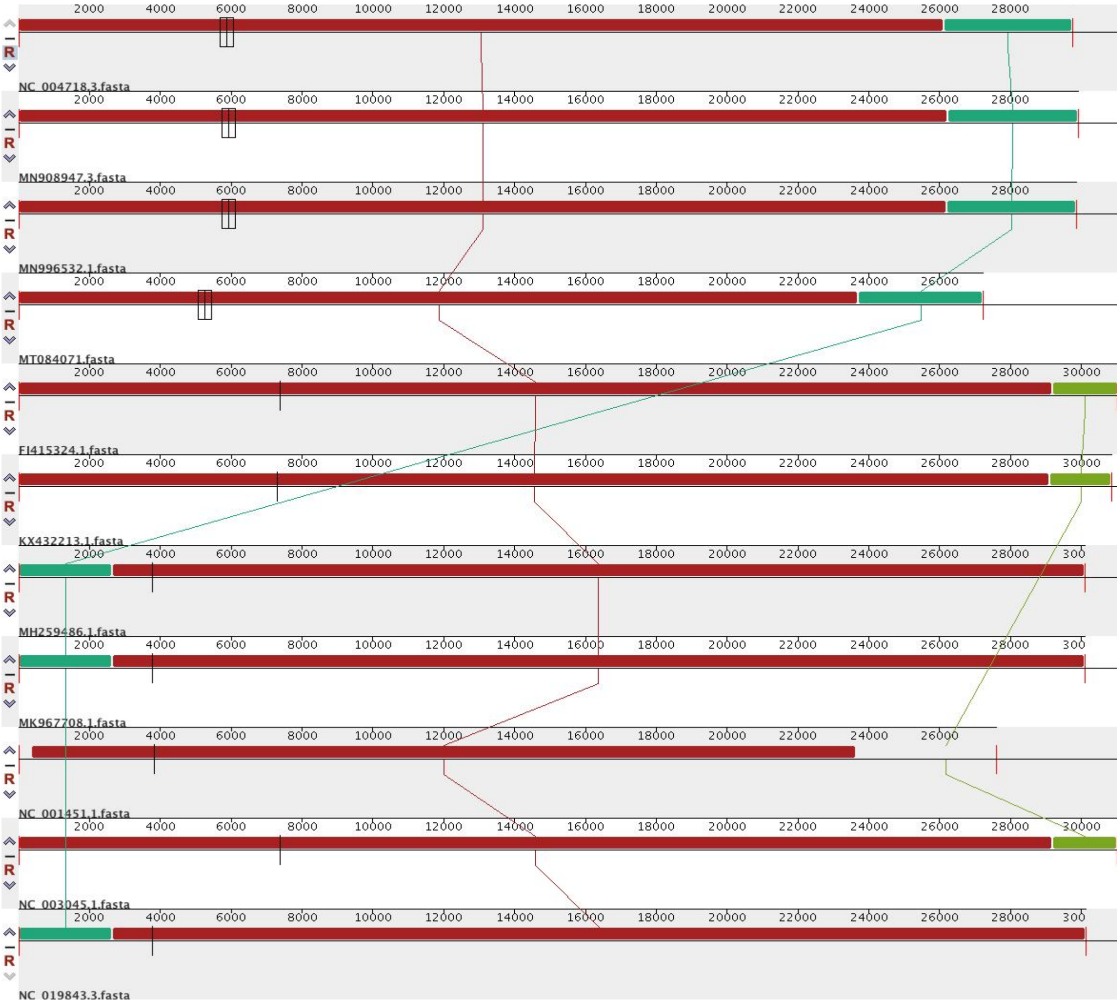


Figure 7: Mauve results for genome alignment of 11 genome sequences of Coronaviruses. The SARS-CoV-2 (NC_004718) was taken as reference genome.

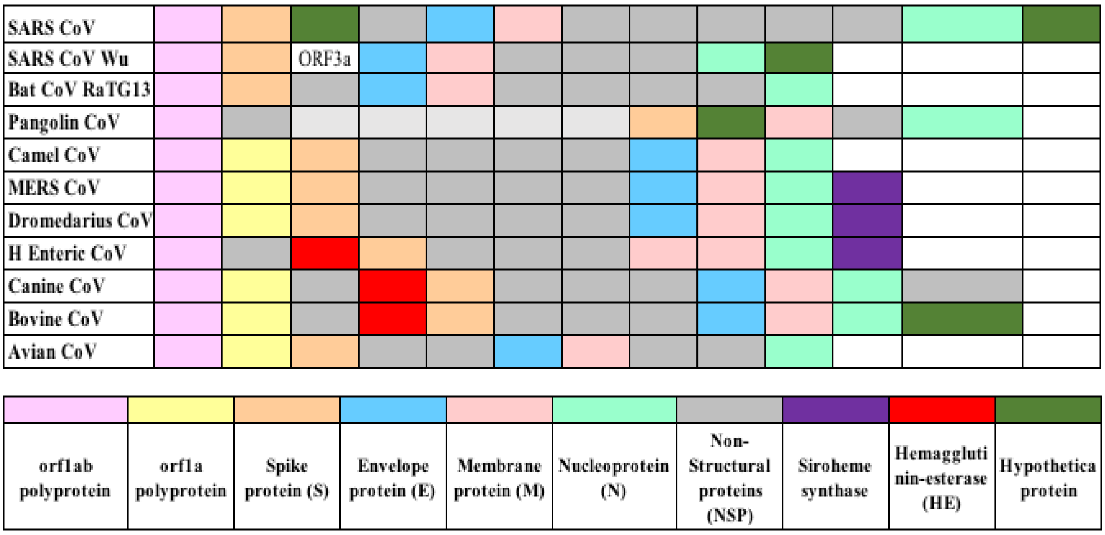
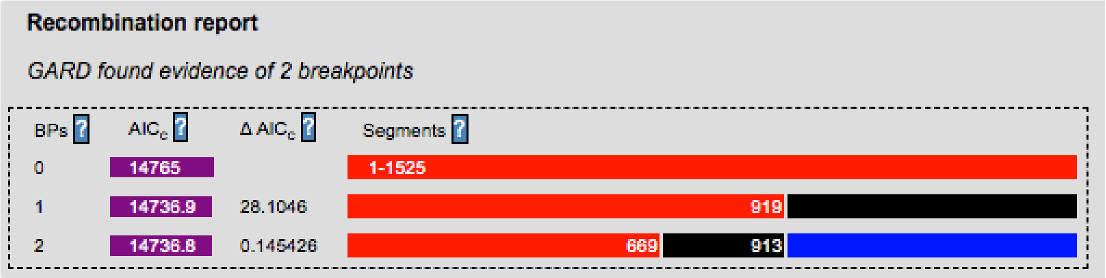


Figure8: Comparison of gene order (ORF3a is required for pore formation by virus in membrane of host cell, predicted by InterPro). orf1ab encodes replicase polyprotein 1 ab.

Comparison of Gene Order: The gene order in different coronavirus genomes was compared and it was found that all the structural genes were conserved in all the coronaviruses (Figure 8) Also, the non-structural proteins (NSPs) were also present. The gene rearrangements as well as presence of some different genes like siroheme synthase can also be seen clearly.

Recombination Analysis: The sequence alignments of N and S genes in 11 CoVs were searched for evidence of potential recombination events and estimate breakpoint locations. The GARD program detected two and three potential recombination breakpoints within nucleoprotein and spike sequences respectively using default values. (Figure 9 & 10).

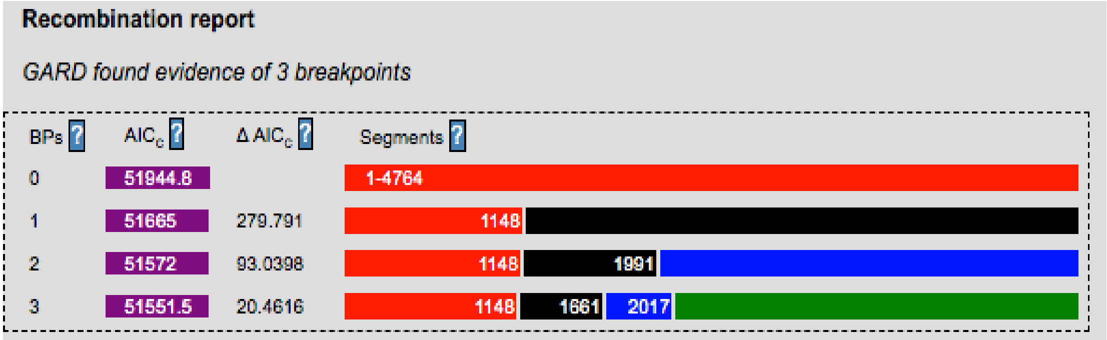


Recombination report

GARD found evidence of 2 breakpoints

BPs ?	AIC_c ?	ΔAIC_c ?	Segments ?
0	14765		1-1525
1	14736.9	28.1046	919
2	14736.8	0.145426	669 913

Figure 9: Recombination report of Nucleoprotein sequences (BP: Breakpoint, Akaike Information Criterion, AICc)



Recombination report

GARD found evidence of 3 breakpoints

BPs ?	AIC_c ?	ΔAIC_c ?	Segments ?
0	51944.8		1-4764
1	51665	279.791	1148
2	51572	93.0398	1148 1991
3	51551.5	20.4616	1148 1661 2017

Figure 10: Recombination report of Spike sequences.

Although GARD analysis detected recombination breakpoint signals at positions 669 and 913 in nucleoprotein sequences and at positions 1148, 1661, 1991 and 2017 in spike sequences, but upon further analysis of these BPs based on the respective p-values, we found that all of these BPs were statistically non-significant breakpoint signals. Thus, it can be concluded that no statistically significant recombination events were detected in nucleoprotein and spike sequences. The insignificant breakpoints may occur frequently due to variation in branch lengths between sequence segments; this could be due to some forms of recombination or to other processes, such as spatial rate variation, heterotachy, etc (http://classic.datamonkey.org/).

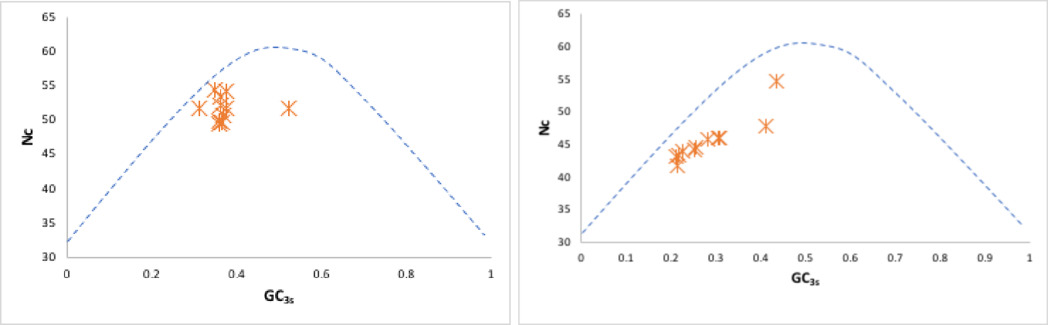
3.2. Codon Usage Analysis

Codon usage of viral genes evolves according to their specific need for different proteins and the viral proteins that are required in large amounts are usually encoded by genes that are optimized to the host codon usage (Tello et al., 2013). In case of viruses, it has been found that attenuated viral vaccines can be effectively developed by replacement of optimized codons with other synonymous codons (Coleman et al., 2008). So, in order to assess the codon usage pattern of coronaviruses and to find out if they also have some optimized codons (which are used more frequently than other synonymous codons), we investigated their 'codon usage bias' by calculating Codon Adaptation Index (CAI) values of the nucleotide sequences of their genomic, nucleoprotein and spike gene sequences (Table 2).

Table 2: Comparison of the CAI

Name of CoV	**CAI values**		
	Genome	**N gene**	**S gene**
SARS-CoV	0.645	0.704	0.664
SARS-CoV Wu	0.676	0.688	0.646
BaT CoV RaTG13	0.675	0.686	0.649
Pangolin CoV	0.646	0.677	0.614
Camel CoV	0.628	0.709	0.597
MERS CoV	0.625	0.706	0.660
Dromedarius CoV	0.635	0.711	0.660
H enteric CoV	0.623	0.707	0.625
Canine CoV	0.614	0.704	0.625
Bovine CoV	0.628	0.705	0.628
Avian CoV	0.624	0.709	0.627

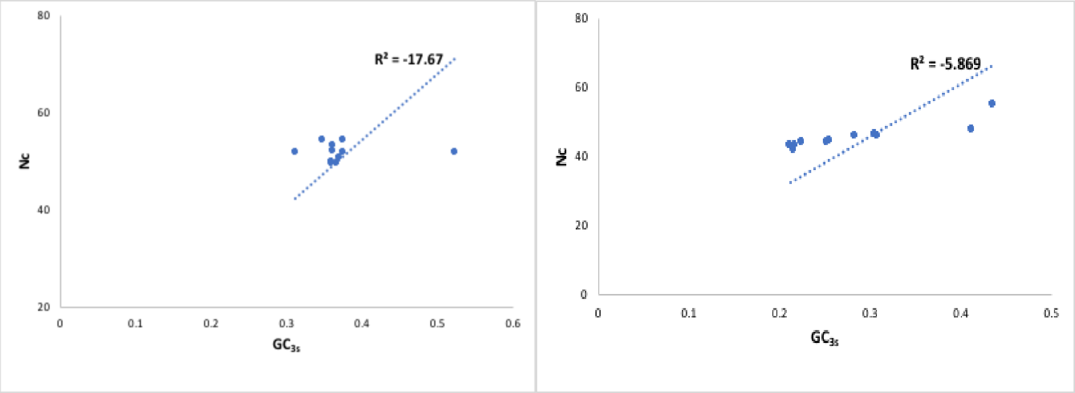
These higher CAI values (>0.6) signified that the N & S genes have a codon usage pattern resembling that in the reference genes (*Homo Sapiens*) with higher expression levels. Further, we constructed ENC–GC3s or ENC plot of N & S genes and found that most of the Nc values lie near but below the standard curve (Nc values of 49.39-54.26 at the GC3s values 0.31-0.52; Nc values of 41.82-54.75 at the GC3s values 0.21-0.43 respectively; Figure10) indicating an involvement of mutation pressure. Thus, construction of ENc plot helps to assess the effective mutational pressure (Sheikh et al., 2020)



(a) (b)

Figure 10: Nc plot of (a) nucleoprotein and (b) Spike proteins. The continuous curve (in dotted blue line) represents the expected curve between GC3s and Nc under random codon usage.

We also found negative correlation between GC3s and Nc (R2 = -17.67) for nucleoprotein genes as well as for spike genes (R2 = -5.86; Figure11) suggesting strong influence of compositional constraints on codon usage bias in these genes of different coronaviruses.



Nucleoprotein Spike

Figure 11: Plot of Nc versus GC3s

3.3. Comparative Sequence analysis of Nucleoprotein & Spike genes

From Table 3 & 4, it is evident that SARS-CoV shared maximum identity with SARS-CoV Wu, Bat and Pangolin CoVs. These results were in consistent to our results obtained from comparative genomics as discussed in earlier sections. Dot plots were also constructed for N & S genes (data not shown here) which further confirmed the close genetic relatedness between these four CoVs.

Table 3: BLASTn of Nucleoprotein sequence of SARS-CoV (NC_004718.3) with other coronaviruses

Name of CoVs	Identities	Alignment score	E-value	Query coverage	Gaps
SARS-CoV Wu	1119/1269	1612	0.0	100%	9/1269(0%)
Bat CoV RaTG13	1115/1269	1594	0.0	100%	9/1269(0%)
Pangolin CoV	1099/1271	1522	0.0	100%	13/1271(1%)
Camel CoV	63/83	60.8	8e-13	24%	0/83(0%)
MERS-CoV	63/83	60.8	8e-13	24%	0/83(0%)
Dromedarius CoV	62/83	56.3	3e-11	24%	0/83(0%)
H-Enteric CoV	11/11	21.1	0.73	0%	0/11(0%)
Canine CoV	11/11	21.1	0.73	0%	0/11(0%)
Bovine CoV	11/11	21.1	0.73	0%	0/11(0%)
Avian CoV	15/17	22.9	0.19	2%	0/17(0%)

Table 4: BLASTn results of Spike protein of SARS-CoV (NC_004718.3) with other coronaviruses

Name of CoVs	Identities	Alignment score	E – value	Query coverage	Gaps
SARS CoV Wu	2782/3735	2378	0.0	96%	96/3735(2%)
Bat CoV RaTG13	2759/3718	2345	0.0	96%	74/3718(1%)
Pangolin CoV	1192/1520	1261	0.0	72%	4/1520(0%)
Camel CoV	72/99	58.1	9e-11	8%	0/99(0%)
MERS-CoV	72/99	58.1	9e-11	9%	0/99(0%)
Dromedarius CoV	72/99	58.1	9e-11	9%	0/99(0%)
H- Enteric CoV	34/39	49.1	5e-08	4%	0/39(0%)
Canine CoV	33/39	44.6	6e-07	3%	0/39(0%)
Bovine CoV	34/39	49.1	5e-08	4%	0/39(0%)
Avian CoV	25/27	41.0	6e-06	8%	0/27(0%)

3.4. Phylogenetic Analysis

In order to analyse the evolutionary relationship among different coronaviruses, phylogenetic trees (Figure12, 13) were reconstructed for N & S genes using neighbour-joining (NJ) method and the resultant tree topologies were evaluated using bootstrap values. We also used Maximum Likelihood method (MLK) and got similar tree topologies (data not shown here). The results indicated that the nucleoprotein genes clustered into 3 groups or monophyletic clades: (1): SARS-CoV (NC_004718) with SARS-CoV Wu, followed by Pangolin and Bat RaTG13 CoVs, (2): Camel, MERS, & Dromedarius CoVs and (3): Bovine, canine and H-enteric CoVs. Also, Avian CoV shared least similarity with all other ingroup taxa. This branching was confirmed by high bootstrap values and further supported our sequence similarity results.

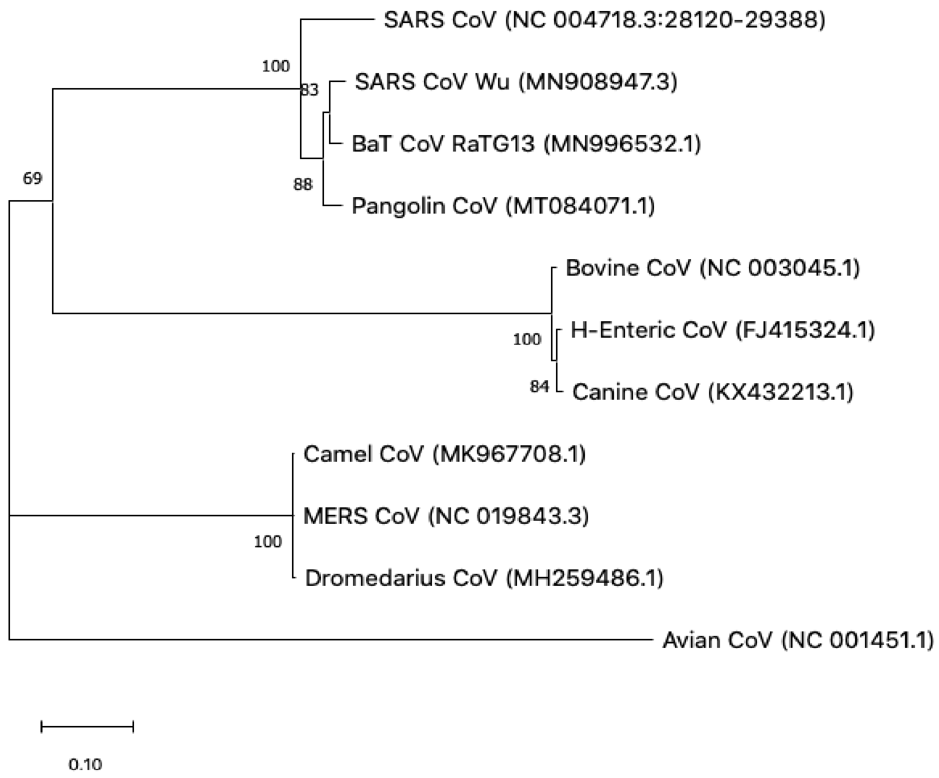


Figure 12: Phylogenetic tree based on nucleotide sequence of N gene in selected CoVs. Bar, 0.10 substitutions per nucleotide position.

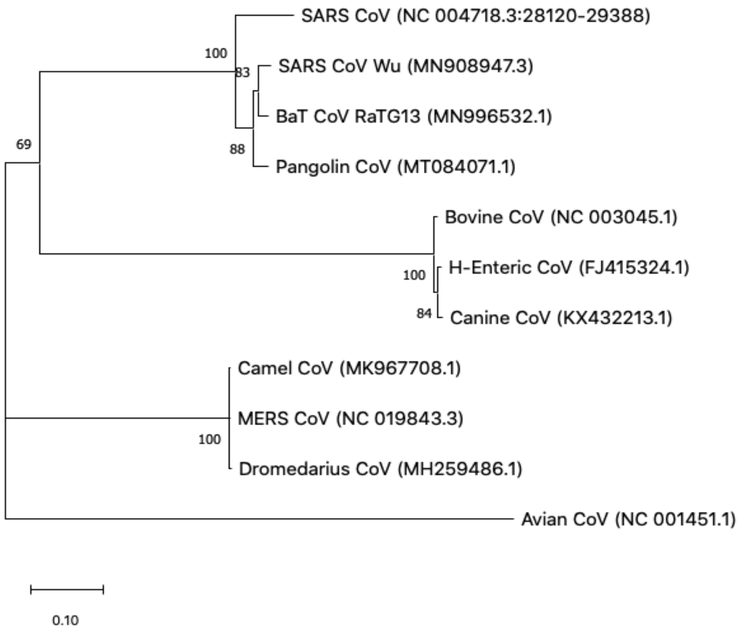


Figure 13: Phylogenetic tree based on nucleotide sequence of Spike gene in selected CoVs. Bar, 1.0 substitutions per nucleotide position.

3.5. Detection of CpG Islands

The distribution of CpG islands in the selected coronaviruses was studied wherein the criteria used for prediction were: island size >100, GC%> 50.0, Obs/Exp > 0.6 (Observed CpG is the number of CpGs present in the sequence, and expected CpG is defined as (number of C * number of G)/length of sequence. Using these criteria, the no. of CpG islands was found to be almost similar (>200) and least in Bat CoV RaTG13 (114). Also, the nucleoprotein genes consisted of only 1 to 2 CpG islands with size ranging from 108-276 bp only. On the other hand, no CpG islands were detected in the sequence of spike gene of selected coronaviruses using the same parameters.

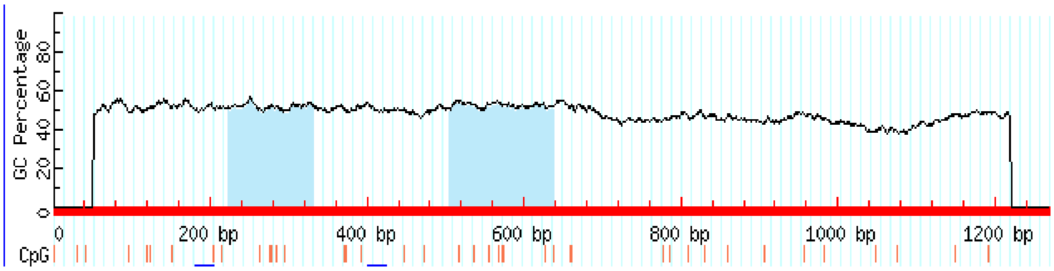


Figure 14: Pictorial representation of two CpG islands (shown as blue bars) detected in gene N of SARS-CoV (NC_004718.3)

Table 5: CpG islands seen in Nucleoprotein genes of selected coronaviruses

Name of CoV	CpG island	Size(bp)	Position
SARS-CoV	Island 1	108	223 - 330
	Island 2	135	504 - 638
SARS CoV Wu	Island 1	205	50-254
	Island 2	110	887-996
BaT CoV RaTG13	Island 1	205	50-254
	Island 2	113	872-984
Pangolin CoV	Island 1	210	50 - 259
Camel CoV	None	--	--
MERS CoV	None	--	--
Dromedarius CoV	Island 1	106	325 - 430
H-Enteric CoV	Island 1	235	339 - 573
Canine CoV	Island 1	235	339 - 573
Bovine CoV	Island 1	276	298 - 573
Avian CoV	Island 1	106	499 - 604

The lower number of CpG islands may be due to the lower GC content than AU content in these CoVs (see section 3.1). Such low CpG islands in genomes of these CoVs may also help in evading human zinc finger antiviral protein (ZAP) mediated immune response, which specifically binds to CpG dinucleotides in viral RNA genomes by its RNA-binding domain (Xia, 2020).

3.6. Identification of Glycosylation sites in Spike Genes

Spike protein is a membrane glycoprotein; hence we studied the variation in glycosylation sites present in spike genes of different coronaviruses since glycosylation is generally correlated with virulence of viruses (Chattopadhyay et al., 2010).

Table 6: Results of glycosylation sites predicted along the sequence of Spike of selected coronaviruses

Name of CoVs	**Positions**	**N-glycosylation sequon***	**Potential Glycosylated sites**
SARS-CoV (NC_004718.3)	29, 65, 73, 109, 118, 119, 158, 227, 269, 318, 330, 357, 589, 602. 691, 699, 783, 1056, 1080, 1116, 1140, 1155, 1176	NYTQ, NVTG, NHTF, NKSQ, NNST, NSTN, NCTF, NITN, NGTI, NITN, NATK, NSTF, NASS, NCTD, NNTI, NFSI, NFSQ, NFTT, NGTS, NNTV, NHTS, NASV, NESL	23
SARS-CoV Wu (MN908947.3)	17, 61, 74, 122, 149, 165, 234, 282, 331, 603, 616, 657, 709, 717, 801, 1074, 1098, 1134, 1158, 1173, 1194	NLTT, NVTW, NGTK, NATN, NKSW, NCTF, NITR, NGTI, NITN, NATR NTSN, NCTE, NNSY, NNSI, NFTI, NFSQ, NFTT, NGTH, NNTV, NHTS, NESL	21
BaT CoV RaTG13	17, 30, 61, 122,149, 165, 234, 282, 331, 343, 370, 603, 616, 657, 705, 713, 797, 1070, 1094, 1130, 1154, 1169, 1190	NLTT, NSST, NVTM, NATM, NKSW, NCTF, NITR, NGTI, NITN, NATT, NSTS, NASN, NCTE, NNSY, NNSI, NFTI, NFSQ, NFTT, NGTH, NNTV, NHTS, NASV, NESL	23
Pangolin CoV	18,31,62, 112, 122, 148, 164, 233, 278, 327, 339, 366, 622, 630, 714, 987, 1011, 1047, 1071	NLTG, NSSQ, NVSM, NTSQ, NATN, NKTW, NCTF, NITK, NGTI, NITN, NATT, NSTS, NNSI, NFTI, NFSQ, NFTT, NGTH, NNTV, NHTS	19
Camel CoV	66, 104, 125, 155, 166, 222, 236, 244, 410, 475, 487, 592, 619, 719, 774, 785, 870, 1160, 1176, 1213, 1225, 1241, 1256, 1277, 1288	NITI, NYSQ, NSTG, NFSY, NHTL, NASL, NCTF, NITE, NLTK, NPTC, NLTT NDTK, NCTA, NSSL, NSSY, NFSF, NLTL, NPTN, NNTR, NIST, NSTG, NVST, NTTL, NESY, NYTY	25
MERS CoV	66, 104, 125, 155, 166, 222, 236, 244, 410, 475, 487, 592, 619, 719, 774, 785, 870, 1160, 1176, 1213, 1225, 1241, 1256, 1277, 1288	NITI, NYSQ, NSTG, NFSD, NHTL, NASL, NCTF, NITE, NLTK, NPTC, NLTT, NDTK, NCTA, NSSL, NSSY, NFSF, NLTL, NPTN, NNTR, NIST, NSTG, NVST, NTTL, NESY, NYTY	25
Dromedarius CoV	66, 104, 125, 155, 166, 222, 236, 244, 410, 475, 487, 592, 619, 719, 774, 785, 870, 1160, 1176, 1213, 1225, 1241, 1256, 1277, 1288	NITI, NYSQ, NSTG, NFSD, NHTL, NASL, NCTF, NITE, NLTK, NPTC, NLTT, NDTK, NCTA, NSSL, NSSY, NFSF, NLTL, NPTN, NNTR, NIST, NSTG, NVST, NTTL, NESY, NYTY	25

Name of CoVs	Positions	N-glycosylation sequon*	Potential Glycosylated sites
H-Enteric CoV	59, 133, 198, 359, 437, 444, 649, 676, 696, 714, 739, 788, 895, 937, 1194, 1224, 1234, 1253, 1267, 1288	NTTL, NTSY, NFTY, NMSS, NVSV, NPST, NATY, NRTF, NSSE, NNTL, NSTS, NDSL, NFSP, NCTG, NNTW, NYTK, NIST, NQTS, NVTF, NHSY	20
Canine CoV	59, 133, 198, 359, 437, 444, 492, 649, 676, 696, 714, 739, 788, 895, 937, 1194, 1224, 1234, 1253, 1267, 1288	NTTL, NTSY, NFTY, NMSS, NVSI, NPSI, NGSL, NATY, NRTF, NSSE, NNTL, NSTS, NDSL, NFSP, NCTG, NNTW, NYTK, NIST, NQTL, NVTF, NHSY	21
Bovine CoV	59, 133, 198, 359, 437, 444, 649, 676, 696, 714, 739, 788, 895, 937, 1194, 1224, 1234, 1253, 1267, 1288	NTTL, NTSY, NFTY, NMSS, NVSV, NPST, NATY, NRTF, NSSE, NNTL, NSTS, NDSL, NFSP, NCTG, NNTW, NYTK, NIST, NQTS, NVTF, NQSY	20
Avian CoV	51, 77, 103, 144, 163, 178, 212, 237, 247, 264, 276, 283, 306, 425, 447, 513, 530, 579, 591, 669, 676, 683, 714, 947, 960, 979, 1014, 1038, 1051, 1074	NISS, NASS, NFSD, NLTV, NLTS, NETI, NGTA, NFSD, NSSL, NTTC, NETG, NPSG, NFSF, NITL, NVTD, NETG, NGTR, NVTE, NLTV, NVST, NISL, NPSS, NCTA, NVTA, NASQ, NGSY, NKTV, NDTK, NYTK, NDSL	30

The sites shown in red depict that all the nine neural networks supported the prediction, * Asn-Xaa-Ser/Thr stretch (where Xaa is any amino acid except Proline).

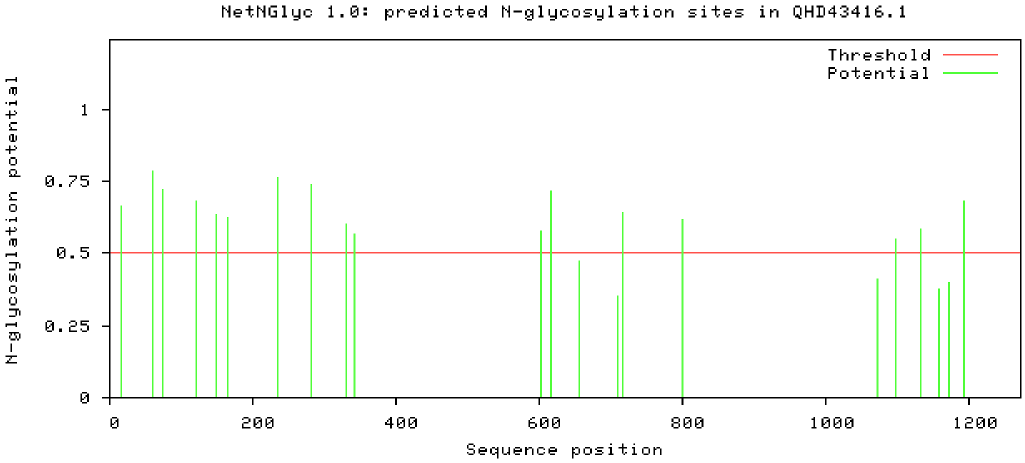


Figure 15: Graphical representation of spike protein of SARS CoV-2.

This analysis predicted a range of 19-30 potential N-linked glycosylation sites in the spike protein of different CoVs (Table 8). These viruses use such heavily glycosylated membrane spike proteins as a means to counteract the host's defence mechanisms (Bagdonaite & Wandall, 2018). Thus, we made an attempt to understand the glycan profile of the spike proteins in different CoVs which may provide further opportunities in order to rationally develop novel therapeutics and vaccines against these viruses.

3.7. Identification of Phosphorylation sites in Nucleoprotein Genes

Nucleoprotein is a phosphoprotein which regulates many important stages in the life cycle of coronaviruses. Thus, we predicted phosphorylation sites in nucleoprotein genes of selected coronaviruses (Table 7) keeping in mind that phosphorylation modifications can be used for phospho regulation of these CoVs as well as help in rational design of live attenuated viruses for use as vaccines (Keck et al., 2015, Noppakunmongkolchai, et al., 2016, Chen et al., 2018).

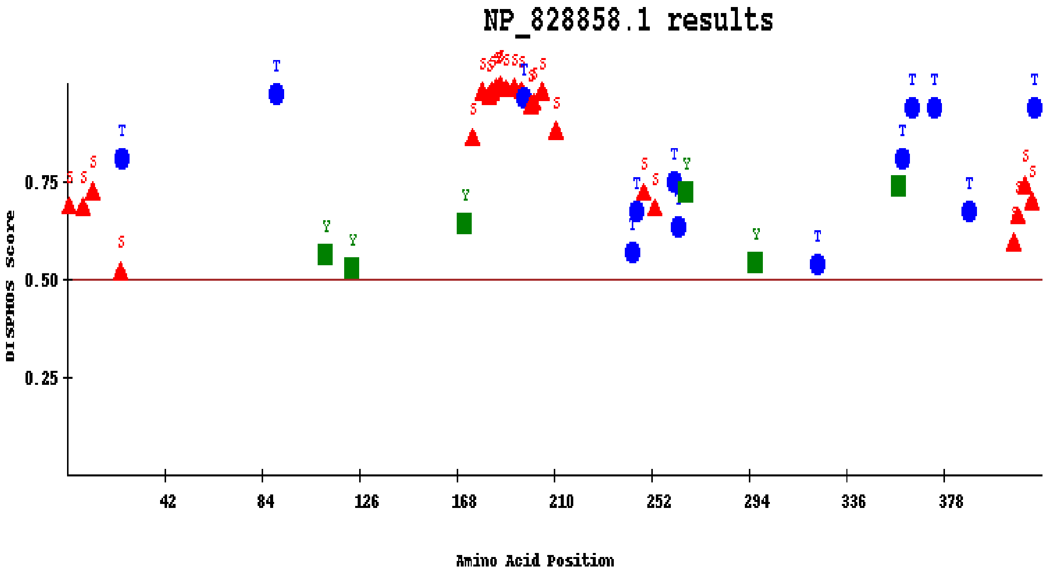


Figure 16: Graphical representation of the potential phosphorylation sites in Nucleoprotein of SARS-CoV (NC_004718.3) at threshold of 0.5.

Table 7: Comparison of phosphorylation sites predicted along the sequence of Nucleoprotein of selected coronaviruses

		Potential Phosphorylated Sites		
		Serine (S)	**Threonine (T)**	**Tyrosine (Y)**
Name of CoV	**Total no. of phosphorylated sites**	**Position**		
SARS-CoV (NC_004718.3)	43	2, 8, 12, 24, 177, 181, 184,185, 187, 188, 189, 191, 195, 198, 202, 203, 207, 213, 251, 256, 411, 413, 416, 419	25, 92, 199, 246, 248, 264, 266, 326, 363, 367, 377, 392, 420	113, 124, 173, 269, 299, 361
SARS CoV Wu (MN908947.3)	44	2, 21, 23, 26, 33, 37, 176, 180, 183, 184, 186, 187, 188, 190, 193, 194, 197, 201, 202, 206, 235, 250, 255, 410, 412, 413, 416	16, 24, 91, 198, 205, 245, 247, 263, 265, 325, 362, 366, 417	172, 268, 298, 360
Bat CoV RaTG13	45	2, 21, 23, 26, 33, 176, 180, 183, 184, 186, 187, 188, 190, 193, 194, 197, 201, 202, 206, 215, 235, 243, 250, 255, 410, 412, 413, 416	16, 24, 91, 198, 205, 245, 247, 263, 265, 325, 362, 366, 417	172, 268, 298, 360
Pangolin CoV	2	--	398, 632	--
Camel CoV	37	3, 24, 169, 171, 172, 173, 176, 177, 179, 182, 183, 185, 186, 187, 190, 192, 195, 200, 204, 256, 375, 380, 391, 401	70, 137, 196, 199, 239, 255, 257, 360, 376, 396, 398, 412	358
MERS-CoV	37	3, 24, 169, 171, 172, 173, 176, 177, 179, 182, 183, 185, 186, 187, 190, 192, 195, 200, 204, 256, 375, 380, 391, 401	70, 137, 196, 199, 239, 255, 257, 360, 376, 396, 398, 412	358
Dromedarius CoV	37	3, 24, 169, 171, 172, 173, 176, 177, 179, 182, 183, 185, 186, 187, 190, 192, 195,	70, 137, 196, 199, 239, 255, 257, 360, 376, 396, 398, 412	358

		Potential Phosphorylated Sites		
		Serine (S)	Threonine (T)	Tyrosine (Y)
Name of CoV	Total no. of phosphorylated sites	Position		
		200, 204, 256, 375, 380, 391, 401		
H-Enteric CoV	47	2, 9, 10, 11, 14, 15, 19, 167, 168, 191, 194, 198, 200, 202, 205, 206, 209, 210, 213, 215, 219, 226, 275, 390, 416, 423, 432, 446	4, 38, 48, 95, 174, 180, 201, 223, 225, 229, 249, 305, 427, 442, 445	186, 187, 380, 441
Canine CoV	47	2, 9, 10, 11, 14, 15, 19, 167, 168, 191, 194, 198, 200, 202, 205, 206, 209, 210, 213, 215, 219, 226, 275, 390, 416, 423, 432, 446	4, 38, 48, 95, 174, 180, 201, 223, 225, 229, 249, 305, 427, 442, 445	186, 187, 380, 441
Bovine CoV	44	2, 9, 10, 11, 14, 15, 19, 167, 168, 191, 194, 198, 200, 202, 205, 206, 209, 210, 213, 215, 219, 226, 275, 390, 416, 432, 446	4, 38, 48, 95, 174, 180, 201, 223, 225, 229, 305, 427, 442, 445	186, 187, 380
Avian CoV	37	3, 29, 54, 125, 127, 145, 165, 168, 172, 173, 177, 181, 185, 190, 192, 212, 340, 342, 343, 344, 352, 379	10, 46, 123, 131, 169, 215, 231, 246, 329, 348, 378	70, 92, 140, 391

3.8. Analysis of physicochemical properties

In our study, more variation in protein length and molecular weights was observed in both nucleoprotein and spike protein sequences in different coronaviruses. Other physicochemical properties which were analysed included isoelectric point and grand average of hydropathicity (GRAVY), results of which are given below.

Table 8: Comparison of the physicochemical properties of Nucleoproteins of the selected coronaviruses

Name of CoVs	Theoretical pI	GRAVY score	Molecular weight (kDa)	Number of amino acids
SARS-CoV (NC_004718.3)	10.11	-1.027	46025.03	422
SARS CoV Wu (MN908947.3)	10.07	-0.971	45625.70	419
Bat CoV RaTG13	10.07	-0.988	45752.84	419
Pangolin CoV**	undefined AA	-1.010	Undefined	419
Camel CoV	10.05	-0.865	45048.28	413
MERS-CoV	10.05	-0.866	45062.31	413
Dromedarius CoV	10.10	-0.864	45070.33	413
H-Enteric CoV	9.62	-0.896	49386.82	448
Canine CoV	9.66	-0.889	49309.74	448
Bovine CoV	9.66	-0.878	49294.75	448
Avian CoV	9.61	-1.034	45032.28	409

**undefined as prediction could not be done due to presence of string of 'Ns' in protein sequence, which are read as 'X' amino acids.

Isoelectric points (pI) of nucleoproteins ranged from 9.6 to 10.1 (Table 8), while that of spike proteins ranged from 5.3 to 7.7. An isoelectric point above 7 indicates a positively charged protein. These observations were also in agreement of the view that RNA molecules are negatively charged and the basic nature of these nucleoproteins help in the electrostatic interactions, which further promote their stability with RNA molecules. GRAVY values of nucleoprotein & spike sequences exhibited a narrow range (-0.864 to -1.034 and -0.221 to -0.011) respectively (Table 9), with less negative values indicating less hydrophilic nature of spike proteins. This may be due to the presence of hydrophilic and extracellular N terminus and a hydrophobic transmembrane segment (TMS), which have been analysed in the next section. On the other hand, GRAVY values for spike proteins of Canine, Bovine and Avian CoVs were observed as less positive, indicating the presence of more hydrophobic residues. Also, the more negative GRAVY values of nucleoproteins indicated the presence of more hydrophilic residues.

Table 9: Comparison of the physicochemical properties of spike proteins of the selected coronaviruses

Name of CoVs	No. of amino acids	Molecular weight	Theoretical pI	GRAVY score
SARS CoV	1255	139109.14	5.56	-0.043
SARS CoV Wu	1273	141178.47	6.24	-0.079
Bat CoV RaTG13	1269	140627.98	6.11	-0.066
Pangolin CoV	1125	123960.98	7.62	-0.221
Camel CoV	1353	149579.30	5.77	-0.077
MERS-CoV	1353	149368.04	5.70	-0.074
Dromedarius CoV	1353	149396.09	5.75	-0.075
H- Enteric CoV	1363	150564.89	5.43	-0.011
Canine CoV	1363	150967.76	5.50	0.017
Bovine CoV	1363	150614.95	5.31	0.005
Avian CoV	1162	128046.70	7.71	0.012

3.9. Hydropathy and Amphipathicity Plots

We analyzed the hydropathy plots of primary sequences of Spike proteins in all the 11 selected coronaviruses and found that these proteins consisted of one transmembrane segment (TMSs), except Avian CoV, which had 3 TMSs (shown as orange colored bars in Figure 17). On the other hand, no TMS were detected in nucleoprotein sequences and were found to contain more hydrophilic areas relative to the hydrophobic areas. Since nucleoprotein is an RNA binding protein and not a membrane protein, so no alpha helices or transmembrane segments were detected in it.

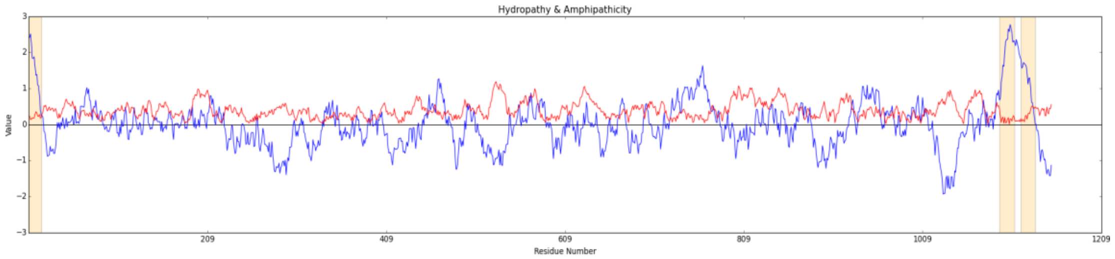


Figure 17: Hydropathy and Amphipathicity plot predictions for Spike protein of Avian-CoV. Here, blue color denotes hydropathy and red color denotes amphipathicity for protein sequences. TMSs are indicated by orange bars.

3.10. Helical Wheel & Helical net Diagrams

We constructed helical wheel diagrams of all the spike as well as nucleoprotein sequences present in different coronaviruses, results of which are given below.

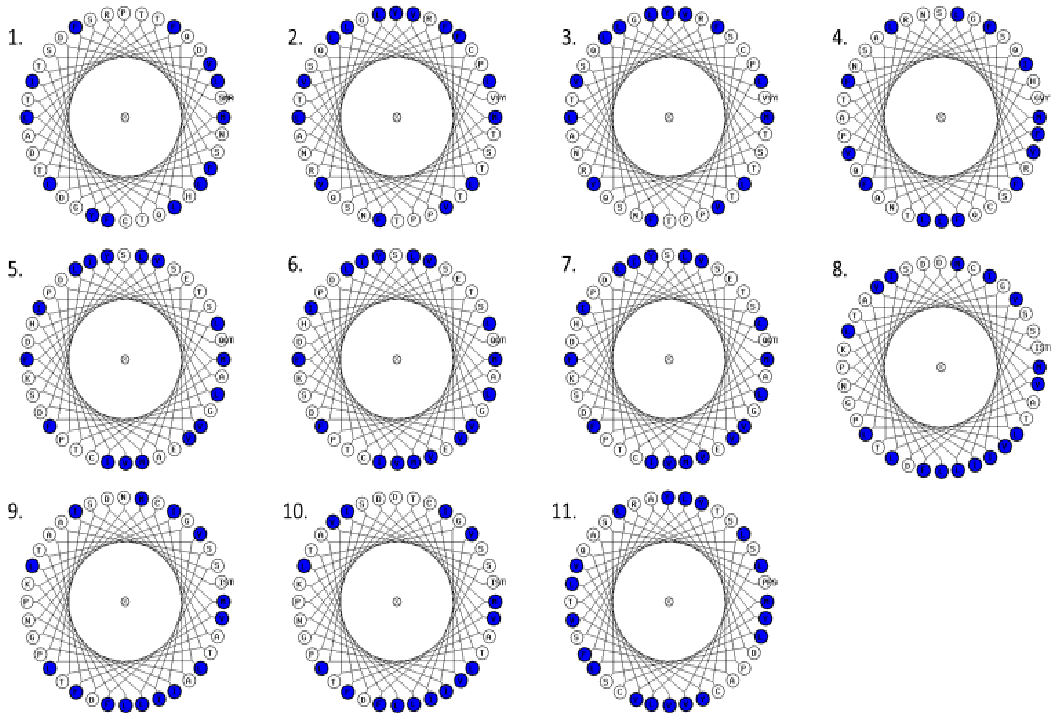


Figure18: Helical wheel diagrams of protein sequences of Spike protein of selected coronaviruses. The hydrophobic amino acid residues are represented by a dark blue colour and hydrophilic residues are not coloured. The hydrophobic residues are comparatively more for Spike protein because of it being a membrane protein with transmembrane segments. Hydrophobic residues seen in protein sequence: F (Phe), I (Ile), L (Leu), V (Val), M (Met), Y (Tyr). Key: 1. SARS CoV (NC_004718.3), 2. SARS-CoV Wu (MN908947.3), 3. BaT CoV RaTG13, 4. Pangolin CoV, 5. Camel CoV, 6. MERS CoV, 7. Dromedarius CoV, 8. H-Enteric CoV, 9. Canine CoV, 10. Bovine CoV, 11. Avian CoV.

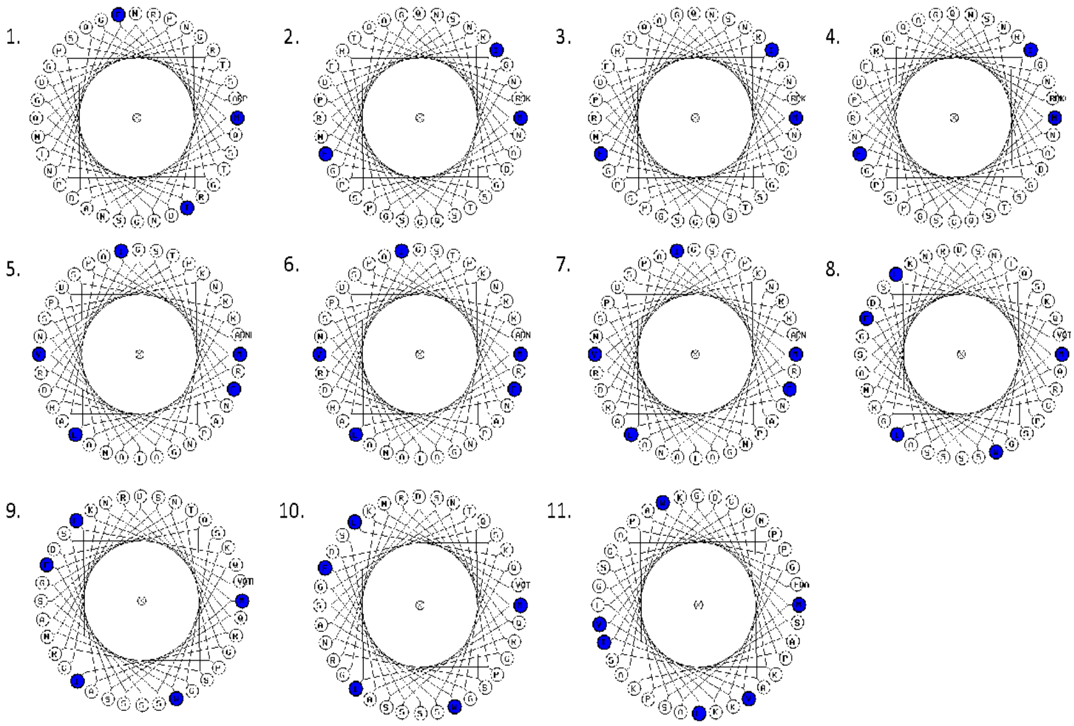
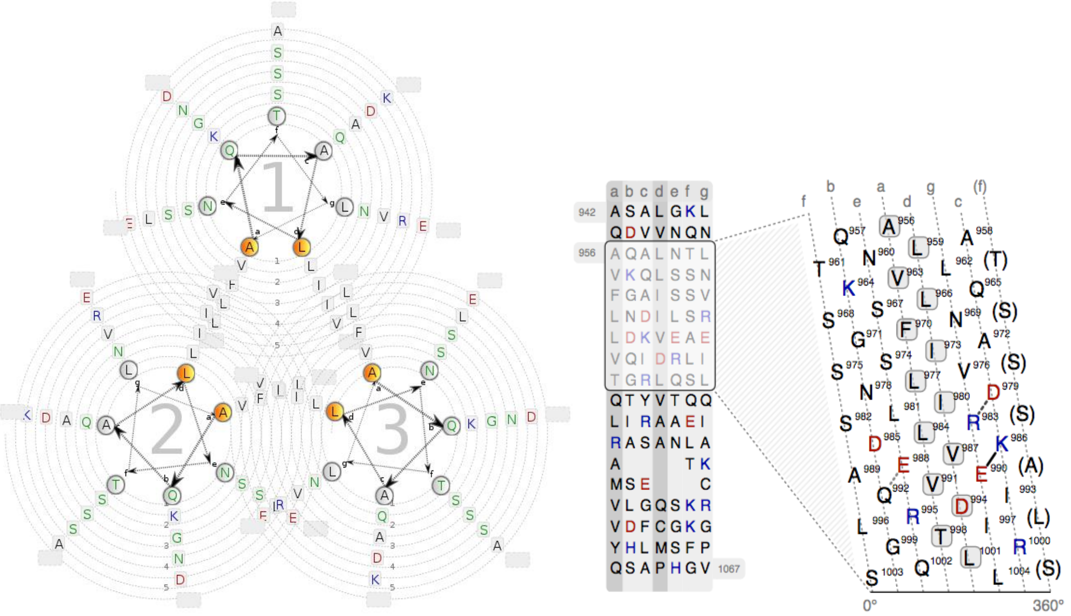


Figure 19: Helical wheel diagrams of protein sequences of Nucleoprotein of selected coronaviruses. (Key similar to Figure14).



(a)

(b)

Figure20: (a) Helical wheel view for 3 heptad repeats in a coiled coil predicted in spike protein sequence of SARS-CoV-2, depicting possible residue interactions. (b) Helical net plots (window-range: 956-1004) showing selected sequence of the spike protein sequence of SARS-CoV-2.

The difference in the amphipathic plot and helical wheel diagrams between transmembrane spike protein and a nucleoprotein was quite marked (Figure 18 & 19), wherein more hydrophilic residues in N proteins can be observed while S proteins were found to contain more hydrophobic residues.

To further explore the nature of TMS, we used Waggawagga (https://waggawagga.motorprotein.de/), an online tool to find out if these TMS were composed of coiled coils or alpha helices. We found that the TMS of spike protein of SARS-CoV-2 was formed of trimeric coiled coils, wherein the charged interactions (salt bridges) between the 3 helical segments could be observed (Figure 20 (a)).

Figure 20 (b) displays the helical net diagram of the same sequence (spike protein of SARS-CoV-2) with one strong, one middle and one weak interaction and Single-Alpha-Helix (SAH) score of 0.0, further confirming the presence of coiled coil regions only. Coiled coils mediate great flexibility in mediating protein-protein interactions with respect to number (dimer, trimer or tetramer), composition and orientation (parallel or antiparallel) of the interacting helical segments (Watkins et al., 2015). Moreover, coiled coils have been used for therapeutic application, for instance, Pimentel et al. (2009) generated a peptide nanoparticle using oligomers of coiled coil fusions for the display of severe acute respiratory syndrome (SARS) virus epitopes

3.11. Detection of Natural Selection

We detected pervasive negative or purifying selection operating on spike and nucleoproteins encoded by different coronaviruses. No sites were identified as positively selected. Notably, the negatively selected amino acid sites may be suitable targets for development of drugs and vaccines because many substitutions at these sites are expected to be intolerable (Suzuki, 2004). The results obtained from Codon-based Z-test and Fisher’s exact test of selection were found to be in favour of rejection of strict-neutrality (dN = dS). Further, we found evidence of pervasive negative selection at 106 sites in spike sequences and at 29 sites in nucleoprotein sequences at p-value threshold of 0.1 using single-likelihood ancestor counting (SLAC) method (Figure 21 & 23). No sites were identified as positively selected. The analysis was based on the models:

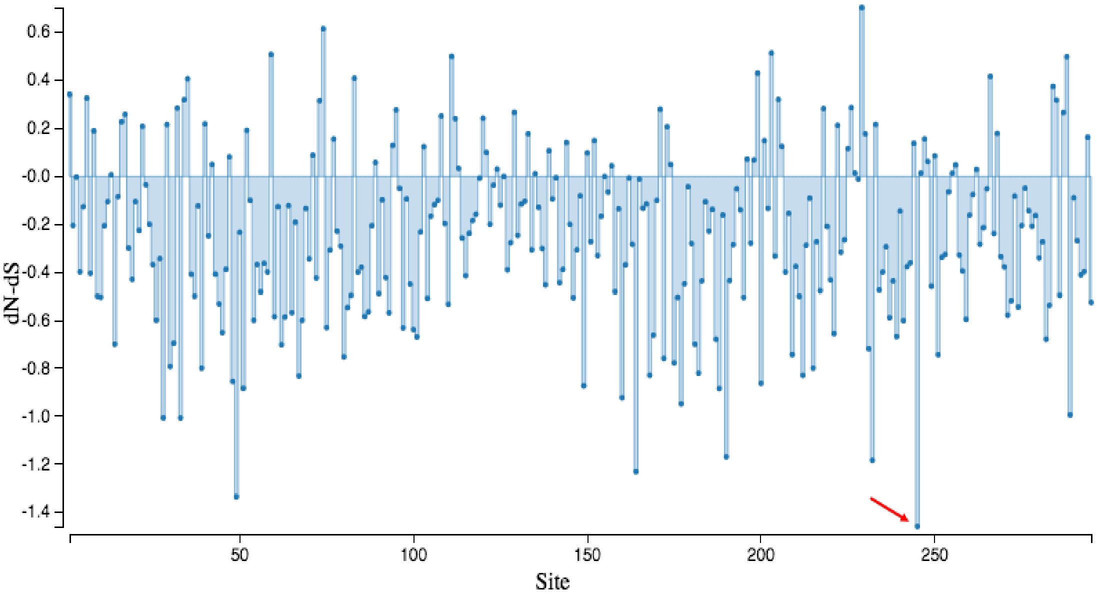


Figure 21: SLAC site graph for Nucleoprotein sequences at p=0.1. The amino acid position 245 (dN-dS = -1.46) is under strongest negative selection (pointed by red arrow).

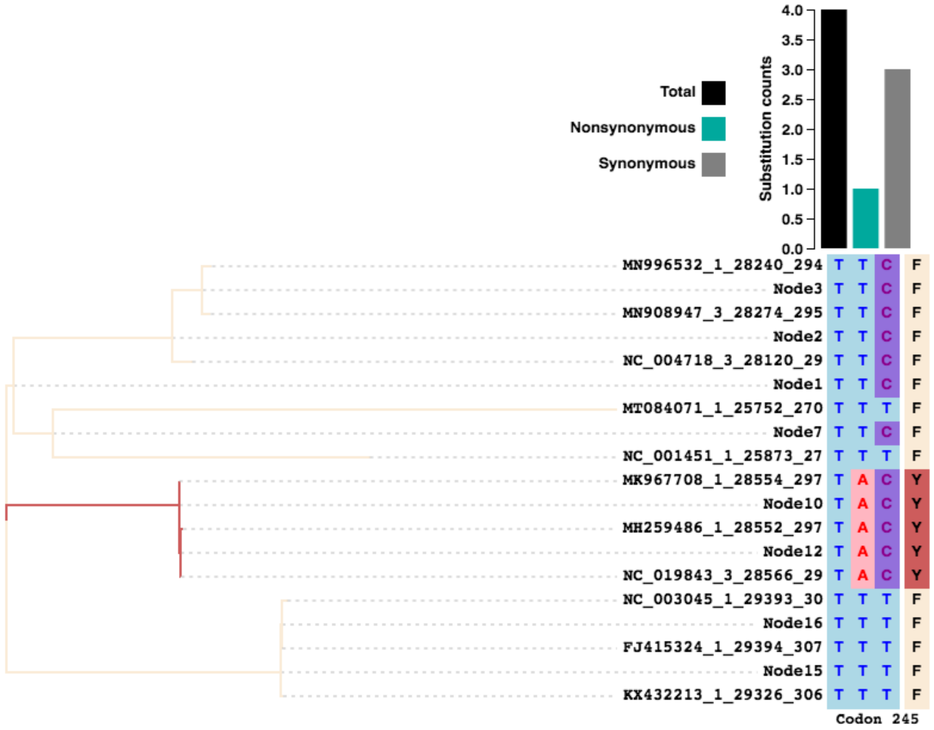

Total
Nonsynonymous
Synonymous
Substitution counts
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
MN996532_1_28240_294 T T C F
Node3 T T C F
MN908947_3_28274_295 T T C F
Node2 T T C F
NC_004718_3_28120_29 T T C F
Node1 T T C F
MT084071_1_25752_270 T T T F
Node7 T T C F
NC_001451_1_25873_27 T T T F
MK967708_1_28554_297 T A C Y
Node10 T A C Y
MH259486_1_28552_297 T A C Y
Node12 T A C Y
NC_019843_3_28566_29 T A C Y
NC_003045_1_29393_30 T T T F
Node16 T T T F
FJ415324_1_29394_307 T T T F
Node15 T T T F
KX432213_1_29326_306 T T T F
Codon 245

Figure22: SLAC phylogenetic Alignment at 245 site of sequence alignment of nucleoprotein sequences.

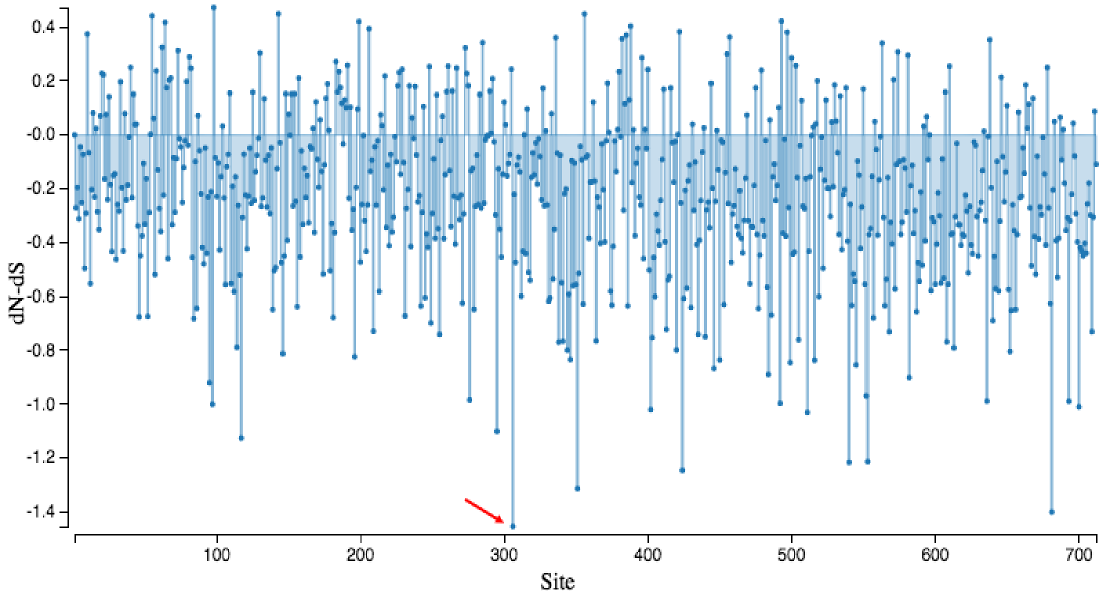


Figure 23: SLAC site graph for spike sequences at p=0.1. The amino acid position 306 (dN-dS = -1.45) is under strongest negative selection (pointed by red arrow).

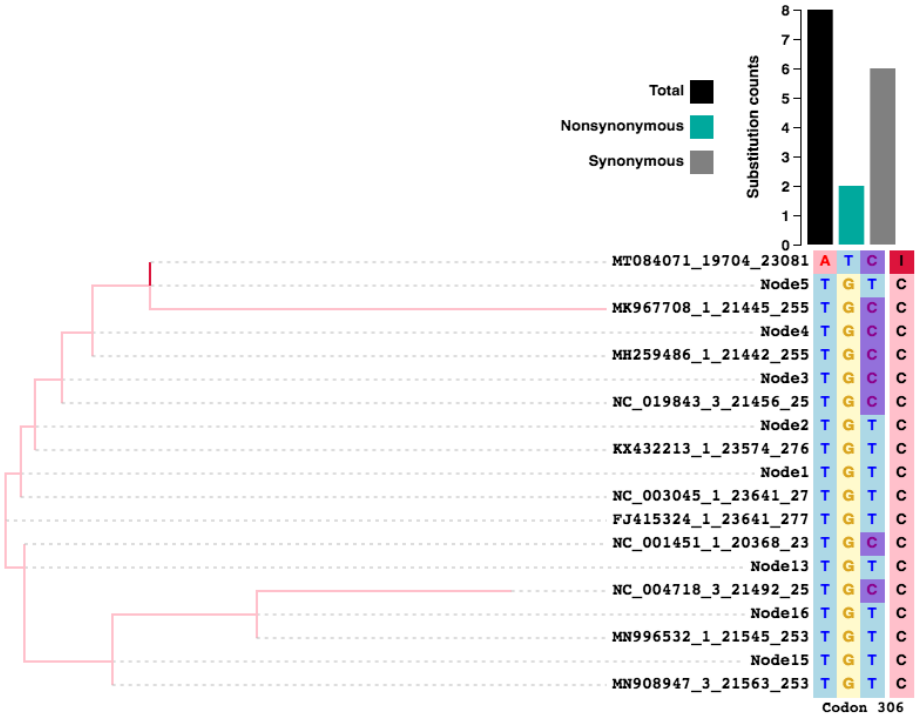


Figure 24: SLAC phylogenetic alignment at 306 site of sequence alignment of spike sequences.

Nucleotide GTR and Global MG94xREV, which gave significant negative values for Log L: -6381.90 & -6159.32 for nucleoprotein genes respectively and -19347.15 & -18746.49 for spike genes respectively, indicating the accuracy of detection of true positives in our results. The analysis was repeated using lower p-value of 0.01, wherein we found evidence of pervasive negative selection at only 2 sites in nucleoprotein sequences and at 21 sites in spike sequences. The most negatively selected site in nucleoprotein and spike sequence alignment was found to be 245 and 306 respectively at p-value of 0.01 and 0.1 (Figure 22 & 24).

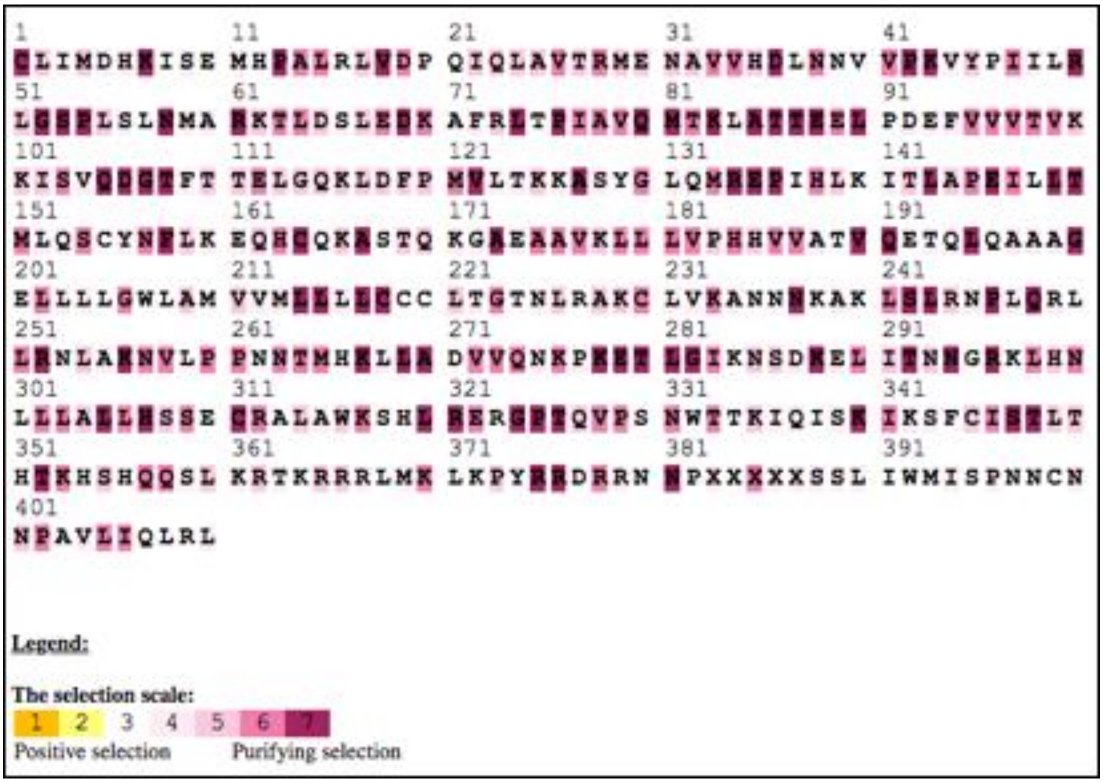


Figure 25: Selecton results for Nucleoprotein gene run on 11 CoV sequences. Purifying selection is colored in shades of magenta.

We also used Selecton Server to detect putative sites under positive or negative selection in nucleoprotein and spike sequences. This tool uses advanced models for detecting positive and purifying selection using a Bayesian inference approach (Stern et al., 2007). In this analysis also, no positively selected sites were found in both N and S sequences. Both the genes were found to be under negative selection. In addition to the sites detected under negative selection by SLAC method, Bayesian inference approach detected additional negatively selected sites in both the genes (Figure 25 & 26).

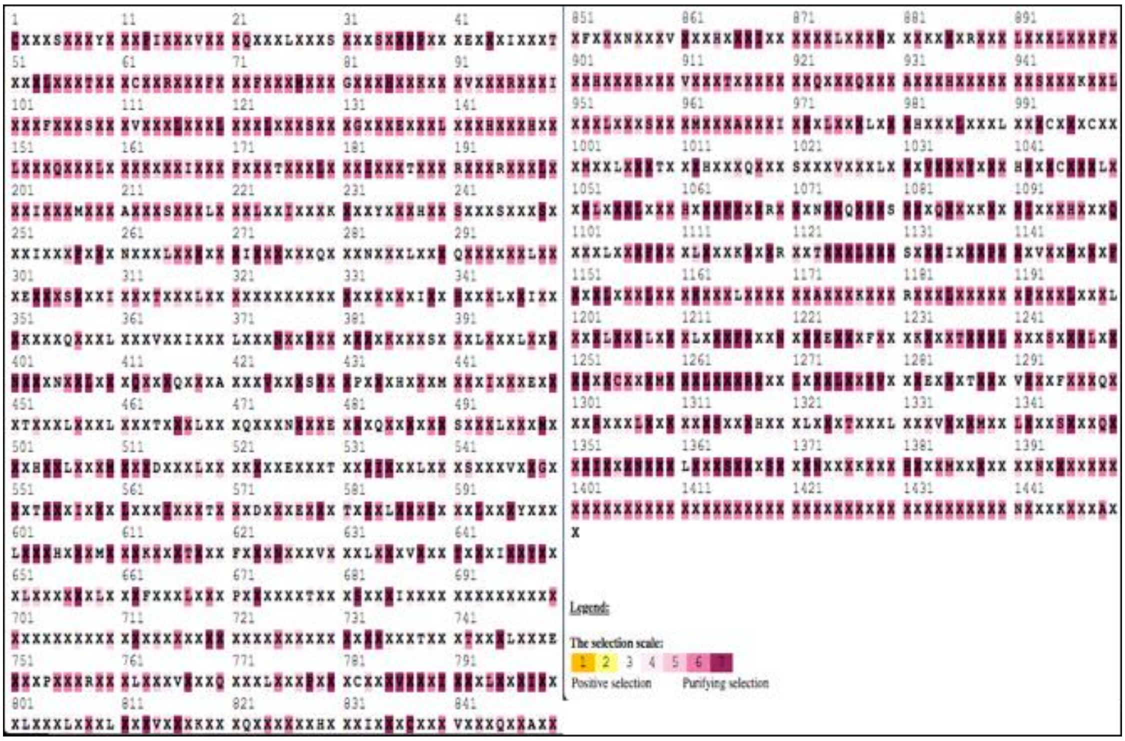


Figure26: Selecton results for Spike gene run on 11 CoV sequences. Purifying selection is colored in shades of magenta.

4. CONCLUSION

In this study, we gained important insights into the genomic features of SARS-CoV-2, the causative organism of coronavirus disease 2019 (Covid-19) using in-silico tools. Comparative genomics provided important evidences of how SARS CoV-2 and other taxonomically related coronaviruses are related to each other at the genetic level. The phylogenetic analyses confirmed the close genetic relationship of SARS CoV-2 with SARS CoV, Bat CoV RaTG13 and Pangolin CoV. Our results showed that the nucleocapsid and spike genes in CoVs are under strong negative evolutionary constraints. The codon usage analysis in the selected CoVs reinforced a fundamental property of most of the RNA viruses, wherein mutation plays a significant role in the evolution of these RNA viruses. So, in this study, we made an attempt to explore the genomes of these coronaviruses with special focus on two important genes (Nucleoprotein and Spike) and laid emphasis on their physico-chemical properties, post translation modifications (PTMs) and presence of coiled coil regions with an expectation that these studied parameters may act as primer for further studies related to development of vaccine targets against SARS-CoV-2.

5. CONFLICT OF INTEREST STATEMENT

The authors declare that there is no conflict of interest.

6. ACKNOWLEDGEMENTS

We are highly thankful to the Centre for Research, Maitreyi College and the Principal, Maitreyi College, University of Delhi for her constant support and encouragement.

7. SOURCE OF FUNDING

Nil

8. REFERENCES

Altschul, S.F., Gish, W., Miller, W., Myers, E. W., & Lipman, D.J. (1990). Basic local alignment search tool. *Journal of Molecular Biology, 215(3)*, 403–410.

Apostolovic, B., Danial, M., & Klok, H.-A. (2010). Coiled coils: attractive protein folding motifs for the fabrication of self-assembled, responsive and bioactive materials. *Chemical Society Reviews*, *39(9)*, 3541.

Artimo, P., Jonnalagedda, M., Arnold, K., Baratin, D., Csardi, G., et al. (2012). ExPASy: SIB bioinformatics resource portal. *Nucleic Acids Research*, *40(W1)*, W597–W603.

Bagdonaite, I., & Wandall, H.H. (2018). Global aspects of viral glycosylation. *Glycobiology, 28(7)*, 443–467.

Cagliani, R., Forni, D., Clerici, M., & Sironi, M. (2020). Computational Inference of Selection Underlying the Evolution of the Novel Coronavirus, Severe Acute Respiratory Syndrome Coronavirus 2. *Journal of Virology, 94(12).* **DOI:** 10.1128/JVI.00411-20.

Chattopadhyay, S., Bagchi, P., Dutta, D., Mukherjee, A., Kobayashi, N., & Chawlasarkar, M. (2010). Computational identification of the post-translational modification sites and the functional family prediction reveals possible moonlighting role of rotaviral proteins. *Bioinformation*, *4(10)*, 448–451.

Chen, L., Keppler, O. T., & Schölz, C. (2018). Post-translational Modification-Based Regulation of HIV Replication. *Frontiers in Microbiology, 9*. https://doi.org/10.3389/fmicb.2018.02131

Coleman, J.R., Papamichail, D., Skiena, S., Futcher, B., Wimmer, E., & Mueller, S. (2008). Virus attenuation by genome-scale changes in codon pair bias. *Science* (New York, N.Y.), *320(5884)*, 1784–1787.

Darling, A.C.E. (2004). Mauve: Multiple Alignment of Conserved Genomic Sequence With Rearrangements. *Genome Research, 14(7),* 1394–1403.

Forni, D., Cagliani, R., Clerici, M., & Sironi, M. (2017). Molecular Evolution of Human Coronavirus Genomes. *Trends in Microbiology, 25(1),* 35–48.

Fung, T. S., & Liu, D. X. (2018). Post-translational modifications of coronavirus proteins: roles and function. *Future Virology, 13(6),* 405–430.

Gupta, A. K., Khan, Md. S., Choudhury, S., Mukhopadhyay, A., Sakshi, Rastogi, A. (2020). CoronaVR: A Computational Resource and Analysis of Epitopes and Therapeutics for Severe Acute Respiratory Syndrome Coronavirus-2. *Frontiers in Microbiology, 11*. https://doi.org/10.3389/fmicb.2020.01858

Gupta, R., & Brunak, S. (2002). Prediction of glycosylation across the human proteome and the correlation to protein function. Pacific Symposium on Biocomputing. *Pacific Symposium on Biocomputing,* 310–322.

Iakoucheva, L.M., Radivojac, P., Brown, C.J., O'Connor, T.R., Sikes, J.G., Obradovic, Z., & Dunker, A.K. (2004). The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Research, 32(3),* 1037–1049.

Keck, F., Ataey, P., Amaya, M., Bailey, C., & Narayanan, A. (2015). Phosphorylation of Single Stranded RNA Virus Proteins and Potential for Novel Therapeutic Strategies. *Viruses, 7(10),* 5257–5273.

Kosakovsky Pond, S.L., Posada, D., Gravenor, M.B., Woelk, C.H., & Frost, S.D.W. (2006). GARD: a genetic algorithm for recombination detection. *Bioinformatics, 22(24),* 3096–3098.

Kumar, R., Verma, H., Singhvi, N., Sood, U., Gupta, V.,.....& Lal, R. (2020). Comparative Genomic Analysis of Rapidly Evolving SARS-CoV-2 Reveals Mosaic Pattern of Phylogeographical Distribution. *MSystems, 5(4)*. **DOI:** 10.1128/mSystems.00505-20.

Kumar, S., Stecher, G., Li, M., Knyaz, C., & Tamura, K. (2018). MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms. *Molecular Biology and Evolution*, *35(6)*, 1547–1549.

Li, F. (2016). Structure, Function, and Evolution of Coronavirus Spike Proteins. *Annual Review of Virology*, *3(1)*, 237-261. https://dx.doi.org/10.1146%2Fannurev-virology-110615-042301

Li, L.-C., & Dahiya, R. (2002). MethPrimer: designing primers for methylation PCRs. *Bioinformatics*, *18(11)*, 1427–1431.

Li, X., Song, Y., Wong, G., & Cui, J. (2020). Bat origin of a new human coronavirus: there and back again. *Science China Life Sciences*, *63(3)*, 461–462.

Lin, S.-M., Lin, S.-C., Hsu, J.-N., Chang, C., Chien, C.-M., Wang, Y.-S., Wu, H.-Y., Jeng, U.-S., Kehn-Hall, K., & Hou, M.-H. (2020). Structure-Based Stabilization of Non-nativeProtein–Protein Interactions of Coronavirus Nucleocapsid Proteins in Antiviral Drug Design. *Journal of Medicinal Chemistry*, *63(6)*, 3131–3141.

McFarlane, A.A., Orriss, G.L., & Stetefeld, J. (2009). The use of coiled-coil proteins in drug delivery systems. *European Journal of Pharmacology*, *625(1–3)*, 101–107.

Noppakunmongkolchai, W., Poyomtip, T., Jittawuttipoka, T., Luplertlop, N., Sakuntabhai, A., Chimnaronk, S., Jirawatnotai, S., & Tohtong, R. (2016). Inhibition of protein kinase C promotes dengue virus replication. *Virology Journal*, *13(1)*. https://doi.org/10.1186/s12985-016-0494-6

Pimentel, T. A. P. F., Yan, Z., Jeffers, S. A., Holmes, K. V., Hodges, R. S., & Burkhard, P. (2009). Peptide Nanoparticles as Novel Immunogens: Design and Analysis of a Prototypic Severe Acute Respiratory Syndrome Vaccine. *Chemical Biology and Drug Design*, *73(1)*, 53–61.

Saier, M. H., Jr, Tran, C. V., & Barabote, R. D. (2006). TCDB: the Transporter Classification Database for membrane transport protein analyses and information. *Nucleic Acids Research*, *34*(Database issue), D181–D186. https://doi.org/ 10.1093/nar/gkj001.

Sharun, K., Sircar, S., Malik, Y.S., Singh, R.K., & Dhama, K. (2020). How close is SARS-CoV -2 to canine and feline coronaviruses? *Journal of Small Animal Practice, 61(8),* 523–526.

Sheikh, A., Al-Taher, A., Al-Nazawi, M., Al-Mubarak, A. I., & Kandeel, M. (2020). Analysis of preferred codon usage in the coronavirus N genes and their implications for genome evolution and vaccine design. *Journal of Virological Methods, 277,* 113806. https://doi.org/10.1016/j.jviromet.2019.113806

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K.,…. & Higgins, D.G. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology, 7,* 539. https://doi.org/10.1038/msb.2011.75

Stern, A., Doron-Faigenboim, A., Erez, E., Martz, E., Bacharach, E., & Pupko, T. (2007). Selecton 2007: advanced models for detecting positive and purifying selection using a Bayesian inference approach. *Nucleic Acids Research, 35*(Web Server), W506–W511.

Suzuki Y. (2004). Negative selection on neutralization epitopes of poliovirus surface proteins: implications for prediction of candidate epitopes for immunization. *Gene, 328,* 127–133.

Tello, M., Vergara, F., & Spencer, E. (2013). Genomic adaptation of the ISA virus to Salmo salar codon usage. *Virology Journal,* 10(1), Article number *223.* https://doi.org/10.1186/1743-422X-10-223

Villard, V., Agak, G. W., Frank, G., Jafarshad, A., Servis, C.,……..& Corradin, G. (2007). Rapid Identification of Malaria Vaccine Candidates Based on α-Helical Coiled Coil Protein Motif. *PLoS ONE, 2(7),* e645. https://doi.org/10.1371/journal.pone.0000645

Watkins, A. M., Wuo, M. G., & Arora, P. S. (2015). Protein–Protein Interactions Mediated by Helical Tertiary Structure Motifs. *Journal of the American Chemical Society, 137(36),* 11622–11630.

Wong, G., Bi, Y. H., Wang, Q. H., Chen, X. W., Zhang, Z. G., & Yao, Y. G. (2020). Zoonotic origins of human coronavirus 2019 (HCoV-19 / SARS-CoV-2): why is this work important? *Zoological Research, 41(3),* 213–219.

Xia, X. (2020). Extreme Genomic CpG Deficiency in SARS-CoV-2 and Evasion of Host Antiviral Defense. *Molecular Biology and Evolution*, *37(9)*, 2699-2705. https://doi.org/10.1093/molbev/msaa094

Zhou, P., Yang, X.-L., Wang, X.-G., Hu, B., Zhang, L.,.......& Shi, Z.-L. (2020). A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*, *579(7798)*, 270–273. https://doi.org/10.1038/s41586-020-2012-7

How to cite this article: Kaur, J., Srishti, & Sharma, S. (2020). Comparative genomic and phylogenetic analysis of spike and nucleocapsid proteins of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and 9 other taxonomically related coronaviruses using in-silico tools. *Vantage: Journal of Thematic Analysis*, *1(2)*: 46-81.
DOI: https://doi.org/10.52253/vjta.2020.v01i02.05