Multi-Omics and Integrated Network Approach to Unveil Evolutionary Patterns, Mutational Hotspots, Functional Crosstalk and Regulatory Interactions in SARS-CoV-2

Vipin Gupta[¥1], Shaiza Haider[¥2], Mansi Verma[¥3], Kalaiarasan Ponnusamy[4], Md. Zubbair Malik[5], Nirjara Singhvi[6], Helianthous Verma[7], Roshan Kumar[8], Utkarsh Sood[9], Princy Hira[10], Shiva Satija[3] and Rup Lal[9*]

[1] PhiXGen Private Limited, Gurugram, Haryana 122001, India

[2] Department of Biotechnology, Jaypee Institute of Information Technology, Noida, sector-62, Uttar Pradesh, India

[3] Department of Zoology, Sri Venkateswara College, University of Delhi, New Delhi-110021, India

[4] School of Biotechnology, Jawaharlal Nehru University, New Delhi, India.

[5] School of Computational and Integrative Sciences, Jawaharlal Nehru University, New Delhi, India.

[6] Department of Zoology, University of Delhi, New Delhi-110007, India

[7] Molecular Biology and Genomics Research Laboratory, Ramjas College, University of Delhi, 110007, India

[8] P.G. Department of Zoology, Magadh University, Bodh Gaya, Bihar-824234, India

[9] The Energy and Resources Institute, Darbari Seth Block, IHC Complex, Lodhi Road, New Delhi-110003, India

[10] Maitreyi College, University of Delhi. Chanakyapuri, New Delhi 110021

*Corresponding Author

[¥] Contributed Equally

Corresponding author Email: ruplal@gmail.com

Abstract

SARS-CoV-2 responsible for the pandemic of the Severe Acute Respiratory Syndrome resulting in infections and death of millions worldwide with maximum cases and mortality in USA. The current study focuses on understanding the population specific variations attributing its high rate of infections in specific geographical regions which may help in developing appropriate treatment strategies for COVID-19 pandemic. Rigorous phylogenetic network analysis of 245 complete SARS-CoV-2 genomes inferred five central clades named a (ancestral), b, c, d and e (subtype e1 & e2) showing both divergent and linear evolution types. The clade d & e2 were found exclusively comprising of USA strains with highest known mutations. Clades were distinguished by ten co-mutational combinations in proteins; Nsp3, ORF8, Nsp13, S, Nsp12, Nsp2 and Nsp6 generated by Amino Acid Variations (AAV). Our analysis revealed that only 67.46 % of SNP mutations were carried by amino acid at phenotypic level. T1103P mutation in Nsp3 was predicted to increase the protein stability in 238 strains except six strains which were marked as ancestral type; whereas com (P5731L & Y5768C) in Nsp13 were found in 64 genomes of USA highlighting its 100% co-occurrence. Docking study highlighted mutation (D7611G) caused reduction in binding of Spike proteins with ACE2, but it also showed better interaction with TMPRSS2 receptor which may contribute to its high transmissibility in USA strains. In addition, we found host proteins, MYO5A, MYO5B & MYO5C had maximum interaction with viral hub proteins (Nucleocapsid, Spike & Membrane). Thus, blocking the internalization pathway by inhibiting MYO-5 proteins which could be an effective target for COVID-19 treatment. The functional annotations of the Host-Pathogen Interaction (HPI) network were found to be highly associated with hypoxia and thrombotic conditions confirming the vulnerability and severity of infection in the patients. We also considered the presence of CpG islands in Nsp1 and N proteins which may confers the ability of SARS-CoV-2 to enter and trigger methyltransferase activity inside host cell.

Introduction

In December 2019, a novel RNA virus, Severe Acute Respiratory Syndrome Corona Virus-2 (SARS-CoV-2), belonging to Coronaviridae family (betacoronavirus), emerged as the reason for the chaos of pneumonia disease also called Covid-19 in Chinese city, Wuhan (Li et al., 2020). Covid-19 was declared as pandemic by WHO on March 11, 2020 (Astuti and Ysrafil, 2020). Major outbreaks were reported in many locations of China, USA, Italy, Spain, Japan,

and South Korea. As of date it has already spread to more than 200 countries of the world surpassing more than 30 thousand deaths and 6 million reported active cases worldwide (https://www.worldometers.info/coronavirus/).

SARS-CoV-2 is a single stranded RNA virus with a genome size ranging from 29.8 kb to 29.9 kb (Khailany et al., 2020). The genomic repertoire of SARS-CoV-2 comprises of 10 open reading frames (ORFs) encoding 27 proteins (Abduljalil and Abduljalil, 2020). ORF1ab encodes for 16 non-structural proteins (Nsp) whereas structural proteins include spike (S), envelope (E), membrane (M), and nucleocapsid (N) proteins (Pyrc et al., 2007; Yang and Leibowitz, 2015). In addition, the genome of SARS-CoV-2 comprises of ORF3a, ORF6, ORF7a, ORF7b, ORF8 and ORF9 genes encoding six accessory proteins, flanked by 5' and 3' UTRs (Khailany et al., 2020). In our previous study (Kumar et al., 2020), a higher mutational rate in the genomes from different geographical locations around the world by accumulation of Single Nucleotide Polymorphisms (SNP) was reported. Even during these early stages of the global pandemic, genomic surveillance has been used to differentiate circulating strains into distinct, geographically based lineages (Forster et al., 2020). However, the ongoing analysis of this global dataset suggests no consolidated significant links between SARS-CoV-2 genome sequence variability, virus transmissibility and disease severity.

It is known that mutations at both genomic and protein level are "Hormonical Orchestra" (Yu et al., 2019) that drives the evolutionary changes, demanding a detailed study of SARS-CoV-2 mutations to understand its successful invasion and infection. The study analyzed that mutational profiles of SARS-CoV-2 isolates show very high mutational rates that show the isolates more virulent, causing significant harm to the hosts (Mandal et., 2020). Thus, in the present study, we selected 245 genomic sequences of SARS-CoV-2 deciphering the phylogenetic relationships, tracing them to SNPs at nucleotide and amino acid (Amino Acid Variation) levels and performing structural re-modelling. Our results revealed the evolutionary relationships among the strains predicting Nsp3 as mutational hotspot for SARS-CoV-2. We further extended the study to understand mechanism of host immunity evasion by Host-Pathogen Interaction (HPI) and confirming their interactions with host proteins by docking studies. We identified sparsely distributed hubs which may interfere and control network stability as well as other communities/modules. This indicated the affinity to attract a large number of low-degree nodes toward each hub, which is a strong evidence of controlling the topological properties of the network by these few hubs (Nafis et al., 2015). We also analyzed the transfer of genomic SNPs to amino acid levels and associations of CpG islands

contributing towards the pathogenicity of SARS-CoV-2. The existence of CpG islands has always been connected with the epigenetic regulation and act as hotspots for methylation (Jones, 2012; Shiraishi et al., 2002; Hoelzer et al., 2008). Here also, the conservancy found in possession of CpG islands towards the extremities of all the genomes considered in the present analysis indicate their importance in evading host immunity. Our study showed an overall depiction of SARS-CoV-2 variations and interactions that eventually may lead to development of rational therapeutic measures and medication against COVID-19.

Material and Methods

Selection of genomes, annotations and phylogeny construction

Publicly available genomes of SARS-CoV-2 viruses were obtained from the NCBI database (https://www.ncbi.nlm.nih.gov/genbank/sars-cov-2-seqs/). Until March 31, 2020 only 375 SARS-CoV-2 genomes were available in the databases. The data was screened for unwanted ambiguous bases using N-analysis program, based on which 245 complete and clean genomes of SARS-CoV-2 were selected for further analysis (Supplementary Info 1). A manually annotated reference database was generated using GenBank file of severe acute respiratory syndrome coronavirus 2 isolate- SARS-CoV-2/SH01/human/2020/CHN (Accession number: MT121215.1) and open reading frames (ORFs) were predicted against the formatted database using prokka (-gcode 1) (Seemann, 2014). Genomic sequences included in the analysis belongs to different countries namely, USA (168), China (53), Pakistan (2), Australia (1), Brazil (1), Finland (1), India (2), Israel (2), Japan (5), Vietnam (2), Nepal (1), Peru (1), South Korea (1), Spain (1), Sweden (1). Whole genomes nucleotide and protein sequences were aligned using mafft (Katoh et al., 2013) at 1000 iterations. The alignments so obtained were processed for phylogeny construction using BioEdit software (Hall et al., 2011). The nucleotide-based phylogeny was annotated and visualized on iTOL server (Letunic and Bork, 2006). While, amino acid-based phylogeny was visualized and annotated using GrapeTree (Zhou et al., 2018).

Genotyping based on SNP/AAV

To detect nucleotide and amino acid variations (AAV) among 245 genomes of SARS-CoV-2, sequence alignment of nucleotide and amino acid, respectively were performed against the reference genome. The change of nucleotide and amino acid was calculated as point variations and were recorded. The interpolation and visualization were plotted using computer programs

in Python. Co-mutation were predicted and clustering was performed using MicroReact (Argimon et al., 2016)

Data and Computer programs:

The genomic analytics is performed using programs in Python and Biopython libraries (Cock et al., 2009). The computer programs and the updated SNP profiles of SARS-CoV-2 isolates are available upon requests.

Construction of the Host-Pathogen Interaction Network of SARS-CoV-2

In order to find the HPI, we subjected SARS-CoV-2 proteins to Host-Pathogen interaction databases such as Viruses.STRING v10.5 (Cook et al., 2018) and HPIDB3.0 (Ammari et al., 2016) for predicting their direct interaction with human as the principal host. The HPI network was constructed and visualized using Cytoscape v3.7.2 (Shannon, et al., 2003). In the constructed Network, proteins with highest degree, which interact with several other signaling proteins in the network indicate a key regulatory role as a hub. In our study, using NetworkAnalyser (Assenov, et al., 2008), plugin of Cytoscape v3.7.2, we identified the hub protein and subjected to functional analysis. The network was functionally annotated using STRINGApp and StringEnrichment app (Doncheva et al., 2019) plugin of Cytoscape using Reactome, GO, InterPRO, KEGG and Pfam databases. This analysis provides an opportunity of a more precise understanding of the biological functions, providing valuable clues for biologists.

Computational structural analysis on wild-type and mutant SARS-CoV-2 proteins

SARS-CoV-2 proteins sequences were retrieved from the NCBI genome database and pairwise sequence alignment of wild-type and mutant proteins were carried out by the Clustal Omega tool (Sievers et al., 2011). The wild-type and mutant homology model of S-protein, NspNsp12 and Nsp13 were constructed using the SWISSMODEL (Waterhouse et al., 2018), whereas the 3D structure of ORF8, ORF3A, Nsp2, Nsp3 and Nsp6 were predicted using Phyre2 server (Kelley et al., 2015). The host proteins (TMPRSS2, RPS6, ATP6V1G1 and MYO5C) 3D structures were generated using the SWISSMODEL and ACE2 structure retrieved from the PDB database (PDB ID: 6M17). These structures were energy minimized by the Chiron energy minimization server (Ramachandran et al., 2011). The effect of the mutation was analyzed using HOPE (Venselaar et al., 2010) and I-mutant (Capriotti et al., 2006). The I-mutant method allows us to predict the stability of the protein due to mutation.

The docking studies for wild and mutant SARS-CoV-2 proteins with host proteins was carried out using PatchDock Server (Schneidman-Duhovny et al., 2005). Structural visualizations and analysis were carried out using pyMOL2.3.5 (Jacobson et al., 2002).

Analysis of CpG regions

SARS-CoV-2 genomes were analysed for the presence of CpG regions that can be targeted for methylation induced gene silencing. To locate the CpG regions, meth primer 2.0 (http://www.urogene.org/methprimer2/) and the CpG Plot (http://www.ebi.ac.uk/Tools/emboss/cpgplot/) programs were used, although some variations were found in both the programs. Both the programs were run on default parameters of a sequence window longer than 100 bp; GC content of ≥50%, and an observed/expected CpG dinucleotide ratio ≥0.60. The presence of common CpG islands was confirmed by performing BLAST using the above reference strain.

Results and Discussion

Phylogenetic relationship between different SARS-CoV-2 strains

In our previous study, we reported a mosaic pattern of phylogenetic clustering of 95 genomes of SARS-COV-2 isolated from different geographical locations (Kumar et al., 2020). Strains belonging to one country were found clustered with distant countries strains but not with the neighboring one. Taking clue from these studies we constructed phylogenetic relatedness of 245 strains of SARS-COV-2 from USA, China, and several other countries including, Spain, Vietnam, Peru, Finland and Pakistan and unravel the significant association of evolutionary patterns among SARS-CoV-2 based on their geographical locations predicting their mosaic phylogenetic arrangements. It was found that the majority of strains from USA were clustered together, but comparatively high divergences were found in strains isolated from China and Japan. Japanese strains were found to be scattered and formed clusters with strains from USA, Pakistan, Vietnam, Taiwan, and China. Even with less number of genomes sequences from Japan, Vietnam and Peru revealed a highly scattered pattern and formed close associations with that of USA and Chinese strains. Strains reported from patients of Taiwan (MT192759), Australia (MT007544), South Korea (MT039890), Nepal (MT072688) and Vietnam (MT192773, MT192772) had travel histories from Wuhan, China (Cheng et al., 2020). However, a strain from Pakistan (MT240479) which clustered with the Japanese strains was found to be isolated from patient having travel history from Iran. Indian strains (MT050439, MT012098) that were isolated from patients who travelled from Dubai, clustered with

Chinese strains. Later, reports confirmed many cases of SARS-CoV-2 in Dubai from China (https://www.newsbytesapp.com/timeline/India/58169/271167/coronavirus-2-positive-cases-detected-in-delhi-telangana). Thus, a clear landscape of phylogenetic relationships could be obtained reflecting mosaic clustering patterns in accordance with the travel history of patients (Figure1A). However, results were in contradiction with the genomic analysis of SARS-COV-2 by Foster et al.,(2020) where they predicted the linear/directive evolution from ancestral node a to node b and c. Whereas we report here both divergent (from ancestral node a to b, c & e) and directive (node c to d) evolution among the SARS-CoV-2 strains (Figure1B and Figure 3B). Since genome-based phylogeny did not highlight the amino acid level changes, thus to ascertain the variations among the SARS-CoV-2 strains at phenotypic level, we constructed whole proteome alignment-based phylogeny, clustered the 245 strains into five major **clades a-e (**Figure 1B). The first cluster, Clade-a had maximum nodes (46), including reference node, and strains from Nepal (MT072688), Pakistan (MT262993), Taiwan (MT192759) along with 15 strains from USA and 27 strains from China. It also had the mutated daughter nodes (highlighted by # in figure 3 B for corresponding nodes) radiating outwards, belonging to China, Finland (MT020781), India (MT012098), Japan (LC534419, LC529905), Taiwan (MT066176), Vietnam (MT192772-3), Brazil (MT126808), Australia (MT007544), South Korea (MT039890) and Sweden (MT093571) along with seven USA strains (Figure 1B). This clade represented the ancestral node as it harbored the oldest known SARS-CoV-2 strain from China and laid down the foundation of rest of the mutated daughter strains worldwide, marking the onset of the divergence in SARS CoV-2. Three significantly diverged network nodes originated from the ancestral clade-a and were marked as clade-b, c and e (Figure 1B). For **Clade-b**, central node included only four strains in which two were from USA (MT184912, MT276328) and one each from Israel (MT276597) and Japan (LC528233). Its major descended radiant belonged to Japan (LC528232, LC534418), Pakistan (MT240479), USA (MT184913, MT184910, MN997409) and China (MT049951, MT226610). It was observed that one of the Chinese strains in clad-b (MT226610) had the longest branch length making the strain very distinct (harboring 25 mutations) by showing exceptionally high rate of evolution. In **Clade-c** lineage, small central node was comprised of Taiwan (MT066175), USA (MT246667, MT233526, MT020881, MT985325, MT020880) and Chinese (MN938384, LR757995) strains. Interestingly one strain each from Spain (MT233523) and India (MT050493) were also found radiating as daughter node from the central one. **Clade-d** lineage, which was originated from clade-c lineage, consisted only of USA strains both in central nodes and radiations. Importantly, 2 strains (MT263416,

MT246471) were found most divergent with varied mutation suggesting the high rate of evolution among USA strains which might be linked with the high pathogenicity among them. **Clade-e** bifurcated into two sub-clads (e1 and e2) by significant set of mutations. Sub-clad-e1 include six strains from USA, one from Israel (MT276598) with radiating nodes from Peru (MT263074) and USA (MT276327); whereas, sub-clad e2 had 32 strains belonging to USA. Thus, formation of five major evolutionary clades and subclades based on the amino acid phylogeny needs attention for identifying the assessment of divergence among SARS-CoV-2 strains. This divergence is a proof of the random evolution of SARS-CoV-2 suggesting network expansion in five clads contradicting to the earlier directed evolution proposed by Foster et al., 2020.

Genotyping and variation estimation

In order to understand the implication of mosaic pattern of transmissions and evolutionary lineage clustering (Clad a-e), we studied the Single Nucleotide Polymorphism (SNP) genotyping from the 245 genome sequences as mutation counts along with their frequency at specific genomic locations. Mutational changes at phenotypic levels were also weighed by assessing Amino Acid Variations (AAV). Interpolations of the SNPs/AAVs data were made by assessing their frequency, genomic positions and type of SNPs/AAVs (Figure 2B), highlighted a large mutational diversity among the virus isolates. We identified a total of 12 SNP types (A>G, A>C, A>T, C>A, C>G, C>T,G>A,G>C, G>T, T>A, T>C, T>G) accounting for mutations at 297 genomic locations (Figure 2A, 2B). Overall pattern of SNPs suggested C>T transition as the most common mutation in the entire genomic sets (Figure2A), however highest frequency was recorded for T>C transitions (Figure 2B). Based on the genomic arbitrators SNP frequencies, we analyzed 14 major locations inside the genomes of SARS-CoV-2 for potential mutation generating different allelic forms for genes (Table 1). The SNP of C>T was first observed at 67^{th} location in 5' UTR region of leader sequence with a frequency of 45 followed by Nsp2 at two locations (885 & 2863) with the frequency of 29 and 44, respectively. Nsp3/PL-PRO and Nsp8 marked the highest frequency of 238 SNP counts of T>C at 5852 and 12299 locations. Another T>C SNP was observed in ORF8 with frequency of 88 at 27973 location. C>T SNP transformation was found in Nsp4 and Nsp12 with the frequency of 88 and 44 at location 8608 and 14234, respectively. Non-structural protein, Nsp13 was strangely found harboring two different SNP (C>T and A>G) at three different locations (17573, 17684, 17886) with a relatively high frequency of 68, 63 and 63

respectively. A>G SNP conversion in S (Spike) protein was found with a frequency of 43. A Low SNP count of G>T transitions were falling in the ORF3a and Nsp6 with frequency of 32 and 21, respectively (Table 1). Though, all SNP counts do not reflect the phenotypic change at protein level and therefore must be estimated at the translation levels for their significant phenotypic effect. Although 297 genomic locations harbored SNPs but their corresponding AAV were found only in 200 genomic locations accounting for 67.34% conversion efficiency. Out of 14 high frequency SNPs, only 9 mutations [Nsp2 (T265I), Nsp3 (S1920P), Nsp6 (L3605F), Nsp12 (P4618L), Nsp13 (P5731L, Y5768C), S (D7611G), Orf3a (Q8327H), Orf8 (L9033S)] were found to reflect at protein level with the highest frequency of 238 in Nsp3 (Table 1). These proteins are known to play various regulatory roles and therefore, mutations at amino acid level can modulate their catalytic activity drastically. Specifically, Nsp3 is the largest and essential component of replication complex in the SARS-CoV-2 genome (Lei et al., 2018) and along with Nsp2 it forms a transcriptional complex in endosome of the infected host cell (Wu et al., 2020). Nsp6 is a multiple-spanning transmembrane protein located into the ER where they induce autophagosomes via an omegosome intermediate (Cottam et al., 2014). Interestingly, the mutation of L3605F causes stiffness in the secondary structure of Nsp6 and leads to low stability of the protein structure in most recent sequences from Asia, America, Oceania and Europe (Benvenuto et al., 2020). Nsp12 and Nsp13 are the key replicative enzymes, which require Nsp6, Nsp7 and Nsp10 as cofactors. Nsp12, RNA dependent RNA polymerase (RdRp) with the presence of the bulkier leucine side chain at location 4618 is likely to create a greater stringency for base pairing to the templating nucleotide, thus modulating polymerase fidelity (Sexton et al., 2016). Nsp13 contains a helicase domain, allowing efficient strand separation of extended regions of double-stranded RNA and DNA (Pachetti et al., 2020). Dual mutations in Nsp13 were reported with profound effect on its activity. P5731L, mutation leads to increased affinity of helicase RNA interaction, whereas Y5768C is a destabilizing mutation increasing the molecular flexibility and leading to decreased affinity of helicase binding with RNA (Begum et al., 2020). Therefore, both the mutations were antagonistic in nature. Thus, ORF1ab polyprotein of SARS-CoV-2 encompasses mutational spectra where signature mutations for Nsp2, Nsp3, Nsp6, Nsp12 and Nsp13 have been predicted. Amino acid mutations in structural proteins S, ORF3a and ORF8 have also been observed with varied frequency of 45, 34 and 89 respectively. The mutation in Spike protein (D7611G) has been reported to outcompete other preexisting subtypes, including the ancestral one. This mutation generates an additional serine protease (Elastase) cleavage site in Spike protein (Bhattacharya et al., 2020) which is

discussed in more details in later sections.ORF3a mutation (Q8327H), is located near TNF receptor associated factor-3 (TRAF-3) regions and has been reported as molecular differences marker in many genomes including Indian SARS- CoV-2 genomes (Hassan et al., 2020) for their delineation. Amino acid change in ORF8 sequence (L9033S) propose that it is preserved (Koyama et al., 2020) therefore it is critical to examine its biological function in SARS-CoV-2.

Our results showed that the mutations (SNPs and AVV) in the virus were not uniformly distributed. Genotyping study annotated few mutations in the SARS-CoV-2 genomes at certain specific locations with high frequency predicting their high selective pressure. Thus, mutations can be predicted as location-specific but not type-specific by SNP count. Highly frequent AAV might be associated with the changes in transmissibility and virulence behavior of the SARS-CoV-2. Therefore, high-frequency AAV mutations in Spike protein, RdRp, helicase and ORF3a are important factors to consider while developing vaccines against the fast-evolving strains of SARS-CoV-2.

Prevalence of Co-mutation in SARS-COV-2 evolution

Interestingly, we observed co-mutations in Nsp13 at locations 5768 (Nsp13_1) and 5731(Nsp13_2) that were prevalent in common 64 genomes, all belonging to USA. The AAV reported above (Table 1) were further analyzed and found occurring in 10 different permutations varying from single to multiple mutated protein combinations. Complete details of these co-mutations combinations are given in Table 2. These co-mutations were mapped over the divergent phylogeny for indicating the evolutionary divergence among the 245 strains. The phylogram (Figure 1B) showed clear divergence of strains from the parent strain due to accumulation of mutations at different level of human to human transmission. We found co-mutations in Nsp3, ORF8, Nsp13, S, Nsp12, Nsp2 and Nsp6 were responsible for the above divergence.

These co-mutations were found linked with lineage **clades a to e,** highlighting their prevalence among them (Figure.1B). In **clade-a,** 40 genomes harbored mutations at only Nsp3 protein while six isolates belonging to USA (MT262993, MT044258, MT159716, MT259248, MT259267) and Pakistan (MT263424) showed no mutation confirming their lineage same as that of the reference/ancestral genome from China. Therefore, Nsp3 marked as first mutational hotspot for accumulating amino acid mutations in SARS-CoV-2. Brazil (MT126808) and USA (MT276331) strains form the descendent from clade-a harbored Nsp3/Nsp6 as first co-mutation. The **clade-b** also had an additional mutation of ORF8 along

with Nsp3 and Nsp6 with three descendant strain from US and China. We observed most distant Chinese strain (MT226610), clustered in **clade-b** and harbored additional 25 AAV making it the highly pathogenic strain in the network as reported above in Figure 1B. The **clade-c** descendant from **clade-a** had a different set of co-mutation with Nsp3-ORF8 proteins. **Clade-d** descended further from **clade-c** had two mutation in Nsp13 (5768/5731) in addition to Nsp3/ORF8 proteins. Two strains from USA in the cluster radiating from **clade-d** harbored additional Nsp6 mutation stating them more divergent with scope of further possible evolution. The next subclade-e1 was found holding another new set of co-mutation of Nsp3/S/Nsp12. Whereas the highest number of co-mutations were found in subclade-e2 with combination of Nsp3/Nsp2/Nsp12/S/ORF3a prevalent in 30 genomes belonging to USA predicting them as active carrier of evolutionary force for SARS-CoV-2 divergence (Figure 3 A & B). Presence of Nsp3 mutation (S1920P) in 238 strains underlined the origin of mutation from reference strain highlighting the first divergence in SARS-CoV-2 strain. In future, more and more genome availability from USA may indicate the evolutionary relationships with these co-mutations. Our result suggested that co-mutations are the major evolutionary force that drives the pathogenicity among the different geographical isolated strains which can responsible for higher and lower order of virulence among them.

The assessment of mutations in SARS-CoV-2 proteins

Amino acid variations were predicted in eight (Nsp2, Nsp3, Nsp6, Nsp12, Nsp13, S, Orf3a, Orf8) SARS-CoV-2 proteins (Table 1). To identify their potential functional role, we carried out the structural analysis of the proteins. Pairwise sequence alignment of wild-type and mutant proteins provided the exact location and changes in amino acids. The GMQE and QMEAN values range from 0.45 to 0.72 and -1.43 to -2.81, respectively. The sequence identity ranges from 34% to 99%, which suggests that the models were constructed with high confidence and best quality (Figure 6). The I-Mutant DDG tool predicts if a mutation can largely destabilize the protein ($\Delta\Delta G<-0.5$ Kcal/mol), largely stabilize ($\Delta\Delta G>0.5$ Kcal/mol) or have a weak effect ($-0.5<=\Delta\Delta G<=0.5$ Kcal/mol). The protein stability analysis showed that all the identified mutations decreased the stability of seven proteins (Nsp2, Nsp6, Nsp12, Nsp13, S, Orf3a, Orf8) except Nsp3 (T1103P) which predicted to increase protein stability (Figure 6 A-H). Further, to explore the role of mutations in SARS-CoV-2 proteins, we carried out HOPE analysis. D614G mutation in S-protein could disturb the rigidity of the protein and due to glycine, hydrophobicity will affect the intra hydrogen bond formation with G594. In ORF8 and Nsp3, the mutation location was not conserved, so it did not affect or damage the protein

function. The mutation (P409L) in Nsp13 was present in the RNA virus helicase C-terminal domain. Since proline is a very rigid amino acid and therefore induce a particular backbone conformation that might be required at this position so this mutation could disturb domain and abolished its function. Mutation L37F (Nsp6) and T85I (Nsp2) were also highly conserved thus could profoundly damage the function of the respective protein. The P227L (Nsp12) mutation was in the RNA binding domain located on the surface of the protein; modification of this residue could disturb interactions with other molecules or other parts of the protein. Conclusively, Nsp3 mutation which appeared in all co-mutation combinations, contributed in increased protein stability among 238 strains could be assigned to their increased pathogenicity. Thus, we attempted to highlight the effects of these mutations in host pathogen interactions.

Modelling of Host-Pathogen Interaction Network and its Functional Analysis

The HPI Network of SARS-CoV-2 (HPIN-SARS-CoV-2) contained 58 edges, 56 nodes, including 5 viral and 51 host proteins (Figure 5). Number of degree (the number of edges per node) calculated based on HPI. The significant existence of few main gene hubs, namely, N, S and M in the network and the attraction of a large number of low-degree nodes toward each hub show strong evidence of controlling the topological properties of the network by these few hubs. N has 37 degrees, S, and M has 16 and 8 degrees, respectively. These viral proteins are the main hubs in the network, which regulate the network. Based on degree distribution, the viral protein N showed highest pathogenicity followed by S and M. N is a highly conserved major structural component of SARS-CoV virion involved in pathogenesis and used as a marker for diagnostic assays (Xia et al., 2020). Another structural protein S (spike glycoprotein), attach the virion to the cell membrane by interacting with host receptor, initiating the infection (Belouzard et al., 2012). The M protein, component of the viral envelope played a central role in virus morphogenesis and assembly via its interactions with other viral proteins (Garoff et al.,1998). Interestingly, we found four host proteins MYO5A, MYO5B, MYO5C and T had a maximum interaction with viral hub proteins. MYO5A, MYO5B, MYO5C interacting with all three (N, S and M) whereas T with two (S and M) viral hub proteins, showed a significant relationship with persistent infections caused by the SARS-CoV-2.

MYO5A, MYO5B and MYO5C proteins are Class V myosin (myosin-5) molecular motor that functions as an organelle transporter (Roland et al., 2011) (Sasaki, et al., 1995). It was found that the presence of myosin protein played a crucial role in coronavirus assembly and budding

in the infected cells (Neuman et al., 2008). These cytoskeletal proteins are of importance during internalization and subsequent intracellular transport of viral proteins. As we know at the entry level of virus, S interacts with host ACE2 receptor that internalizes the virus into the endosomes of the host cell inducing conformational changes in the S glycoprotein (Belouzard et al., 2012). It was found that inhibition of MYO5A, MYO5B, and MYO5C was efficient in blocking the internalization pathway, thus this target can be used for the development of a new treatment for SARS-CoV-2 (Dewerchin et al., 2014). Patients suffering from COVID-19 undergo two major condition in the severe stage, thrombotic phenomenon and hypoxia, that are acting as silent killers (Bikdeli et al., 2020; Negri et al., 2020 https://doi.org/10.1101/2020.04.15.20067017). Hypoxia, condition where oxygen level of the body reduces drastically results in the elevated expression of T protein in the body (Shao et al., 2015; Yoon et al., 2006). T protein (Brachyury/TBXT) is transcription factor involved in regulating genes required for mesoderm formation and differentiation thus playing an important role in pathogenesis. The detailed functional analysis of HPIN-SARS-CoV-2 was mapped on the radiological findings from the COVID-19 severely infected patients and non-survivors. It was reported that the levels of fibrin-degrading proteins, fibrinogen and D-dimer protein were 3-4 folds higher as compared to healthy individual. Therefore, reflecting coagulation activation from infection/sepsis, cytokine storm and impending multiple organs failure (Tang et al., 2020; Shi et al., 2020; Han et al., 2020, Li et al., 2020). In our network, we found 24 proteins (ANGPTL1, TNN, FGL2, ANGPTL6, TNC, FCN3, FCN2, ANGPTL4, FGB, FGA, ANGPT2, ANGPTL5, FGG, TNR, ANGPTL3, FCN1, FIBCD1, ANGPTL2, ANGPTL7, ANGPT4, MFAP4, FGL1, TNXB and ANGPT1) are associated with the above etiology (Figure 5 C). We also found the interaction of SMAD family proteins and SUMO1 with N protein, which may lead to inhibition of apoptosis of infected lung cells (Zhao et al, 2008). The interactome study reveals a significant role of identified host proteins in viral budding and related symptoms of COVID-19.

The mutation in SARS-CoV-2 proteins inhibit viral penetration into host

In order to validate the effect of phenotypic variation (AAV), significant host proteins interactions from HPIN-SARS-CoV-2 were considered for *in silico* docking studies. Docking of S-Protein (wild type and mutant) with ACE2, TMPRSS2 and one of myosin proteins (MYO5C) were analyzed. Recent studies have shown that SARS-CoV-2 uses the ACE2 for entry and the serine protease TMPRSS2 for S protein priming (Wrapp et al., 2020). The poly-proteins (Nsp12, Nsp13, Nsp2, Nsp3 and Nsp6) of ORF1A and ORF1AB were docked with

RPS6 and ATP6V1G1 host proteins. The docking results showed that mutant S-protein could not bind efficiently with ACE2 and MYO5C, whereas mutation slightly promotes the binding with TMPRSS2 (Table 3, Figure 6 and Figure 5C). TMPRSS2 have been detected in both nasal and bronchial epithelium by immunohistochemistry (Bertram et. al., 2012), reported to occur largely in alveolar epithelial type II cells which are central to SARS-CoV-2 pathogenesis (Furong et al., 2020). The wild-type S-protein form 16 hydrogen bonds and 1058 non-bonded contacts with ACE2; whereas the mutant protein forms 12 hydrogen bond and 738 non-bonded contacts (Figure 6). This result suggests that D614G mutation in S-protein could affect viral entry into the host. Similarly, mutations present in the Nsp12, Nsp13, Nsp2, Nsp3 and Nsp6 of SARS-CoV-2 could inhibit the interaction with RPS6, but these mutations promote the binding with ATP6V1G1 expect Nsp6 (L3605F). The RPS6 contributes to control cell growth and proliferation (Chauvin et al., 2014), so a loss of interaction with RPS6 could probably inhibit the production of viruses. Overall structural and interactome analyses suggests that identified mutations (Nsp2 (T265I), Nsp3 (S1920P), Nsp6 (L3605F), Nsp12 (P4618L), Nsp13 (P5731L, Y5768C), S (D7611G)) in SARS-CoV-2 might play an important role in modifying the efficacy of viral entry and its pathogenesis. However, these observations required critical revaluation as well as experimental work to confirm the *in-silico* results.

Regulation of SARS-CoV-2 pathogenicity by CpG island

The genotyping analysis that we performed showed high frequency rate (45) of SNP at 5'UTR region (Table 1) and recent study also suggested that suppression of GC content could play a vital role in specific antiviral activities (Xia, 2020). As seen in SNP analysis, the common transitions of C>T and G>A alter the GC content of the SARS-CoV-2 (Table 1). This directed the analysis towards understanding the role of CpG island which is involved in silencing of transcription and down regulation of viral replication (Vivekanandan et al., 2010). Viral infections upregulate host DNA methyltransferase genes (DNMTs), and their overexpression leads to methylation of host CpG islands along with the viral CpGs (Vivekanandan et al., 2010). Since increased frequency of CpG motifs can serve as Pathogen-associated molecular pattern (PAMP) or Damage-associated molecular pattern (DAMP) which are potent inducers of strong innate immune responses (Barber, 2011; Frieman, 2008). Thus, CpG island profiling and their importance of existence in SARS-CoV-2 genomes was proceeded. We found that CpG islands were consistently present in two regions of the genome at the positions 285-385 nucleotides (101 bp) and 28,324-28,425 nucleotides (102 bp). The results were consistent in

all 245 genomes analyzed in the present study with 100% conservancy in 237 genome sequences (Figure 7 A).

In the remaining 8 genomes, five genomes (MT246474.1 (G to A substitution at 354th position with respect to reference genome); MT276329.1, MT276330.1 and MT276598.1 (C to T substitution at 313th position) and MT246455.1 (G to T substitution at 332nd position)) showed point mutation in 5' CpG island; whereas three genomes (MT159718.1 (C to T substitution at 28409th position); MT159717.1 and MT184911.1 (G to T substitution at 28378th position)) showed point mutation in 3'CpG end (Figure 7 D). Interestingly, all these sequences belong to USA. On further locating CpG island positions with respect to proteins, it was found that these two CpG islands were located at two prime locations within the genome, one in Nsp1 (Figure 7 B), and another within Nucleocapsid (N) protein (Figure 7 C). Previously, it was reported that both the proteins interacted with 5' UTR region playing crucial roles in viral replication and gene expressions (Guan et al., 2012; Yang and Leibowitz, 2015; Galan et al., 2005). Most pivotal role of N protein revolves around encapsulation of viral gRNA which leads to formation of ribonucleoprotein complex (RNP), which is a vital step in assembly of viral particles (Cong et al., 2017).

Nsp1 protein in coronaviruses plays a regulatory role in transcription and viral replication (Cong et al., 2017). It is known to interact with 5' UTR of host cell mRNA to induce its endonucleolytic cleavage (Huang et al., 2011; Narayanan et al., 2015), thus inhibiting host gene expression (Kamitani et al., 2009). It also plays an important role in blocking IFN-dependent antiviral signaling pathways leading to dysregulation of host immune system (Kamitani et al., 2006; Wathelet et al., 2007; Law et al., 2007). CpG sites can be targeted by Zinc Finger Antiviral Proteins which can mediate antiviral restriction through CpG motif detection (Bick et al, 2003; Liu et al., 2015; Chiu et al., 2018). Apart from this, CpG oligodeoxynucleotides (ODNs) are known to act as adjuvants and are already established as a potent stimulator for host immune system (Campbell, 2017; Becker, 2005; Yuan, 2017; Singh et al., 2016; Yu et al., 2018). Moreover, recent studies conducted on influenza A genome and Zika virus genome has shown that by increasing the CpG dinucleotides in viral genome, impairment of viral infection is observed (Gaunt et al., 2016; Trus et al., 2020). Our result showed that the presence of conserved CpG islands in Nps1 and N protein across 245 genomes of SARS-CoV-2 indicated their role in pathogenesis and can be targeted by Zinc Finger Antiviral Proteins or exploited to design CpG-recoded vaccines.

Conclusions

The genomic and proteomic survey of 245 SARS-CoV-2 strains reported from subset of population of different countries reflected global transmission during the outbreak of COVID-19. The viral phylogenetic network with five clads (a-e) provided a landscape of the current stage of epidemic where major divergence was observed in USA strains. From this we propose genotypes linked to geographic clads in which signature SNP can be used to track and monitor the epidemic. Demarcation of co-mutation in the SARS-CoV-2 strains by assessing co-mutations also highlighted the evolutionary relationships among the viral proteins. Our results suggested that co-mutation are indicative of AAV based induced pathogenicity leading to multiple mutations embedded in few genomes. However, co-mutations are still in evolutionary process and more combinations can be predicted with a large dataset. High-frequency AAV mutations were present in the critical proteins, including the Nsp2, Nsp3, Nsp6, Nsp12, Nsp13, S, Orf3a, Orf8 which could be considered for designing a vaccine. Comparative analysis of proteins from wild and mutated strains showed positive selection of mutation in Nsp3 but not in rest of the mutants. HPI model can be used as the fundamental basis for structure-guided pathogenesis process inside host cell. The interactome study showed MYO-5 proteins as a key host partner and also highlighted the key role of N, S ad M viral proteins for conferring SARS-CoV-2 pathogenicity. The mutation in the S protein could affects the viral entry by loose binding with ACE2. The presence of CpG islands in N and Nsp1 protein could play a critical role in pathogenesis regulation. Based on our multi-omics approach: genomics, proteomics, interactomics and structural biology; provided an opportunity for better understanding of COVID-19 pandemic and can be considered in ongoing vaccine development programs.

Authors Contribution

RL, VG, SH, MV conceived and designed the study. VG, HV, SH, NS, KP, MZM performed the analysis and develop figures. VG, SH, MV, KP wrote the manuscript and RL, RK, HV, US, PH, SS help in shaping manuscript.

Conflict of Interest

Authors declare no conflict of interest

Acknowledgements

VG acknowledge Phixgen Pvt. Ltd. for research fellowship. MV, SS acknowledge Dr. P. Hemalatha Reddy, Principal, Sri Venkateswara College, University of Delhi for her constant support and encouragement. RL and US also acknowledge The National Academy of Sciences, India, for support under the NASI-Senior Scientist Platinum Jubilee Fellowship Scheme. NS acknowledge Council of Scientific and Industrial Research (CSIR), New Delhi for doctoral fellowships. KP thanks Hub of Bioinformatics for providing support. SH would like to thank Jaypee Institute of Information Technology, Noida India for providing support. HV would like to thank Ramjas College, University of Delhi, Delhi for providing support. RK acknowledges Magadh University, Bodh Gaya for providing support. MZM acknowledge Department of Health Welfare, Government of India under young scientist scheme for financial support. PH would like to thank Maitreyi College, University of Delhi, Delhi for providing support.

References

- Abduljalil, J. M. & Abduljalil, B. M. Epidemiology, genome, and clinical features of the pandemic SARS-CoV-2: a recent view. *New Microbes New Infect* **35**, 100672-100672, doi:10.1016/j.nmni.2020.100672 (2020).
- Ammari, M. G., Gresham, C. R., McCarthy, F. M. & Nanduri, B. HPIDB 2.0: a curated database for host–pathogen interactions. *Database* **2016**, doi:10.1093/database/baw103 (2016).
- Assenov, Y., Ramírez, F., Schelhorn, S.-E., Lengauer, T. & Albrecht, M. Computing topological parameters of biological networks. *Bioinformatics* **24**, 282-284, doi:10.1093/bioinformatics/btm554 %J Bioinformatics (2007).
- Astuti, I. & Ysrafil. Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2): An overview of viral structure and host response. *Diabetes Metab Syndr* **14**, 407-412, doi:10.1016/j.dsx.2020.04.020 (2020).
- Barber, G. N. Cytoplasmic DNA innate immune pathways. **243**, 99-108, doi:10.1111/j.1600-065X.2011.01051.x (2011).

- Becker, Y. CpG ODNs treatments of HIV-1 infected patients may cause the decline of transmission in high risk populations - a review, hypothesis and implications. *Virus Genes* **30**, 251-266, doi:10.1007/s11262-004-5632-2 (2005).
- Begum, F., Banerjee, A. K., Tripathi, P. P. & Ray, U. Two mutations P/L and Y/C in SARS-CoV-2 helicase domain exist together and influence helicase RNA binding. *bioRxiv*, 2020.2005.2014.095224, doi:10.1101/2020.05.14.095224 (2020).
- Begum, F. *et al.* Analyses of spike protein from first deposited sequences of SARS-CoV2 from West Bengal, India. *bioRxiv*, 2020.2004.2028.066985, doi:10.1101/2020.04.28.066985 (2020).
- Belouzard, S., Millet, J. K., Licitra, B. N. & Whittaker, G. R. Mechanisms of Coronavirus Cell Entry Mediated by the Viral Spike Protein. **4**, 1011-1033 (2012).
- Benvenuto, D. *et al.* Evolutionary analysis of SARS-CoV-2: how mutation of Non-Structural Protein 6 (NSP6) could affect viral autophagy. *Journal of Infection*, doi:https://doi.org/10.1016/j.jinf.2020.03.058 (2020).
- Bertram, S. *et al.* Influenza and SARS-coronavirus activating proteases TMPRSS2 and HAT are expressed at multiple sites in human respiratory and gastrointestinal tracts. *PLoS One* **7**, e35876-e35876, doi:10.1371/journal.pone.0035876 (2012).
- Bhattacharyya, C. *et al.* Global Spread of SARS-CoV-2 Subtype with Spike Protein Mutation D614G is Shaped by Human Genomic Variations that Regulate Expression of TMPRSS2 and MX1 Genes. *bioRxiv*, 2020.2005.2004.075911, doi:10.1101/2020.05.04.075911 (2020).
- Bick, M. J. *et al.* Expression of the Zinc-Finger Antiviral Protein Inhibits Alphavirus Replication. *Journal of Virology* **77**, 11555, doi:10.1128/JVI.77.21.11555-11562.2003 (2003).
- Bikdeli, B. *et al.* COVID-19 and Thrombotic or Thromboembolic Disease: Implications for Prevention, Antithrombotic Therapy, and Follow-up. *Journal of the American College of Cardiology*, 27284, doi:10.1016/j.jacc.2020.04.031 (2020).
- Campbell, J. D. Development of the CpG Adjuvant 1018: A Case Study. *Methods Mol Biol* **1494**, 15-27, doi:10.1007/978-1-4939-6445-1_2 (2017).
- Capriotti, E., Calabrese, R. & Casadio, R. Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. *Bioinformatics* **22**, 2729-2734, doi:10.1093/bioinformatics/btl423 %J Bioinformatics (2006).

- Chauvin, C. *et al.* Ribosomal protein S6 kinase activity controls the ribosome biogenesis transcriptional program. *Oncogene* **33**, 474-483, doi:10.1038/onc.2012.606 (2014).
- Cheng, S.-C. *et al.* First case of Coronavirus Disease 2019 (COVID-19) pneumonia in Taiwan. *Journal of the Formosan Medical Association* **119**, 747-751, doi:https://doi.org/10.1016/j.jfma.2020.02.007 (2020).
- Chiu, H.-P. *et al.* Inhibition of Japanese encephalitis virus infection by the host zinc-finger antiviral protein. *PLOS Pathogens* **14**, e1007166, doi:10.1371/journal.ppat.1007166 (2018).
- Cong, Y., Kriegenburg, F., de Haan, C. A. M. & Reggiori, F. Coronavirus nucleocapsid proteins assemble constitutively in high molecular oligomers. *Scientific Reports* **7**, 5740, doi:10.1038/s41598-017-06062-w (2017).
- Cong, Y. *et al.* Nucleocapsid Protein Recruitment to Replication-Transcription Complexes Plays a Crucial Role in Coronaviral Life Cycle. *Journal of virology* **94**, e01925-01919, doi:10.1128/JVI.01925-19 (2020).
- Cottam, E. M., Whelband, M. C. & Wileman, T. Coronavirus NSP6 restricts autophagosome expansion. *Autophagy* **10**, 1426-1441, doi:10.4161/auto.29309 (2014).
- Dewerchin, H. L., Desmarets, L. M., Noppe, Y. & Nauwynck, H. J. Myosins 1 and 6, myosin light chain kinase, actin and microtubules cooperate during antibody-mediated internalisation and trafficking of membrane-expressed viral antigens in feline infectious peritonitis virus infected monocytes. *Vet Res* **45**, 17-17, doi:10.1186/1297-9716-45-17 (2014).
- Forster, P., Forster, L., Renfrew, C. & Forster, M. Phylogenetic network analysis of SARS-CoV-2 genomes. *Proceedings of the National Academy of Sciences* **117**, 9241, doi:10.1073/pnas.2004999117 (2020).
- Frieman, M., Heise, M. & Baric, R. SARS coronavirus and innate immunity. *Virus Research* **133**, 101-112, doi:https://doi.org/10.1016/j.virusres.2007.03.015 (2008).
- Galán, C., Enjuanes, L. & Almazán, F. A point mutation within the replicase gene differentially affects coronavirus genome versus minigenome replication. *Journal of virology* **79**, 15016-15026, doi:10.1128/JVI.79.24.15016-15026.2005 (2005).
- Garoff, H., Hewson, R. & Opstelten, D. J. Virus maturation by budding. *Microbiol Mol Biol Rev* **62**, 1171-1190 (1998).

- Gaunt, E. *et al.* Elevation of CpG frequencies in influenza A genome attenuates pathogenicity but enhances host response to infection. *Elife* **5**, e12735-e12735, doi:10.7554/eLife.12735 (2016).
- Guan, B.-J., Su, Y.-P., Wu, H.-Y. & Brian, D. A. Genetic evidence of a long-range RNA-RNA interaction between the genomic 5' untranslated region and the nonstructural protein 1 coding region in murine and bovine coronaviruses. *Journal of virology* **86**, 4631-4643, doi:10.1128/JVI.06265-11 (2012).
- Han, H. *et al.* Prominent changes in blood coagulation of patients with SARS-CoV-2 infection. *Clin Chem Lab Med*, doi:10.1515/cclm-2020-0188 (2020).
- Hassan, S. S. *et al.* On spatial molecular arrangements of SARS-CoV2 genomes of Indian patients. *bioRxiv*, 2020.2005.2001.071985, doi:10.1101/2020.05.01.071985 (2020).
- Hoelzer, K., Shackelton, L. A. & Parrish, C. R. Presence and role of cytosine methylation in DNA viruses of animals. *Nucleic Acids Research* **36**, 2825-2837, doi:10.1093/nar/gkn121 %J Nucleic Acids Research (2008).
- Huang, C. *et al.* SARS Coronavirus nsp1 Protein Induces Template-Dependent Endonucleolytic Cleavage of mRNAs: Viral mRNAs Are Resistant to nsp1-Induced RNA Cleavage. *PLOS Pathogens* **7**, e1002433, doi:10.1371/journal.ppat.1002433 (2011).
- Jacobson, M. P., Friesner, R. A., Xiang, Z. & Honig, B. On the Role of the Crystal Environment in Determining Protein Side-chain Conformations. *Journal of Molecular Biology* **320**, 597-608, doi:https://doi.org/10.1016/S0022-2836(02)00470-9 (2002).
- Jones, P. A. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature Reviews Genetics* **13**, 484-492, doi:10.1038/nrg3230 (2012).
- Kamitani, W., Huang, C., Narayanan, K., Lokugamage, K. G. & Makino, S. A two-pronged strategy to suppress host protein synthesis by SARS coronavirus Nsp1 protein. *Nat Struct Mol Biol* **16**, 1134-1140, doi:10.1038/nsmb.1680 (2009).
- Kamitani, W. *et al.* Severe acute respiratory syndrome coronavirus nsp1 protein suppresses host gene expression by promoting host mRNA degradation. *Proc Natl Acad Sci U S A* **103**, 12885-12890, doi:10.1073/pnas.0603144103 (2006).
- Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. & Sternberg, M. J. E. The Phyre2 web portal for protein modeling, prediction and analysis. *Nature Protocols* **10**, 845-858, doi:10.1038/nprot.2015.053 (2015).

- Khailany, R. A., Safdar, M. & Ozaslan, M. Genomic characterization of a novel SARS-CoV-2. *Gene Reports* **19**, 100682, doi:https://doi.org/10.1016/j.genrep.2020.100682 (2020).
- Koyama, T., Platt, D. & Parida, L. *Variant analysis of COVID-19 genomes*. (2020).
- Krinner, S. *et al.* CpG domains downstream of TSSs promote high levels of gene expression. *Nucleic Acids Research* **42**, 3551-3564, doi:10.1093/nar/gkt1358 %J Nucleic Acids Research (2014).
- Law, A. H. Y., Lee, D. C. W., Cheung, B. K. W., Yim, H. C. H. & Lau, A. S. Y. Role for Nonstructural Protein 1 of Severe Acute Respiratory Syndrome Coronavirus in Chemokine Dysregulation. *Journal of virology* **81**, 2537-2537, doi:10.1128/jvi.02744-06 (2007).
- Lei, J., Kusov, Y. & Hilgenfeld, R. Nsp3 of coronaviruses: Structures and functions of a large multi-domain protein. *Antiviral Res* **149**, 58-74, doi:10.1016/j.antiviral.2017.11.001 (2018).
- Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* **23**, 127-128, doi:10.1093/bioinformatics/btl529 %J Bioinformatics (2006).
- Li, Q. *et al.* Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus-Infected Pneumonia. *The New England journal of medicine* **382**, 1199-1207, doi:10.1056/NEJMoa2001316 (2020).
- Li, T., Lu, H. & Zhang, W. Clinical observation and management of COVID-19 patients. *Emerging Microbes & Infections* **9**, 687-690, doi:10.1080/22221751.2020.1741327 (2020).
- Liu, C.-H., Zhou, L., Chen, G. & Krug, R. M. Battle between influenza A virus and a newly identified antiviral activity of the PARP-containing ZAPL protein. *Proceedings of the National Academy of Sciences*, 201509745, doi:10.1073/pnas.1509745112 (2015).
- Mandal, S., Singh, R.S., Sharma, S.K., Malik, M.Z. and Singh, R.B. Complexity in SARS-CoV-2 genome data: Price theory of mutant isolates. *bioRxiv*, doi.org/10.1101/2020.05.04.077511 (2020).
- Nafis S, Kalaiarasan P, Brojen Singh RK, Husain M, Bamezai RN. Apoptosis regulatory protein-protein interaction demonstrates hierarchical scale-free fractal network. Brief Bioinform. 2015;16(4):675-699. doi:10.1093/bib/bbu036

- Narayanan, K., Ramirez, S. I., Lokugamage, K. G. & Makino, S. Coronavirus nonstructural protein 1: Common and distinct functions in the regulation of host and viral gene expression. *Virus Res* **202**, 89-100, doi:10.1016/j.virusres.2014.11.019 (2015).
- Neuman, B. W. *et al.* Proteomics analysis unravels the functional repertoire of coronavirus nonstructural protein 3. *Journal of virology* **82**, 5279-5294, doi:10.1128/JVI.02631-07 (2008).
- NHS Press Conference, Feb. 4 2020 - National Health Commission (NHC) of the People's Republic of China
- Pachetti, M. *et al.* Emerging SARS-CoV-2 mutation hot spots include a novel RNA-dependent-RNA polymerase variant. *Journal of Translational Medicine* **18**, doi:10.1186/s12967-020-02344-6 (2020).
- Pyrc, K., Berkhout, B. & van der Hoek, L. The Novel Human Coronaviruses NL63 and HKU1. *Journal of Virology* **81**, 3051, doi:10.1128/JVI.01466-06 (2007).
- Qi, F., Qian, S., Zhang, S. & Zhang, Z. Single cell RNA sequencing of 13 human tissues identify cell types and receptors of human coronaviruses. *Biochemical and Biophysical Research Communications* **526**, 135-140, doi:https://doi.org/10.1016/j.bbrc.2020.03.044 (2020).
- Ramachandran, S., Kota, P., Ding, F. & Dokholyan, N. V. Automated minimization of steric clashes in protein structures. *Proteins* **79**, 261-270, doi:10.1002/prot.22879 (2011).
- Roland, J. T. *et al.* Rab GTPase–Myo5B complexes control membrane recycling and epithelial polarization. *Proceedings of the National Academy of Sciences* **108**, 2789, doi:10.1073/pnas.1010754108 (2011).
- Sasaki, H. *et al.* Myosin-actin interaction plays an important role in human immunodeficiency virus type 1 release from host cells. *Proc Natl Acad Sci U S A* **92**, 2026-2030, doi:10.1073/pnas.92.6.2026 (1995).
- Schneidman-Duhovny, D., Inbar, Y., Nussinov, R. & Wolfson, H. J. PatchDock and SymmDock: servers for rigid and symmetric docking. *Nucleic acids research* **33**, W363-W367, doi:10.1093/nar/gki481 (2005).
- Sexton, N. R. *et al.* Homology-Based Identification of a Mutation in the Coronavirus RNA-Dependent RNA Polymerase That Confers Resistance to Multiple Mutagens. *Journal of virology* **90**, 7415-7428, doi:10.1128/JVI.00080-16 (2016).

- Shi, H. *et al.* Radiological findings from 81 patients with COVID-19 pneumonia in Wuhan, China: a descriptive study. *Lancet Infect Dis* **20**, 425-434, doi:10.1016/S1473-3099(20)30086-4 (2020).
- Shiraishi, M., Sekiguchi, A., Oates, A. J., Terry, M. J. & Miyamoto, Y. HOX gene clusters are hotspots of de novo methylation in CpG islands of human lung adenocarcinomas. *Oncogene* **21**, 3659-3662, doi:10.1038/sj.onc.1205453 (2002).
- Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* **7**, 539-539, doi:10.1038/msb.2011.75 (2011).
- Singh, S. M. *et al.* Characterization of Immune Responses to an Inactivated Avian Influenza Virus Vaccine Adjuvanted with Nanoparticles Containing CpG ODN. *Viral Immunology* **29**, 269-275, doi:10.1089/vim.2015.0144 (2016).
- Sharma, S., Singh, I., Haider, S., Malik, M.Z., Ponnusamy, K. and Rai, E.. ACE Homo-dimerization, Human Genomic variants and Interaction of Host Proteins Explain High Population Specific Differences in Outcomes of COVID19. bioRxiv. doi.org/10.1101/2020.04.24.050534 (2020).
- Tang, X. *et al.* On the origin and continuing evolution of SARS-CoV-2. *National Science Review*, doi:10.1093/nsr/nwaa036 (2020).
- Trus, I. *et al.* CpG-Recoding in Zika Virus Genome Causes Host-Age-Dependent Attenuation of Infection With Protection Against Lethal Heterologous Challenge in Mice. **10**, doi:10.3389/fimmu.2019.03077 (2020).
- Venselaar, H., Te Beek, T. A. H., Kuipers, R. K. P., Hekkelman, M. L. & Vriend, G. Protein structure analysis of mutations causing inheritable diseases. An e-Science approach with life scientist friendly interfaces. *BMC Bioinformatics* **11**, 548-548, doi:10.1186/1471-2105-11-548 (2010).
- Vivekanandan, P., Daniel, H. D., Kannangai, R., Martinez-Murillo, F. & Torbenson, M. Hepatitis B virus replication induces methylation of both host and viral DNA. *J Virol* **84**, 4321-4329, doi:10.1128/jvi.02280-09 (2010).
- Waterhouse, A. *et al.* SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic acids research* **46**, W296-W303, doi:10.1093/nar/gky427 (2018).
- Wathelet, M. G., Orr, M., Frieman, M. B. & Baric, R. S. Severe acute respiratory syndrome coronavirus evades antiviral signaling: role of nsp1 and rational design of an

attenuated strain. *Journal of Virology* **81**, 11620-11633, doi:10.1128/jvi.00702-07 (2007).

- Wrapp, D. *et al.* Structural Basis for Potent Neutralization of Betacoronaviruses by Single-Domain Camelid Antibodies. *Cell* **181**, 1004-1015.e1015, doi:https://doi.org/10.1016/j.cell.2020.04.031 (2020).
- Wu, C. *et al.* Analysis of therapeutic targets for SARS-CoV-2 and discovery of potential drugs by computational methods. *Acta Pharmaceutica Sinica B*, doi:https://doi.org/10.1016/j.apsb.2020.02.008 (2020).
- Xia, X. Extreme Genomic CpG Deficiency in SARS-CoV-2 and Evasion of Host Antiviral Defense. *Molecular Biology and Evolution*, doi:10.1093/molbev/msaa094 (2020).
- Yang, D. & Leibowitz, J. L. The structure and functions of coronavirus genomic 3′ and 5′ ends. *Virus Research* **206**, 120-133, doi:https://doi.org/10.1016/j.virusres.2015.02.025 (2015).
- Yoon, D. *et al.* Hypoxia-inducible Factor-1 Deficiency Results in Dysregulated Erythropoiesis Signaling and Iron Homeostasis in Mouse Development. **281**, 25703-25711, doi:10.1074/jbc.M602329200 (2006).
- Yu, C.-H., Qin, Z., Martin-Martinez, F. J. & Buehler, M. J. A Self-Consistent Sonification Method to Translate Amino Acid Sequences into Musical Compositions and Application in Protein Design Using Artificial Intelligence. *ACS Nano* **13**, 7471-7482, doi:10.1021/acsnano.9b02180 (2019).
- Yu, P. *et al.* A CpG oligodeoxynucleotide enhances the immune response to rabies vaccination in mice. *Virol J* **15**, 174-174, doi:10.1186/s12985-018-1089-1 (2018).
- Yuan, F. *et al.* Immunoprotection induced by CpG-ODN/Poly(I:C) combined with recombinant gp90 protein in chickens against reticuloendotheliosis virus infection. *Antiviral Res* **147**, 1-10, doi:https://doi.org/10.1016/j.antiviral.2017.04.019 (2017)

Tables

Table 1: Common SNP and AAV mutations occurring in SARS CoV-2 genomes

CDS	Point Mutation	Position	Frequency		Amino Acid-R	Variant	Position	Frequency

5'UTR	C→ T	67	45				
Nsp2	C→ T	885	29	T	I	265	31
Nsp2	C→ T	2863	44				
Nsp3/PL-PRO	T→ C	5852	238	S	P	1920	238
Nsp4	C→ T	8608	88				
Nsp6	G→ T	10909	21	L	F	3605	21
Nsp8	T→ C	12299	238				
Nsp12 (RdRp)	C→ T	14234	44	P	L	4618	46
Nsp13 (Hel)	C→ T	17573	63	P	L	5731	64
Nsp13 (Hel)	A→ G	17684	63	Y	C	5768	64
Nsp13 (Hel)	C→ T	17886	68				
S	A→ G	23232	43	D	G	7611	45
Orf3a	G→ T	25392	32	Q	H	8327	34
Orf8	T→ C	27973	88	L	S	9033	89

783

784 **Table 2**: Co-mutations combinations and genomic location identified in different proteins of

785 SARS-COV-2

Positions	**Variation(s)**	**(Co)Mutations**	**Mutated protein**	**Descendants**
1920	S>P	Nsp3	1	87
5768/5731/1920/9033	Y>C/P>L/S>P/L>S	Nsp13_1/Nsp13_2/Nsp3/ORF8	4	62
1920/9033	S>P/L>S	Nsp3-ORF8	1	22
4618/7611/8327/1920/265	P>L/D>G/Q>H/S>P/T>I	nsp12/S/ORF3a/Nsp3/Nsp2	5	30
4618/8327/1920/265	P>L/Q>H/S>P/T>I	Nsp12/ORF3a/Nsp3/Nsp2	4	1
4618/7611/8327/1920	P>L/D>G/Q>H/S>P	Nsp12/S/ORF3a/Nsp3	4	3
3605/1920	L>F/S>P	Nsp6/Nsp3	2	16
3605/1920/9033	L>F/S>P/L>S	Nsp6/Nsp3/ORF8	3	3
5768/5731/3605/1920/9033	Y>C/P>L/L>F/S>P/L>S	Nsp13_1/Nsp13_2/Nsp6/Nsp3/ORF8	5	2

4618/7611/1920	P>L/D>G/S>P	Nsp12/S/Nsp3	3	12

Table 3. *In silico* docking analysis of SARS-CoV-2 proteins with Human proteins

SARS CoV-2	Host Protein	Wild type score	Mutant score	Difference*
S-Protein	ACE2	18296	17722	574
S-Protein	TRMPSS2	20284	21180	-896
S-Protein	MYO5C	18538	17390	1148
Nsp13	RPS6	17772	15750	2022
Nsp13	ATP6V1G1	14432	20242	-5810
Nsp12	RPS6	16570	15750	820
Nsp12	ATP6V1G1	17150	20242	-3092
Nsp6	RPS6	19336	17736	1600
Nsp6	ATP6V1G1	17614	16022	1592
Nsp3	RPS6	22888	21866	1022
Nsp3	ATP6V1G1	20760	21070	-310
Nsp2	RPS6	22584	19540	3044
Nsp2	ATP6V1G1	18402	18592	-190

Figure Legends

Figure 1. Phylogenetic network of 245 SARS-CoV-2 genomes. (A) Nucleotide based phylogenetic analysis of SARS-CoV-2 isolates using the Maximum Likelihood method based on the Tamura-Nei model, (B) Amino acid based phylogenomic analysis. Circle areas are proportional to the number of taxa. The map is diverged into 5 major clade (a-e) representing variation in the genomes at amino-acid level. The colored circle represents the country of origin of each isolate.

Figure 2. Distribution of SNP (A, B) and AAV (C, D) mutations of SARS-CoV-2 isolates from the globe. (A) Frequency based plot of 12 possible SNP mutations across 245 genomes, (B) Frequencies of the single SNP mutations with locations on the genome, (C) AAV based mutations across the genomes, (D) Top 9 AAV mutations holding highest frequencies among 245 genomes and their respective positions. The nucleotide and amino-acid positions are based on the reference genome of SARS-CoV-2.

Figure 3. (A) AAV based phylogenetic map of 245 SARS-CoV-2 genomes. Node color represents co-mutational combinations. The formation of each clade is well correlated with the mutational combinations (n=10). (B) Genetic separation among the SARS-CoV-2 strains showing divergent evolution (from ancestral node a to b, c & e) and directive evolution (node c to d) with adjoining daughter nodes represented by #.

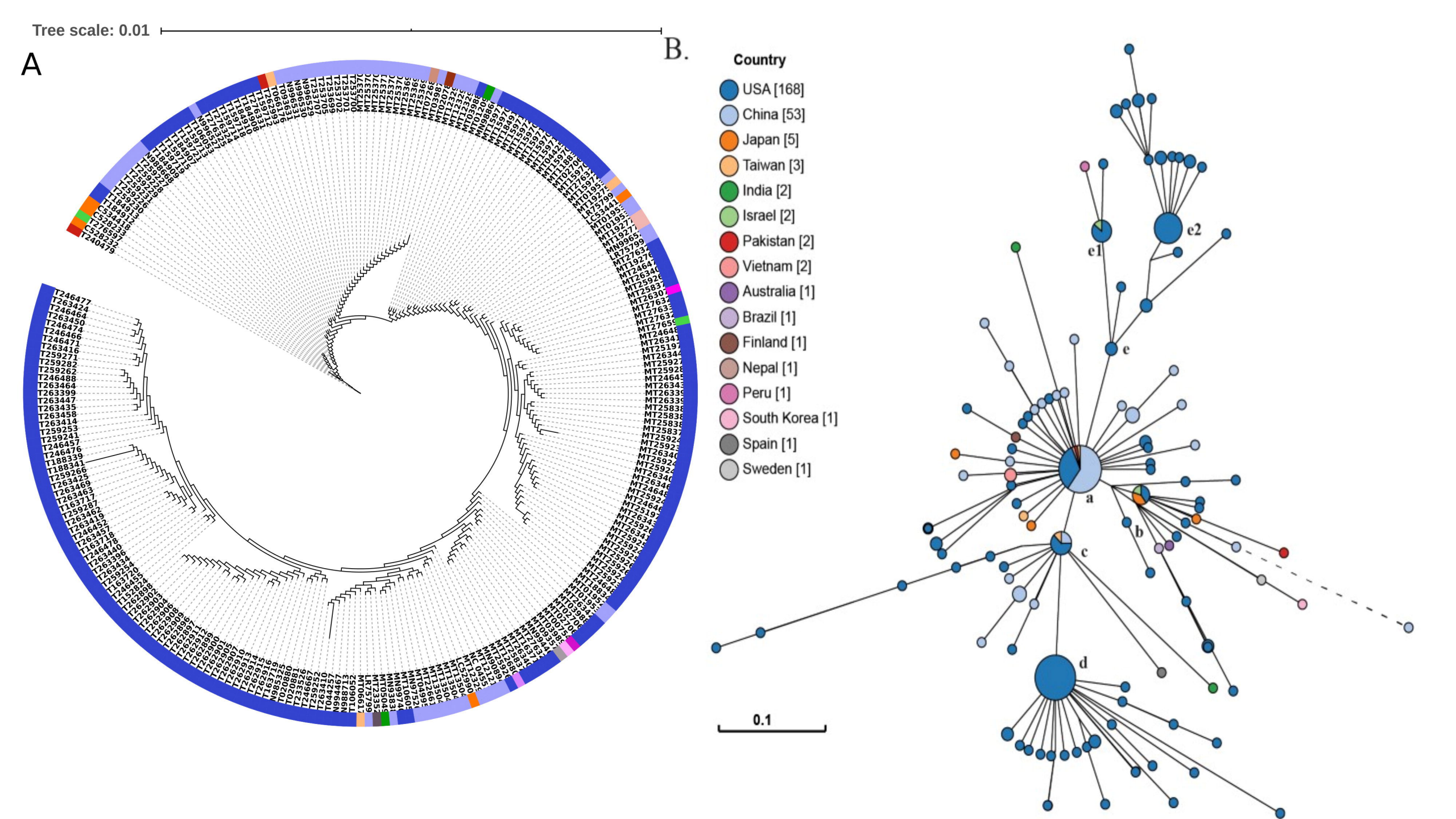
Figure 4. 3-D structure prediction of SARS-CoV-2 proteins harboring mutations at different locations to check for its stability in the cell. Structure are predicted using SwissModel and Phyre2 servers.

Figure 5. (A) Host-pathogenic interaction map of SARS-CoV-2 and human proteins. Red triangles represent viral proteins found to be directly interacting with the human proteins, whereas the pink nodes denote the human proteins. Four major hubs were identified (green) found interacting with maximum viral proteins. (B) Number of degree (the number of edges per node) calculated based on HPI. The significant existence of few main gene hubs, namely, N, S and M in the network and the attraction of a large number of low-degree nodes toward each hub show strong evidence of controlling the topological properties of the network by these few hubs. N has 37 degrees, S, and M has 16 and 8 degrees, respectively. These viral proteins are the main hubs in the network, which regulate the network. (C) Functional enrichment of the human proteins (grey nodes) found to be directly interacting with viral proteins (pink nodes). Color bar around the nodes represent their functionality role in human body.

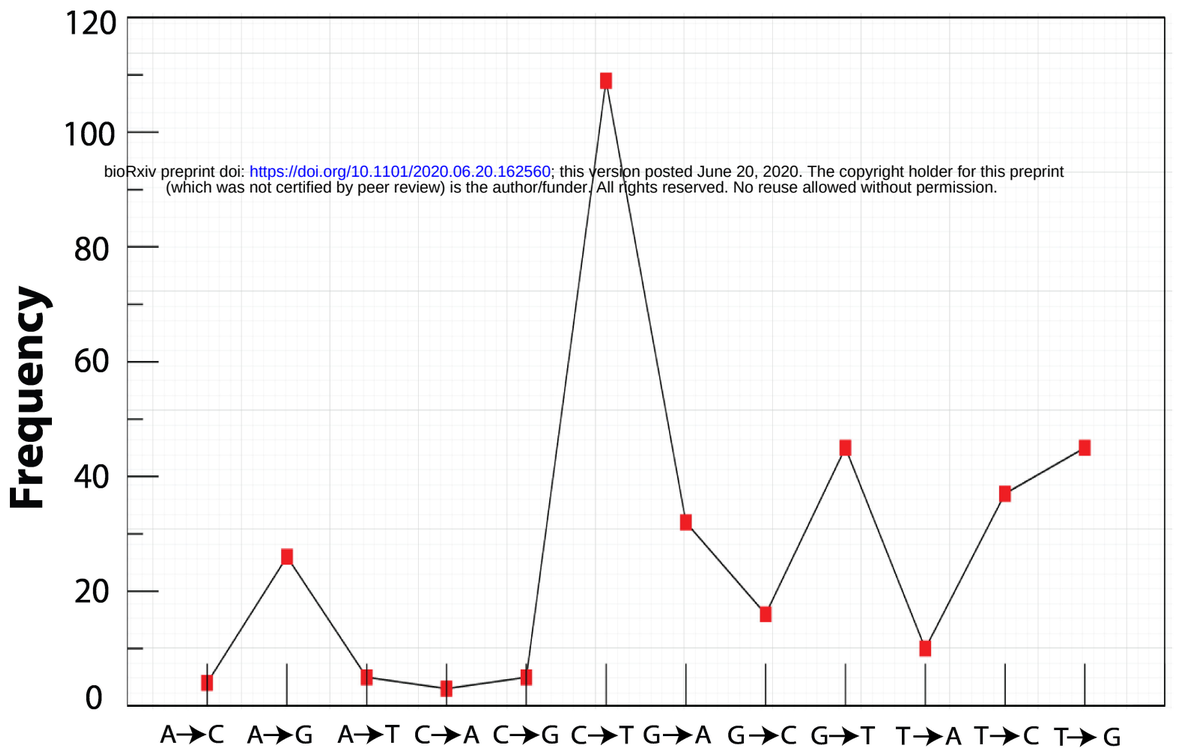
Figure 6. *In-silico* receptor-ligand docking analysis for mutated S protein (D7611G) from SARS-CoV-2 and ACE2 protein present in human. B & C represents amino-acid interactions between wild type and mutated Spike protein with ACE2 receptor.

Figure 7. (A) Detection of two CpG islands in Wuhan_Hu-1 complete genome sequence (Accession number: MT121215.1), marked by blue arrows with their respective positions. (B) One of the CpG island was found to be located towards the 5’ end of the genome, in ORF1ab. (C) Another CpG island was found towards the 3’ end of the genome, located in ORF9 coding for N protein. (D) Five strains of USA showing point mutations in CpG island 1 located on 5’ end of genome (at positions 313, 332 and 354 with respect to reference genome) and Three sequences showing substitutions in CpG island 2 at 28367, 28378 and 28409 positions respectively.

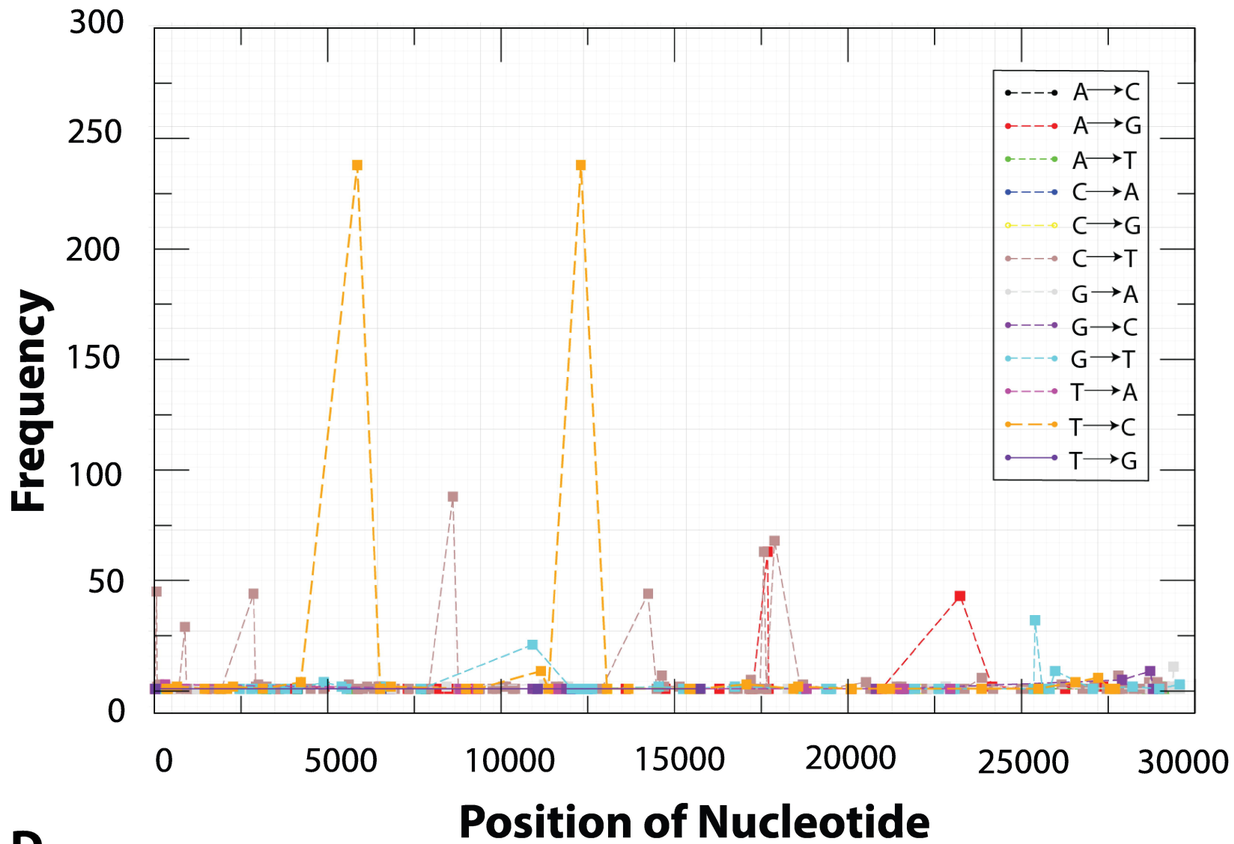
`


Tree scale: 0.01
A
B.
Country
USA [168]
China [53]
Japan [5]
Taiwan [3]
India [2]
Israel [2]
Pakistan [2]
Vietnam [2]
Australia [1]
Brazil [1]
Finland [1]
Nepal [1]
Peru [1]
South Korea [1]
Spain [1]
Sweden [1]
0.1
a
b
c
d
e
e1
e2

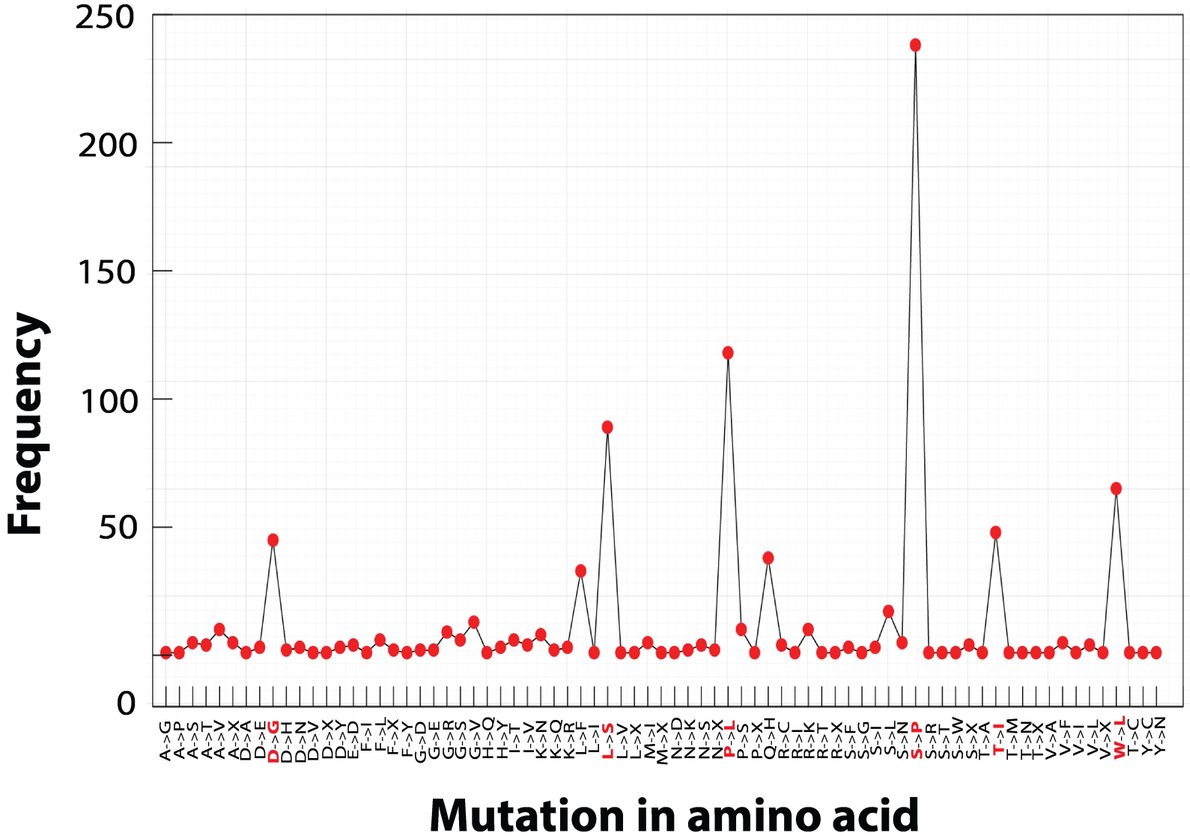
A.



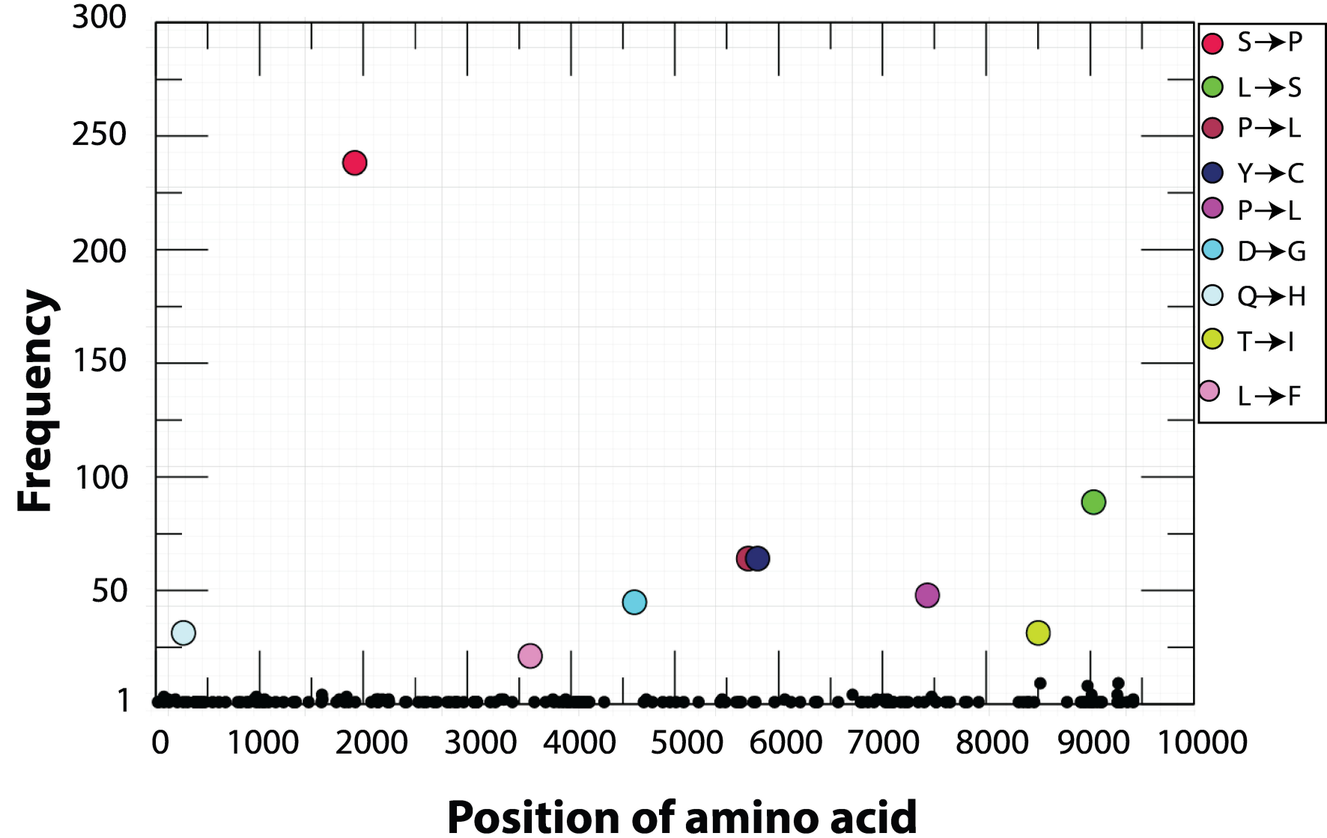
B.



C.

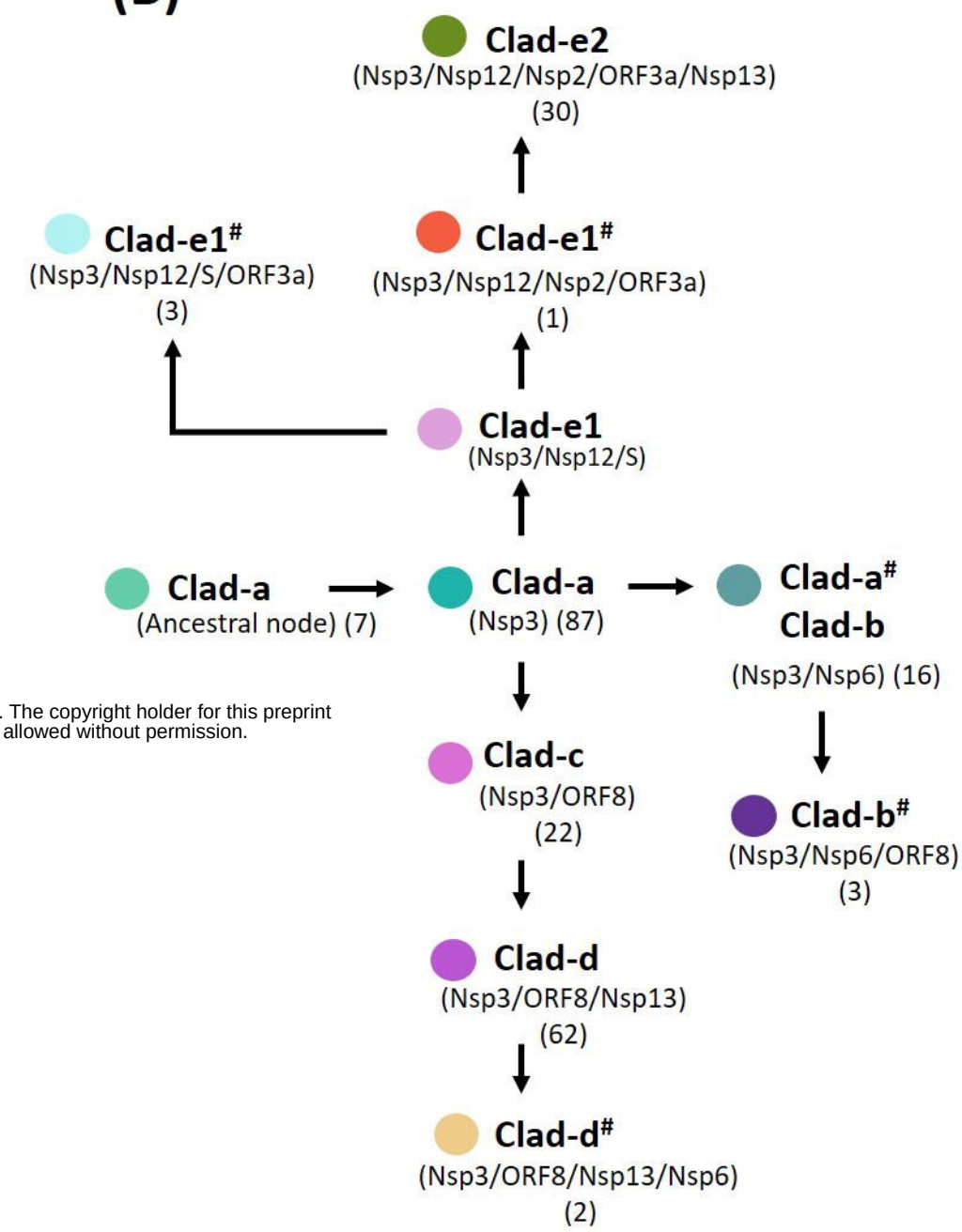


D.



(A)

(B)



#- Radiating daughter node

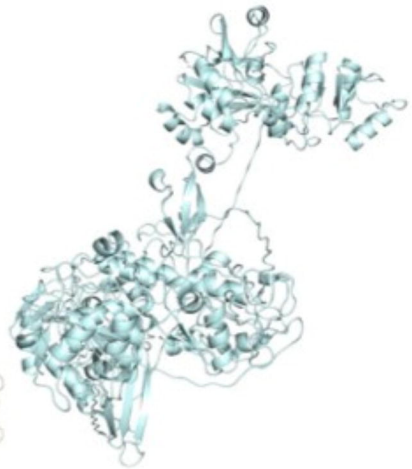
Country Legend

- USA
- China
- Australia
- Brazil
- India
- Japan
- Pakistan
- South Korea
- Finland
- Israel
- Taiwan
- Sweden
- Nepal
- Peru
- Spain
- Vietnam

Country

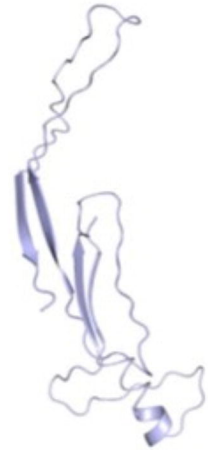
Mutations

S-Protein

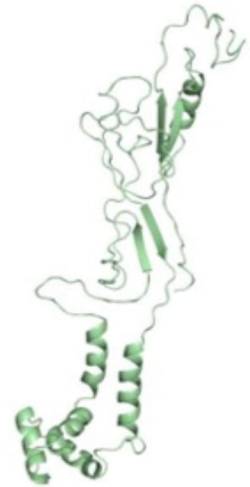


NSP3

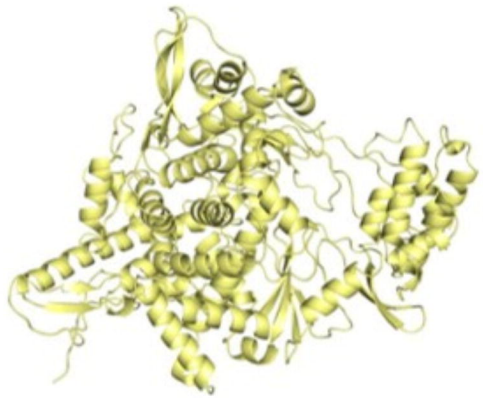
NSP6



ORF8

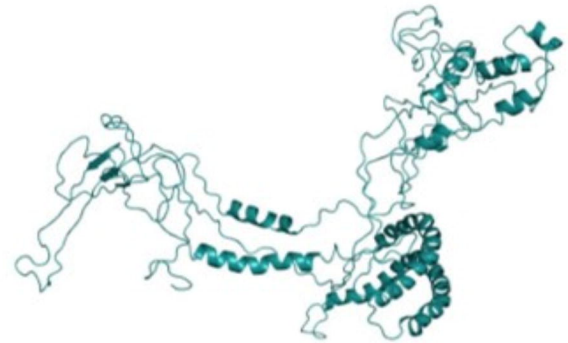


ORF3A

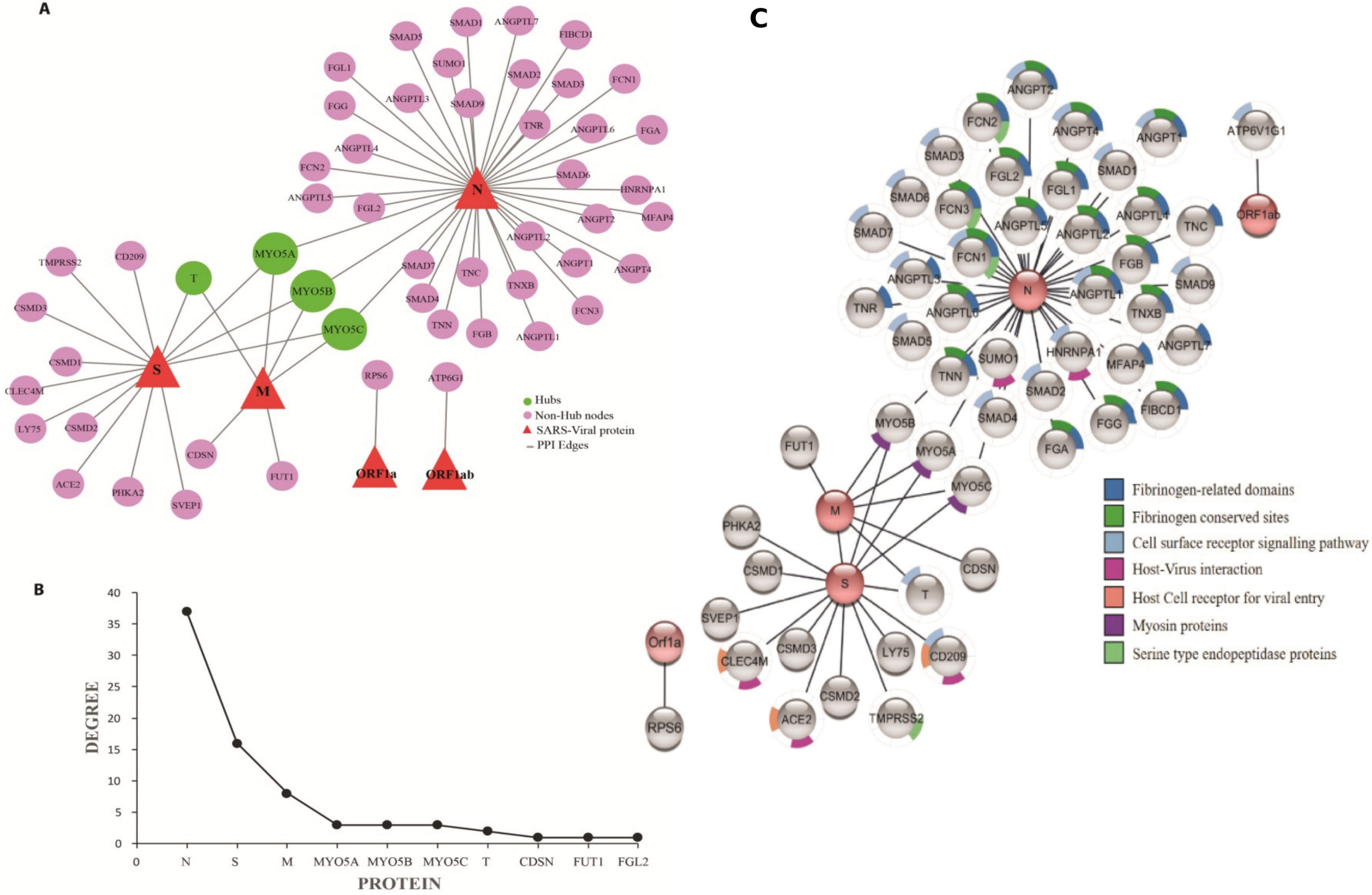


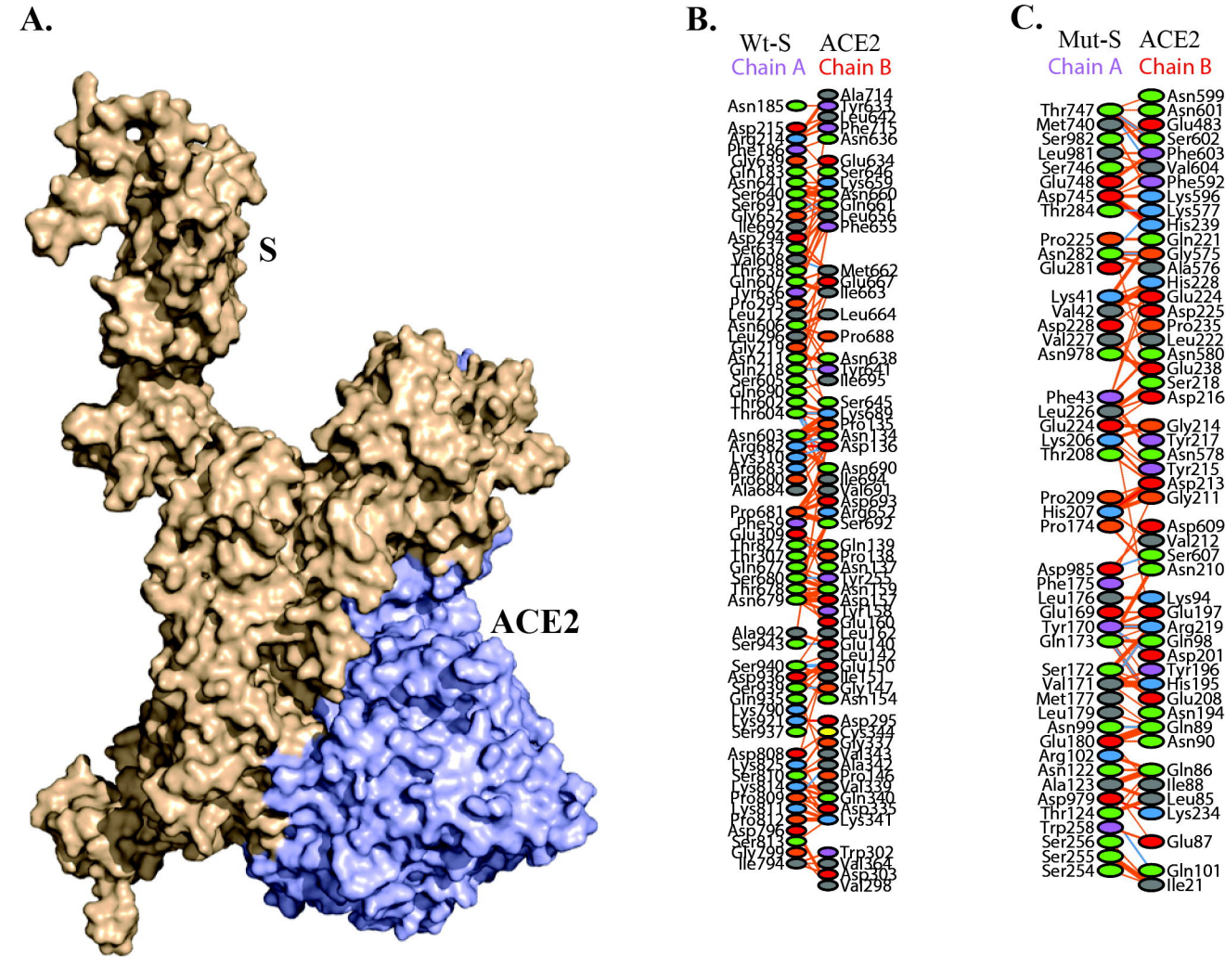
NSP12

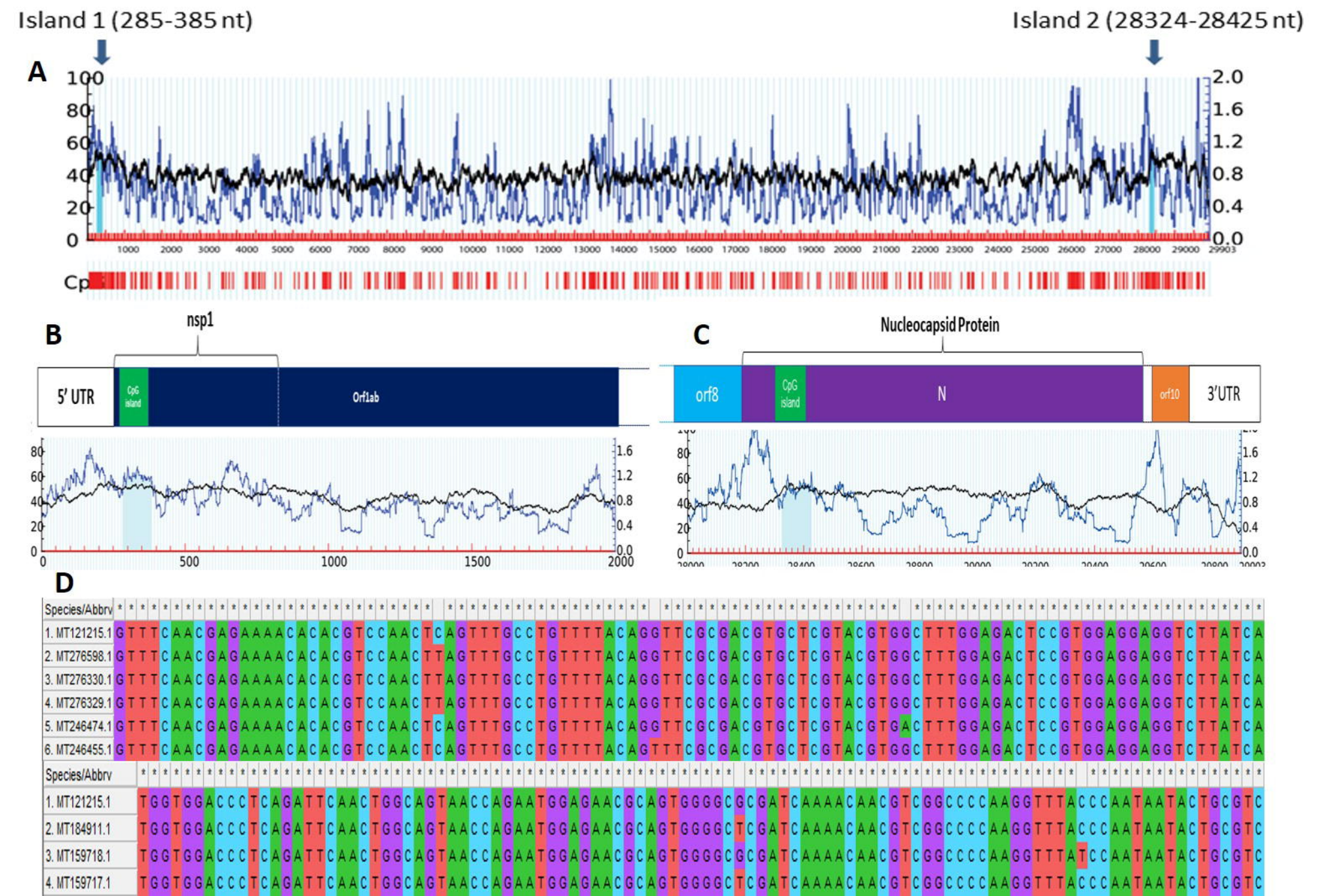
NSP13



NSP2


A
SMAD5
SMAD1
ANGPTL7
FIBCD1
SUMO1
FGL1
SMAD2
SMAD3
FCN1
ANGPTL3
SMAD9
FGG
TNR
ANGPTL6
FGA
ANGPTL4
FCN2
SMAD6
HNRNPA1
ANGPTL5
FGL2
N
ANGPT2
MFAP4
ANGPTL2
TMPRSS2
CD209
MYO5A
SMAD7
TNC
ANGPT1
ANGPT4
T
MYO5B
TNXB
CSMD3
SMAD4
FCN3
MYO5C
TNN
FGB
ANGPTL1
CSMD1
S
M
RPS6
ATP6G1
CLEC4M
LY75
CSMD2
CDSN
ORF1a
ORF1ab
ACE2
PHKA2
FUT1
SVEP1
Hubs
Non-Hub nodes
SARS-Viral protein
PPI Edges
B
DEGREE
40
35
30
25
20
15
10
5
0
0
N
S
M
MYO5A
MYO5B
MYO5C
T
CDSN
FUT1
FGL2
PROTEIN
C
ANGPT2
FCN2
ANGPT4
ANGPT1
ATP6V1G1
SMAD3
FGL2
SMAD1
FGL1
SMAD6
FCN3
ANGPTL5
ANGPTL4
TNC
ORF1ab
SMAD7
ANGPTL2
FCN1
FGB
ANGPTL3
N
ANGPTL1
SMAD9
TNR
ANGPTL6
TNXB
SMAD5
HNRNPA1
ANGPTL7
SUMO1
MFAP4
TNN
SMAD2
FIBCD1
MYO5B
SMAD4
FGG
FUT1
MYO5A
FGA
MYO5C
M
PHKA2
CDSN
CSMD1
S
T
SVEP1
Orf1a
CLEC4M
CSMD3
LY75
CD209
CSMD2
ACE2
TMPRSS2
RPS6
Fibrinogen-related domains
Fibrinogen conserved sites
Cell surface receptor signalling pathway
Host-Virus interaction
Host Cell receptor for viral entry
Myosin proteins
Serine type endopeptidase proteins


A.
S
ACE2
B.
Wt-S
ACE2
Chain A
Chain B
Asn185
Asp215
Arg214
Phe186
Gly639
Gln183
Asn641
Ser640
Ser691
Gly652
Ile692
Asp294
Ser637
Val308
Thr638
Gln607
Tyr636
Pro295
Leu272
Asn606
Leu296
Gly219
Asn211
Gln218
Ser605
Gln690
Thr602
Thr604
Asn603
Arg682
Lys310
Arg683
Pro684
Ala684
Pro681
Phe59
Glu309
Thr827
Thr307
Gln677
Ser680
Thr678
Asn679
Ala942
Ser943
Ser940
Asp936
Ser939
Gln935
Lys790
Lys921
Ser937
Asp808
Lys810
Ser813
Lys814
Pro809
Lys811
Pro812
Asp796
Ser795
Gly799
Ile794
Ala714
Tyr633
Leu642
Phe715
Asn636
Glu634
Ser646
Lys659
Asn660
Gln661
Leu656
Phe655
Met662
Glu667
Ile663
Leu664
Pro688
Asn638
Tyr641
Ile695
Ser645
Lys689
Pro135
Asn134
Asp136
Asn690
Val694
Val691
Asp693
Arg652
Ser692
Gln139
Pro138
Asn137
Tyr255
Asn159
Asp157
Tyr158
Glu160
Leu161
Glu140
Leu142
Glu150
Ile151
Gly147
Asn154
Asp295
Cys344
Gly337
Val343
Ala341
Pro146
Val339
Gln340
Asp335
Lys341
Trp302
Val364
Asp303
Val298
C.
Mut-S
ACE2
Chain A
Chain B
Thr747
Met740
Ser982
Leu981
Ser746
Glu748
Asp745
Thr284
Pro225
Asn282
Glu281
Lys41
Val42
Asp228
Val227
Asn978
Phe43
Leu226
Glu224
Lys206
Thr208
Pro209
His207
Pro174
Asp985
Phe175
Leu176
Glu169
Tyr170
Gln173
Ser172
Val171
Met177
Leu179
Asn99
Glu180
Arg102
Asn122
Ala123
Asp979
Thr124
Trp258
Ser256
Ser255
Ser254
Asn599
Asn601
Glu483
Ser602
Phe603
Val604
Phe592
Lys596
Lys577
His239
Gln221
Gly575
Ala576
His228
Glu224
Asp225
Pro235
Leu222
Asn580
Glu238
Ser218
Asp216
Gly214
Tyr217
Asn578
Tyr215
Asp213
Gly211
Asp609
Val212
Ser607
Asn210
Lys94
Glu197
Arg219
Gln98
Asp201
Tyr196
His195
Glu208
Asn194
Gln89
Asn90
Gln86
Ile88
Leu85
Lys234
Glu87
Gln101
Ile21


Island 1 (285-385 nt)
Island 2 (28324-28425 nt)
A
Cp
B
nsp1
5′ UTR
CpG island
Orf1ab
C
Nucleocapsid Protein
orf8
CpG island
N
orf10
3′UTR
D
Species/Abbrv
1. MT121215.1
2. MT276598.1
3. MT276330.1
4. MT276329.1
5. MT246474.1
6. MT246455.1
Species/Abbrv
1. MT121215.1
2. MT184911.1
3. MT159718.1
4. MT159717.1