Using the nucleocapsid protein to investigate the relationship between SARS-CoV-2 and closely related bat and pangolin coronaviruses

Noah Schuster[1*]

[1]Department of Biology, DePauw University, Greencastle, Indiana

[*]Correspondence: noahschuster_2021@depauw.edu; +1 (219)-336-5808

Abstract

An initial outbreak of coronavirus disease 2019 (COVID-19) in China has resulted in a massive global pandemic causing well over 9,200,000 cases and 470,000 deaths worldwide. The virus responsible, SARS-CoV-2, is known to have a close association with Bat-CoV RaTG13 and Pangolin-CoV MP789. The nucleocapsid protein can serve as a decent model for determining phylogenetic and evolutionary relationships in coronaviruses. Therefore, this present study uses the nucleocapsid gene and protein to further investigate the relationship between SARS-CoV-2 and closely related bat and pangolin coronaviruses. Sequence and phylogenetic analyses have revealed the nucleocapsid gene and protein in SARS-CoV-2 are closely related to those found in Bat-CoV RaTG13 and Pangolin-CoV MP789. Evidence of recombination was detected in the N gene and a double amino acid insertion was observed in the N-terminal region as well. Homology models for the N-Terminal Domain (NTD) revealed similar structure but distinct electrostatic surfaces and topological variations in the β-hairpin that are likely reflective of specific adaptive functions. Collectively, this data helps to strengthen the relationship among SARS-CoV-2, Bat-CoV RaTG13, and Pangolin-CoV MP789 and provide additional insights into the structure and adaptive nature of the nucleocapsid protein. Furthermore, this data will help enhance our understanding of the history behind SARS-CoV-2 and how it developed prior to its emergence.

Introduction

In December 2019, an outbreak of an unknown disease occurred in Wuhan, China.[1] Soon after, this illness was named coronavirus disease 2019 (COVID-19) and defined as a respiratory infection caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), with patients displaying symptoms that manifested as severe cases of pneumonia.[1,2] Initial cases showed a close association with the local animal market in Wuhan, which contained both live and dead animals, suggesting that a spillover event occurred from an unidentified animal source into the human population.[2] Following its emergence in China, SARS-CoV-2 has managed to spread across six continents and result in over 9,100,000 confirmed cases and 470,000 deaths worldwide.[3] As of now, there is still no available vaccine to combat the disease, and while the number of cases and deaths continue to grow, there is a huge need to continue research that studies every aspect of the virus, including its phylogeny and evolution.

Coronaviruses (CoVs) are diverse single-stranded positive-sense RNA viruses with genomes ~30 kilobases in length and belong to the subfamily Coronavirinae in the family Coronaviridae of the order Nidovirales. Furthermore, there are four recognized and distinct genera in Coronavirinae: *Alphacoronavirus*, *Betacoronavirus*, *Deltacoronavirus*, and *Gammacoronavirus*.[4] Additionally, these four genera then split into a variety of distinct subgenera (https://talk.ictvonline.org/ictv-reports/). Officially, SARS-CoV-2 has been ranked within the genus *Betacoronavirus* and the subgenus *Sarbecovirus*, which are the same groups that contain severe acute respiratory syndrome coronavirus (SARS-CoV).[5] An overall 96% genome identity revealed the closest relative to SARS-CoV-2 was Bat-CoV RaTG13, followed by an 88% genome identity with bat-SL-CoVC45 and bat-SL-CoVZXC21.[6] In relation to SARS-CoV, only 79% genome identity was found to indicate a more genetically distant

relationship exists between each virus.[6,7] Recently, CoVs that were isolated from pangolins were found to have a 90–91% genome identity with SARS-CoV-2.[8,9] In the spike protein, the receptor-binding domain (RBD) in SARS-CoV-2 held greater similarity with pangolin CoVs in six vital RBD residues unlike in Bat-CoV RaTG13 which only shares a single residue.[9] Therefore, it has been hypothesized that recombination occurred during the evolution of pangolin CoVs and Bat-CoV RaTG13, contributing to the development of SARS-CoV-2.[8,9,10]

Although phylogeny for SARS-CoV-2 is largely established, the complete evolutionary history for the virus is still in question. A recent investigation that focused on the molecular evolution of SARS-CoV-2 indicated majority of the open-reading frames undergo purifying selection, and there exists little evidence of positive selection.[11] Furthermore, bats are natural reservoirs that possess a variety of CoV species.[12,13] With similarity to Bat-CoV RaTG13 found in horseshoe bats (*Rhinolophus affinis*), SARS-CoV-2 likely has its origins in bats as well. With bats serving as the natural reservoir, the identity of an intermediate host is still largely unresolved. A range of conflicting studies have come out to suggest, but not definitely prove, that snakes, turtles, and/or Malayan pangolins (*Manis javanica*) could potentially be the intermediate host for the virus.[8,9,14,15] Even if one of these animals is the intermediate host, SARS-CoV-2 has yet to be isolated from any of them.[16]

In the remaining one-third of all CoV genomes, there exist several open-reading frames near the 3'-terminus that encode four main structural proteins: the nucleocapsid (N), matrix (M), envelope (E), and spike (S) proteins.[4,17] The N protein is multifunctional and involved in a variety of processes such as genome encapsidation and circularization, transcription, RNA packaging, genome transport to viral budding sites, repressing interferon production, cell cycle inhibition, and blocking RNA interference.[17-20] Overall, the N protein possesses two highly conserved domains:

the N-Terminal Domain (NTD) and C-Terminal Domain (CTD).[20,21] Between these domains is an intrinsically disordered but flexible region known as the Linker Region (LKR) that is rich in serine and arginine residues.[21] In particular, the NTD is divergent in sequence and length; however, it has a β-sheet core along with a highly electropositive groove and β-hairpin extension that binds RNA to neutralize the negative charges found in the phosphate-sugar backbone.[21] For majority of CoVs their NTDs are structurally similar; however, electrostatic surface comparisons reveal very uniquely arranged charge distributions, suggesting that CoV NTDs utilize distinct modes of RNA-binding.[21,22] When interacting with RNA, the NTD functions as scaffolding to assist in genome packaging and virion assembly.[22,23] Since both domains in the N protein are large and well-conserved, this protein functions as a decent model for establishing phylogenetic relationships, calculating divergence times, and determining if any potential recombination events occurred among specific CoVs.[12]

Currently, there is only a limited amount of research focusing exclusively on the N protein in SARS-CoV-2. Herein, this study relies on a bioinformatic and homology modeling approach in which both N gene and protein sequences are used to further investigate the relationship shared between SARS-CoV-2 and a set of closely related bat and pangolin CoVs.

Materials and Methods

1. Sequence Acquisition and Identity Analysis

Nucleocapsid gene and protein sequences in SARS-CoV-2, pangolin CoVs, and bat CoVs were retrieved from the NCBI RefSeq, Nucleotide, and Protein databases. Percent identity was calculated by MatGAT v2.01 to determine the shared identity for each sequence type.[24] In determining values for amino acid identity, a PAM250 scoring matrix was used.

2. Phylogenetic and Recombination Analysis

Sequences were aligned using MUSCLE on the MEGA-X v10.1.7 software.[25,26] The default alignment settings were retained, except clustering methods which were both changed to neighbor joining. In addition, nucleotide sequences were aligned by codons. An individual gene tree and protein tree were constructed using maximum likelihood methods and visualized on MEGA-X.[27] While constructing both trees, uniform rates and partial deletion were used. The heuristic model was nearest-neighbored-interchange and the initial tree was automatically made with the default option NJ/BioNJ.

Based on the protein tree, one sequence was chosen from each split. Using those sequences, nucleotide and amino acid identities for the NTD and CTD were calculated by MatGAT v2.01 to evaluate the identity shared among each domain. In determining values for amino acid identity, a PAM250 scoring matrix was used. Gene sequences were aligned using MUSCLE, after which the alignment was used by SimPlot v3.5.1 to generate a nucleotide similarity plot in detecting for potential recombination events.[28] The similarity plot was computed with a Kimura (2-parameter) distance model and similarity scores were calculated using a 250 bp sliding window and step of 10 bp, along with a 50% consensus threshold.

3. NTD Homology Modeling

Homology models were built using the NTD in SARS-CoV (PDB: 2OFZ) and carried out using the MODELLER v9.23 software.[29,30] Ten models were built for each sequence and the model presenting with the lowest discrete optimized protein energy (DOPE) was chosen. Models were then evaluated using Verify-3D, ERRAT, and PROCHECK on the SAVES v5.0 online sever (https://servicesn.mbi.ucla.edu/SAVES/). Verify-3D measures the compatibility between a model and its own amino acid sequence, calculating scores ≥ 80% to indicate ideal compatibility.[31,32] ERRAT measures the quality of non-bonded interactions between atoms, producing scores ≥ 90 to indicate ideal quality.[33] PROCHECK generates a Ramachandran plot to verify torsion angles and determines if residues are properly built; furthermore, at least ≥ 90% of these residues should fall within the favorable region.[34,35]

To ensure a model had the most stable conformation, regions that presented with either low or moderate scoring were subjected to energy minimization using UCSF Chimera.[36] First, hydrogen atoms were added. Then, charges were assigned using the parameters designated for standard residues in the AMBER ff14SB option. Lastly, the steepest descent and conjugate gradient technique was employed with 100 and 10 steps, respectively. The stereochemical quality of the model was determined again with Verify-3D, ERRAT, and PROCHECK. The model was reliable if it presented with improved values than previously determined by each program.

4. Comparing NTD Models

Once NTD models were constructed, they were superimposed on UCSF Chimera to determine root-mean-square deviation (RMSD) values.[37] Parameters used for superimposition were defined using the default Needleman-Wunsch algorithm and BLOSUM62 scoring matrix. If

the RMSD between two models was considerably lower in relation to any other pair, then that pair was viewed as structurally similar. Structural and topological variations in secondary structures, loops, and/or coils were also considered. Lastly, electrostatic surface potentials were determined by UCSF Chimera and visualized according to columbic surface coloring schemes.

Results

Database searching acquired a total of 10 nucleocapsid gene and protein sequences. Among these sequences, six were from pangolin CoVs, three were from bat CoVs, and the remaining one was from SARS-CoV-2 (Table S1). To ensure clarity and specificity, the isolate name was used to differentiate between the multiple pangolin and bat CoVs in the dataset. Notably, five of the six pangolin CoVs were all isolated between 2017–2018.

The SARS-CoV-2 N gene shares a high sequence identity with Bat-CoV RaTG13 (97%) and Pangolin-CoV MP789 (96%); furthermore, N genes found in pangolin CoVs isolated between 2017–2018, along with bat-SL-CoVZXC21 and bat-SL-CoVZC45, were found to share lower identities at 90–91% (Table S2). Similarly, the SARS-CoV-2 N protein shares a high sequence identity with Bat-CoV RaTG13 (99%) and Pangolin-CoV MP789 (97%); moreover, the N proteins in pangolin CoVs isolated between 2017–2018, along with bat-SL-CoVZXC21 and bat-SL-CoVZC45, were also found to share lower identities at 93–94% (Table S2).

In particular, the alignment revealed that within the N-terminal region of the N protein in SARS-CoV-2, Pangolin-CoV MP789, and Bat-CoV RaTG13, each presented with a double amino acid insertion consisting of residues arginine and asparagine (Figure 1). This insertion was observed in bat-SL-CoVZXC21 and bat-SL-CoVZC45 as well; however, bat-SL-CoVZXC21 had two serine residues while bat-SL-CoVZC45 had arginine and serine residues. Noticeably, all of the pangolin CoVs isolated between 2017–2018 held a conserved group of arginines in the position next to where the double insertion did not occur (Figure 1). For remaining CoVs in the alignment, this same position contained a conserved set of glutamine residues. This region was one of the

more significant differences; however, other conserved amino acid groups for pangolin CoVs isolated between 2017–2018 were observed throughout the alignment.

Phylogenetic analysis suggests the N gene in SARS-CoV-2 is genetically closer to the N gene found in Bat-CoV RaTG13, followed by a more distant relationship to the N gene found in Pangolin-CoV MP789 (Figure 2A). In addition, the N genes found in pangolin CoVs that were isolated between 2017–2018 formed a polytomy while bat-SL-CoVZXC21 and bat-SL-CoVZC45 functioned as sister groups that were distant from all other N genes in the tree (Figure 2A). Phylogenetic analysis of the N protein revealed the same groupings of CoVs as presented in the gene tree; however, the topology is different (Figure 2B). The protein tree shows bat-SL-CoVZXC21 and bat-SL-CoVZC45 in closer relation to SARS-CoV-2, whereas both CoVs in the gene tree held a lesser relationship to SARS-CoV-2. Conversely, the gene tree has pangolin CoVs isolated between 2017–2018 in closer relation to SARS-CoV-2, unlike the protein tree where they have a lesser relation to SARS-CoV-2 (Figure 2).

Based on the phylogenetic positions of each N protein, one representative sequence was selected from each split in the tree for additional analysis. From this, N proteins that belonged to SARS-CoV-2, Bat-CoV RaTG13, Pangolin-CoV MP789, bat-SL-CoVZXC21, and Pangolin-CoV GX-P5L were chosen. It was found that the NTD in SARS-CoV-2 is 100% identical to the NTD in Bat-CoV RaTG13 while the CTD in SARS-CoV-2 is 99% identical to the CTD in Bat-CoV RaTG13 (Table 1). The N protein domains in Pangolin-CoV MP789 shared a very high identity with SARS-CoV-2 at 98% for the NTD and CTD as well (Table 1). In fact, all of the selected N protein sequences expressed high levels of identity for the NTD and CTD with the lowest values being only 94% and 96%, respectively (Table 1). Nucleotide sequences

corresponding to each domain followed a similar trend as well. However, it is interesting that the NTD holds greater nucleotide identity while the CTD holds less (Table 1).

A similarity plot reveals that sequence identities are not homogenous among N genes found within SARS-CoV-2, Bat-CoV RaTG13, Pangolin-CoV MP789, bat-SL-CoVZXC21, and Pangolin-CoV GX-P5L. The complete span of the N gene found in Bat-CoV RaTG13 shares higher nucleotide identity with Pangolin-CoV MP789 and SARS-CoV-2 than with either bat-SL-CoVZXC21 or Pangolin-CoV GX-P5L (Figure 3). However, nucleotides 485–565 found in Pangolin-CoV MP789, bat-SL-CoVZXC21, and Bat-CoV RaTG13 share nearly the same identity to each other and with SARS-CoV-2 (Figure 3). Interestingly, that same region does not possess any shared identity alongside Pangolin-CoV GX-P5L. As with the percent identities, this similarity plot also reveals how the NTD and CTD in Bat-CoV RaTG13 is most similar to the domains in SARS-CoV-2 (Figure 3).

Once homology models were constructed, they were refined and validated to meet appropriate energy and stereochemical parameters (Table S3). Models were then superimposed to yield RMSD values (Figure 4A; Table 2). From this, the NTDs in Pangolin-CoV MP789 and Bat-CoV RaTG13 were structurally similar to the NTD found in SARS-CoV-2; furthermore, the NTDs found in Pangolin-CoV GX-P5L and bat-SL-CoVZXC21 each showed greater similarity towards the NTD found in Pangolin-CoV MP789.

In looking at secondary structure, the β-hairpin (amino acids 46–55) is where a majority of topological diversity can be found. The hairpin found in SARS-CoV-2 extends furthest from the protein core while in Pangolin-CoV GX-P5L the hairpin extends the shortest distance from the core (Figure 4B). Additionally, the hairpin in Bat-CoV RaTG13 appears as an

intermediate that does not extend as far or as short when compared to hairpins in SARS-CoV-2 and Pangolin-CoV GX-P5L (Figure 4B). As for the hairpins in Pangolin-CoV MP789 and bat-SL-CoVZXC21, they are more similar (Figure 4B). Interestingly, β-hairpin structure ultimately varies in topology despite homogeneity in amino identity, except at residue 47 in Pangolin-CoV GX-P5L which is valine instead of isoleucine (Figure S1).

Three additional variations were observed. First, residues 76–78 in Pangolin-CoV MP789 show a relaxed coil; however, it is not clear why this region is different as amino acids in this region are the same when compared to the CoV NTDs that do not present with this feature (Figure 4C; Figure S1). Second, residues 81–82 in Bat-CoV RaTG13 show another relaxed coil. At residue position 81, the amino acid found was aspartic acid and is shared only with SARS-CoV-2 while the other CoV NTDs possess glutamic acid (Figure S1). Even though amino acid identity for this region is shared with SARS-CoV-2, only Bat-CoV RaTG13 displayed this relaxed coil (Figure 4D). Third, in the NTD for Pangolin-CoV MP789 residues 88–91 presented with yet another relaxed coil (Figure 4E). At residue 88, all other CoV NTDs have threonine while Pangolin-CoV MP789 possesses asparagine (Figure S1). Besides a single residue difference, the remainder of this region holds the same amino acid identity.

Electrostatic potential maps revealed similar but still innately visible electrostatic surface patterns for each CoV NTD (Figure S2). Interestingly, residue 33 within the 3/10-helices of the NTD for Pangolin-CoV GX-P5L and bat-SL-CoVZXC21 is lysine (Figure S1). Because of this extra positively charged amino acid, a small region of positive charge is visible on the top portion of their NTDs (Figure S2). This was not observed in any other NTD; however, the likely function and possible advantage of this residue was not investigated further.

Discussion

Based on nucleotide and amino acid identity, the N gene and protein in SARS-CoV-2 share highest identity with the N gene and protein in Bat-CoV RaTG13; although, the N gene and protein found in Pangolin-CoV MP789 share nearly identical values. Each phylogenetic tree has Pangolin-CoV MP789 diverging prior to the SARS-CoV-2 and Bat-CoV RaTG13 sister group. This does not imply SARS-CoV-2 rose directly from either Pangolin-CoV MP789 or Bat-CoV RaTG13, but this is evidence of a close phylogenetic relationship. Additionally, it is unknown whether or not the common ancestors which gave rise to these lineages of CoVs are still in existence. At most, this data is suggestive of an intermingled bat and pangolin origin for SARS-CoV-2; however, the scope of this relationship is largely unclear. Both the N genes and proteins found among pangolin CoVs isolated between 2017–2018, bat-SL-CoVZXC21, and bat-SL-CoVZC45 expressed lower identities suggesting a slightly weaker relation to SARS-CoV-2. Perhaps these groups of viruses served as precursors that contributed to the evolutionary development of the SARS-CoV-2, Bat-CoV RaTG13, and Pangolin-CoV MP789 lineages. Alternatively, there is possibly an unsampled and circulating CoV in nature that is a missing link in the evolutionary history for SARS-CoV-2. Furthermore, the NTD and CTD in the SARS-CoV-2 N protein share highest identity with Bat-CoV RaTG13, followed by close identity with Pangolin-CoV MP789 as well. Slightly higher amino acid conservation than nucleotide conservation was observed for the entire N protein and both domains. Even though the shared nucleotide identities were not extremely lower, this does suggest enough time has passed for silent mutations and/or genetic drift to begin taking minor effects on the N gene.

At the start of the N protein in SARS-CoV-2, Pangolin-CoV MP789, Bat-CoV RaTG13, bat-SL-CoVZXC21, and bat-SL-CoVZC45 present with a double amino acid insertion. Though

naturally occurring mutations are possible, a recombination event could have occurred in these CoVs. Even so, the sequences used in this study provide little evidence as to how this would occur. As previously suggested, perhaps an unsampled and widely circulating CoV in nature is responsible for introducing this type of insertion. Next to this insertion are conserved glutamine residues as well. Altogether, these unique features help reinforce a shared evolutionary history among these CoVs; however, the extent of this history is not fully known. In respect to the entire N protein, the N-terminal region is disordered and functions in RNA-binding.[15] The advantage of this insertion and conserved glutamine residues within the N-terminal region is not clear, but it might help to further promote effective RNA-binding. For example, the insertions found in SARS-CoV-2, Pangolin-CoV MP789, Bat-CoV RaTG13, and bat-SL-CoVZXC21 possess arginine which is a positively charged amino acid that likely interacts with RNA. If so, this could have greater implications on the pathogenicity of CoVs possessing this insertion, such as in SARS-CoV-2. Experimental analysis might be an essential tool in determining the relative importance of this double insertion.

Recombination analysis resulted in minimal evidence that recombination has occurred in the N gene. However, the N genes in Pangolin-CoV MP789 and Bat-CoV RaTG13 might be composed of a small fragment originally found in bat-SL-CoVZXC21. Considering that bat-SL-CoVZXC21 and Bat-CoV RaTG13 circulate among *Rhinolophus affinis*, perhaps a recombination event originated in that species and then made way into *Manis javanica*, resulting in a second recombination event between bat-SL-CoVZXC21 and Pangolin-CoV MP789. With bat-SL-CoVZXC21 serving as the source, this path of recombination is possible given the N gene is more conserved among Pangolin-CoV MP789 and Bat-CoV RaTG13. A recent study of the S gene in Pangolin-CoV MP789 suggested this specific gene was constructed by fragments

found in either bat-SL-CoVZXC21 or bat-SL-CoVZC45, and from Bat-CoV RaTG13.[8] Therefore, evidence of these CoVs having undergone recombination at some point is apparent and that nucleotides 485–565 for the N gene in SARS-CoV-2 was introduced from potential recombination events among Pangolin-CoV MP789, Bat-CoV RaTG13, and bat-SL-CoVZXC21.

Structurally, the NTD in Pangolin-CoV MP789 and Bat-CoV RaTG13 are similar to the NTD in SARS-CoV-2, but not with each other. Among these, the region which displays the greatest variation is the β-hairpin. A recent study elucidated the crystal structure for the NTD in SARS-CoV-2, and researchers discovered that when compared to the NTDs in SARS-CoV, MERS-CoV and HCoV-OC43, the β-hairpin presented structural and topological variations.[38] From this, it is understood these differences likely result in adaptive features that determine how specific CoV N proteins will bind to their own RNA genomes. The varying electrostatic surface potentials help support this concept. With varying distributions of positive charge, the mode of binding for the NTD will vary among N proteins. For the NTDs found in SARS-CoV-2, Pangolin-CoV MP789, and Bat-CoV RaTG13 a structural relationship does exist. However, electrostatic surface maps suggest unique modes of RNA-binding. Given that structures for the NTDs in Bat-CoV RaTG13 and Pangolin-CoV MP789 share greater similarity to the NTD in SARS-CoV-2 than with each other, it is likely these CoVs influenced the evolution leading up to the structural features observed in the SARS-CoV-2 N-NTD and how it binds to the genome. Also, an extended β-hairpin in SARS-CoV-2 might allow for a larger binding pocket that can accommodate RNA possessing higher ordered structure. Discussed by Kang *et al*. 2020, they describe the N-terminal tail in the NTD as flexible and extending outward; therefore, this feature and the β-hairpin might function together in allowing for a binding site that contains RNA of higher ordered structure.[38] This could have further implications on how the genome is packaged when

forming a helical nucleocapsid to make a complete virion. Collectively, variations in β-hairpin topology could be reflective of how an N protein's NTD interacts with a specified RNA genome.

Another recent study that identified epitopes in the SARS-CoV-2 N protein found the N protein in pangolin CoV failed to align with two of those epitope sequences.[39] From this, separations exist between functional adaptation and molecular evolution. There are structural variations between NTDs in SARS-CoV-2 and Pangolin-CoV MP789; despite their N proteins having a close evolutionary history. The structural differences in the NTD are likely reflective of adaptive functions in each N protein; therefore, it is likely there are features inherent only to the SARS-CoV-2 N protein that require further investigation. An example of such a specific feature would be the extended β-hairpin.

In regard to the flexible coils in Pangolin-CoV MP789 and Bat-CoV RaTG13, these features might also be inherent to their respective NTDs. Based on the data used to verify homology models, it seems these features are not errors as each model presented with ideal values for energy and stereochemical quality. For example, only asparagine is found at residue 88 in Pangolin-CoV MP789; therefore, this could affect spatial arrangements of subsequent amino acids and cause a looser coil to form. Though the data is not shown, attempts at correcting these regions were done but resulted in lower stereochemical quality. Subsequently, in order to preserve a valid structure these flexible coils were not tampered with and kept as they were. Experimentally derived crystal structures for the NTD in these CoVs could help corroborate these findings, along with their potential functions.

It is clear CoVs originate in bats and can spillover into the human population through an intermediate host, such as with palm civets transmitting SARS-CoV.[13] Unfortunately, this study

does not provide evidence that pangolins are intermediate hosts for SARS-CoV-2. However, by using the N gene and protein, this study reinforces the existence of a close phylogenetic relationship among SARS-CoV-2, Bat-CoV RaTG13, and Pangolin-CoV MP789. Although the evolutionary history is similar, the N proteins in these CoVs seem to have adaptive functions specific to the CoVs they associate with. It is also evident that NTDs in these N proteins exhibit slight topological variation, primarily in the β-hairpin. Continuing surveillance of CoVs in bats and pangolins can help collect additional sequence information that can make sense of the relationship between these CoVs. Ultimately, this would allow us to better understand the complete history of SARS-CoV-2 and what it could imply for the future should another novel CoV emerge causing a disease more dangerous than COVID-19.

Acknowledgments

I want to thank all the previous scientists who have spent countless amounts of time researching this virus and their dedication to both understanding and fighting this ongoing pandemic. I also give thanks to Nataly Basora Mateo, Sabrina Krause, and Aziza Shemet Pitcher for providing their own encouragement and insights on this project, along with their helpful feedback during the initial construction of the manuscript.

Conflict of Interests

The author has declared that no conflicts of interest exist.

Tables and Figures

	N-Terminal Domain (NTD)		C-Terminal Domain (CTD)	
	Nucleotide (%)	**Amino Acid (%)**	**Nucleotide (%)**	**Amino Acid (%)**
Pangolin-CoV MP789	97.9	97.6	94.6	98.3
Pangolin-CoV GX-P5L	93.9	94.4	90.7	96.6
bat-SL-CoVZXC21	93.9	96.0	88.4	96.6
Bat-CoV RaTG13	98.9	100.0	95.5	99.2

Table 1. Nucleotide and amino acid identities for the NTD and CTD when set-up against SARS-CoV-2.

	1	**2**	**3**	**4**
1. SARS-CoV-2	**-**	**-**	**-**	**-**
2. Bat CoV RaTG13	**0.389**	**-**	**-**	**-**
3. bat-SL-CoVZXC21	**0.356**	**0.504**	**-**	**-**
4. Pangolin-CoV GX-P5L	**0.422**	**0.462**	**0.375**	**-**
5. Pangolin-CoV MP789	**0.329**	**0.429**	**0.200**	**0.289**

Table 2. Matrix displaying the RMSD values shared across 126 amino acids in each NTD homology model. '-' Empty Box.

Species/Abbrv	*	*	*	*	*	*	*					*	*	*	*	*	*	*	*	*	*	*	*
1. SARS-CoV-2	M	S	D	N	G	P	Q	N	Q	R	N	A	P	R	I	T	F	G	G	P	S	D	S
2. Pangolin-CoV_MP789	M	S	D	N	G	P	Q	N	Q	R	N	A	P	R	I	T	F	G	G	P	S	D	S
3. Pangolin-CoV_GX-P5L	M	S	D	N	G	P	Q	N	R	-	-	A	P	R	I	T	F	G	G	P	S	D	S
4. Pangolin-CoV_GX-P2V	M	S	D	N	G	P	Q	X	R	-	-	A	P	R	I	T	F	G	G	P	S	D	S
5. Pangolin-CoV_GX-P5E	M	S	D	N	G	P	Q	N	R	-	-	A	P	R	I	T	F	G	G	P	S	D	S
6. Pangolin-CoV_GX-P1E	M	S	D	N	G	P	Q	N	R	-	-	A	P	R	I	T	F	G	G	P	S	D	S
7. Pangolin-CoV_GX-P4L	M	S	D	N	G	P	Q	N	R	-	-	A	P	R	I	T	F	G	G	P	S	D	S
8. Bat-CoV_RaTG13	M	S	D	N	G	P	Q	N	Q	R	N	A	P	R	I	T	F	G	G	P	S	D	S
9. bat-SL-CoVZXC21	M	S	D	N	G	P	Q	N	Q	S	S	A	P	R	I	T	F	G	G	P	S	D	S
10. bat-SL-CoVZC45	M	S	D	N	G	P	Q	N	Q	R	S	A	P	R	I	T	F	G	G	P	S	D	S

Figure 1. Screen capture of the MUSCLE alignment for the start of the N protein. The red box indicates the insertion region present in SARS-CoV-2, Pangolin-CoV MP789, Bat-CoV RaTG13, bat-SL-CoVZXC21, and bat-SL-CoVZC45. Furthermore, this box shows a conserved set of arginine (R) residues in Pangolin-CoVs isolated between 2017–2018 while remaining CoVs show a conserved set of glutamine (Q) residues.

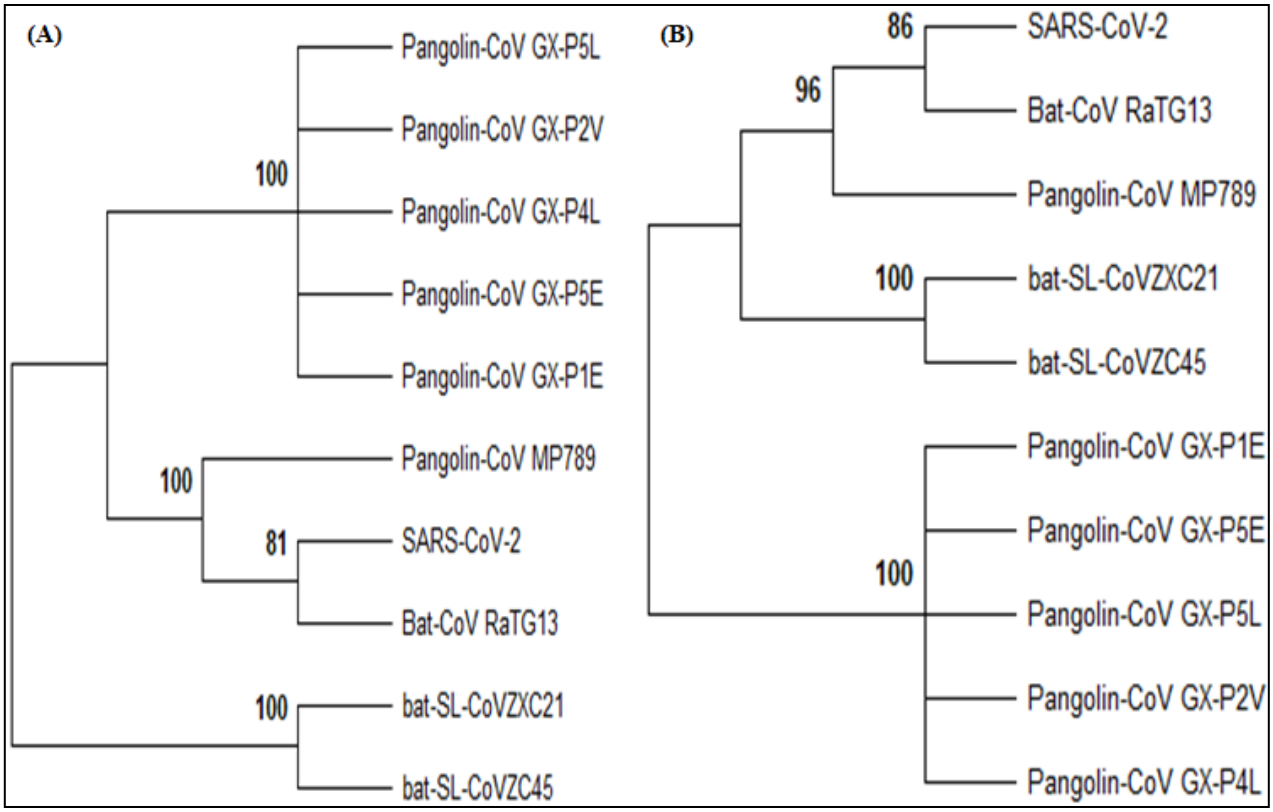

Figure 2. (A) Phylogenetic analysis of N gene sequences depicting evolutionary relationships observed among SARS-CoV-2, Pangolin-CoVs, and closely related Bat-CoVs. The phylogeny was inferred using a Jukes-Cantor substitution model. A log likelihood of -3096.80 was determined with bootstrap values calculated out of 500 replicates and nodes with < 70.00% support collapsed. (B) Phylogenetic analysis of N protein sequences depicting evolutionary relationships observed among SARS-CoV-2, Pangolin-CoVs, and closely related Bat-CoVs. The phylogeny was inferred using a JTT matrix-based substitution model. A log likelihood of -1560.88 was determined with bootstrap values calculated out of 500 replicates and nodes with < 70.00% support collapsed.

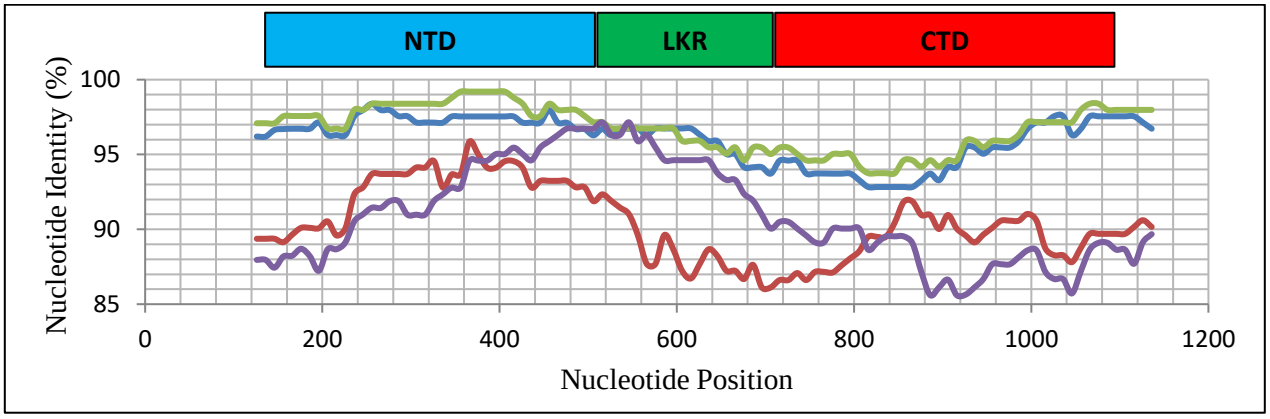

Figure 3. Similarity plot based on the N gene sequence from SARS-CoV-2. Subject sequences, as depicted in the plot, are from Bat-CoV RaTG13 (green), bat-SL-CoVZXC21 (purple), Pangolin-CoV MP789 (blue), and Pangolin-CoV GX-P5L (red). Nucleotide positions 1–125 and 1,160–1,260 are not shown.

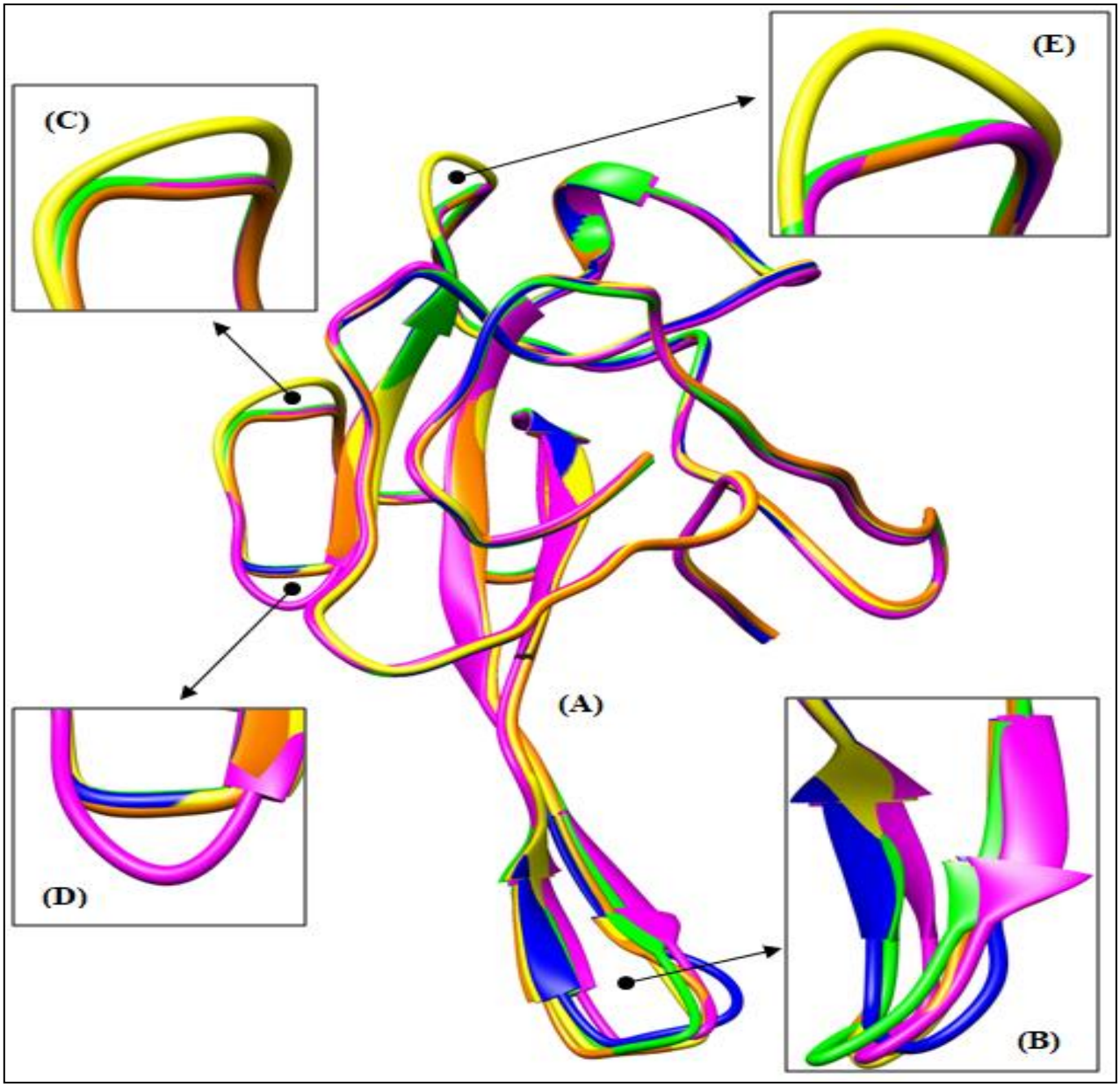

Figure 4. (A) Superimposition of CoV NTDs: SARS-CoV-2 (green), Pangolin-CoV MP789 (yellow), bat-SL-CoVZXC21 (orange), Pangolin-CoV GX-P5L (blue), and Bat-CoV RaTG13 (magenta). (B) The topological variations in the β-hairpin. (C) First flexible coil in Pangolin-CoV MP789. (D) The flexible coil in Bat-CoV RaTG13. (E) Second flexible coil in Pangolin-CoV MP789.

References

[1]He F, Deng Y, and Li W. (2020) "Coronavirus disease 2019: what we know?" *Journal of Medical Virology*. 2020. Pg 1-7.

[2]Brüssow H. (2020) "The Novel Coronavirus – A Snapshot of Current Knowledge." *Microbial Biotechnology*. 13.3. Pg 607-612.

[3]WHO: World Health Organization. (2020) "Coronavirus disease 2019 (COVID-19) Situation Report – 156." Coronavirus disease (COVID-19) situation reports. World Health Organization.

[4]Chen Y, Liu Q, and Guo D. (2020) "Emerging coronaviruses: Genome structure, replication, and pathogenesis." *Journal of Medical Virology*. 92. Pg 418-423.

[5]Gorbalenya AE, Baker SC, Baric RS, *et al.* (2020) "The species *Severe acute respiratory syndrome-related coronavirus*: classifying 2019-nCoV and naming it SARS-CoV-2." Nature Microbiology. 5. Pg 536–544.

[6]Lu R, Zhao X, Li J, *et al*. (2020) "Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding." *The Lancet*. 395. Pg 565-574.

[7]Wu A, Peng Y, Huang B, *et al*. (2020) "Genome Composition and Divergence of the Novel Coronavirus (2019-nCoV) Originating in China." *Cell Host & Microbe*. 27.3. Pg 325–328.

[8]Liu P, Jiang J-Z, Wan X-F, *et al*. (2020) "Are pangolins the intermediate host of the 2019 novel coronavirus (SARS-CoV-2)?" *PLoS Pathogens*. 16.5.

[9]Wong MC, Sara JC, Ajami NJ, *et al*. (2020) "Evidence of recombination in coronaviruses implicating pangolin origins of nCoV-2019." *bioRxiv*.

[10]Andersen KG, Rambaut A, Lipkin WI, *et al.* (2020) "The proximal origin of SARS-CoV-2." *Nature Medicine*. 26. Pg 450–452.

[11]Cagliani R, Forni D, Clerici M, *et al*. (2020) "Computational inference of selection underlying the evolution of the novel coronavirus, SARS-CoV-2." *Journal of Virology*. Pg. 411-420.

[12]Vijaykrishna D, Smith GJ, Zhang JX, *et al*. (2007) "Evolutionary insights into the ecology of coronaviruses. *Journal of Virology*. 81.8. Pg 4012–4020.

[13]Cui J, Li F, and Shi Z. (2019) "Origin and evolution of pathogenic coronaviruses." *Nature Reviews Microbiology*. 17. Pg 181–192.

[14]Ji W, Wang W, Zhao X, *et al*. (2020) "Cross-species transmission of the newly identified coronavirus 2019-nCoV. *Journal of Medical Virology*." 92. Pg 433–440.

[15]Liu Z, Xiao X, Wei X, *et al*. (2020). "Composition and divergence of coronavirus spike proteins and host ACE2 receptors predict potential intermediate hosts of SARS-CoV-2." *Journal of Medical Virology*. 92. Pg 595-601.

[16]Zheng J. (2020). "SARS-CoV-2: an Emerging Coronavirus that Causes a Global Threat." *International Journal of Biological Sciences*. 16.10. Pg 1678–1685.

[17]Tok TT and Tatar G. (2017) "Structures and Functions of Coronavirus Proteins: Molecular Modeling of Viral Nucleoprotein." *International Journal of Virology & Infectious Disease*. 2.1.

[18]Gui M, Liu X, Guo D, *et al*. (2017) "Electron microscopy studies of the coronavirus ribonucleoprotein complex." *Protein & Cell*. 8.3. Pg 219–224.

[19]Lo CY, Tsai TL, Lin CN, *et al*. (2019) "Interaction of coronavirus nucleocapsid protein with the 5'- and 3'-ends of the coronavirus genome is involved in genome circularization and negative-strand RNA synthesis." *The FEBS Journal*. 286.16. Pg 3222-3239.

[20]Neuman BW and Buchmeier MJ. (2016) "Supramolecular Architecture of the Coronavirus Particle." *Advances in Virus Research*. 96. Pg 8–9.

[21]McBride R, van Zyl M, and Fielding BC. (2014) "The coronavirus nucleocapsid is a multifunctional protein." *Viruses*. 6.8. Pg 2991–3018.

[22]Saikatendu KS, Joseph JS, Subramanian V, *et al*. (2007) "Ribonucleocapsid formation of severe acute respiratory syndrome coronavirus through molecular action of the N-terminal domain of N protein." *Journal of Virology*. 81.8. Pg 3913–3921.

[23]Huang Q, Yu L, Petros AM, *et al*. (2004) "Structure of the N-terminal RNA-binding domain of the SARS CoV nucleocapsid protein." *Biochemistry*. 43.20. Pg 6059-6063.

[24]Campanella JJ, Bitincka L, and Smalley J. (2003) "MatGAT: An application that generates similarity/identity matrices using protein or DNA sequences." *BMC Bioinformatics*. 4.29.

[25]Edgar RC. (2004) "MUSCLE: multiple sequence alignment with high accuracy and high throughput." *Nucleic Acids Research*. 32.5. Pg. 1792–1797.

[26]Kumar S, Stecher G, Li M, *et al*. (2018) "MEGA X: Molecular Evolutionary Genetics Analysis across computing platforms." *Molecular Biology and Evolution* 35. Pg 1547-1549.

[27]Whelan S and Goldman N. (2001) "A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach." *Molecular Biology and Evolution.* 18. Pg 691-699.

[28]Lole KS, Bollinger RC, Paranjape RS, *et al*. (1999) "Full-length human immunodeficiency virus type 1 genomes from subtype C-infected seroconverters in India, with evidence of intersubtype recombination." *Journal of Virology*. 73.1. Pg 152-160.

[29]Fiser A and Šali A. (2003) "MODELLER: generation and refinement of homology-based protein structure models." *Methods in Enzymology*. 374. Pg 463-493.

[30]Eswar N, Webb B, Marti-Renom MA, *et al*. (2006) "Comparative protein structure modeling using Modeller." *Current Protocols in Bioinformatics*. 5. Pg 1-47.

[31]Bowie JU, Lüthy R, and Eisenberg D. (1991) "A method to identify protein sequences that fold into a known three-dimensional structure." *Science*. 253.5016. Pg 164–170.

[32]Lüthy R, Bowie JU, and Eisenberg D. (1992) "Assessment of protein models with three-dimensional profiles." *Nature*. 356.6364. Pg 83–85.

[33]Colovos C and Yeates TO. (1993) "Verification of protein structures: patterns of nonbonded atomic interactions." *Protein Science*. 2.9. Pg 1511–1519.

[34]Laskowski RA, Rullmann JAC, MacArthur MW, *et al*. (1996) "AQUA and PROCHECK-NMR: Programs for checking the quality of protein structures solved by NMR." *Journal of Bimolecular NMR*. 8. Pg 477–486.

[35]Laskowski RA, MacArthur MW, Moss DS, *et al*. (1993) "PROCHECK: a program to check the stereochemical quality of protein structures." *Journal of Applied Crystallography*. 26. Pg 283-291.

[36]Pettersen EF, Goddard TD, Huang CC, *et al*. (2004) "UCSF Chimera-a visualization system for exploratory research and analysis." *Journal of Computational Chemistry*. 25.13. Pg 1605-1612.

[37]Meng EC, Pettersen EF, Couch GS, *et al*. (2006) "Tools for integrated sequence-structure analysis with UCSF Chimera." *BMC Bioinformatics*. 7.339.

[38]Kang S, Yang M, Hong Z, *et al*. (2020) "Crystal structure of SARS-CoV-2 nucleocapsid protein RNA binding domain reveals potential unique drug targeting sites." *Acta Pharmaceutica Sinica B.* [published online ahead of print, 2020 Apr 20].

[39]Tilocca B, Soggiu A, Sanguinetti M, *et al*. (2020) "Comparative computational analysis of SARS-CoV-2 nucleocapsid protein epitopes in taxonomically related coronaviruses." *Microbes and Infection*. *22*.4-5. Pg 188–194.