 doi:10.20944/preprints202006.0165.v2

Article

Conserved Sequence Features in the Spike Protein Provide Evidence Suggesting the Origin of SARS-CoV-2 (COVID-19)-Related Viruses by Recombination between SARS virus and Another Sarbecovirus

Radhey S. Gupta[1*], Bijendra Khadka[1]

1 Department of Biochemistry and Biomedical Sciences, McMaster University, Hamilton, Ontario CA L8N 3Z5

* Correspondence: gupta@mcmaster.ca

Abstract: Both SARS-CoV-2 (COVID-19) and SARS coronaviruses (CoVs) are members of the subgenus *Sarbecovirus*. To understand the origin of SARS-CoV-2, protein sequences from sarbecoviruses were analyzed to identify highly-specific molecular markers consisting of conserved inserts or deletions (termed CSIs) in the spike (S) and nucleocapsid (N) proteins that are specific for either particular clusters/lineages of these viruses or are commonly shared by specific lineages. Three novel CSIs in the N-terminal domain of the spike protein S1-subunit (S1-NTD) are uniquely shared by the SARS-CoV-2, BatCoV-RaTG13 and most pangolin CoVs, distinguishing this cluster of viruses (SARS-CoV-2r) from all others. In the same positions, where these CSIs are found, related CSIs are also present in two other sarbecoviruses (viz. CoVZXC21 and CoVZC45 forming CoVZC cluster), which form an out group of the SARS-CoV-2r cluster. These three CSIs are not found in the SARS-CoVs. However, both SARS and SARS-CoV-2r CoVs contain two large CSIs in the C-terminal domain of S1 (S1-CTD), which binds the human ACE-2 receptor, that are absent in the CoVZC cluster of CoVs. These results indicate that while the S1-NTD of the SARS-CoV-2r viruses possesses the sequence characteristics of the CoVZC cluster of CoVs, their S1-CTD resembles the SARS viruses. Thus, the spike protein of SARS-CoV-2r viruses has likely originated from a recombination event between the S1-NTD of the CoVZC viruses and the S1-CTD of SARS viruses. This inference is also supported by the amino acid sequence similarity of the S1-NTD and S1-CTD from SARS-CoV-2 compared to the CoVZC and SARS CoVs. We also present evidence that one of the pangolin-CoV_MP789, whose receptor-binding domain is most similar to the SARS-CoV-2, is also derived by a recent recombination between the S1-NTD of the CoVZC CoVs and the S1-CTD of a SARS-CoV-2 related virus. Several other identified CSIs are specific for others clusters of sarbecoviruses including a clade consisting of bat SARS-CoVs (BM48-31/BGR/2008 and SARS_BtKY72). Structural mappings studies show that the identified CSIs are located within surface-exposed loops and form distinct patches on the surface of the spike protein. These surface loops/patches are predicted to interact with other host components and play important role in the biology/pathology of SARS-CoV-2 virus. Lastly, the CSIs specific for the SARS-CoV-2r clade provide novel means for development of new diagnostic and therapeutic targets for these viruses.

Keywords: Conserved signature indels (CSIs) specific for SARS and SARS-CoV-2-related viruses. Molecular markers distinguishing different clades of *Sarbecovirus*, Evolutionary relationships between SARS and SARS-CoV-2-related viruses, Origin of SARS-CoV-2 and Pangolin CoV_MP789 viruses, Novel sequence and structural features of spike and nucleocapsid proteins. Genetic recombination.

 doi:10.20944/preprints202006.0165.v2

1. Introduction

The current worldwide pandemic (COVID-19) is caused by a novel coronavirus (CoV) designated as SARS-CoV-2 (1). SARS-CoV-2 is the third coronavirus in the past two decades, responsible for a serious outbreak and health threat (2-5). The other two outbreaks were caused by coronaviruses now known as Severe Acute Respiratory Syndrome (SARS) and Middle East Respiratory Syndrome (MERS) (4, 6, 7). However, unlike the SARS and MERS, whose overall health impact was limited, SARS-CoV-2 or "COVID-19 virus" (alternate term used here for SARS-CoV-2) has now infected >23 million people worldwide resulting in >800,000 deaths (https://coronavirus.jhu.edu/). In view of the propensity of some CoVs to cause serious outbreaks, it is of much importance to understand the evolution of disease-causing CoVs and explore the genetic differences that distinguish these viruses from other relatively benign coronaviruses.

Coronaviruses are a large group of viruses that are a part of the subfamily *Coronavirinae* (4, 6). Most of these viruses have been isolated or originated from bats or avian species, which are natural reservoirs for these viruses (4, 8-11). Based on their phylogenetic branching and genomic structures, the viruses from the subfamily *Coronavirinae* have been divided into four genera viz. *Alphacoronavirus, Betacoronavirus, Gammacoronavirus,* and *Deltacoronavirus* (4, 6). Of these four genera, only a few viruses from *Alpha-* and *Beta-coronavirus* genera infect humans and cause respiratory illness (4, 6). Members of the *Alphacoronavirus* lineage cause only mild disease in humans, viz. HCoV-NL63, HCoV-229E. Betacoronaviruses, however, cause severe respiratory illnesses in humans and are responsible for different coronavirus epidemics viz. SARS, MERS and SARS-CoV-2 (2, 4-6, 12). Phylogenetic studies indicate that the genus *Betacoronavirus* is made up of four separate clusters commonly referred to as A, B, C and D, which are now recognized as distinct subgenera with the names *Embecovirus* (A), *Sarbecovirus* (B), *Merbecovirus* (C) and *Nobecovirus* (D), respectively (4, 6, 8, 12). Based on phylogenetic studies, while both SARS-CoV and SARS-CoV-2 and many other bat-CoVs and bat-SARS-related (SARSr-CoV) are a part of the subgenus *Sarbecovirus* (2, 4, 8, 12), the MERS-CoV groups within the subgenus *Merbecovirus* (6, 8). Thus, from the viewpoint of understanding the origin and evolution of COVID-19 virus, it is important to determine how SARS-CoV-2 differs from SARS and other bat CoVs within the subgenus *Sarbecovirus* (2, 12-16). The genome sequence of SARS-CoV-2 is most closely related (96% whole genome identity) to a bat CoV (BatCov-RaTG13) (2, 16), followed by 91.02% identity to a virus (pangolin CoV_MP789) from pangolins (14, 16, 17). In contrast, it exhibits lower (80-88%) sequence identity to the SARS-CoV (2, 12, 16). These results provide evidence suggesting that SARS-CoV-2 is derived from bat/pangolin CoVs (2, 16). Sequence comparisons have also identified a 12 nucleotide (4 aa) insertion in the spike (S) protein of SARS-CoV-2, which creates a polybasic furin cleavage site at the boundary of the S1 and S2 subunits/domains (14, 16). Additionally, sequence comparison studies show that the sequence of the receptor binding domain (RBD) (located within the S1-CTD) of the spike protein from SARS-CoV-2 is more similar to a pangolin CoV_MP789 than to the RBD from Bat-RaTG13-CoV (15, 16). In addition, in the sequence alignments of the RBD, Lu et al. (12) and Zhang et al. (14) have described 1-2 large indels (inserts or deletions) in the SARS and SARS-CoV-2 viruses, but their evolutionary significance or specificities are unclear. While the insights provided by these studies are important, further studies on understanding the differences between SARS-CoV-2 and other CoVs from the subgenus *Sarbecovirus* will be very useful.

Genome sequences are now available for many *Betacoronavirus* strains/isolates providing an extensive resource for understanding the evolution of SARS-CoV-2. A detailed study can thus be undertaken on identifying novel molecular features that are unique to this virus, allowing for a better understanding of the evolution of this virus and its properties. Genome sequences provide a means for carrying out different types of genetic and biochemical studies. Of these studies, one important class of molecular markers which have proven very useful for evolutionary and biochemical studies is comprised of conserved signature indels (insertions/deletions) (CSIs) in genes/proteins that are uniquely shared by a given group of species/viruses (18-22). The CSIs that are useful for evolutionary studies are generally of fixed lengths, present at specific positions in particular genes/proteins, and are flanked on both sides by conserved regions to ensure that they

 doi:10.20944/preprints202006.0165.v2

constitute reliable molecular markers or characteristics (18-22). The CSIs in gene/protein sequences generally result from rare genetic changes. Due to the discrete nature of these genetic changes, the presence or absence of CSIs in different lineages is generally not affected by factors that limit the reliability of inferences from phylogenetic trees (23-26). Because of the above characteristics, CSIs have provided important means for understanding evolutionary relationships and for the demarcation of specific groups of organisms in molecular terms (18, 21, 22, 24, 26, 27). Although the CSIs have been widely used for understanding the evolutionary relationships and for molecular demarcation of prokaryotic and eukaryotic organisms, this approach has not yet been used for examining the evolutionary relationships amongst viruses.

We describe here the results of our phylogenetic studies and a CSI-identification based approach on sarbecoviruses genomes to understand the origin and novel molecular features of SARS-CoV-2 and other sarbecoviruses. Our analyses have identified several novel CSIs in the spike (S) and nucleocapsid (N) proteins, which serve to clearly demarcate distinct clusters of sarbecoviruses. The spike protein of CoVs is composed of two subunits (resulting from proteolytic cleavage), of which the S1 subunit is responsible for binding to the host cell receptor(s) while the S2 subunit mediates membrane fusion and virus internalization of the virus (4, 28, 29). A number of CSIs identified here that are found in the N-terminal domain of the S1-subunit (S1-NTD) are distinctive characteristics of a clade (SARS-CoV-2r cluster) consisting of SARS-CoV-2 as well as a few closely related bat and pangolin CoVs. Interestingly, in the same positions in the S1-NTD, CSIs (related in sequences) are also present in another cluster of sarbecoviruses (CoVZC cluster). On the other hand, the C-terminal domain of S1 (S1-CTD) contains a number of other CSIs that are commonly shared by SARS-CoV-2r viruses and the SARS viruses, but which are lacking in the CoVZC cluster of viruses. These results suggest that the spike protein from the SARS-CoV-2r viruses is chimeric, originating from a recombination event between the S1-NTD of the CoVZC cluster of CoVs and the S1-CTD of SARS viruses. Our results also indicate that one of the pangolin CoV_MP789, whose RBD exhibits maximal similarity to the SARS-CoV-2 virus (14, 17, 30, 31), is also derived by a genetic recombination between two different sarbecoviruses. We also briefly discuss the structural and functional significance of the identified CSIs and their utilities for development of novel diagnostic and therapeutic targets.

2. Materials and Methods

2.1 Construction of phylogenetic trees

For the construction of phylogenetic trees, protein sequences for the spike (S) and RNA-dependent RNA polymerase (RdRp) proteins from representative viral species/isolates from different lineages of BetaCoVs were retrieved from the NCBI genome database. As the focus of this study is on the viruses from the B-lineage of *Betacoronavirus* (i.e. *Sarbecovirus*), more sequences were used for this group in comparison to the other lineages/subgenera of Beta-CoVs. Multiple sequence alignments (MSA) were created using ClustalW algorithm from the MEGA6 software package (32). Poorly aligned regions from the sequence alignments were removed using the Gblocks_0.91b program (33). Maximum-likelihood phylogenetic trees based on the resulting sequence alignments were inferred based on the JTT matrix-based model using MEGA6 (32). All positions with less than 95% site coverage were not considered during analysis. The percentage of trees in which the associated taxa clustered together is shown next to different branches. Pairwise sequence similarity matrix of protein sequences were determined used Clustal Omega program (34).

2.2 Identification of CSIs in Protein Sequences

Multiple sequence alignments of S, N and RdRp proteins, created as described above, were examined manually. The sequences were examined to identify insertions or deletions (indels), which were specifically found in some or all viruses from the subgenus *Sarbecovirus* and which were flanked by at least 4-5 conserved amino acid residues within the neighbouring 40-50 residues (19, 21, 22). The indels which were not flanked by conserved regions were not further considered, as they do

 doi:10.20944/preprints202006.0165.v2

not provide reliable molecular characteristics. Furthermore, as the focus of this study is on the *Sarbecovirus* subgenus of Beta-CoVs, indels in conserved regions which were specific for CoVs from other subgenera of Beta-CoVs were not further investigated. Query sequences encompassing the indel and its flanking 40-50 amino acids were collected for all potential CSIs. Afterwards, another detailed BLASTp search was carried out on these query sequences using the NCBI non-redundant database. All significant hits obtained from these BLASTp searches were examined in order to determine the group specificities of the identified CSIs. Signature files for the CSIs were created using SIG_CREATE and SIG_STYLE programs described in our earlier work (19) that are available on the GLEANS (www.gleans.net) server. Sequence information in different figures is shown for only a limited number of strains from different groups/lineages of sarbecoviruses. However, unless otherwise specified, the described CSIs are specific for the indicated clusters and they are also present in other strains of CoVs from these clusters.

2.3 Homology modelling of Proteins to Map the Locations of the CSIs in Protein Structures

The structural locations of the identified CSIs were mapped in protein structures using the experimentally solved structures and by creating homology models for spike proteins using different available structures as templates (35, 36). Homology models for the N-terminal domain (NTD) of the spike protein for SARS-Cov-2/Wuhan-Hu-1 (Acc. no: YP_009724390) and BatSARS-like-CoVZC45 (Acc. no: AVP78031) were created based on the available experimental structure of SARS-Cov-2 spike protein (PDB: 6vsb) (36). Cyro-EM based structure for the receptor-binding domain (RBD) of SARS-Cov-2 spike protein (PDB: 6m17, Chain E) was utilized to create a homology model of the RBD for BAT SARS-CoVZC45 to map the structural location of the CSI in SARS-Cov-2/Wuhan-Hu-1. Homology modeling was carried out using the MODELLER v9.11 program (37) and their stereochemical properties were assessed as described in our earlier work (22, 38).

3. Results

3.1 Phylogenetic Analysis of Betacoronaviruses

The evolutionary relationships amongst the major lineages of coronaviruses have been described previously (4, 6-8, 31). Recent studies have also examined the evolutionary relationship of SARS-CoV-2 to other closely related viruses from the subgenus *Sarbecovirus* (2, 12, 14, 31). However, as many genome sequences are now available for strains/isolates from SARS-CoV and SARS-CoV-2 related viruses, we have also constructed a phylogenetic tree(s) for *Betacoronaviruses* to serve as a reference point in our work on identification of CSIs. Phylogenetic trees were constructed based on both spike and RdRp proteins. The tree for the spike protein is shown in Figure 1. A more detailed tree containing information for other strains/isolates from the subgenus *Sarbecovirus* is provided in Figure S1. The tree based on the RdRp sequences, which is very similar to that reported by Zhang et al. (14), is provided as Figure S2. The tree shown in Figure 1 confirms that within the Beta-CoVs four main clusters corresponding to the four subgenera are observed (4, 6). The clusters corresponding to these four subgenera are marked in the tree along with their commonly known clade designations viz. clades A, B, C and D. These four subgenera are separated from each other by long-branches and supported by 100% bootstrap scores.

Of these four lineages, the subgenus *Sarbecovirus* contains both SARS-CoV, SARS-CoV-2 and related viruses, and its members form a tight cluster in the tree (Figure 1). Although the interrelationships of different viral strains/isolates within this subgenus are not resolved (as indicated by the low bootstrap scores for most nodes), it is possible to draw some inferences based on this tree.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 August 2020 doi:10.20944/preprints202006.0165.v2

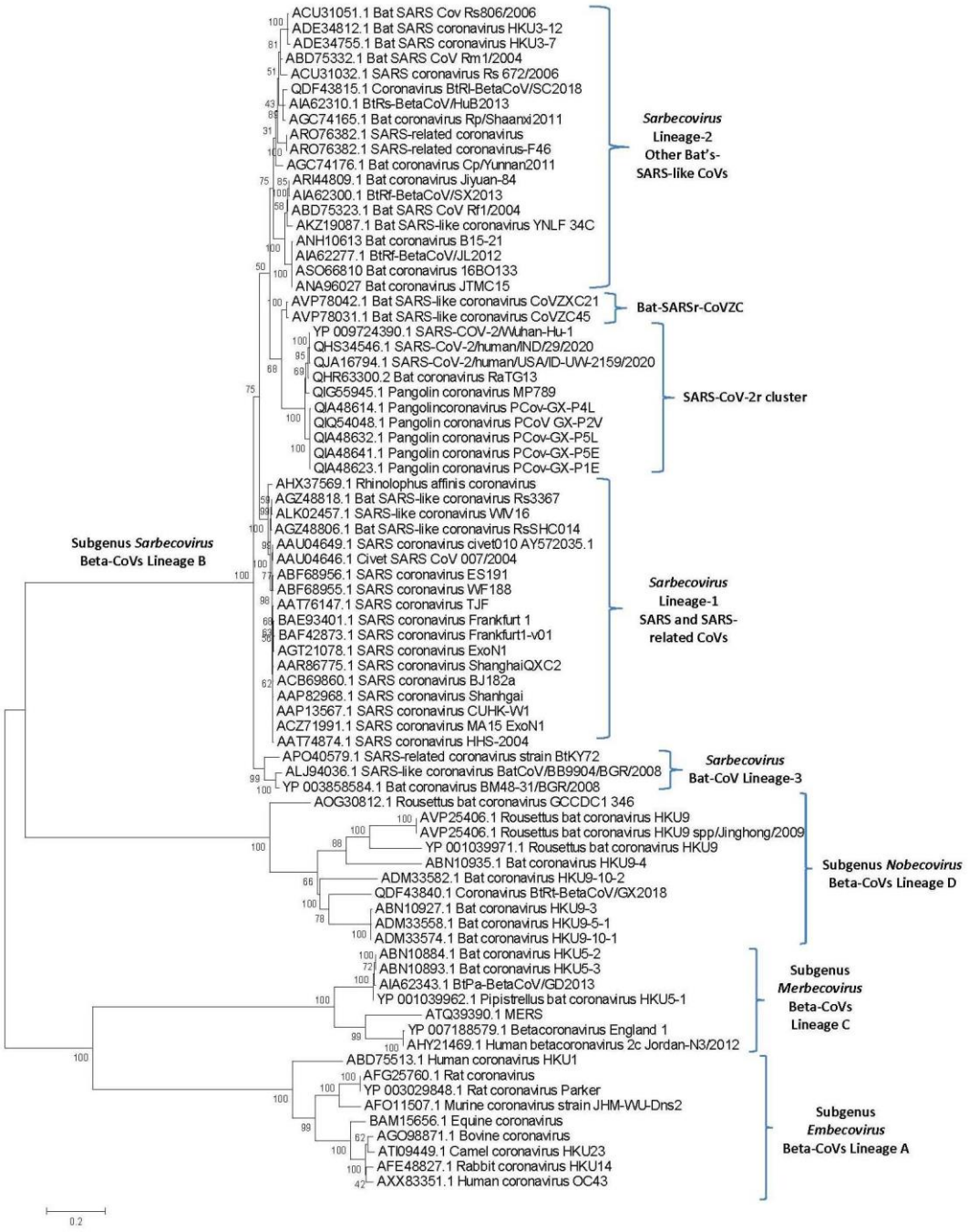

Figure 1. A maximum-likelihood distance tree based on sequence alignment of the spike protein from representative viruses/strains of the genus *Betacoronavirus*. The tree was constructed as described in the Methods section and the % bootstraps for different branches are indicated on the nodes. The clades corresponding to different subgenera within the *Betacoronavirus* as well as other clusters and lineages within the subgenus *Sarbecovirus* are labeled. A tree showing sequence information for additional CoVs strains from the subgenus *Sarbecovirus* is provided as Figure S1.

Some inferences that can be drawn from the tree shown in Figure 1 are as follows: (i) In accordance with earlier studies (2, 14, 31), the SARS-CoV-2 strains group reliably with the BatCoV-RaTG13 and pangolin CoVs and within this clade BatCoV-RaTG13 exhibits a closer relationship to the SARS-CoV-2 than the pangolin CoVs. We will be referring to this clade as the SARS-CoV-2r (related) cluster. (ii) Two bat SARS-like coronavirus strains viz. CoVZXC21 and CoVZC45 form an out group of the SARS-Cov-2r cluster, indicating a closer relationship of these viral strains to this cluster than the other Bat SARS-like CoVs. A close relationship of these two

 doi:10.20944/preprints202006.0165.v2

viruses to the SARS-CoV-2r cluster is also seen in trees based on whole genomes in recent studies (2, 12, 14). The clade for these two viral strains is marked in the tree as Bat-SARSr-CoVZC cluster and it will be referred to simply as the CoVZC cluster. (iii) Within the subgenus *Sarbecovirus*, SARS and SARS-like CoVs form a separate clade (marked as *Sarbecovirus* lineage B-1) distinct from the clade encompassing SARS-CoV-2r cluster, CoVZC cluster and a cluster grouping the other bat-SARS-CoVs. We have labeled this latter clade/cluster consisting of other bat-SARS-CoVs as *Sarbecovirus* lineage B-2. Of the clades/clusters indicated above, while the SARS-CoV-2r cluster has strong statistical support, the nodes separating other clusters have low bootstrap scores and they are separated by short branches. (iv) Three CoV strains viz. SARS-like CoV-BtkY72, BatCoV/BB9904/BGR/2008 and Bat CoV BM48-31/BGR/2008 formed a separate deeper-branching cluster. A distinct (and generally deeper) branching of these bat-SARS-like CoVs has also been observed in other studies (2, 12, 14). We have designated this clade as the *Sarbecovirus* lineage-3 in our work. The tree based on RdRp (Figure S2) showed even more limited resolution among the sarbecoviruses. However, some of the relationships inferred from Figure 1, viz. distinctness of the SARS-CoV-2r as well as a separate grouping of the CoVs from *Sarbecovirus* lineage-3, are also supported by this tree. The tree shown in Figure 1, despite its limitations, provides us with a useful reference point for discussing and interpreting the evolutionary significance of different CSIs identified in this work.

3.2 Identification and Characteristics of Conserved Signature Indels in Spike and Nucleocapsid Proteins

The branching of species in phylogenetic trees is influenced by large numbers of variables including differences in evolutionary rates between the examined taxa, regions of sequences that are included or excluded in tree construction, order of sequence alignment, choice of the outgroup, evolutionary model for tree construction, lateral gene transfers, long-branch attraction effect etc. (19, 24, 25, 39) and hence it is often not resolved. Furthermore, the viruses (specifically RNA viruses) are known to undergo frequent genetic recombination (7, 8, 10, 31, 40, 41), making it more difficult to resolve their evolutionary history by means of phylogenetic analysis. Thus, it is important to examine the evolutionary relationships among CoVs by means of other sequence-based approaches that do not rely on phylogenetic tree construction. As noted in the introduction, CSIs in protein sequences, which are uniquely shared by a given group of organisms, provide an important tool and class of molecular markers that have proven very useful for identifying evolutionary relationships (18, 21, 22, 24, 26, 42-45). CSIs in genes/protein sequences result from rare genetic changes. Therefore, when a CSI of a definite length is present at a specific position within a given group of viruses (organisms), its most parsimonious explanation is that the genetic change giving rise to the CSI occurred in a common ancestor of the indicated group and was then retained by the other members of that group (18, 21, 22, 24, 26, 42). Furthermore, due to the discrete molecular nature of CSIs and their presence *within* a conserved region of the genes/proteins, the presence or absence of these molecular characteristics is generally not affected by most variables that can confound inferences that are based on phylogenetic trees. Considering these characteristics, the shared presence of CSIs by a given clade of viruses/organisms provides reliable evidence, independent of the phylogenetic trees, of the common ancestry and relatedness of that group of species (18, 21, 22, 26). Our analyses of the S- and N- proteins have identified several informative CSIs whose descriptions and evolutionary significance are described below.

A key question concerning SARS-CoV-2 is how it has evolved from other CoVs that are members of the subgenus *Sarbecovirus*. Based on genome sequence similarity and phylogenetic branching, SARS-CoV-2 is most closely related to a bat CoV (RaTG13) followed by a pangolin CoV. However, there is no molecular characteristic known that is specifically shared by the viruses from the SARS-CoV-2r cluster. Our analyses have identified several CSIs that provide strong evidence of a specific relationship of the viruses within this cluster and provide insights concerning the evolution of the SARS-CoV-2r cluster of viruses. In Figure 2, we present partial sequence alignments of two conserved regions from the S1-N-terminal domain (S1-NTD) where CSIs of specific lengths distinguishing a number of different lineages of sarbecoviruses are found. In the sequence alignment

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 August 2020 doi:10.20944/preprints202006.0165.v2

shown in Figure 2A, a 6 aa insert in the S-protein (boxed and labeled ❶) is specifically present in all members of the SARS-CoV-2r cluster (i.e. SARS-CoV-2 strains, Bat-CoV-RaTG13, as well as most pangolin CoVs, except the pangolin-CoV_MP789. In the same position where this 6 aa insert is found, a 3 aa insert (marked ❷) is also present in the two viruses from the CoVZC cluster, as well as in pangolin CoV_MP789. As noted earlier, the two CoVs from the CoVZC cluster form an out group of the SARS-CoV-2r cluster. However, in view of the branching of pangolin CoV_MP789 with members of the SARS-CoV-2r cluster (Figure 1), the absence of the 6 aa CSI specific for the SARS-CoV-2r clade in this strain and its sharing of the 3 aa CSI specific for the CoVZC cluster is unexpected.

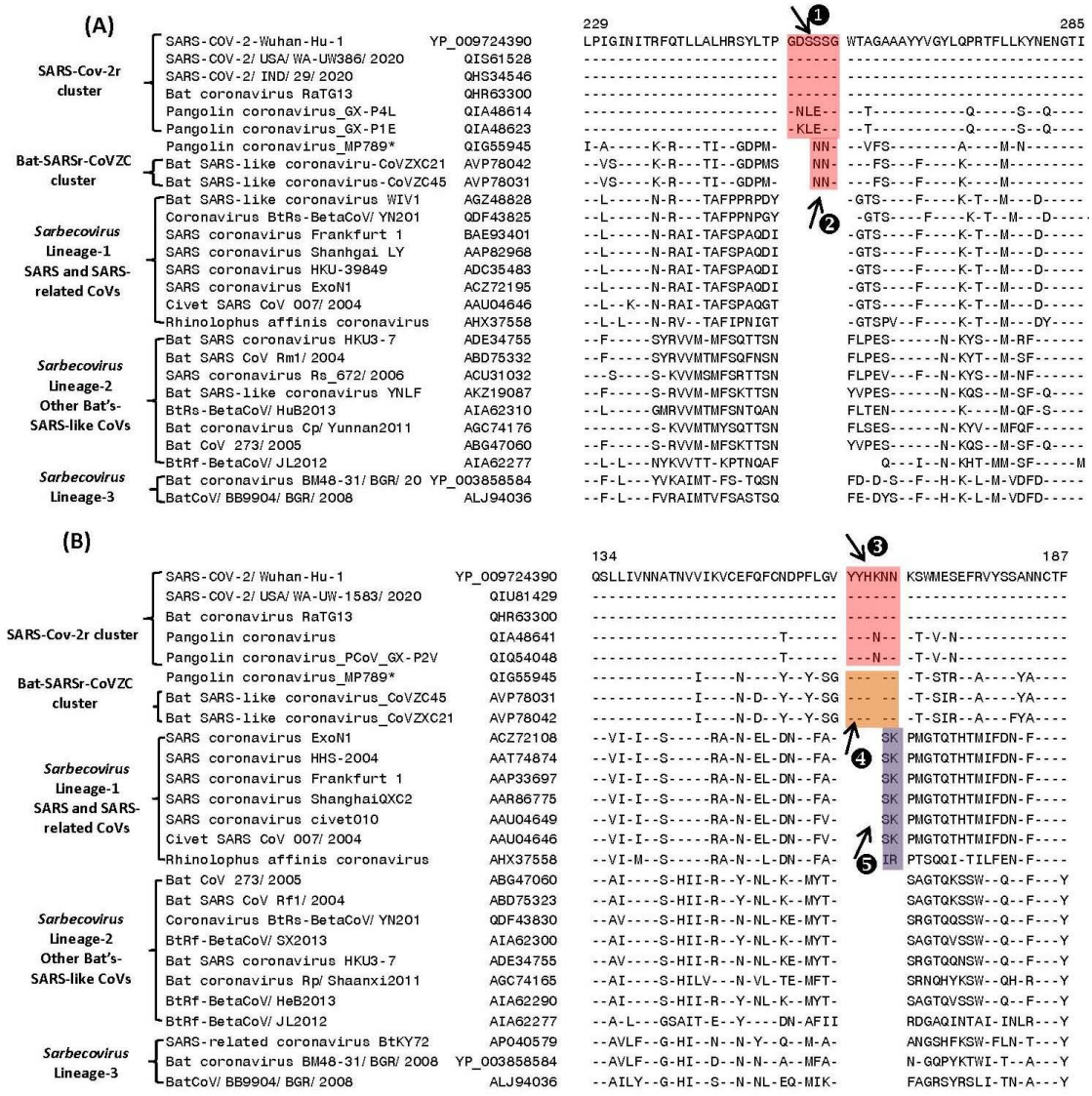

Figure 2. Partial sequence alignments of two conserved regions from the S1-N-terminal domain showing a number of CSIs that are specific for different clades/lineages of sarbecoviruses. (A) This panel shows a 6 aa insert (❶) in a conserved region that is present in different viruses from the SARS-CoV-2r cluster, except pangolin-CoV_MP789. In the same position, a 3 aa insert (❷) is commonly shared by the two CoVs from the CoVZC cluster and pangolin-CoV_MP789. Panel (B) shows a 6 aa insert (❸) in the S1-NTD that is specific for the SARS-CoV-2r cluster except pangolin-CoV_MP789. The CoVs from the CoVZC cluster and pangolin-CoV_MP789 contain a 5 aa insert (❹) in this position. The 2 aa CSI present in this region (labeled as ❺), is specific for the B-1 lineage of *Sarbecovirus* comprised of SARS-CoV and related viruses. Dashes (-) in these alignments denote identity with the amino acid shown in the top sequence. The numbers on the top indicate the

 doi:10.20944/preprints202006.0165.v2

position within the indicated proteins. The * for the pangolin CoV_MP789 indicates that its sequence characteristics are anomalous.

Figure 2B shows a sequence alignment from another conserved region of the S1-NTD where CSIs of specific lengths are present in the same position in a number of clusters of sarbecoviruses. In this case, a 6 aa insert (marked as ❸) is again commonly shared by all viruses from the SARS-CoV-2r cluster except pangolin-CoV_MP789, which contains a shorter 5 aa insert. Interestingly, this 5 aa insertion (labeled as ❹) is also a commonly shared characteristics of the two CoVs (viz. CoVZXC21 and CoVZC45) from the CoVZC cluster. Importantly, in this case the sequence of the 5 aa insert in the CoVZC strains and pangolin-CoV_MP789 is identical to that found in other members of the SARS-CoV-2r cluster, except that it is shorter by 1 aa residue. The distribution of these two CSIs (i.e. ❸ and ❹) in different CoVs strains strongly indicates that the pangolin-CoV_MP789 strain, despite its branching with the SARS-CoV-2r cluster in Figure 1, is more closely related to the CoVZC cluster of CoVs in the S1-N-terminal domain than to the other pangolin CoVs. Furthermore, as noted above, since the viruses from the CoVZC cluster form an immediate outgroup of the SARS-CoV-2r cluster (Figure 1), the shared presence of related CSIs (i.e. ❸ and ❹) in these two CoVs clusters is most parsimoniously explained by postulating that a 5 aa insertion initially occurred in a common ancestor of the CoVZC strains (and pangolin-CoV_MP789). This was then followed by a subsequent genetic change leading to 1 aa insertion within the 5 aa insert in a common ancestor of the SARS-CoV-2r cluster of CoVs. In addition to the CSIs ❸ and ❹, the sequence region shown in Figure 3B also contains a 2 aa CSI (labeled as ❺), which is specific for the B-1 lineage of *Sarbecoviruses* that is comprised of SARS-CoV and related viruses.

In Figure 3, we present partial sequence alignments from two conserved regions of the nucleocapsid and spike proteins containing several additional CSIs that also support the deduced evolutionary relationships between the members of the SARS-CoV-2r and the CoVZC clusters of CoVs. In the partial sequence alignment of N-protein shown in Figure 3A, a 2 aa deletion (boxed and labeled ①) is present in a highly conserved region that is commonly shared by all members of the SARS-CoV-2r cluster as well as the two CoVs comprising the CoVZC cluster. The shared presence of this CSI in all of the CoVs from the SARS-CoV-2r and the CoVZC clusters, but not in any other CoVs, provides strong evidence that these two groups of CoVs are specifically related to each other and that the genetic change leading to this CSI occurred in a common ancestor of these CoVs. In addition to this CSI, two species from the *Sarbecovirus* lineage-3 also contain a 1 aa deletion (marked ②) in the same position.

In the sequence alignment shown in Figure 3B, a 7 aa insertion in a conserved region within the S1-N-terminal domain (labeled ❻) is commonly shared by the SARS-CoV-2 strains, BatCov-RaTG13 and one of the pangolin CoVs. However, in contrast to this pangolin CoV strain (GX-P2V), all other pangolin strains (except MP789) contain a 5 aa insert and their sequences are very similar (nearly identical). As the sequence for the strain GX-P2V is very similar to those of the other pangolin CoVs, but it contains some ambiguity (marked by X), the presence of a 7 aa insert in this strain is surprising. Due to the unexpected nature of this indel, it needs to be verified to exclude the possibility that the presence of a 7 aa insert in this strain is not due to a sequencing error. Nonetheless, as all other pangolin CoVs contain a shorter 5 insert in this position, the results for this CSI indicate that in this sequence region SARS-CoV-2 is more closely related to the BatCoV-RaTG13, in comparison to the pangolin CoVs.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 August 2020 doi:10.20944/preprints202006.0165.v2

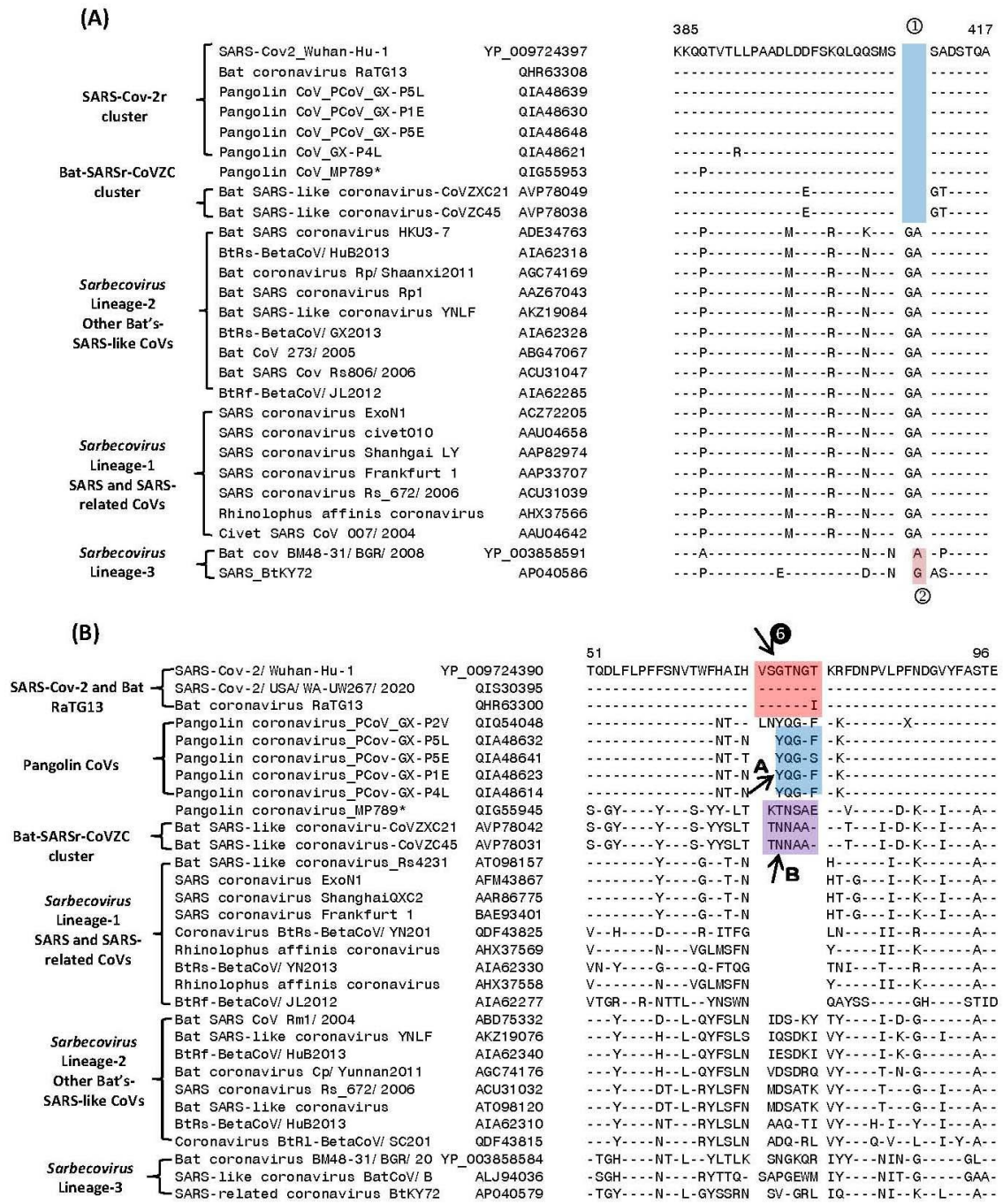

Figure 3. Partial sequence alignments of (A) nucleocapsid and (B) spike protein showing some CSIs that are specific for certain clusters/lineages of sarbecoviruses. Panel (A) shows a 2 aa deletion (①) in the N-protein that is specific for the SARS-CoV-2r and CoVZC clusters of CoVs. The viruses from *Sarbecovirus* lineage-3 contain only a 1 aa deletion (②) in this position. (A) The CSI indicated as ❻ marks a region within the S1-NTD where a 7 aa insertion is present in SARS-CoV-2 strains and bat CoV RaTG13. Most pangolin homologs have a 5 aa insert (highlighted in blue and marked **A** in this position, whereas CoVs from the CoVZC cluster and pangolin-CoV_MP789 contain a 6 aa insert (highlighted in purple and marked **B** in this position).

Interestingly, the two bat CoVs from the CoVZC cluster as well as the pangolin CoV_MP789 contain a 6 aa insert in this position rather than the 5 aa or 7 aa inserts found in the SARS-CoV-2r and CoVZC clusters, respectively. This result again supports the inference from the CSIs marked ❷ and ❹ (Figure 2) that the pangolin-CoV_MP789 in its S1-N-terminal domain is specifically related to the CoVZC cluster of CoVs. In the sequence alignment shown in Figure 3B, although the CoVs from

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 August 2020 doi:10.20944/preprints202006.0165.v2

Sarbecovirus lineages 2 and 3 also contain 6-7 aa insertions, these appears to be of independent origin. Importantly, the CoVs from the B-1 lineage of *Sarbecoviruses* comprising of SARS-CoV and related viruses do not contain an insertion in this position.

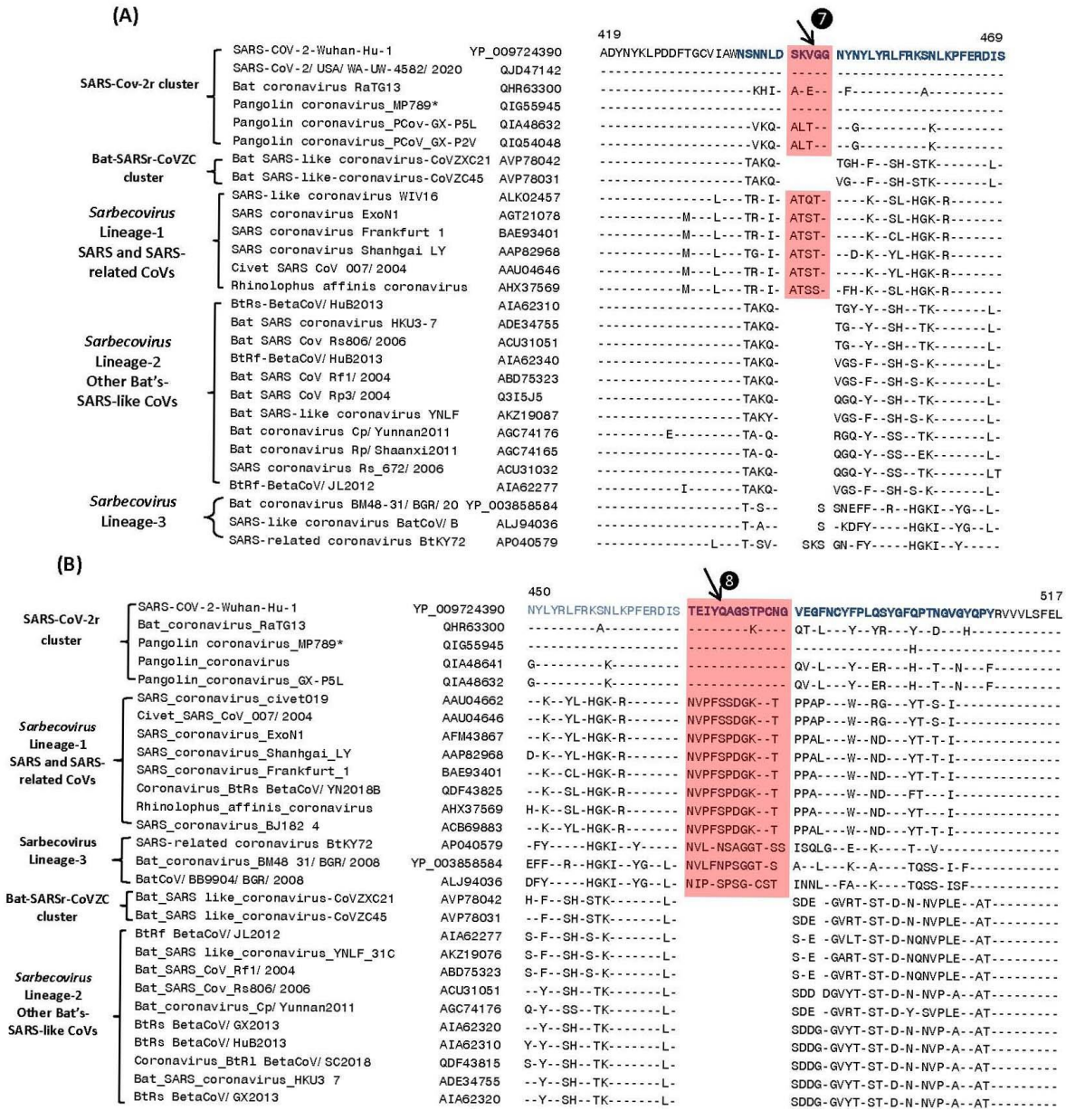

Figure 4. Partial sequence alignment of two conserved regions (overlapping) from the S1-CTD encompassing the receptor binding domain (RBD) of the spike protein, where two CSIs, panel (A) (❼) and panel (B) ❽) are mainly found in the SARS and SARS-CoV-2r clusters of viruses. The sequence region comprising the RBD [11] is shown in dark blue whereas the overlapping parts of the sequence are marked in light blue. Dashes (-) in the sequence alignments denote identity with the amino acid shown in the top line.

In Figure 4, we show sequence information for two other CSIs that are present in the C-terminal domain of the S1 subunit. Although these insertions have been described previously (12, 14, 16), their evolutionary significance or specificity has not been clear from earlier work. Both these CSIs are present in close proximity to each other within the RBD of the spike protein (sequence of the RBD is shown in dark blue with overlapping sequence (residues 450-469) colored light blue). Figure 4A shows a 5 aa insert (marked ❼) that is commonly shared by members of the SARS-CoV-2r clade as well as different SARS-like CoVs that are part of the *Sarbecovirus* lineage-1 (see Figure 1). However,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 August 2020 doi:10.20944/preprints202006.0165.v2

this CSI is not found in members of the CoVZC cluster or other bat-SARS-related CoVs comprising the *Sarbecovirus* lineages-2 and 3. Figure 4B shows another CSI, in this case a 13 aa insertion in the spike protein (marked ❽), which is also commonly shared by members of the SARS-CoV-2r clade as well as different SARS-like CoVs of *Sarbecovirus* lineage-1. In addition, this CSI is also present in *Sarbecovirus* lineage-3, but it is lacking in members of the CoVZC cluster as well *Sarbecovirus* lineages-2. Furthermore, it is of interest to note that both these CSIs are present in all of the pangolin homologs including pangolin CoV_MP789, but they are lacking in the two strains from the CoVZC cluster of CoVs. Additionally, the amino acid sequence of pangolin CoV_MP789 in the RBD region is highly similar to the SARS-CoV-2 sequence and it differs from the sequence of the SARS-CoV-2-Wuhan-Hu1 in only 1 aa residue (14, 16). The significances of these observations in the origin of SARS-CoV-2r clusters of viruses as well as the pangolin CoV_MP789 virus are discussed in later sections.

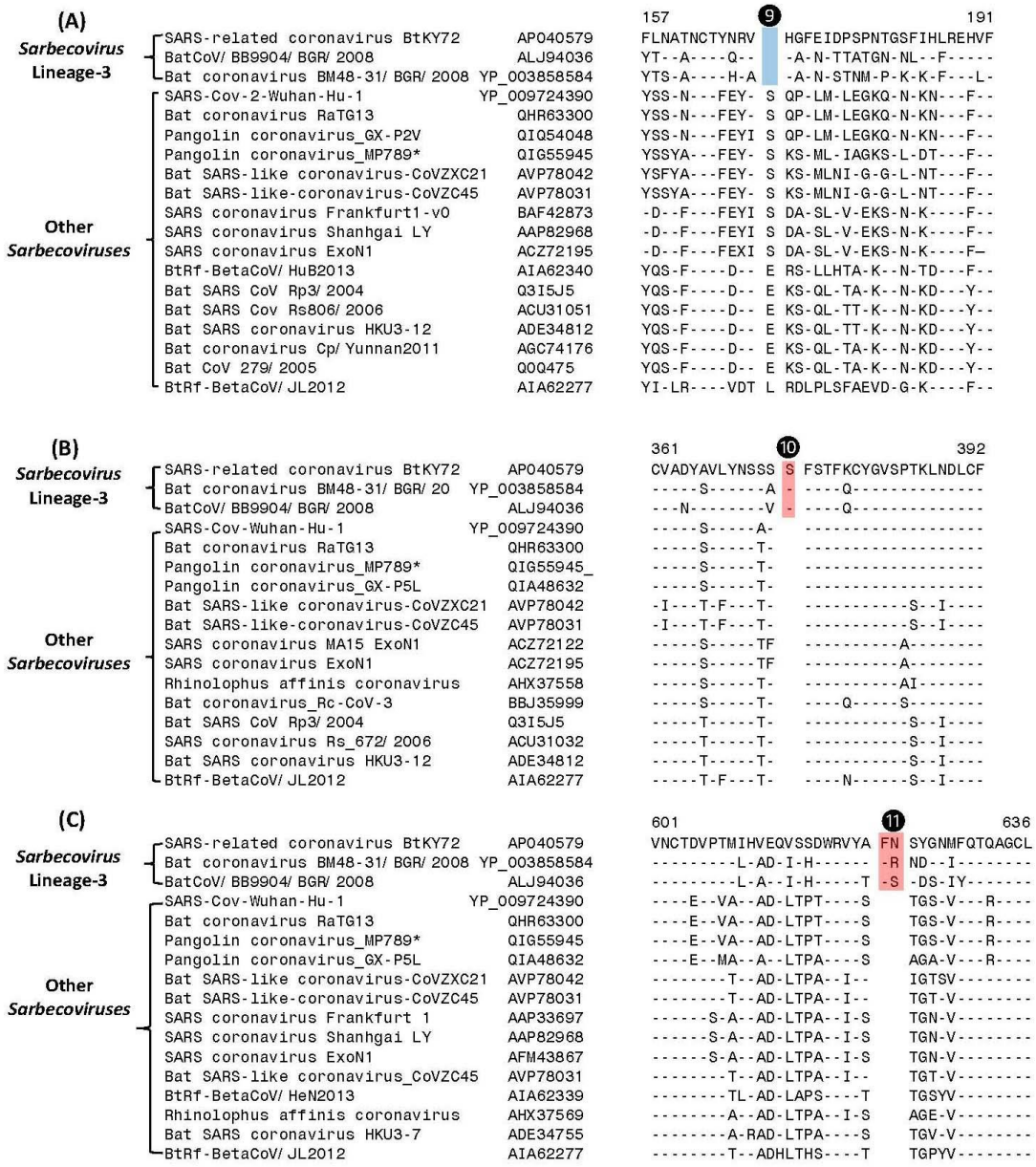

Figure 5. Excerpts from the sequence alignment for three regions of the spike protein depicting three CSIs those are specific for the CoVs from *Sarbecovirus* lineage-3.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 August 2020 doi:10.20944/preprints202006.0165.v2

In addition to these CSIs, we have also identified several other CSIs in the S- and N- proteins that are uniquely shared by members of the *Sarbecovirus* lineages-3, which branches separately from all other sarbecoviruses. Sequence information for 3 of these CSIs found in the S-protein is shown in Figure 5. Of these, the two CSIs shown in Figure 5A and 5B are present in the S1 subunit, whereas the CSI in Figure 5C is located within the S2 subunit of the spike protein. In all of these cases, the highlighted CSIs (marked ❾, ❿ and ⓫) are only found in the S-protein homologs from this particular lineage of *Sarbecovirus* but not in any other CoVs. In Figure 6 we present sequence alignments of two other conserved regions from the S- and N-proteins, each of which contains two different CSIs.

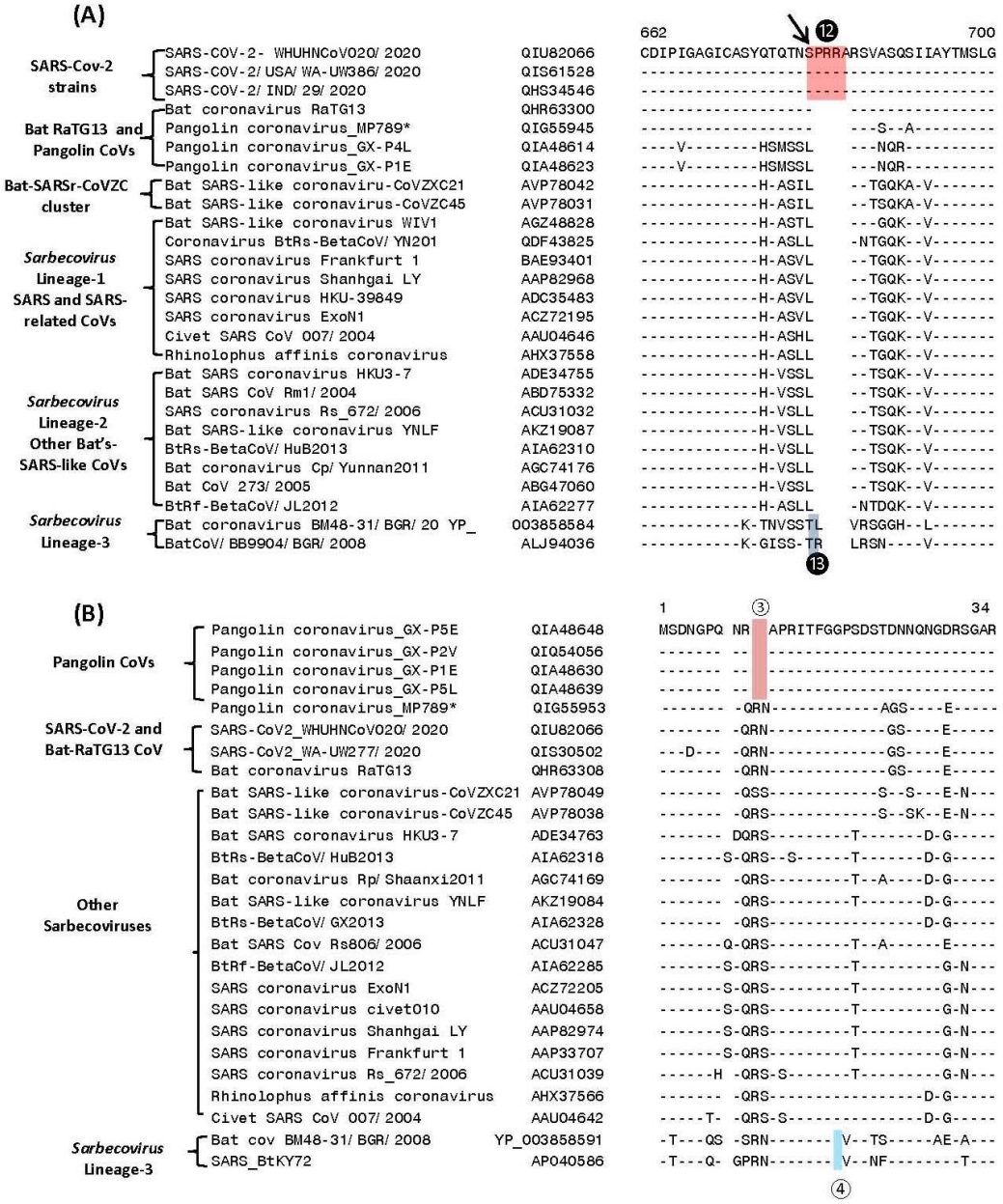

Figure 6. Partial sequence alignments of (A) spike proteins and (B) nucleocapsid protein showing a number of CSIs (highlighted) that are specific for different clades/lineages of sarbecoviruses. The panel (A) shows a 4 aa CSI (⓬) specific for the SARS-CoV-2 identified in earlier work (14, 16). A 1 aa insertion (marked ⓭) is also present in the same position in CoVs from *Sarbecovirus* lineage-3. (B) In the sequence alignment of N-protein, a 2aa deletion (③) is highlighted that is specific for pangolin CoVs. A 1 aa deletion (④) specific for *Sarbecovirus* lineages-3 is also present in a neighboring region.

 doi:10.20944/preprints202006.0165.v2

In Figure 6A, where a partial sequence alignment is shown for the S-protein, the larger 4 aa CSI (marked ❾), described previously (14, 16, 41), is specific for the SARS-CoV-2 homologs. In the same position, where this CSI is located, a 1 aa insertion (marked ❿) is also present in members of the *Sarbecovirus* lineage-3. Figure 6B shows partial sequence alignment from the N-terminal region of the N-protein. Of the two highlighted CSIs that are present in this region, a 2 aa deletion (marked ⑥) is specific for the pangolin-homologs. However, while this 2 aa deletion is present in all other pangolin homologs, it is absent in the protein homolog from Pangolin-CoV-MP789. This provides further evidence concerning the unusual nature of this pangolin-CoV strain. In addition to this CSI, the sequence alignment shown in Figure 6B also contains another CSI consisting of 1 aa deletion (marked ⑦) that is specific for the *Sarbecovirus* lineage-3.

3.3 Sequence Similarity Studies on SARS-CoV-2 and Pangolin-CoV_MP789

The CSIs that we have identified for different clades of sarbecoviruses show contradictory results for two sets of CoVs. The first anomaly concerns the pangolin-CoV-strain_MP789, which branches with the SARS-CoV-2r cluster in phylogenetic trees (see Figure 1). However, for all of the CSIs present in the S1-NTD (viz. ❷, ❹ and ❻A), the different CSIs in the pangolin-CoV_MP789 are identical to those found in the CoVZC cluster of CoVs, whereas the other pangolin-CoVs share the characteristics of the SARS-CoV-2r cluster of viruses (CSIs ❶, ❸). In contrast to this, for the two CSIs in the S1-CTD (❼and ❽), the pangolin CoV_MP789 displays characteristics similar to those seen for other pangolin CoVs and the SARS-CoV-2r cluster of viruses, and its behavior differs from the CoVZC cluster of CoVs. The second anomaly concerns the distribution characteristics of the CSIs that are related to the SARS-CoV-2r cluster of CoVs. Within the S1-NTD, a number of CSIs are found that are specific for the SARS-CoV-2r cluster of viruses. In the same positions, where these CSIs are found, the CoVZC cluster of CoVs also contains similar but related CSIs. On the other hand, the two CSIs that are present in the S1-CTD (❼and ❽), they are commonly shared by the SARS-CoV-2r cluster as well as by SARS and SARS-like CoVs (*Sarbecovirus* lineage-1), but they are absent in the CoVZC cluster of CoVs.

To understand these contradictory observations, we have determined the pairwise amino acid similarity of the SARS-CoV-2 as well as pangolin CoV_MP789 to the representative CoVs from CoVZC cluster, SARS-CoVs and also other pangolin viruses. These determinations were done for the entire sequence length of the spike protein as well as its specific sequence regions including the S1-NTD, S1-CTD and the entire C-terminal region (excepting S1-NTD). The results of these analyses are presented in Table 1.

As seen from Table 1, for the entire spike protein, the highest similarity of pangolin-CoV_MP789 is observed for the SARS-CoV-2 (90.51%). This value is lower than the similarity of SARS-CoV-2 to Bat-CoV-RaTG13 (97.7%) or other pangolin CoV strains such as GX-P2V (92.4%). However, for the S1-NTD region (aa 1-318), the highest similarity of pangolin-CoV_MP789 is observed for the Bat-SARS-like-CoV-CoVZC45 (85.94%), whereas all other CoVs including SARS-CoV-2, Bat-CoV-RaTG13, SARS CoV-ExoN1 and other pangolin CoVs exhibit much lower similarity (52-68%). A more dramatic difference in sequence similarity is observed when the sequence comparison is made for S1-CTD region (aa 319-540) or for the entire C-terminal region of the spike protein (i.e. from aa 320-1265). In both these cases, pangolin-CoV_MP789 showed the highest similarity to the SARS-CoV-2 (96.86% for S1-CTD and 98.1% for the C-terminal region, which includes the S2 subunit, respectively). In comparison, much lower similarity is observed for the Bat-SARS-like-CoV-CoVZC45 (Table 1). Of particular interest is the observed sequence similarity of the pangolin-CoV_MP789 for the S1-CTD region or for the entire C-terminal region to the SARS-CoV-2 virus. This sequence similarity is noticeably higher than that observed for any other CoVs including Bat-CoV-RaTG13. Recently, Liu et al. (31) have also reported that the spike protein of pangolin-CoV_MP789 (now designated as pangolin-CoV-2020) exhibits higher similarity in

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 August 2020 doi:10.20944/preprints202006.0165.v2

S1-NTD to Bat-CoV-ZC45 and Bat-CoV-ZXC21, whereas its C-terminal sequence (including the S1-CTD) is more closely related to the Bat-CoV-RaTG13 and SARS-CoV-2 viruses.

Table 1: Amino Acid Identity (%) for Different Regions of the Spike Protein from Representative CoVs

	Spike Protein (Full length)					
	SARS-CoV-2 Wuhan	Bat RaTG13	SARS-ExoN1	Pangolin MP789	Pangolin GX-P2V	Bat-SARS-like CoVZC45
SARS-COV-2/Wuhan-Hu-1	100					
Bat coronavirus RaTG13	97.71	100				
SARS-ExoN1	77.14	77.60	100			
Pangolin CoV_MP789	90.51	89.55	76.60	100		
Pangolin CoV GX-P2V	92.43	93.05	77.22	87.57	100	
Bat SARS-like CoVZC45 (CoVZC cluster)	81.93	81.75	77.62	87	81.35	100
	S1- N-terminal Region/Domain (S1-NTD) (aa 1-318)					
SARS-COV-2/Wuhan-Hu-1	100					
Bat coronavirus RaTG13	98.45	100				
SARS-ExoN1	54.45	54.90	100			
Pangolin CoV_MP789	68.34	68.65	52.15	100		
Pangolin CoV GX-P2V	87.65	88.27	54.25	67.40	100	
Bat SARS-like CoVZC45 (CoVZC cluster)	66.77	67.08	50.50	85.94	64.84	100
	S1-C-terminal Region/Domain (includes Receptor binding domain) (aa 319-540)					
SARS-COV-2/Wuhan-Hu-1	100					
Bat coronavirus RaTG13	90.13	100				
SARS-ExoN1	76.13	77.64	100			
Pangolin CoV_MP789	96.86	89.69	77.04	100		
Pangolin CoV GX-P2V	86.1	87.44	76.74	87.44	100	
Bat SARS-like CoVZC45 (CoVZC cluster)	71.08	70.59	76.04	70.59	70.59	100
	C-terminal Sequence Spike Protein (aa 320-1265)					
SARS-COV-2/Wuhan-Hu-1	100					
Bat coronavirus RaTG13	97.46	100				
SARS-ExoN1	89.81	89.47	100			
Pangolin CoV_MP789	98.10	96.61	89.30	100		
Pangolin CoV GX-P2V	94.07	94.50	89.64	94.29	100	
Bat SARS-like CoVZC45 (CoVZC cluster)	87.04	86.61	92.19	87.26	86.72	100

Sequence similarity for different regions of representative CoVs from the SARS-CoV-2r, SARS-ExoN1 virus, a virus from the CoVZC cluster (Bat SARS-like CoVZC45) and two pangolin viruses including pangolin CoV_MP789. Other CoVs from these clusters (including Bat SARS-like CoVZXC21) showed similar results as shown here for the chosen representative(s) from these clades.

Table 1 also shows the results of sequence similarity determination for the SARS-CoV-2 versus either CoVZ45 or the SARS-Exo-N1 CoVs for different regions of the spike protein. At the whole protein level, SARS-CoV-2 exhibits 81.93% sequence identity to the CoVZC45 virus vs. 77.14% identity to the SARS-ExoN1 virus. However, for the N-terminal domain (S1-NTD) the sequence similarity of the SARS-CoV-2 to the CoVZC45 was much higher (66.77%) in comparison to the SARS-ExoN1 virus (54.45%). Interestingly, results opposite to these were obtained when the sequence similarity determination are made for the S1-CTD or the entire C-terminal region of the spike protein. For both these regions, the SARS-CoV-2 homolog exhibited higher sequence similarity to the SARS-ExoN1 (76.13%) than to the CoVZC45 virus (71.08%). Results similar those of the SARS-CoV-2 were observed for the Bat-CoV RaTG13 (Table 1). These results again indicate that the S1- N-terminal domain of SARS-CoV-2r viruses is more closely related to the CoVZC cluster of CoVs, while the S1-CTD and the rest of the C-terminal region (S2 subunit) bear greater resemblance to the SARS (ExoN1) CoVs.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 August 2020 doi:10.20944/preprints202006.0165.v2

We have also created a multiple sequence alignment of the spike protein from pangolin-CoV_MP789 and representative CoVs from SARS-CoV-2r and CoVZC clusters (Figure S3). This sequence alignment shows that within the S1-NTD (aa 1- 320), in >80% of the polymorphic sites, the pangolin-CoV_MP789 contains the same amino acid residues as the CoVs from the CoVZC cluster. In contrast, in the remainder of the S-protein (aa 330-1265), the polymorphic replacements in pangolin-CoV_MP789 are predominantly (>90%) the same as seen in SARS-CoV-2. These results further suggest that the S1-NTD of the spike protein from pangolin-CoV_MP789 bears a close resemblance to the CoVZC cluster of CoVs, though the rest of its sequence is strikingly similar to the SARS-CoV-2. It is of interest that in the sequence of pangolin CoV_MP789, a Ssp1 restriction site is present between aa 327-328, i.e. near the junction of the S1-NTD and S1-CTD domains. However, the possible significance of this restriction site, if any, is unclear.

3.4 Localizations of the CSIs in Protein Structures

We have also examined the locations of the identified CSIs in the three-dimensional structure of spike protein. For these studies, using the available structures for the spike proteins (PDB IDs: 6acc, 6m17, 6vsb), homology models were created for the spike protein from a number of different CoVs including CoVZC-45, SARS ExoN1 and SARS-CoV-2 (Wuhan-Hu-1) to localize and visualize the locations of different identified CSIs (Figure 7 and Figure S4).

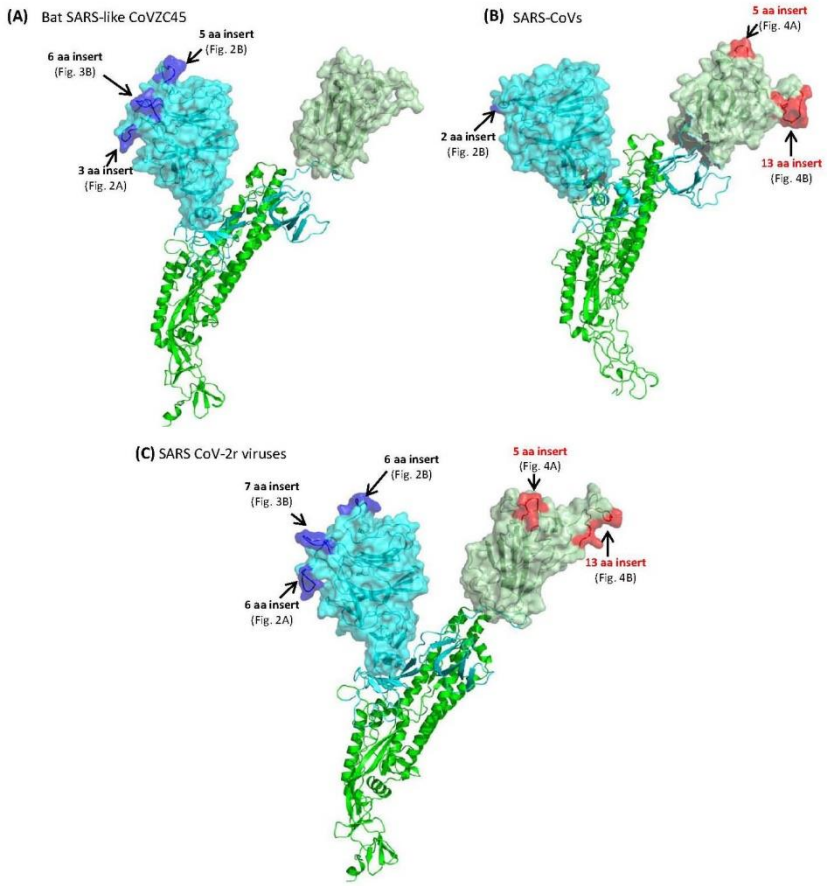

Figure 7. Mapping the surface locations of some of the identified CSIs in the spike protein by homology modeling. (A) Homology models of the spike protein from BatSARS-like-CoVZC45 based on the available structures of SARS-CoV-2 (PDB IDs: 6vsb, and 6m17) and SARS-CoV-2 spike proteins (PDB ID: 6vsb). (B) A cartoon and surface representation of a Cryo-EM structure of the SARS CoV Spike protein (PDB ID: 6acc). (C) A cartoon and surface representation of a spike protein from SARS-CoV-2. The S1-NTD and S1-CTD domains are homology models based on the experimental structure of the SARS-CoV-2 spike protein (PDB ID: 6vsb, and 6m17). In all of these models, the S1-NTD and S1-CTD domains are shown in cyan and dull green color, respectively, whereas the S2 subunit is shown in bright green color. The identified CSIs in the S1-NTD and S2-CTD domains of these CoVs are labeled and shown in blue and red colors, respectively.

 doi:10.20944/preprints202006.0165.v2

In Figure 7, the panel A shows a homology model for the spike protein from Bat SARS-like CoVZC45 virus. The three CSIs that are present in the N-terminal domain of S1 are shown in blue and all of them are located in surface-exposed loops/patches on the S1-NTD of the protein. In CoVZC45 virus, no CSI was identified in the C-terminal domain of S1. In panel B, we show a homology model for the SARS-CoV ExoN1 protein. This protein contains two CSIs (5 aa and 13 aa) in the S1-CTD and a small 2 aa CSI in the S1-NTD. Similar to the localization of CSIs in the CoVZC45 virus, all of the CSIs in SARS-CoV ExoN1 are also present in surface-exposed loops on the protein. The panel C shows a homology model for the spike protein from a SARS-Cov-2r cluster of viruses (Wuhan-Hu-1). These viruses contain all of the CSIs that are present in S1-NTD of CoVZC45 virus as well as the two CSIs that are found in the S1-CTD of SARS ExoN1 virus. The sequence and structural characteristics of different CSIs in the N- and C-terminal domains indicates that the viruses from the SARS-CoV-2r cluster are chimeric in origin with their S1-NTD domain resembling the CoVZC45 virus and the S1-CTD related to the SARS viruses.

Both SARS-Cov-1 (or SARS) and SARS-Cov-2 viruses bind to the human ACE2 receptor (2, 13, 35, 46). Structural and biochemical studies indicate that the amino acid residues L455, F486, Q493, S494, N501, and Y505 from the RBD of SARS-CoV-2 are major determinants in the binding of spike protein with the human ACE2 (13, 46). Of the two CSIs that are present in the S1-CTD, the 13 aa CSI (❽) lies in between the residues that are critical for binding to the ACE2 receptor and the 5 aa CSI (❼) is also adjoining to the receptor-binding region. We have used homology modeling to examine how the structure of the RBD domain might be altered by the presence or absence of these CSIs. In Figure 8A, we show a cartoon representation of the RBD-ACE2 complex from SARS-CoV-2 virus (PDB ID: 6m17), where the identified 5 aa CSI and 13 aa CSI are highlighted in red and labelled and the RBD domain and ACE2 receptor are shown in cyan and magenta colors.

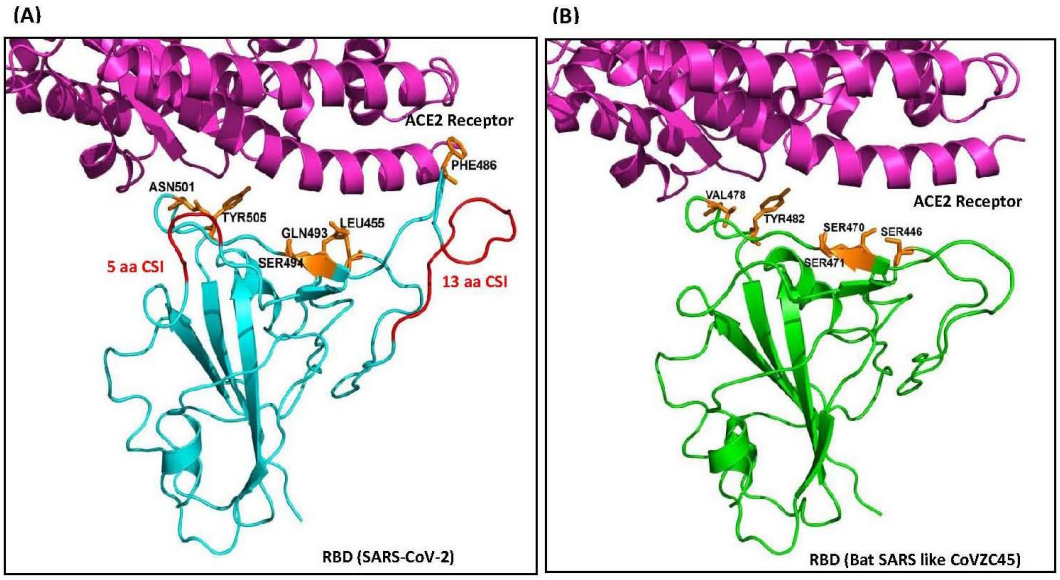

Figure 8. Homology modeling to examine the effects of CSIs (❼and ❽) in the S1-CTD on the binding of spike proteins to the ACE2 receptor. (A) A cartoon representation of RBD-ACE2 complex (PDB ID: 6m17), where RBD domain and ACE2 receptor are shown in cyan and magenta colors, respectively, and the described 5 aa and 13 aa CSIs are highlighted in red and labelled. **(B)** A superimposition of the homology model of RBD of the spike protein (shown as the green cartoon) from Bat SARS-like-CoVZC45, which lacks both 5 aa and 13 aa CSIs, into the RBD domain of RBD-ACE2 complex (PDB ID: 6m17). The amino acid residues in the RBD domain of SARS-Cov-2 that are indicated to be important in its binding to the ACE-2 receptor are shown as orange sticks and marked in **(A)** and the positions of the corresponding residues in SARS-like-CoVZC45 RBD are also labeled.

 doi:10.20944/preprints202006.0165.v2

As has been shown in earlier studies (13, 46), the critical residues in the RBD of SARS-CoV-2 are properly juxtaposed to interact with the ACE2 receptor. The residue Phe486 in SARS-CoV-2, which is located at the end of an extended loop, is reported to form a strong van der Waals contact with residue (Met82) from ACE2 receptor (35). In Figure 8B, we show a superimposition of the homology model of the RBD of spike protein (shown as green cartoon) from Bat SARS-like-CoVZC45 (Accession no: AVP78031), which lack both the 5 aa and 13 aa CSIs that are present in the RBD domain. One important difference one sees in this case is that due to the absence of the 13 aa CSI, the extended loop where the residue Phe486 is found is lacking in this protein structure. Recently, Zhou et al. (41) have also used homology modeling to examine the interaction of the RBD from a Bat-derived CoV-RmYN02, which lacks both the CSIs ❼ and ❽, with the human ACE2 receptor (13, 46). Their results also indicate that the absence of CSI ❽, which is shown as two separate indels in their study, will affect the binding of RBD to the ACE2 receptor (41). Thus, the presence of these CSIs in CoVs plays an important role in their ability to interact with the ACE2 receptor.

The present work has also identified multiple other CSIs that are specific for the CoVs from the deeper-branching *Sarbecovirus*-Lineage-3. The locations of these CSIs in the structure of the spike protein have been mapped and the results from these studies are shown in Figure S4. Similar to the other CSIs in the spike protein, all of these CSIs are also located in surface-exposed loops of the spike protein. These results are in accordance with the results from earlier studies, where most of the studied CSIs have been found in surface-exposed loops of proteins (22, 38, 47, 48), which are known to play important roles in mediating novel protein-protein or protein-ligands interactions (38, 48-50).

Discussion

Both SARS and SARS-CoV-2 viruses are members of the subgenus *Sarbecovirus* (2, 4, 5, 12). The viruses within this subgenus show limited genetic divergence. Due to this, the interrelationships among different sarbecoviruses which, in addition to the SARS and SARS-CoV-2, include a number of other bat/pangolin-CoVs that are not known to infect humans, remain poorly understood (2, 12, 14, 15). Although the evolutionary relationships among sarbecoviruses have been examined in a number of studies by constructing phylogenetic trees, in all of these studies sarbecoviruses form a tightly-linked cluster showing limited resolution among different branches (2, 12, 14). Furthermore, as noted in the Results section, the branching of species in phylogenetic trees is influenced by large numbers of variables including, but not limited to, genetic recombination, which is indicated to be common among the CoVs (7, 8, 10, 31, 40, 41, 47, 48). All of these factors make it difficult to draw reliable inferences regarding the evolutionary relationships among sarbecoviruses by means of phylogenetic analysis. Nonetheless, in all of the phylogenetic studies the SARS-CoV-2 strains form a strongly supported clade with a BatCov-RaTG13 and several CoV isolates from pangolin (2, 12, 14, 17, 31, 47, 48). Nevertheless, the relationship and origin of this cluster (i.e. SARS-CoVr) of CoVs to the other viruses within the subgenus *Sarbecovirus* remains unclear (2, 12, 14, 47).

In the present work we have examined the evolutionary relationships among sarbecoviruses using a sequence-based approach that does not depend upon phylogenetic tree construction. This approach has proven very useful in clarifying several key important relationships which could not be resolved/understood by phylogenetic means (18, 19, 24-26). In this approach, sequences of different proteins are examined to identify inserts or deletions in conserved regions (i.e. CSIs) that are uniquely shared by a specific group of organisms/viruses, providing independent means for assessing the interrelationships among these species. The molecular markers such as CSIs, which are specific for a particular group of organisms, represent rare genetic changes that provide in molecular terms, a reliable means for identification/demarcation of a given clade of organisms. Further, the shared presence of these molecular markers by specific clades provide important means for establishing the evolutionary relationships among different groups of organisms/viruses (18, 19, 24-26). In the present work we have identified many CSIs in the spike and nucleocapsid proteins that reliably demarcate a number of distinct clades of sarbecoviruses and also support specific

 doi:10.20944/preprints202006.0165.v2

explanations about the origin and evolution of SARS-CoV-2r viruses and a pangolin CoV _MP789, which is most closely related to the SARS-CoV-2 virus in its receptor binding domain (14, 16, 31). Some of the key findings from our work and their evolutionary significances are summarized below.

1. We have identified 3 CSIs in the spike protein (❶, ❸ and ❻), where CSIs of specific lengths are only present in all/most viruses from the SARS-CoV-2r cluster (viz. SARS-CoV-2, BatCov-RaTG13 and pangolin CoVs (except strain MP789). These CSIs provide strong evidence, independently of the phylogenetic trees, that the SARS-CoV-2 virus is specifically related to the other viruses in this cluster. The described CSIs provide reliable means for distinguishing this clade from all other coronaviruses.
2. We also describe here 3 CSIs (①, ❸ and ❹), where either identical or similar (but smaller) CSIs are present in the same position in the SARS-CoV-2r cluster of CoVs and two bat SARS-like viruses from the CoVZC cluster (viz. CoVZXC21 and CoVZC45). In phylogenetic trees, the viruses from the CoVZC cluster consistently and form an out group of the SARS-CoV-2r cluster (Figure 1 and (2, 12, 14). Thus, the genetic changes responsible for these three CSIs have likely originated in a common ancestor of these two groups of CoVs.
3. Three CSIs present in the S1-NTD region (❷, ❹ and ❻B) are uniquely shared by the two viruses from the CoVZC cluster and one of the pangolin CoVs strain_MP789, (recently designated as pangolin-CoV-2020)(31). The presence of these three CSIs in the pangolin-CoV_MP789 spike protein and the absence in this pangolin strain of the CSIs (❶, ❸ and ❻A) that are specific for the SARS-CoV-2r cluster (or other pangolin homologs), indicates that the sequence of the spike protein from pangolin CoV_MP789 is unusual.
4. Two CSIs present in the S1-CTD (Fig. 4, ❼ and ❽) are commonly shared by all of the viruses from the SARS-CoV-2r cluster (including pangolin-CoV-MP789) and by the *Sarbecovirus* lineage-1 comprising of the SARS and SARS-related viruses. However, both these CSIs are absent in the CoVZC cluster of CoVs. The shared presence of these two CSIs in the SARS and SARS-CoV-2r viruses indicates that in the S1-CTD region where these CSIs are found, the sequence characteristics of the SARS-CoV-2r cluster, including the pangolin CoV_MP789, are similar to the SARS and SARSr viruses.

The findings noted above and other results obtained in this study have revealed two important anomalies in the sequence characteristics of the spike protein from two groups of CoVs, providing important insights into the origin/evolution of SARS-CoV-2r viruses and the pangolin CoV_MP789 virus. In Figure 8, we summarize the nature of these anomalous observations and their implications regarding the origin of the spike protein from these CoVs.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 26 August 2020 doi:10.20944/preprints202006.0165.v2

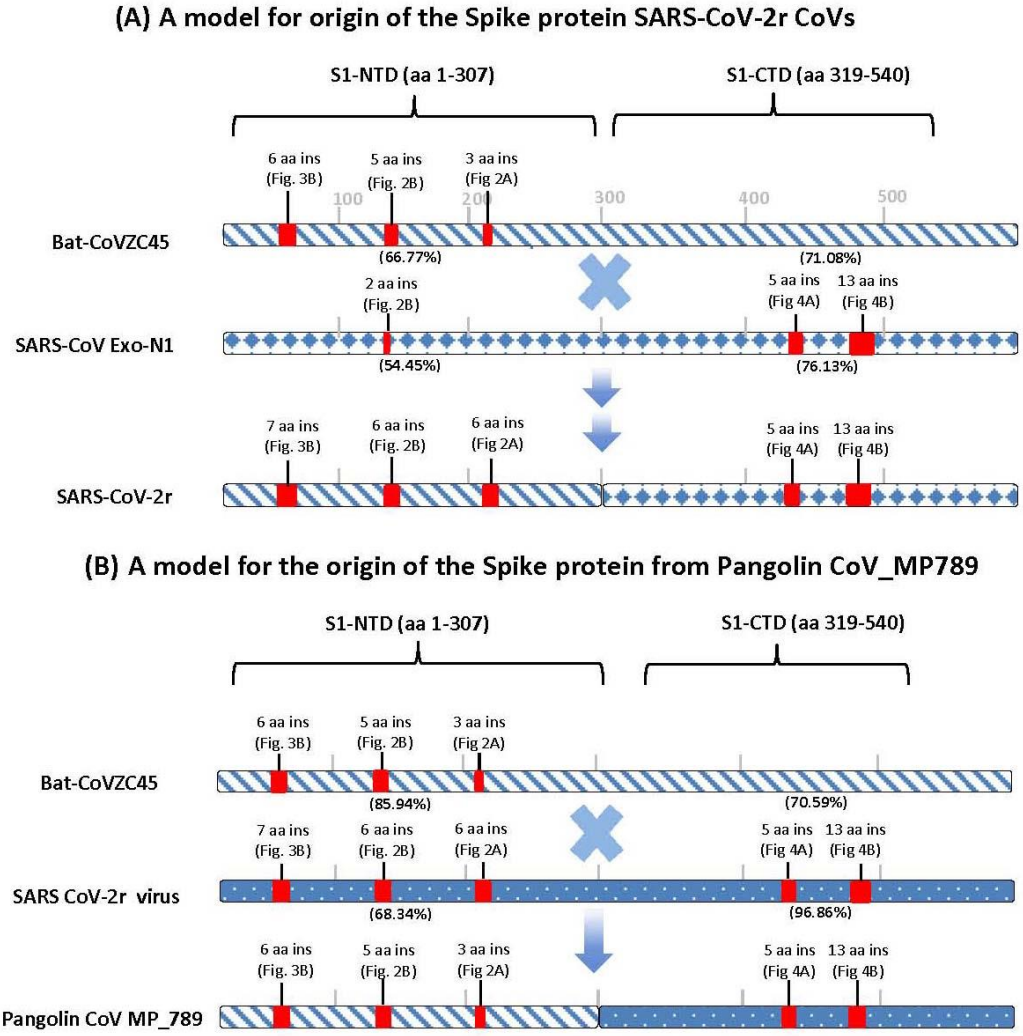

Figure 8. A conceptual diagram summarizing the sequence characteristics of specific lineages of sarbecoviruses and their implications regarding the origin of SARS-CoV-2r and Pangolin CoV_MP789 viruses. (A) The sequence characteristics and distribution pattern of different CSIs in the S1-NTD and S1-CTD domains of SARS-CoVZC45, SARS and SARS-CoV-2r cluster of CoVs and the inference based on them regarding the origin of SARS-CoV-2r viruses. The numbers below the lines in parenthesis indicate the % amino acid sequence identity of the indicated region to the SARS-CoV-2 Wuhan virus. (B) Sequence characteristics of the S1-NTD and S1-CTD domains from SARS-CoVZC45, SARS-CoV-2r and pangolin-CoV-MP789 and a model based upon them for the origin of the pangolin-CoV_MP789 virus. The X in (A) and (B) indicates a genetic recombination event. The numbers below the lines in parenthesis indicate the % amino acid sequence identity of the indicated region to the pangolin-CoV-MP789 virus. The arrows indicate the evolutionary pathway.

The first set of anomalous observations concerns the sequence characteristics of the spike protein from SARS-CoV-2r CoVs (Figure 8A). As noted above, in the S1-NTD region, both SARS-CoV-2r and CoVZC cluster of CoVs contain a number of CSIs in the same position that are related in sequence. Of these CSIs, the sequences of the CSIs ❸ and ❹ are identical except that the CSI ❹ is 1 aa shorter. Most of these CSIs are not found in the S1-NTD of SARS and SARS-related viruses. As the CSIs in protein sequences result from rare genetic changes, and the CoVZC cluster of CoVs form an out group of the SARS-CoV-2r cluster (Figure 1), the shared presence of these three CSIs in the SARS-CoV-2r and CoVZC cluster of CoVs indicates that the genetic changes responsible for these CSIs occurred originally in a common ancestor of these two groups of CoVs. These observations suggest that the S1-NTD region of the SARS-CoV-2r viruses, where these CSIs are found, is derived from CoVZC cluster of CoVs. This inference is also supported by the observation that the S1-NTD

 doi:10.20944/preprints202006.0165.v2

region of the SARS-CoV-2r viruses exhibits maximal sequence similarity (66.77%) to the CoVZC cluster of CoVs. In contrast to these sequence characteristics of the SARS-CoV-2r CoVs, these viruses share two prominent CSIs in the S1-CTD (❼and ❽), that are either exclusively or mainly found in all of SARS and SARS-like CoVs. However, both these CSIs are absent in the CoVZC cluster of CoVs. The shared presence of these two highly conserved CSIs, which are part of the receptor-binding domain of these viruses, by both SARS and SARS-CoV-2r viruses strongly indicates that the S1-CTD region of the SARS-CoV-2r viruses has originated from a SARS-like virus. This inference is corroborated by the observation that the S1-CTD region of SARS-CoV-2 virus exhibits highest sequence similarity (76.13%) to SARS virus and much lower sequence similarity to the CoVZC CoVs. The simplest explanation to account for these contrasting sequence characteristics of the S1-NTD and S1-CTD regions of the SARS-CoV-2r viruses is that this cluster of virus have originated by a recombination event between the S1-NTD region of a CoVZC cluster of CoVs and the C-terminal region (including the S1-CTD) of a SARS-like virus. As the above noted sequence characteristics are commonly shared by all of the SARS-CoV-2r viruses, the postulated recombination event took place in a common ancestor of the SARS-CoV-2r cluster of viruses. This recombination event was followed by other genetic changes in the CSI regions and elsewhere which differentiate the SARS-CoV-2r viruses from the CoVZC and SARS clusters of CoVs. The origin of the SARS-CoV-2r viruses from SARS-related viruses by a recombination event explains why both these groups of viruses, despite their not branching together in phylogenetic trees (Figure 1), contain a similar RBD which in both cases binds to the same human receptor (2, 13, 35, 46-48).

The second set of anomalous observations concerns the pangolin CoV_MP789, which branches with the SARS-CoV-2r cluster in phylogenetic trees (see Figure 1). For all of the CSIs present in the S1-NTD (viz. ❶, ❷, ❸, ❹ and ❻A), the sequence characteristics of the CSIs in pangolin-CoV_MP789 are identical to those found in the CoVZC cluster of CoVs and differ from those found in other pangolin-CoVs as well as other members of the SARS-CoV-2r cluster of viruses. However, this pangolin strain shares the two CSIs present in the S1-CTD (❼ and ❽), which are shared characteristics of the SARS-CoVs and the SARS-CoV-2r cluster of CoVS, but which are absent in the CoVZC cluter of viruses. Furthermore, the amino acid sequences of the CSIs ❼ and ❽ in pangolin CoV_MP789 are identical to those seen for the SARS-CoV-2 virus. The contrasting nature of the pangolin CoV_MP789 sequence is also apparent from the results of pair-wise amino acid sequence similarity on different regions of the spike protein. The S1-NTD region of pangolin CoV_MP789 shows highest similarity (85.94%) to the CoVZC cluster of viruses, while its S1-CTD domain is highly similar (96.86%) to the SARS-CoV-2 virus and shows much lower similarity (68.34%) to the CoVZC45 viruses. To account for the distribution patterns of the identified CSIs and the results of sequence similarity studies, we are postulating that the spike protein from pangolin CoV-MP789 is chimeric resulting from a recombination event between the S1-NTD region from a CoVZC cluster of virus and the C-terminal region of a virus that is most closely related to the SARS-CoV-2 (Figure 8B). In a recent study, Liu et al. (31) have also independently concluded that the pangolin CoV_MP789 (designated as pangolin-CoV-2020 in their study) has resulted from a recombination event involving these two groups of CoVs. In contrast to the recombination event postulated in Figure 8A, which occurred in a common ancestor of the SARS-CoV-2r cluster of CoVs, the recombination event giving rise to pangolin CoV_MP789 is indicated to be more recent. The S1-CTD region of the virus that was involved in this latter recombination event exhibits much higher sequence similarity (96.86%) to the SARS-CoV-2/Wuhan-Hu-1 strain than that observed for the same region for Bat Cov-RaTG13 virus (90.13%), which based upon its overall sequence similarity is indicated to be the most closely related to SARS-CoV-2 (Table 1) (2, 16). Based on this observation, it is expected that pangolin/bat species may harbor yet another CoV that is most similar to the human SARS-CoV-2 in both its S1-NTD and S1-CTD regions.

It is also important to consider the possible functional roles/significances of the identified CSIs. Extensive earlier work on other CSIs provides evidence that the rare genetic changes represented by them are functionally important (38, 49-52). Furthermore, most of the studied CSIs in proteins are

 doi:10.20944/preprints202006.0165.v2

present in surface-exposed loops that are involved in or are predicted to play important roles in mediating novel protein-protein or protein-ligand interactions (38, 49, 53, 54). In the present work, most of the CSIs that we have identified are in the spike protein, which plays a pivotal role in the functioning of CoVs by enabling binding of the virus to its cellular receptors and the subsequent fusion of the virus with the host cell membrane (13, 29, 35, 46). The sequence of the S-protein is highly conserved among sarbecoviruses and the main regions where differences are observed among different clades/lineages of CoVs are where the identified CSIs are found. Thus, the changes represented by the described CSIs signify the sites of major evolutionary changes within the S-protein. Hence, they are predicted to be important in the host tropisms of these viruses and as well as in human infectivity (12, 14).

All of the CSIs identified in the present work in spike protein are located as surface-exposed loops/patches on this protein. Of these CSIs the CSIs ❼ and ❽, which are commonly shared by the SARS and SARS-CoV-2 viruses, are located within the receptor binding domain (RBD) of the S-protein (residues 423-494) and form a significant portion of the RBD (10-12, 45). Based on recent studies (13, 46), a number of amino acids from the RBD of SARS-CoV-2 (viz. L455, F486, Q493, S494, N501, and Y505), which are in close proximity of these CSIs, play critical role in the binding of spike protein from SARS and SARS-CoV-2 to the human ACE2 receptor. Thus, the genetic changes giving rise to these two CSIs play central roles in the ability of the SARS and SARS-CoV-2 viruses to bind to the ACE2 receptor, which facilitates human infection. In the present work, we have also identified three CSIs in the N-terminal domain of spike protein that are specific for the SARS-CoV-2r cluster of viruses. These CSIs also form novel surface-exposed loops (lobes) in the S1-NTD domain of the S-protein (Figure 7). Based on earlier work, both the S1-NTD as well as the S1-CTD domains of CoVs are known to be involved in the binding of the spike protein to different host receptors (4, 28). It is therefore predicted that the novel surface-exposed loops/patches formed by the three SARS-CoV-2r-specific CSIs on the S1-NTD domain of the spike protein should also play important roles in the virulence or other properties of the SARS-CoV-2 viruses. The surface-exposed loops formed by these CSIs may enable novel and specific interactions with other surface-exposed proteins/components in the host cell. Consequently, this could enhance the binding/entry of virus in human cells, as well as greater transmissibility and virulence of the SARS-CoV-2 virus. Thus, it is of importance to test/confirm the functional roles of these CSIs by means of experimental studies.

Our analysis has identified multiple CSIs in the S- and N- proteins that are uniquely found in the *Sarbecovirus* lineage-3. These results suggest that this lineage is undergoing rapid evolutionary changes and its members should possess novel functional characteristics. The members of this lineage also contain the 13 aa insertion (❽) (Figure 4B) in the S1-CTD that is a part of the RBD and is involved in making contact with the ACE2 receptor. Hence, it should be of interest to investigate the members of this lineage for their ability to bind to the ACE2 receptor and its infectivity for human cells.

Lastly, the molecular markers identified in this work, due to their specificity for SARS-CoV-2r viruses and other specific clades of sarbecoviruses provide important means for developing novel diagnostic tests for the identification of different lineages of sarbecoviruses. These tests can be based on different commonly used techniques viz. PCR-based, q-PCR-based, pyrosequencing, immunological or antibody-based methods, MALDI-TOF, aptamer-based methods, as well as rapid *in silico* identification of the viruses in genomic and metagenomic sequences by means of BLAST searches. In earlier work, the CSIs have been used for developing novel and highly-specific diagnostic tests for a number of important bacterial pathogens (19, 23, 55). Additionally, the identified CSIs in the spike protein, due to their predicted functional importance, also provide potential targets for development of novel therapeutics targeting these viruses (27, 56).

Supplementary Materials: The following are available online at www.mdpi.com/xxx/s1, **Figure S1.** A maximum likelihood tree based on spike protein sequences showing the branching of different strains of sarbecoviruses; **Figure S2**. A maximum likelihood tree based on RNA dependent RNA polymerase protein

sequences showing branching of different strains of sarbecoviruses; **Figure S3**. A multiple sequence alignment of the spike protein from pangolin CoV_MP789 and representative CoV strains from SARS-CoV-2 and CoVZC clusters showing the chimeric nature of pangolin CoV_MP789 sequence. **Figure S4**. Homology model of the spike protein from SARS-related coronavirus BtKY72 showing the locations of three CSIs specific for the *Sarbecovirus*-lineage 3

Author Contributions: For research articles with several authors, a short paragraph specifying their individual contributions must be provided. The following statements should be used "Conceptualization, RSG. methodology, BK and RSG; software, BK and RSG; validation, BK and RSG.; formal analysis, BK and RSG.; investigation, BK and RSG; resources, RSG; data curation, BK and RSG; writing—original draft preparation, RSG; writing—review and editing, RSG and BK; visualization, BK and RSG.; supervision, RSG; project administration, RSG; funding acquisition, RSG. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Discovery Grant number RGPIN-2019-06397 from the Natural Science and Engineering Research Council of Canada awarded to Radhey S. Gupta.

Acknowledgments: We thank Drs. Herb Schellhorn and Anjalee Gupta for their reading of the manuscript and many helpful comments to enhance the clarity of the presented work.

Conflicts of Interest: The authors declare no conflict of interest.

Reference List

1. Gorblalenya, A. E. e. al. C. S. G. o. t. I. C. o. T. o. V. (2020) The species Severe acute respitatory syndrome-related coronavirus: classifying 2019-nCoV and naming ti SARS-CoV-2. *Nature Microbiol* **5**, 536-544.
2. Zhou, P., Yang, X. L., Wang, X. G., Hu, B., Zhang, L., Zhang, W., Si, H. R., Zhu, Y., Li, B., Huang, C. L. *et al.* (2020) A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature* **579**, 270-273.
3. Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., Zhao, X., Huang, B., Shi, W., Lu, R. *et al.* (2020) A Novel Coronavirus from Patients with Pneumonia in China, 2019. *N. Engl. J Med.* **382**, 727-733.
4. Cui, J., Li, F. & Shi, Z. L. (2019) Origin and evolution of pathogenic coronaviruses. *Nat. Rev. Microbiol* **17**, 181-192.
5. Peeri, N. C., Shrestha, N., Rahman, M. S., Zaki, R., Tan, Z., Bibi, S., Baghbanzadeh, M., Aghamohammadi, N., Zhang, W. & Haque, U. (2020) The SARS, MERS and novel coronavirus (COVID-19) epidemics, the newest and biggest global health threats: what lessons have we learned? *Int J Epidemiol.*
6. Forni, D., Cagliani, R., Clerici, M. & Sironi, M. (2017) Molecular Evolution of Human Coronavirus Genomes. *Trends Microbiol* **25**, 35-48.
7. Holmes, E. C. & Rambaut, A. (2004) Viral evolution and the emergence of SARS coronavirus. *Philos. Trans. R Soc. Lond B Biol. Sci.* **359**, 1059-1065.
8. Wong, A. C. P., Li, X., Lau, S. K. P. & Woo, P. C. Y. (2019) Global Epidemiology of Bat Coronaviruses. *Viruses* **11**, 174.
9. Quan, P. L., Firth, C., Conte, J. M., Williams, S. H., Zambrana-Torrelio, C. M., Anthony, S. J., Ellison, J. A., Gilbert, A. T., Kuzmin, I. V., Niezgoda, M. *et al.* (2013) Bats are a major natural reservoir for hepaciviruses and pegiviruses. *Proc. Natl. Acad. Sci. U. S A* **110**, 8194-8199.
10. Anthony, S. J., Johnson, C. K., Greig, D. J., Kramer, S., Che, X., Wells, H., Hicks, A. L., Joly, D. O., Wolfe, N. D., Daszak, P. *et al.* (2017) Global patterns in coronavirus diversity. *Virus Evol* **3**, vex012.

11. Zheng, J. (2020) SARS-CoV-2: an Emerging Coronavirus that causes a Global Threat. *Int. J. Biol. Sci.* **16**, 1678-1685.
12. Lu, R., Zhao, X., Li, J., Niu, P., Yang, B., Wu, H., Wang, W., Song, H., Huang, B., Zhu, N. *et al.* (2020) Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. *Lancet* **395**, 565-574.
13. Wan, Y., Shang, J., Graham, R., Baric, R. S. & Li, F. (2020) Receptor Recognition by the Novel Coronavirus from Wuhan: an Analysis Based on Decade-Long Structural Studies of SARS Coronavirus. *J Virol.* **94**.
14. Zhang, T., Wu, Q. & Zhang, Z. (2020) Probable Pangolin Origin of SARS-CoV-2 Associated with the COVID-19 Outbreak. *Curr. Biol.* **30**, 1346-1351.
15. Zhang, Y. Z. & Holmes, E. C. (2020) A Genomic Perspective on the Origin and Emergence of SARS-CoV-2. *Cell* **181**, 223-227.
16. Andersen, K. G., Rambaut, A., Lipkin, W. I., Holmes, E. C. & Garry, R. F. (2020) The proximal origin of SARS-CoV-2. *Nat. Med.* **26**, 450-452.
17. Xiao, K., Zhai, J., Feng, Y., Zhou, N., Zhang, X., Zou, J. J., Li, N., Guo, Y., Li, X., Shen, X. *et al.* (2020) Isolation of SARS-CoV-2-related coronavirus from Malayan pangolins. *Nature*.
18. Baldauf, S. L. & Palmer, J. D. (1993) Animals and fungi are each other's closest relatives: congruent evidence from multiple proteins. *Proc. Natl. Acad. Sci. U. S A* **90**, 11558-11562.
19. Gupta, R. S. (2014) in *Methods in Microbiology*

New Approaches to Prokaryotic Systematics, eds. Goodfellow, M., Sutcliffe, I., & Chun, J. (Academic Press, pp. 153-182.

20. Gupta, R. S. (2016) Molecular signatures that are distinctive characteristics of the vertebrates and chordates and supporting a grouping of vertebrates with the tunicates. *Mol. Phylogenet. Evol* **94**, 383-391.
21. Sharma, R. & Gupta, R. S. (2019) Novel Molecular Synapomorphies Demarcate Different Main Groups/Subgroups of Plasmodium and Piroplasmida Species Clarifying Their Evolutionary Relationships. *Genes (Basel)* **10**, 490.
22. Khadka, B., Chatterjee, T., Gupta, B. P. & Gupta, R. S. (2019) Genomic Analyses Identify Novel Molecular Signatures Specific for the Caenorhabditis and other Nematode Taxa Providing Novel Means for Genetic and Biochemical Studies. *Genes (Basel)* **10**, 739.
23. Wong, S. Y., Paschos, A., Gupta, R. S. & Schellhorn, H. E. (2014) Insertion/deletion-based approach for the detection of Escherichia coli O157:H7 in freshwater environments. *Environ. Sci. Technol* **48**, 11462-11470.
24. Gupta, R. S. (2016) Impact of genomics on the understanding of microbial evolution and classification: the importance of Darwin's views on classification. *FEMS Microbiol Rev.* **40**, 520-553.
25. Rokas, A. & Holland, P. W. (2000) Rare genomic changes as a tool for phylogenetics. *Trends Ecol. Evol* **15**, 454-459.
26. Springer, M. S., Stanhope, M. J., Madsen, O. & de Jong, W. W. (2004) Molecules consolidate the placental mammal tree. *Trends Ecol. Evol* **19**, 430-438.
27. Gupta, R. S. (2018) Impact of Genomics on Clarifying the Evolutionary Relationships amongst Mycobacteria: Identification of Molecular Signatures Specific for the Tuberculosis-Complex of Bacteria with Potential Applications for Novel Diagnostics and Therapeutics. *High Throughput.* **7**, 31.

28. Li, F. (2016) Structure, Function, and Evolution of Coronavirus Spike Proteins. *Annu. Rev. Virol.* **3**, 237-261.
29. Grove, J. & Marsh, M. (2011) The cell biology of receptor-mediated virus entry. *J Cell Biol.* **195**, 1071-1082.
30. Lam, T. T., Jia, N., Zhang, Y. W., Shum, M. H., Jiang, J. F., Zhu, H. C., Tong, Y. G., Shi, Y. X., Ni, X. B., Liao, Y. S. *et al.* (2020) Identifying SARS-CoV-2-related coronaviruses in Malayan pangolins. *Nature* **583**, 282-285.
31. Liu, P., Jiang, J.-Z., Wan, X.-F., Hua Y, Li, L., Zhou, J., Wang, X., Hou, F., Chen, J., Zou, J. *et al.* (2020) Are pangolins the intermediate host of the 2019 novel Coronavirus (SARS-CoV-2)? *PLOS Pathogens* **16**, e1008421.
32. Tamura, K., Stecher, G., Peterson, D., Filipski, A. & Kumar, S. (2013) MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Mol. Biol. Evol* **30**, 2725-2729.
33. Talavera, G. & Castresana, J. (2007) Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst Biol.* **56**, 564-577.
34. Sievers, F. & Higgins, D. G. (2014) Clutal Omega, Accurate Alignment of Very Large Numbers of Sequences. *Methods Mol. Biol.* **1079**, 105-116.
35. Yan, R., Zhang, Y., Li, Y., Xia, L., Guo, Y. & Zhou, Q. (2020) Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2. *Science* **367**, 1444-1448.
36. Wrapp, D., Wang, N., Corbett, K. S., Goldsmith, J. A., Hsieh, C.-L., Abiona, Q., Graham, B. S. & McLellan, J. S. (2020) Cryo-EM structure of the 2019-nCoV spike in the perfusion conformation. *Science* **367**, 1260-1263.
37. Eswar, N., Webb, B., Marti-Renom, M. A., Madhusudhan, M. S., Eramian, D., Shen, M. Y., Pieper, U. & Sali, A. (2007) Comparative protein structure modeling using MODELLER. *Curr. Protoc. Protein Sci.* **Chapter 2**, Unit.
38. Khadka, B. & Gupta, R. S. (2017) Identification of a conserved 8 aa insert in the PIP5K protein in the Saccharomycetaceae family of fungi and the molecular dynamics simulations and structural analysis to investigate its potential functional role. *Proteins* **85**, 1454-1467.
39. Baldauf, S. L. (2003) Phylogeny for the faint of heart: a tutorial. *Trends Genet.* **19**, 345-351.
40. Woo, P. C. Y., Huang, Y., Lau, S. K. P. & Yuen, K.-Y. (2020) Coronavirus Genomics and Bioinformatics Analysis. *Viruses* **2**, 1804-1820.
41. Zhou, H., Chen, X., Hu, T., Li, J., Song, H., Liu, Y., Wang, P., Liu, D., Yang, J., Holmes, E. C. *et al.* (2020) A Novel Bat Coronavirus Closely Related to SARS-CoV-2 Contains Natural Insertions at the S1/S2 Cleavage Site of the Spike Protein. *Curr. Biol.* **30**, 2196-2203.
42. Gupta, R. S., Lo, B. & Son, J. (2018) Phylogenomics and Comparative Genomic Studies Robustly Support Division of the Genus Mycobacterium into an Emended Genus Mycobacterium and Four Novel Genera. *Front Microbiol* **9**, 67.
43. Bapteste, E. & Philippe, H. (2002) The potential value of indels as phylogenetic markers: position of trichomonads as a case study. *Mol. Biol. Evol.* **19**, 972-977.
44. Gupta, R. S. (1998) Protein phylogenies and signature sequences: A reappraisal of evolutionary relationships among archaebacteria, eubacteria, and eukaryotes. *Microbiol. Mol. Biol. Rev.* **62**, 1435-1491.
45. Rivera, M. C. & Lake, J. A. (1992) Evidence that eukaryotes and eocyte prokaryotes are immediate relatives. *Science* **257**, 74-76.

46. Shang, J., Ye, G., Shi, K., Wan, Y., Luo, C., Aihara, H., Geng, Q., Auerbach, A. & Li, F. (2020) Structural basis of receptor recognition by SARS-CoV-2. *Nature* **581**, 221-224.
47. Boni, M. F., Lemey, P., Jiang, X., Lam, T. T., Perry, B. W., Castoe, T. A., Rambaut, A. & Robertson, D. L. (2020) Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. *Nat. Microbiol*.
48. Li, X., Giorgi, E. E., Marichannegowda, M. H., Foley, B., Chun, X., Kong, X.-P., Chen, Y., Gnanakaran, B., Korber, B. & Gao, F. (2020) Emergence of SARS-CoV-2 through Recombination and Strong Purifying Selection. *Sci. Adv.* **6**.
49. Singh, B. & Gupta, R. S. (2009) Conserved inserts in the Hsp60 (GroEL) and Hsp70 (DnaK) proteins are essential for cellular growth. *Mol. Genet. Genomics* **281**, 361-373.
50. Alnajar, S., Khadka, B. & Gupta, R. S. (2017) Ribonucleotide reductases from Bifidobacteria contain multiple conserved indels distinguishing them from all other organisms: *In silico* analysis of the possible role of a 43 aa bifidobacteria-specific insert in the Class III RNR homolog. *Front. Microbiol.* **8**, Article 1409.
51. Gupta, R. S., Nanda, A. & Khadka, B. (2017) Novel molecular, structural and evolutionary characteristics of the phosphoketolases from bifidobacteria and Coriobacteriales. *PLoS. ONE.* **12**, e0172176.
52. Hassan, F. M. N. & Gupta, R. S. (2018) Novel Sequence Features of DNA Repair Genes/Proteins from Deinococcus Species Implicated in Protection from Oxidatively Generated Damage. *Genes (Basel)* **9**.
53. Akiva, E., Itzhaki, Z. & Margalit, H. (2008) Built-in loops allow versatility in domain-domain interactions: lessons from self-interacting domains. *Proc. Natl. Acad. Sci. U. S. A* **105**, 13292-13297.
54. Hashimoto, K. & Panchenko, A. R. (2010) Mechanisms of protein oligomerization, the critical role of insertions and deletions in maintaining different oligomeric states. *Proc. Natl. Acad. Sci. U. S. A* **107**, 20352-20357.
55. Ahmod, N. Z., Gupta, R. S. & Shah, H. N. (2011) Identification of a *Bacillus anthracis* specific indel in the yeaC gene and development of a rapid pyrosequencing assay for distinguishing *B. anthracis* from the *B. cereus* group. *J. Microbiol. Methods* **87**, 278-285.
56. Nandan, D., Lopez, M., Ban, F., Huang, M., Li, Y., Reiner, N. E. & Cherkasov, A. (2007) Indel-based targeting of essential proteins in human pathogens that have close host orthologue(s): discovery of selective inhibitors for Leishmania donovani elongation factor-1alpha. *Proteins* **67**, 53-64.