 doi:10.20944/preprints202006.0165.v1

Article

Evolutionary Origin of SARS-CoV-2 (COVID-19 virus) and SARS viruses through the identification of Novel Protein/DNA Sequence Features Specific for Different Clades of Sarbecoviruses

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 and its relation to other viruses, protein sequences from sarbecoviruses were analyzed to identify conserved inserts or deletions (termed CSIs) demarcating either particular clusters/lineages of sarbecoviruses or those shared by specific lineages shedding light on their interrelationships. We report several clade-specific CSIs in the spike (S) and nucleocapsid (N) proteins that reliably demarcate distinct sarbecoviruses clades providing important insights into the origin and evolution of SARS-CoV-2. Two CSIs in the N-terminal domain (NTD) of S-protein are uniquely shared by SARS-CoV-2, BatCoV-RaTG13 and most pangolin CoVs (SARS-CoV-2r cluster); another CSI supports a closer relationship of SARS-CoV-2 to BatCov-RaTG13. Three additional CSIs in the NTD are specific for two Bat-SARS-like CoVs (viz. CoVZXC21 and CoVZC45; CoVZC cluster) which form an outgroup of the SARS-CoV-2r cluster. Interestingly, one of the pangolin-CoV-MP789 also shares these CSIs but lack the CSIs specific for the SARS-CoV-2r cluster. The N-terminal sequence (aa 1-320) of the S-protein for pangolin-CoV-MP789 shows highest similarity (85.94%) to the CoVZC cluster, while its C-terminal region including the receptor binding domain (RBD) is most similar (97-98% identity) to the SARS-CoV-2 virus. These observations indicate that the spike protein sequence for the strain MP789 is of chimeric origin. Multiple CSIs described here also distinguish two bat SARS-CoVs strains (BM48-31/BGR/2008 and SARS_BtKY72) from all others. Our work also clarifies that two large CSIs (5 aa and 13 aa) found in the RBD of S-protein are mainly specific for the SARS and SARS-CoV-2r clusters of CoVs. The surface loops formed by these CSIs are predicted to be important in the binding of S-protein with the human ACE-2 receptor. Lastly, we have mapped the locations of different CSIs in the structure of the S-protein. These studies reveal that the three CSIs specific for the SARS-CoV-2r cluster form distinct surface-exposed loops/patches on the S-protein. As the surface-exposed loops play important roles in mediating novel interactions, the novel lobes/patches formed by the SARS-CoV-2-specific CSIs in the spike protein are predicted to play important roles in the interaction of this protein with other surface-exposed components in the host cells thereby enhancing the binding/infectivity of this virus to humans.

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

 doi:10.20944/preprints202006.0165.v1

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 >7.0 million people worldwide resulting in >400,000 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 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-CoV, MERS-CoV 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 of 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), whereas 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). Genome sequence of SARS-CoV-2 indicates that it is most closely related (96% whole genome identity) to a bat CoV (BatCov-RaTG13) (2, 16), followed by 91.02% identity to a virus from pangolins (14, 16). 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). Sequence comparison studies also show that the sequence of the receptor binding domain (RBD) of S-protein from SARS-CoV-2 is more similar to a pangolin homolog than to the S-protein from Bat-RaTG13-CoV (15, 16). In addition, in the sequence alignments of the RBD from S-protein, Lu et al. (12) and Zhang et al. (14) have described 1-2 large indels (inserts or deletions) that are present in the SARS and SARS-CoV-2 viruses, but their evolutionary significance was unclear. While the insights provided by these studies are important, further studies on understanding the differences between SARS-CoV-2 and other viruses 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, whose discovery is facilitated by genome sequences, is comprised of conserved signature indels (insertions/deletions) (CSIs) in genes/proteins that are uniquely shared by a given group of species/viruses (17-21). 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

 doi:10.20944/preprints202006.0165.v1

by conserved regions to ensure that they constitute reliable predictors (17-21). The CSIs in gene/protein sequences generally result from rare genetic changes. Due to the rare and 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 (22-25). Due to the above characteristics, the CSIs have provided important means for understanding evolutionary relationships and for the demarcation of specific groups of organisms in molecular terms (17, 20, 21, 23, 25, 26). The genetic changes represented by CSIs are also indicated to be functionally important for the organisms (18, 27, 28). Further, due to the presence of these changes in conserved regions, they also provide a novel means for development of diagnostic tests (22, 29). 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 coronaviruses genomes to understand the origin and novel molecular features of SARS-CoV-2. Our analyses have identified several novel CSIs in the spike (S) and nucleocapsid (N) proteins. Some of these CSIs are uniquely shared by the SARS-CoV-2, BatCov-RaTG13 and pangolin CoVs (SARS-CoV-2r cluster). Several other CSIs, identified in this work, serve to clearly demarcate distinct clusters (or sub-lineages) of *Sarbecovirus*. Our results also indicate that one of the pangolin CoV (MP789), which shows highest similarity to the S-protein of SARS-CoV-2 in the receptor-binding domain (RBD) is of a chimeric origin, raising interesting questions about this strain. In addition, some of the described CSIs are commonly shared by the SARS-CoV-2r cluster and other specific lineages of *Sarbecovirus* providing important insights into the origin and evolution of SARS-CoV-2. The results presented here show that the identified CSIs provide novel and reliable molecular means for distinguishing SARS-Cov-2 and related viruses from each other and the evolutionary and functional significance of these CSIs are discussed in this work.

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 (30). Poorly aligned regions from the sequence alignments were removed using the Gblocks_0.91b program (31). Maximum-likelihood phylogenetic trees based on the resulting sequence alignments were inferred based on the JTT matrix-based model using MEGA6 (30). 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 (32).

2.2 Identification of CSIs in Protein Sequences

Multiple sequence alignments of S, N and RdRp proteins, created as described above, were examined manually 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 (18, 20, 21). The indels which were not flanked by conserved regions were not further considered as they do 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 investigated in this study. 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 (18) 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 from these clusters.

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

To map the locations of the identified CSIs in protein structures, homology models were created for spike proteins using different available structures as templates. A homology model for the N-terminal domain (NTD) of the S-subunit of the spike protein from SARS-Cov-2/Wuhan-Hu-1 (Acc no: YP_009724390) was created based on the available experimental structure of SARS-Cov-2 spike protein (PDB: 6vsb) (33). Cyro-EM based structure for the receptor binding domain (RBD) of SARS-Cov-2 spike protein (PDB: 6m17, Chain E) was used to map the location of the CSIs in this virus. A homology model of the RBD for BAT SARS-CoVZC45 was also created using the experimental structure of SARS-Cov-2 RBD (PDB ID: 6m17, Chain E). The homology model of NTD of spike protein from SARS-related coronavirus BtKY72 (Acc no: APO40579) was based on the available experimental structure of SARS coronavirus BJ012 spike protein (PDB: 5x58). Homology modeling was carried out using the MODELLER v9.11 program (34) and their stereochemical properties were assessed as described in our earlier work (21, 27).

3. Results

3.1 Phylogenetic Analysis of Betacoronaviruses

The evolutionary relationships amongst the major lineages of coronaviruses have been described previously (4, 6-8). Recent studies have also examined the evolutionary relationship of SARS-CoV-2 to other closely related viruses from the subgenus *Sarbecovirus* (2, 12, 14). 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 and 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 as 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 were generally not resolved (as indicated by the low bootstrap scores for most nodes), it is possible to draw some inferences based on this tree. These include: (i) In accordance with the results from earlier studies, 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 CoV (2, 14). 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 outgroup of the SARS-Cov-2r cluster, indicating a closer relationship of these viral strains to this cluster than other Bat SARS-like CoVs. A close relationship of these two 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. (iii) Within the subgenus *Sarbecovirus*, SARS and SARS-like CoVs form

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 June 2020 doi:10.20944/preprints202006.0165.v1

a separate clade (marked as *Sarbecovirus* lineage B-1) from the clade encompassing SARS-CoV-2r cluster, CoVZC cluster and a cluster grouping the other bat-SARS-CoVs. We have labeled this later clade/cluster consisting of other bat-SARS-CoVs as *Sarbecovirus* lineage B-2.

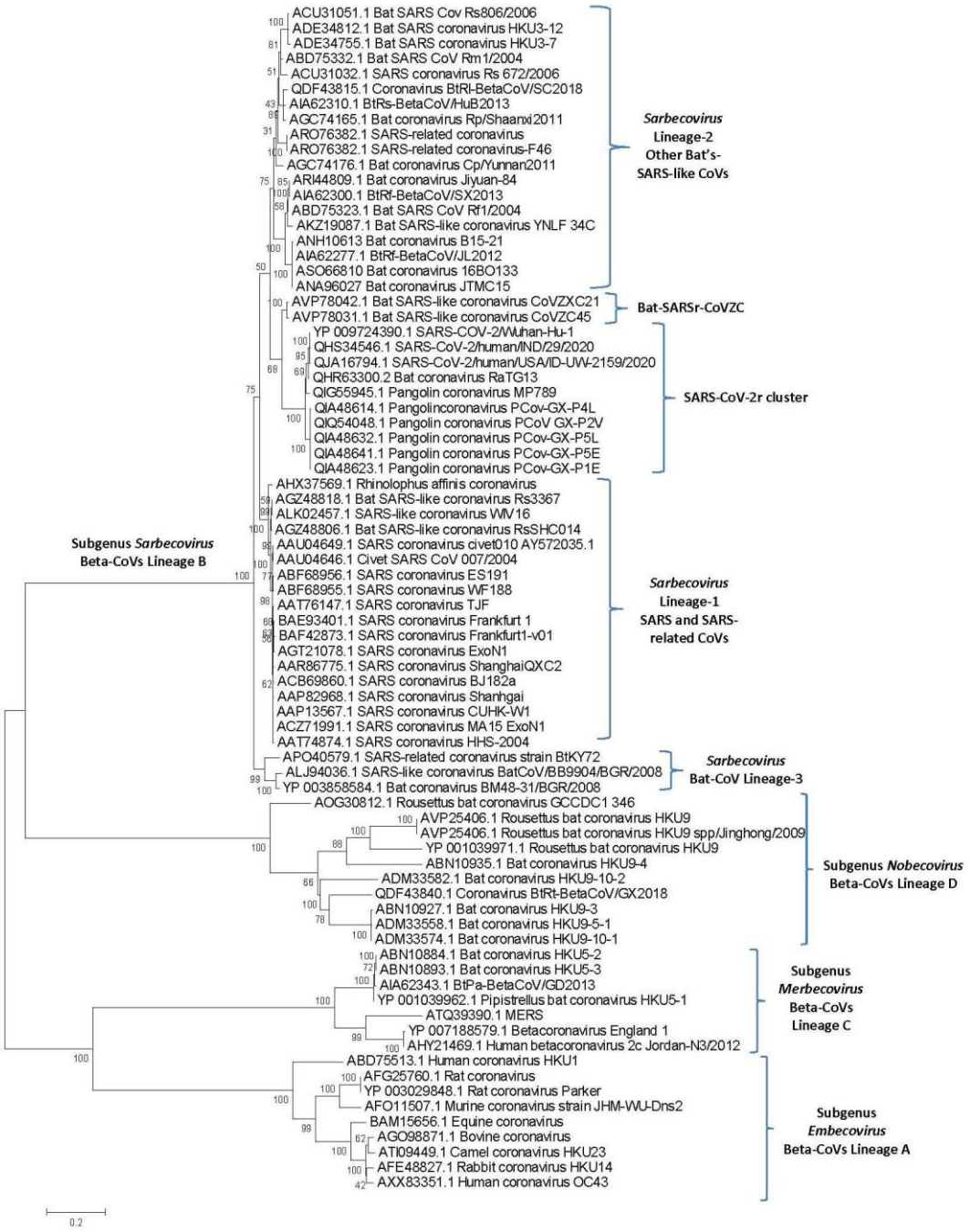

Figure 1. A maximum-likelihood distance tree based on sequence alignment of the spike protein from representative viruses/strains from the genus *Betacoronavirus*. The tree was constructed as described in the Methods section and the % bootstraps for different branches are indicated on the nodes. The 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.

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

 doi:10.20944/preprints202006.0165.v1

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 a useful reference point for interpreting the evolutionary significance of the 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 but not limited to differences in evolutionary rates between different 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, genetic recombination, long-branch attraction effect etc. (35-38) and hence it is often not resolved. Thus, it is important to examine and confirm the inferences from phylogenetic trees by means of other independent approaches not based 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 been proven very useful for identifying evolutionary relationships (17, 20, 21, 23, 25, 39-42). CSIs in genes/proteins sequences result from rare genetic changes and when a CSI of a definite length is present at a specific position within a given group of viruses (organisms), the 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 members of that group (17, 20, 21, 23, 25, 39). 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 many variables that can confound inferences based on phylogenetic trees. Considering these characteristics, the shared presence of CSIs by a given clade of viruses/organisms provide reliable evidence, independent of the phylogenetic trees, of the common ancestry and relatedness of that group of species (43-46). 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 relatively benign CoVs. 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 further insights concerning the evolution of the SARS-CoV-2r cluster from other closely-related viruses. In Figure 2, we present partial sequence alignments from two conserved regions of the S protein where CSIs of specific lengths distinguishing a number of different lineages of *Sarbecovirus* are found. The dashes (-) in these as well as all other sequence alignments shown denote identity with the amino acids on the top line. In the sequence alignment 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 one of the pangolin-CoV strains (MP789). As this CSI is absent in all other *Sarbecovirus*, including members of the deeper branching *Sarbecovirus* lineage-3, this CSI likely originated in a common ancestor of the SARS-CoV-2r cluster (excepting pangolin-CoV_MP789). In the same position where this 6aa insert is found, a 3 aa insert (marked ❷) is present in the two viruses from the CoVZC cluster as well as in the protein homolog from pangolin-CoV_MP789. As noted earlier and seen in Figure 1, the two bat SARS-like CoVs from the CoVZC cluster form an immediate out group of the SARS-CoV-2r cluster. In view of the branching of pangolin-CoV_MP789 with other members of the SARS-CoV-2r cluster in Figure 1, the absence the 6 aa CSI specific for this clade in this viral strain and its sharing of the 3 aa CSI that is specific for the CoVZC cluster is

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 June 2020 doi:10.20944/preprints202006.0165.v1

surprising as these results suggest that this pangolin strain is more closely related to the CoVZC cluster of CoVs.

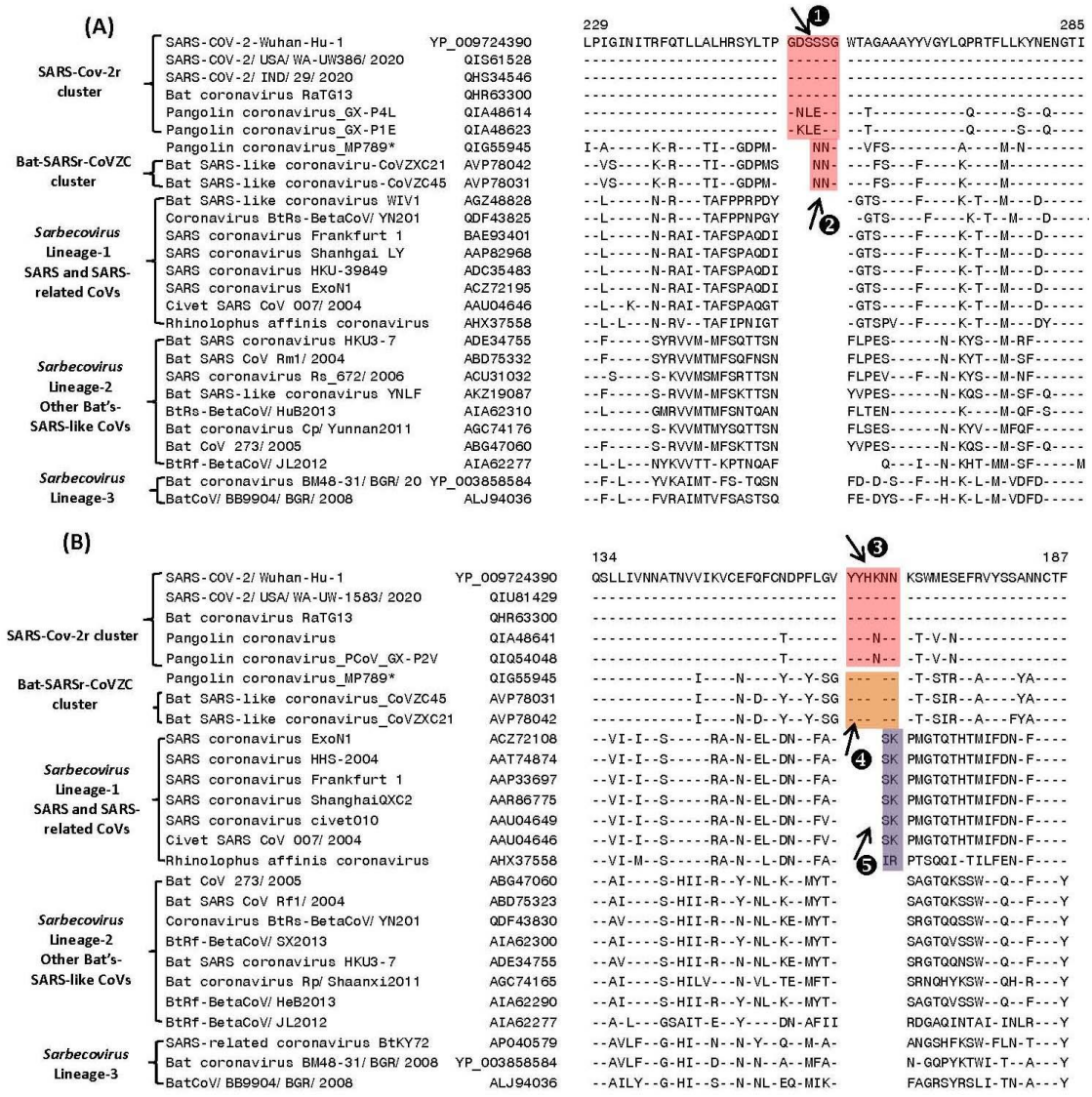

Figure 2. Partial sequence alignments of two conserved regions from the N-terminal domain of spike proteins showing a number of CSIs (highlighted) that are specific for different clades/lineages of *Sarbecovirus*. (A) This panel shows a 6 aa insert (❶) in a conserved region that is present in different viruses from the SARS-CoV-2r cluster, except pangolin-CoV_MP789. In the same position, a 3 aa insert (❷) is commonly shared by CoVs from the CoVZC cluster and pangolin-CoV_MP789. Panel (B) shows a 6 aa insert (❸) in the N-protein that is specific for the SARS-CoV-2r cluster of viruses 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 *Sarbecoviruses* that is comprised of SARS-CoV and related viruses. Dashes (-) in these alignments as well as all other sequence alignments denote identity with the amino acid shown in the top sequence. The numbers on the top indicate the position of these sequences within the indicated proteins. The * for the pangolin CoV-MP789 sequence indicate that its characteristics are anomalous and they are discussed later.

Figure 2B shows sequence alignment from another conserved region of the spike protein from N-terminal domain where CSIs of specific lengths are present in the same position in a number of identified clusters of the sarbecoviruses. In this case, a 6 aa insert (marked as ❸) again shows similar distribution as the CSI ❶ in Figure 2A as this CSI is commonly shared by all viruses from the

 doi:10.20944/preprints202006.0165.v1

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 bat CoVs (viz. CoVZXC21 and CoVZC45CoV) from the CoVZC cluster. Importantly, in this case the sequence of the 5 aa insert in the CoVZC strains and pangolin-CoV_MP789 is very similar (or identical) to that found in other members of the SARS-CoV-2r cluster, except that it is lacking one amino acid residues. The distribution of these two CSIs (i.e. ❸ and ❹) in different CoV strains again strongly indicates that the pangolin-CoV_MP789 strain, despite its branching with the SARS-CoV-2r cluster, is more closely related to the CoVZC cluster of CoVs in this sequence region of the spike protein. The anomalous sequence characteristics of the pangolin-CoV_MP78 are discussed in detail later. It is noteworthy, that in addition to the shared presence of these two CSIs, the sequence surrounding the CSIs ❷ and ❹ are also very similar in CoVZC strains and pangolin-CoV_MP789, providing further evidence of a specific relationship between these CoVs in this region of the spike protein. As the viruses from the CoVZC cluster form an immediate outgroup of the SARS-CoV-2r cluster, the presence of related CSIs but of different lengths (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 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 excluding pangolin-CoV_MP789. 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 specifically found in the B-1 lineage of *Sarbecoviruses* that is comprised of SARS-CoV and related viruses. This 2 aa CSI serves to distinguish the SARS-CoV and related viruses from all other clusters/lineages of sarbecoviruses.

In Figure 3, we present partial sequence alignments from two conserved regions of the N and S proteins showing several additional CSIs that also support and confirm the deduced evolutionary relationships between the members of the SARS-CoV-2r and SARSr-CoVZC clusters of coronaviruses. In the partial sequence alignment of N-protein shown in Figure 3A, a 2 aa deletion (boxed and labeled ①) present in a highly conserved region is commonly shared by all members of the SARS-CoV-2r cluster as well as in the two bat SARSr-CoVs comprising the CoVZC cluster. The shared presence of this CSI in all of the CoVs from the SARS-CoV-2r and SARSr-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 N-terminal domain (NTD) of the spike protein (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 and appears to be artifactual as it is the only sequence that is inconsistent with other CSIs Thus, the sequence of the GX-P2V should be re-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 strongly indicate that 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: 14 June 2020 doi:10.20944/preprints202006.0165.v1

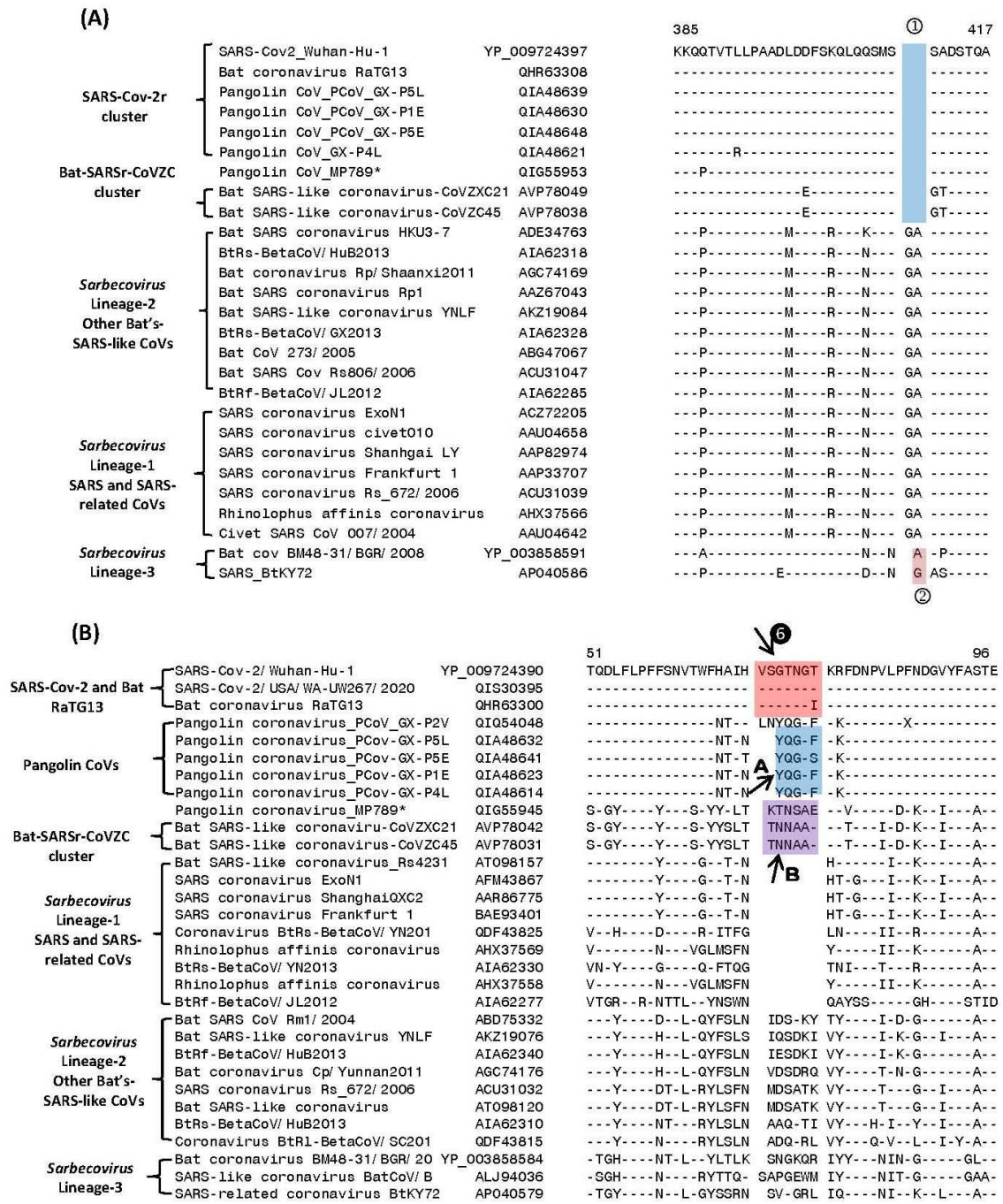

Figure 3. Partial sequence alignments of (A) nucleocapsid and (B) spike protein showing some CSIs that are specific for some clusters/lineages of *Sarbecovirus*. Panel (A) shows a 2 aa deletion (①) in the N-protein that is specific for the SARS-CoV-2r and CoVZC clusters of CoVs. The viruses from *Sarbecovirus* lineage-3 contain only a 1 aa deletion (②) in this position. (A) The CSI indicated as ❻ marks a region 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 (viz. CoVZXC21 and CoVZC45CoV) from the CoVZC cluster as well as the pangolin CoV strain MP789 contain a 6 aa insert in this position rather than the 5 aa or 7 aa acid inserts and the amino acid sequences of these inserts differ from other CoVs. These results again support the inference from the CSIs ❷ and ❹ (Figure 2) that the pangolin-CoV_MP789 in its N-terminal domain is specifically related to the CoVZC cluster of CoVs. In the sequence alignment

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 June 2020 doi:10.20944/preprints202006.0165.v1

shown in Figure 3B, the CoVs from other lineages of *Sarbecovirus* (lineages 2 and 3) also contain a 6-7 aa insert, which based on its sequence is of independent origin. However, the presence of these inserts does not impact the interpretation and significance of the CSIs discussed above.

In Figure 4, we show sequence information for two other CSIs in the spike protein, whose presence was noted in earlier studies (12, 14, 16), but their evolutionary significance and specificity was not clearly specified.

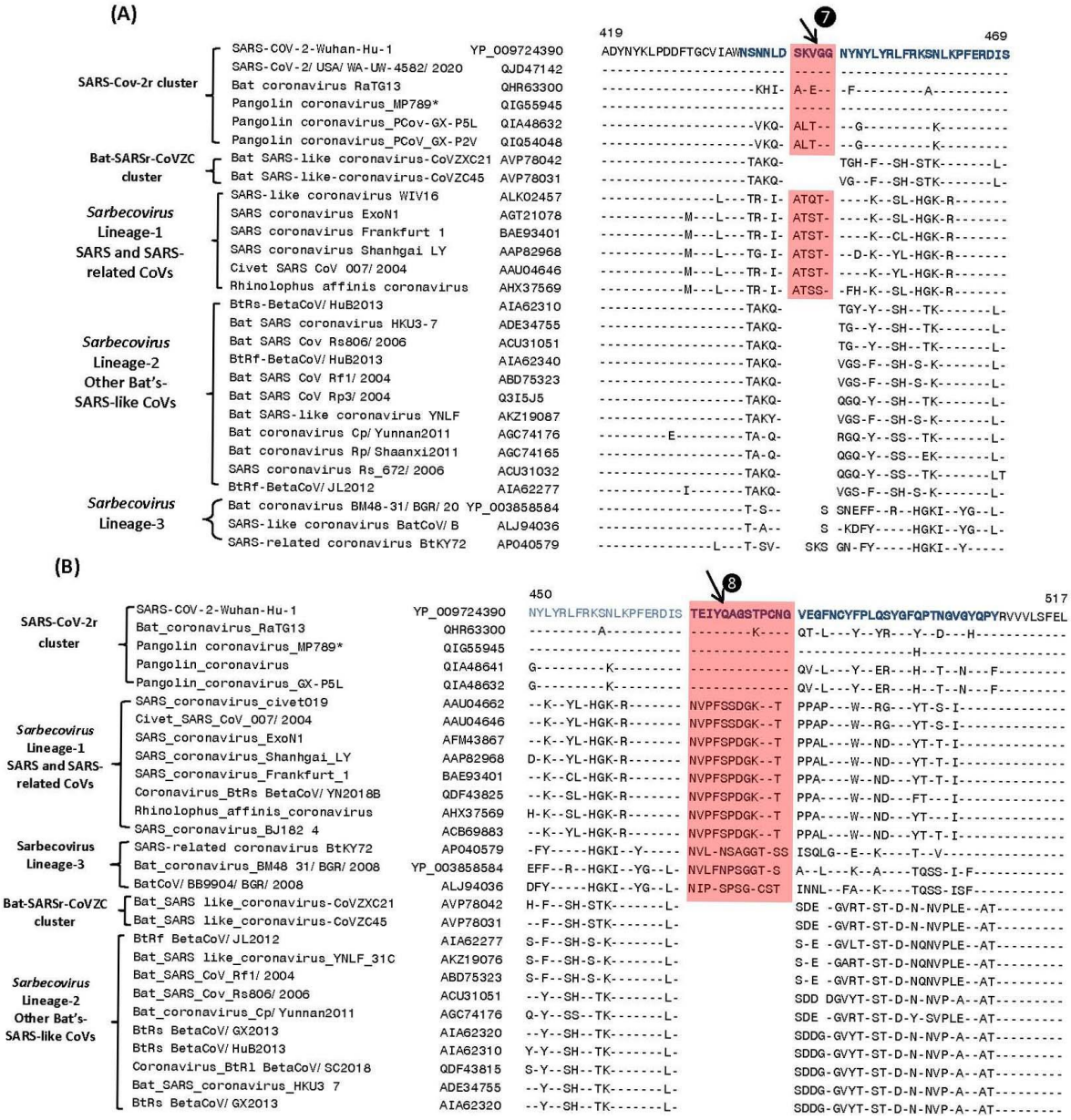

Figure 4. Partial sequence alignment of two conserved regions (overlapping) encompassing the receptor binding domain (RBD) of S-protein, where two large CSIs (❼ and ❽) are mainly found in the SARS and SARS-CoV-2r cluster of viruses. The sequence region that comprises the RBD [11] is shown in dark blue whereas the overlapping parts of the sequence are marked in light blue color. Dashes (-) in sequence alignments denote identity with the amino acid shown in the top sequence.

Both these CSIs are present in close proximity of 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, this CSI is not found in members of the CoVZC cluster or other

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 June 2020 doi:10.20944/preprints202006.0165.v1

bat-SARS-related CoVs comprising the *Sarbecovirus* lineages-2 and 3. Figure 4B shows another large CSI, in this case a 13 aa insertion in the spike protein (marked ❽), which is lacking in members of the CoVZC cluster as well *Sarbecovirus* lineages-2, but is commonly shared by all other sarbecoviruses (see Figure 1). Furthermore, it is also of much interest that both these CSIs are present in all of the pangolin homologs including pangolin-CoV_MP789, but are lacking in the two strains from CoVZC cluster. Additionally, the amino acid sequence of pangolin-CoV_MP789 in this region is highly similar to the SARS-CoV-2 sequence and there is only one amino acid residue that differs between the two sequences (14, 16). The significance of these observations in the evolution of SARS and SARS-CoV-2r clusters of viruses as well as pangolin-CoV_MP789 is discussed in later sections.

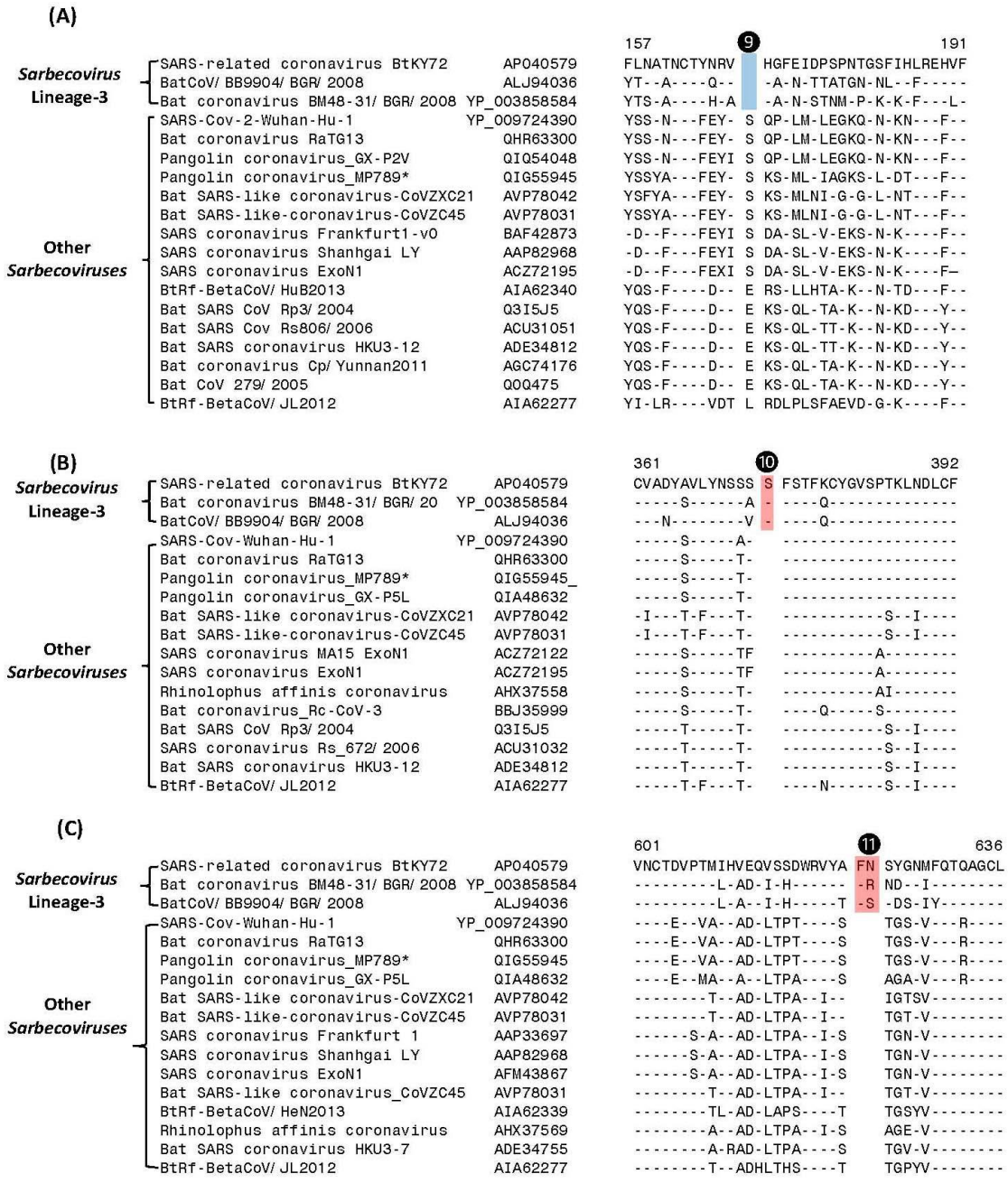

Figure 5. Excerpts from sequence alignment for three separate regions of the spike protein depicting three CSIs, which are specific for the CoVs from *Sarbecovirus* lineages-3.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 June 2020 doi:10.20944/preprints202006.0165.v1

We have also identified multiple 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. 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.

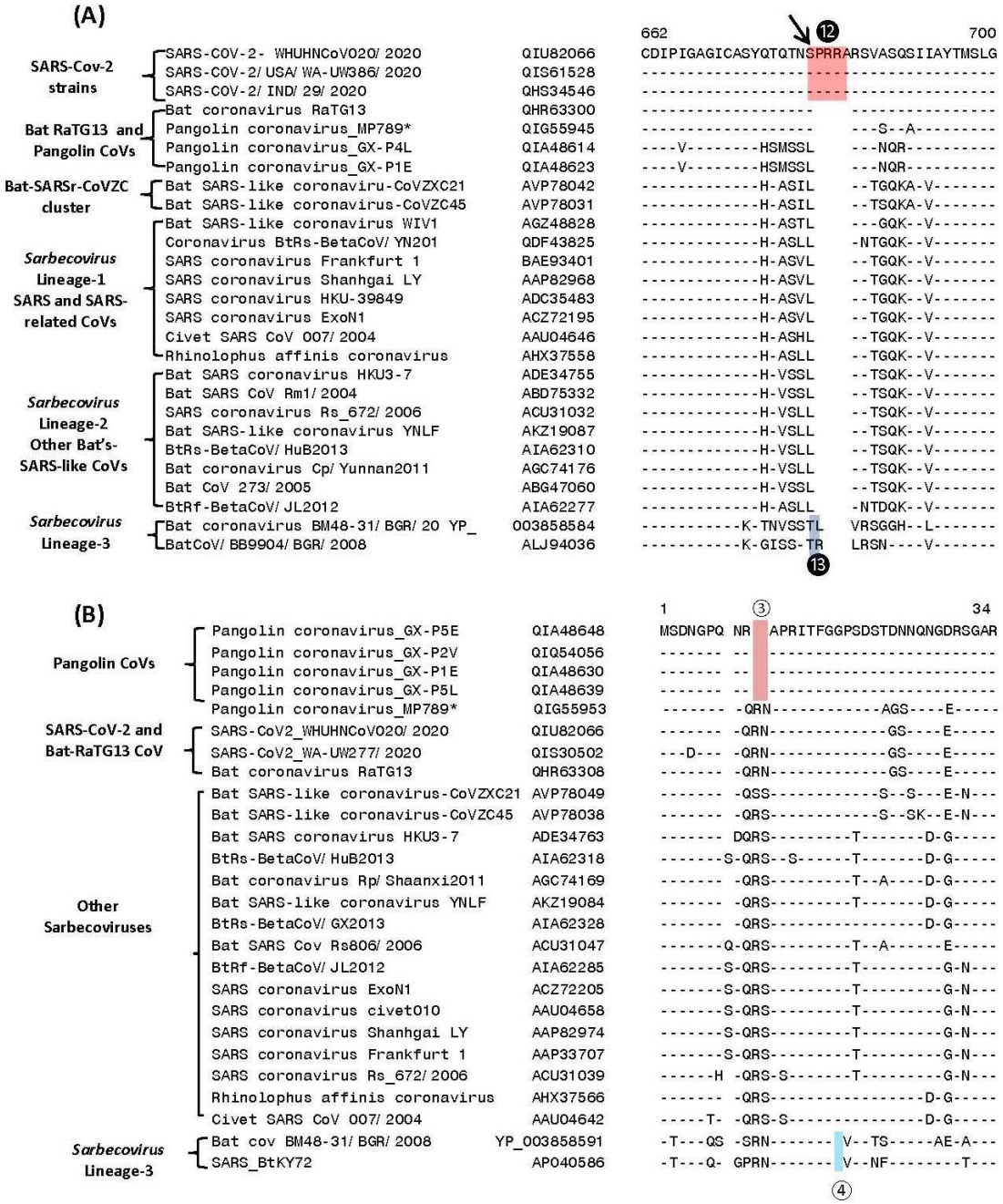

Figure 6. Partial sequence alignments of (A) spike proteins (B) nucleocapsid protein showing a number of CSIs (highlighted) that are specific for different clades/lineages of sarbecoviruses. The panel (A) shows a 4 aa CSI (⓬) specific for the SARS-CoV-2 identified in earlier work [14]. A 1 aa insertion (marked ⓭) is also present in the same position in CoVs from *Sarbecovirus* lineages-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.v1

In Figure 6A, where partial sequence alignment is shown for the S-protein, the larger 4 aa CSI (marked ❾), previously described previously (14, 16), 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* lineages-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 N-protein homolog from Pangolin-CoV-MP789. This provides further evidence suggesting the anomalous nature/annotation of this pangolin CoV strain (MP789), which is inconsistent with other pangolin-CoV. In addition to this CSI, the sequence alignment shown in Figure 6B also contains another CSIs consisting of 1 aa deletion (marked ⑦) that is specific for the lineages-3 of *Sarbecovirus*.

3.3 Sequence Similarity Studies on Pangolin-CoV_MP789

The results presented show that the pangolin-CoV-strain_MP789, which branches with the SARS-CoV-2r cluster in phylogenetic trees (see Figure 1), exhibits anomalous behavior in comparison to the other pangolin CoV strains regarding sharing of CSIs that are specific for different groups/clusters of sarbecoviruses. For all of the CSIs present in the N-terminal region of the spike protein viz. (❷, ❹ and ❻**A**), this strain displays characteristics that are more similar to the CoVZC cluster of CoVs, whereas the other pangolin CoVs share the characteristics of the SARS-CoV-2r cluster of viruses (CSIs ❶, ❸). Additionally, this strain also does not share the CSIs that are specific for the other pangolin CoV strains (viz. ❻**A** and ③). On the other hand, for the two CSIs that are found in the RBD domain that are commonly shared by SARS and SARS-CoV-2r cluster of viruses (❼and ❽), but not found in the CoVZC cluster of CoVs, both these CSIs are present in pangolin-CoV_MP789 as well as in other pangolin CoVs. Furthermore, in the RBD domain, the amino acid sequence of the pangolin-CoV_MP789 displays highest similarity to the SARS-CoV-2. To understand the significance of these observations, we have determined the pairwise amino acid similarity of the pangolin-CoV_MP789 to the SARS-CoV-2r and CoVZC clusters of CoVs. These measurements were done for the entire spike protein and sequence regions constituting its N-terminal domain, RBD as well as the entire C-terminal region. 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 N-terminal 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 and other pangolin CoVs exhibit much lower similarity (67-68%). A more dramatic result is observed when the sequence comparison is made for the RBD (aa 319-540) or for the entire C-terminal region of the spike protein (i.e. from aa 320-1265). In both of these cases, pangolin-CoV_MP789 exhibits the highest similarity to the SARS-CoV-2 (96.86% for RBD and 98.6% for the C-terminal), whereas much lower similarity is observed for the Bat-SARS-like-CoV-CoVZC45 (Table 1). Of particular interest is the observation that the observed similarity of pangolin-CoV_MP789 for the RBD or for the entire C-terminal region to SARS-CoV-2 is higher than that observed for any other CoVs including Bat-CoV-RaTG13. Liu et al. (47) have also reported that the spike protein of pangolin-CoV_MP789 (designated as pangolin-CoV-2020) exhibits higher similarity in the N-terminal region to Bat-CoV-ZC45 and Bat-CoV-ZXC21, whereas its C-terminal sequence is more closely related to Bat-CoV-RaTG13 and SARS-CoV-2 viruses.

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). In this sequence alignment, we have highlighted the positions where a given sequence polymorphism is common to pangolin-CoV_MP789 and CoVZC cluster of viruses (highlighted in

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 June 2020 doi:10.20944/preprints202006.0165.v1

yellow) as well as other positions, where polymorphic substitutions are common to pangolin-CoV_MP789 and SARS-CoV-2 viruses. The results from this comparison/analysis shows that for the first 325 amino acids, in nearly all of the polymorphic sites (>80%), the pangolin-CoV_MP789 contains the same amino acid residues as in CoVZC cluster of CoVs, whereas in the remainder of the protein (aa 330-1265) the polymorphic replacements in pangolin-CoV_MP789 are predominantly (>90%) the same as seen in SARS-CoV-2. Thus, while the first 325 aa in the sequence of spike protein from pangolin-CoV_MP789 bears a close resemblance to the CoVZC cluster of CoVs, the rest of the sequence is most similar to the SARS-CoV-2.

Table 1

Amino Acid Similarity Matrix (% Identity) for Different Regions of Spike Protein

Spike Protein (Full length)					
	SARS-CoV-2 Wuhan	Bat RaTG13	Pangolin MP789	Pangolin GX-P2V	Bat-SARS-like CoVZC45
SARS-COV-2/Wuhan-Hu-1	100				
Bat coronavirus RaTG13	97.71	100			
Pangolin CoV MP789	90.51	89.55	100		
Pangolin CoV GX-P2V	92.43	93.05	87.57	100	
Bat SARS-like CoVZC45	81.93	81.75	87	81.35	100
N-terminal Region (domain)-Spike Protein (aa 1-318)					
SARS-COV-2/Wuhan-Hu-1	100				
Bat coronavirus RaTG13	98.45	100			
Pangolin CoV MP789	68.34	68.65	100		
Pangolin CoV GX-P2V	87.65	88.27	67.40	100	
Bat SARS-like CoVZC45	66.77	67.08	85.94	64.84	100
Receptor-binding Domain Spike Protein (aa 319-540)					
SARS-COV-2/Wuhan-Hu-1	100				
Bat coronavirus RaTG13	90.13	100			
Pangolin CoV MP789	96.86	89.69	100		
Pangolin CoV GX-P2V	86.1	87.44	87.44	100	
Bat SARS-like CoVZC45	71.08	70.59	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			
Pangolin CoV MP789	98.10	96.61	100		
Pangolin CoV GX-P2V	94.07	94.50	94.29	100	
Bat SARS-like CoVZC45	87.04	86.61	87.26	86.72	100

Sequence similarity for different regions of Pangolin-CoV-MP789 for the SARS-CoV-2 and a CoV from the CoVZC cluster (Bat SARS-like CoVZC45) is highlighted.

We have examined the nucleotide sequence of the spike protein from pangolin-CoV_MP789 for any sequence feature that might be present near the junction (near aa 320-330) where a shift in similarity from CoVZC-like sequence to SARS-CoV-2 like sequence occurs. Figure 7 below shows the nucleotide sequence of the S-protein from pangolin-CoV_MP789 from aa 317-336. This sequence was examined for any restriction enzyme site(s) and a Ssp1 restriction site (marked with arrow) is present between aa positions 327-328, which is exactly where the transition from one type of sequence to the other occurs. The significance of this observation will be explained in the Discussion section.

 doi:10.20944/preprints202006.0165.v1

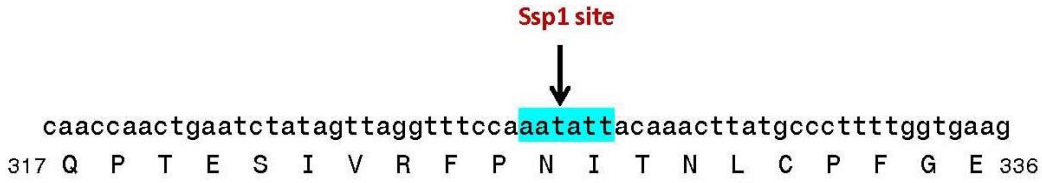

Figure 7. Nucleotide and amino acid sequence of the spike protein from pangolin-CoV_MP789 surrounding the site where sequence changes from CoVZC-like to SARS-CoV-2 like sequence.

3.4 Localizations of the CSIs in Protein Structures

We 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: 6vsb, 6m17, 5x58), homology models were created for both the N-terminal and RBD domain of the spike protein S1 subunit from a number of different clusters of sarbecoviruses (Figure 8 and Figure S4). In the top panels, we show the homology models for the NTD (panel A) and receptor binding domain RBD (panel B) for the S1 subunit of the spike protein from SARS-Cov-2/Wuhan-Hu-1 virus, which contains several of the described CSIs. In the NTD three different CSIs (❶, ❸ and ❻) are found that are specific for the SARS-CoV-2r cluster of CoVs. As seen from Figure 8A, all three of these CSIs are located in surface-exposed loops of the spike protein and they form novel lobes or patches on its surface. Similarly, within the RBD of the spike protein, two large CSIs (5aa and 13 aa) are present (❼ and ❽), which are primarily found in the SARS-CoV-2r cluster of viruses and the *Sarbecovirus* lineage-1 of CoVs, which encompasses SARS and some SARS-related viruses. As seen from Figure 8B, these two CSIs in the RBD also form prominent surface-exposed loops in the protein structure. For a number of CSIs in the S-protein that are specific for the *Sarbecovirus*-Lineage-3, their locations in the structure of the spike protein were also mapped and 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 of earlier studies, where most of studied CSIs have been found in surface-exposed loops of proteins (21, 27, 48, 49), which are known to play important roles in mediating novel protein-protein or protein-ligands interactions (27, 28, 49, 50).

As shown in Figures 4 and 8B, two of the large CSIs (inserts) that are specific for the *Sarbecovirus* lineage-1 and SARS-CoV-2r clusters of CoVs, are located within the RBD of the SARS and SARS-CoV-2 spike protein. This has also been noted in earlier studies (10-12, 45). It is also now known that the human ACE2 protein serves as the cellular receptor for both SARS-Cov-1 and SARS-Cov-2 viruses (2, 13, 51, 52). These studies also reveal 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 RBD of S-protein to human ACE2 (13, 51). As the 13 aa CSI lies in between the residues that are critical for binding to the ACE2 receptor and the 5 aa CSI 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 lack of these CSIs, which are not found in several other lineages/clusters of CoVs. In Figure 8C, we show a cartoon representation of RBD-ACE2 complex from SARS-CoV-2 virus (PDB ID: 6m17), where the identified 5 aa CSI and 13 aa CSI are highlighted red and labelled and the RBD domain and ACE2 receptor are shown as cyan and magenta. As reported earlier, the critical residues in the RBD 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, has been suggested to result in the formation of a stronger van der Waals contact with residue (Met82) from ACE2 receptor (52). 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 lacks both 5 aa and 13 aa CSIs, into the RBD domain of RBD-ACE2 complex (PDB ID: 6m17, chain E). One important difference seen in the RBD structure of this protein 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. As these CSIs form a significant part of the spike protein's RBD, the absence of these CSIs in CoVs that are lacking them is expected to have an important effect on their ability to interact with the ACE2 receptor.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 June 2020 doi:10.20944/preprints202006.0165.v1

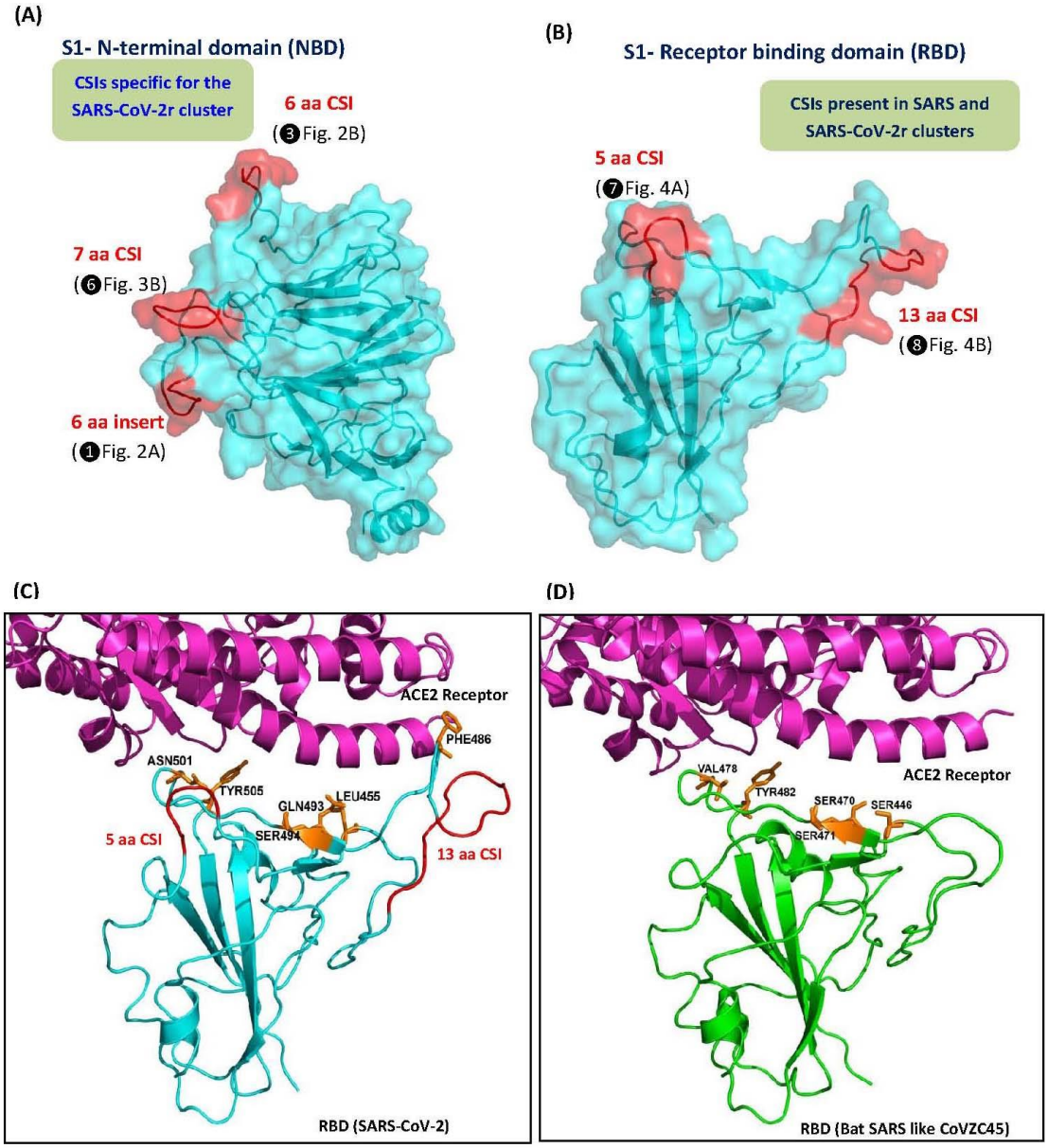

Figure 8. Mapping the surface locations of some of the CSIs described in this work using the available structures and homology models of the spike proteins. **(A)** The left panel shows a homology model of S1 N-terminal domain (NTD) from SARS-Cov-2/Wuhan-Hu-1 based on experimental structure of SARS-Cov-2 spike protein (PDB ID: 6vsb), and **(B)** a cyro-EM based structure of SARS-Cov-2 receptor binding domain (RBD) (PDB ID: 6m17, chain E). Both these domains are shown as cyan cartoon and transparent surface. The locations of the identified CSIs in NTD and RBD are highlighted red and labelled. **(C)** 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 red and labelled. **(D)** Superimposition of the homology model of RBD of spike protein (shown as the green cartoon) from Bat SARS-like-CoVZC45 (Acc no: AVP78031) (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 **(C)** and the positions of the corresponding residues in SARS-like-CoVZC45 RBD are also labeled.

Discussion

 doi:10.20944/preprints202006.0165.v1

The coronaviruses responsible for the SARS and COVID-19 epidemics are both members of the subgenus *Sarbecovirus* within the subfamily *Coronavirinae* (2, 4, 5, 12). The viruses within the *Coronavirinae* subfamily exhibit enormous diversity in terms of their genomic structures, natural hosts, pathogenicity as well as cellular receptors (4, 6, 10). However, the viruses within the subgenus *Sarbecovirus* show limited genetic divergence and the interrelationships amongst different members, which include SARS-CoV, SARS-CoV-2 as well as other bat-CoVs not known to infect humans, are not reliably resolved (2, 12, 14, 15). Thus, in terms of understanding the origin of SARS-CoV-2 (and also SARS-CoV) from other closely-related viruses that are relatively benign, it is important to understand the evolutionary relationships amongst different sarbecoviruses. Recently, phylogenetic trees have been constructed for sarbecoviruses based on different datasets (2, 12, 14). In all of these studies, the sarbecoviruses form a tightly-linked cluster with limited resolution of different branches. While all of these studies show that the SARS-CoV-2 strains form a distinct and strongly supported clade with a BatCov-RaTG13 and a virus from pangolin, the relationship of this cluster (i.e. SARS-CoVr) of viruses to other viruses within the subgenus *Sarbecovirus* remain unclear (2, 12, 14). Furthermore, while the SARS-CoV-2 has clearly originated from the other two viruses in the SARS-CoVr cluster, aside from the distinct grouping of these viruses in phylogenetic trees, no other known characteristic is uniquely shared by the viruses from this cluster.

Thus, the aim of this study was to examine the evolutionary relationships among sarbecoviruses using a sequence-based approach that does not depend upon construction of phylogenetic trees. This approach has proven useful in clarifying several key important relationships which could not identified by phylogenetic means (17, 18, 23-25). In this approach sequences of different proteins are examined to identify inserts or deletions (indels) in conserved regions that are uniquely shared by a specific group of organisms, providing independent means for assessing the interrelationships among these species. As noted earlier, the molecular markers such as CSIs, due to their specificity for a particular group of organisms, represent synapomorphic characteristics and provide a reliable means for identification/demarcation of different clades of organisms in molecular terms and establishing evolutionary relationships among them (17, 18, 23-25). The results from these studies have identified many CSIs in the spike and nucleocapsid proteins that reliably demarcate a specific clade of sarbecoviruses and support a specific explanation for s the origin and evolution of SARS and SARS-CoV-2 viruses. A summary diagram showing the clade specificities of the identified CSIs is presented in Figure 9 and their evolutionary significance and key derived inferences 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 additional evidence that the SARS-CoV-2 is specifically related to these two viruses and the described CSIs provide reliable means for distinguishing this clade from all other coronaviruses. Further, the CSI in Fig. 3B, where a 7 aa insert is only found in the SARS-CoV-2 and BatCov-RaTG13 (except a sequence from pangolin CoV which is likely anomalous) provide evidence that BatCov-RaTG13 virus is more closely-related to SARS-CoV-2 than the pangolin CoVs. The genetic changes responsible for these CSIs likely occurred in a common ancestor of the SARS-CoV-2r cluster or in an ancestor of SARS-CoV-2 and BatCov-RaTG13 viruses. These inferences are also supported by the branching of these viruses in phylogenetic trees. Further, the species distributions of the identified CSIs also strongly suggest that these CSIs represent inserts in these lineages rather than deletions in all other sarbecoviruses. We have also identified two CSIs (❻**A** and ③) which are specific for the pangolin CoVs (except strain MP789) distinguishing them from the SARS-CoV-2 and BatCov-RaTG13 viruses.

2. We describe here 2 CSIs (① 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

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 June 2020 doi:10.20944/preprints202006.0165.v1

viruses (viz. CoVZXC21 and CoVZC45). In phylogenetic trees, these two bat viruses consistently form an outgroup of the clade comprising of the SARS-CoV-2r cluster (Fig. 1 and (2, 12, 14)). The genomes of these viruses show 88% identity to the SARS-CoV-2 genome and only BatCov-RaTG13 and pangolin CoVs exhibit higher similarity to the SARS-CoV-2 (11, 12, 47). Thus, based on their branching position, whole genome sequence identity and the shared presence of these two CSIs, these two clusters of viruses have shared a common ancestor exclusive of other sarbecoviruses.

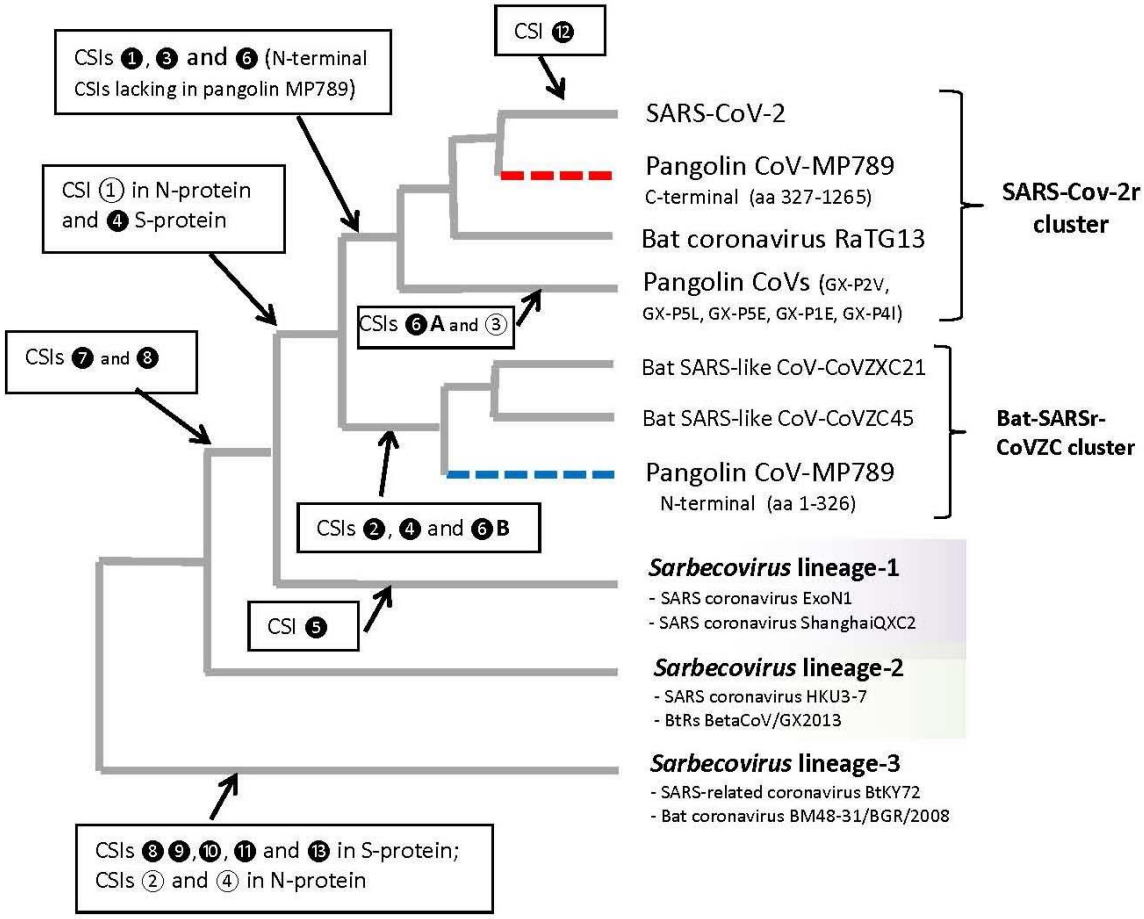

Figure 9. The evolutionary relationships among the CoVs of the subgenus *Sarbecovirus* based on branching in a spike protein phylogenetic tree and the specificities of different identified CSIs. The dashed lines for Pangolin-CoV-MP789 indicates its chimeric nature, where the N-terminal fragment (labeled blue) is most similar to the CoVZC cluster of CoVs, while C-terminal (75%) sequence (labeled red) shows high degree of sequence similarity to the SARS-CoV-2 virus. The arrow marks the evolutionary stages where the genetic changes responsible for different CSIs are postulated to have occurred.

2. Three CSIs were identified in this work (❷, ❹ and ❻B) that are uniquely shared by the two bat-SARS-like viruses from the CoVZC cluster and the pangolin-CoV-MP789. The presence of these three CSIs in the pangolin-CoV-MP789 spike protein and the absence in this pangolin protein of the CSIs (❶, ❸ and ❻) that are specific for the SARS-CoV-2r cluster (which are shared by other pangolin homologs), brought to our attention the anomalous nature of the spike protein sequence from pangolin-CoV-MP789. The results presented here strongly indicate that the spike protein sequence from this strain is chimeric, where the N-terminal sequence region (aa 1-325), where all of the CSIs exhibiting anomalous species distribution (❶, ❷, ❸, ❹ and ❻B) are found, is most closely related (85.94% identity) to the CoVZC cluster of viruses, whereas the remainder of the protein (aa 326-1265) exhibits highest

sequence similarity (98.10%) to the SARS-CoV-2 virus. The inference that pangolin-CoV-MP789 spike protein (gene) is a chimera derived from viruses from these two sources is also strongly supported by the amino acids that are found in polymorphic positions in the N-terminal region versus the C-terminal region (Figure S3). The chimeric nature of the pangolin-CoV-MP789 spike protein also explains the earlier observations (14, 16, 47, 53) that while the genome of pangolin-CoV-MP789 exhibits lower overall sequence identity (91.02%) to the SARS-CoV-2 in comparison to the BatCov-RaTG13 virus (96.0 %), in the RBD domain of the S-protein, its sequence closely corresponds to SARS-CoV-2, while the sequence from BatCov-RaTG13 shows multiple changes.

The sequence heterogeneity for the N- and C-terminal region of the spike protein from pangolin-CoV_MP789 (referred to as pangolin-CoV-2020) was also reported by Liu et al. (47), but they attributed this anomaly to a recombination event. In this context, an important observation made in this work is that in the nucleotide sequence of the S-protein gene, where a shift in amino acid sequence from CoVZC-like to the SARS-CoV-2-like occurs, a Ssp1 restriction site (AATATT) is found. The presence of this restriction enzyme site, which makes a blunt-ended cut, at the predicted junction of the two types of sequences, lends strong credence to our interpretation that the spike protein gene sequence in pangolin-CoV-MP789 is of a chimeric nature. As the genome for pangolin-CoV_MP789 was assembled from metagenomic reads (47), it is likely that the fragments from two different viruses were assembled into a single contig to give rise to the gene for this chimeric protein.

3. The chimeric composition of the pangolin-CoV-MP789 spike protein also highlights another important observation. Our results show that the C-terminal 75% of this protein (aa 320-1265) exhibits higher similarity to the SARS-CoV-2 sequence (98.10%) than that observed for the BatCov-RaTG13 virus (97.46%). The difference in sequence similarity between pangolin-CoV-MP789 and BatCov-RaTG13 is much greater in the RBD (aa 319-540), where these viruses exhibit 96.86% and 90.13% sequence identity to the SARS-CoV-2, respectively. These results indicate that the sequence of the C-terminal 75% of the spike protein from pangolin-CoV-MP789 is most closely related to the SARS-CoV-2 virus than any other virus including BatCov-RaTG13. The N-terminal end sequence of the virus, whose C-terminal 75% sequence is present in pangolin-CoV-MP789, remains to be identified. However, such a virus is expected to be a close relative of the SARS-CoV-2 and it if of much interest to identify and characterize this virus. It is predicted that the N-terminal region of the S-protein from this novel virus will be similar to the SARS-CoV-2r cluster of viruses and it will contain the characteristic signature CSIs (❶, ❸ and ❻) that are present in this region. The presence of these CSIs should also help in the identification of this N-terminal sequence region from metagenomic sequence pool. Another point which is of interest is that the sequence for pangolin-CoV-MP789 virus was isolated from pangolins captured/seized between March and July 2019. Although the sequence of pangolin-CoV_MP789 (or pangolin-CoV-2020) was assembled from pangolin samples, all of the genomic sequences examined in this study were associated with bat coronaviruses (47). This suggests that although this virus is indicated to be of pangolin origin, it is more likely a bat virus. This inference is supported by the absence in this virus of the pangolin-specific CSIs in S- and N-protein and its higher degree of similarity to the bat CoVs in comparison to the pangolin CoVs. If our presumption that pangolin-CoV-MP789 is a bat virus is correct, then it will argue against the possibility that pangolin species served as an intermediary in the transmission of this virus from bats to humans (14-16, 47).

5. Our results also provide strong evidence that two bat SARS-CoVs from *Rhinolophus* sp. from Bulgaria (strain BM48-31/BGR/2008) and Kenya (strain SARS_BtKY72) form a separate lineage within the subgenus *Sarbecovirus*. The distinctness of this lineage (*Sarebecovirus* Lineage-3) is strongly supported by multiple identified CSIs which are specific for this group (see Figure 9). In addition, we have also identified one CSI (❺) that is specifically present in

 doi:10.20944/preprints202006.0165.v1

SARS- and SARS-CoV related viruses, distinguishing this group/clade of viruses from other sarbecoviruses.

6. Lastly, our work also clarifies the viral group specificities of the two CSIs consisting of 5 aa and 13 aa inserts (Fig. 4, ❼ and ❽) that are found in the RBD domain of the spike protein. Of these two CSIs, the 5 aa CSI (❼) is specific for the viruses from the SARS-CoV-2r cluster and *Sarbecovirus* lineage-1 (includes SARS- and SARSr viruses), whereas the 13 aa CSI (❽), in addition to the SARS-CoV-2r cluster and *Sarbecovirus* lineage-1 is also shared by the members of the *Sarbecovirus* lineage-3. However, both these CSIs are not found in the bat viruses from CoVZC cluster and *Sarbecovirus* lineage-2, which branches in between the SARS-CoV-2r cluster and *Sarbecovirus* lineage-3 (see Figures 1 and 9). Thus, the shared presence of these CSIs in the observed viral groups cannot be explained by a commonly shared evolutionary history of these viral groups, but it is more likely due to some non-specific mechanism such as recombination, which is indicated to be common among RNA viruses (7, 8, 10, 54). As both these CSIs are located in close proximity of each other, the shared presence of these CSIs in the indicated clusters of sarbecoviruses could be explained by one or more genetic recombination event(s) involving this region of the S-protein gene.

The molecular markers identified in this work, in addition to their utility in reliably demarcating specific viral clades and lineages of *Sarbecovirus* provide evidence as to how these different groups/lineages are related (Figure 9). Further, as all of the described CSIs are present in (or flanked by) conserved regions, the sequence regions where these CSIs are found provide important means for developing novel diagnostic tests for the identification of different lineages of sarbecoviruses and distinguishing among them. 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 the important bacterial pathogens *Bacillus anthracis* and *Escherichia coli* O157/H7 (18, 22, 29). Additionally, the CSIs in protein sequences due to the predicted importance for protein function also provide new drug targets for the CSI-containing organisms (26, 55).

The question can now be considered as to what might be the functional significances of the identified CSIs in different lineages of coronaviruses/sarbecoviruses. It is important in this context that most of the prominent CSIs identified in this study are in the spike (S) protein. The S-protein is one of the major proteins of coronaviruses, which form surface spikes that mediate the binding of virus to its cellular receptors and subsequent fusion with the membrane of host cell (13, 51, 52, 56). The sequence of the S-protein shows high degree of conservation among the sarbecoviruses and the main regions where differences are observed among different clades/lineages of these viruses are where the identified CSIs are found. Thus, the genetic events leading to these CSIs have served as one of the main mechanisms for the occurrence of major evolutionary changes within the S-protein. Thus, the changes represented by the identified CSIs are predicted to play important roles in the host tropisms of these viruses and as well as human infectivity (12, 14).

The spike protein is post-translationally cleaved into two subunits, an S1-subunit comprising of the N-terminal half of the protein, which binds to the receptor and the S2-subunit, which forms the stack of the spike and mediates membrane fusion (13, 51, 52, 56). All of the CSIs identified in our work in S-protein are in the S1-subunit. The importance of these CSIs in the functioning of the spike protein becomes clear upon examining their locations in the structure of the spike protein (Figure 7). As noted earlier, the SARS and SARS-CoV-2 viruses, which are responsible for the SARS and COVID-19 epidemics, differ from most other sarbecoviruses (that are not known to infect human) in containing the 5 aa and 13 aa inserts (❼ and ❽), that are commonly shared by these two viruses. Both these CSIs are located within the RBD of the S-protein (residues 423-494) and they form a significant portion of the RBD. Thus, the viruses which do not contain these CSIs will be missing a

significant part of the RBD that binds to the ACE2 receptor, and thus they are predicted to be deficient in binding to the ACE2 receptor. Thus, the genetic changes giving rise to these two CSIs have likely played a critical role in the ability of these viruses to bind to the ACE2 receptor and thereby infect humans. In Figure 7, we also show the structural localization of three other CSIs, which differentiates the SARS-CoV-2r cluster of viruses from all other sarbecoviruses including SARS- CoV-1. These three CSIs are also found to form novel surface-exposed loops (lobes) in the structure of the S-protein. Extensive earlier work on other CSIs provide evidence that that the surface-exposed loops formed by CSIs play important roles in mediating novel protein-protein or protein-ligand interactions that are required for the cellular/biological function of the CSI-containing organisms (27, 28, 50, 57). Thus, it is predicted that these three CSIs, which differentiates the SARS-CoV-2r cluster of viruses from other sarbecoviruses should also play important roles in the virulence or other properties of SARS-CoV-2 viruses. It is possible that the novel surface-exposed patches present on the spike protein of SARS-CoV-2r viruses are enabling specific interactions with other surface-exposed proteins or components in the host cell, which could serve as co-receptors for these viruses, thereby enhancing the binding and entry of virus in human cells. It should be noted that membrane co-receptors are known for several other enveloped viruses that are internalized by receptor-mediating endocytosis (56, 58). Thus, the presence of these novel sequence features in SARS-CoV-2r viruses is predicted to play important roles in the host-tropism as well as greater transmissibility of SARS-CoV-2 virus and it will be important to test/confirm these predictions by means of experimental studies.

Our results also show that other members of the SARS-CoV-2r cluster (viz. bat RaTG13 and pangolin CoVs) contain all of the novel sequence features of SARS-CoV-2 virus, except the 4 aa insert present at the junction of S1/S2 subunits (14, 16). Thus, these viruses should also be able to bind to the ACE2 receptor and they are likely to be pathogenic to humans. Further, our analysis has identified several novel sequence features that distinguish the *Sarbecovirus* lineage-3 from all other CoVs. The identification of multiple CSIs in the S- and N- proteins that are uniquely found in the *Sarbecovirus* lineage-3 indicates that this lineage is undergoing rapid evolution and its members should possess novel functional characteristics. The members of this lineage also contain the 13 aa insertion (❽) that is present in the RBD and involved in making contacts with the ACE2 receptor. Thus, 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.

Supplementary Materials: The following are available online at www.mdpi.com/xxx/s1, **Figure S1:** Figure S1. A maximum likelihood tree based on spike protein sequences showing 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 MP789 sequence. Figure S4.). Homology model of N-terminal domain (NTD) of 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. et al. (2020) The species Severe *acute respitatory syndrome*-related coronavirus: classifying 2019-nCoV and naming it 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**.

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. 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.

18. 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.

19. 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.

20. 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**.

21. 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**.

22. 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.

23. 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.

24. Rokas, A. & Holland, P. W. (2000) Rare genomic changes as a tool for phylogenetics.*Trends Ecol. Evol* **15**, 454-459.

25. 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.

26. 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**.

27. 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.

28. 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.

29. 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.

30. 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.

31. Talavera, G. & Castresana, J. (2007) Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments.*Syst Biol.* **56**, 564-577.

32. Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Soding, J. *et al.* (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega.*Mol. Syst. Biol.* **7**, 539.

33. 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.

34. 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.

35. Gupta, R. S. (2014) in *Methods in Microbiology New Approaches to Prokaryotics Systematics*, eds. Goodfellow M, Sutcliffe IC, & Chun J (Elsevier, London), pp. 153-182.

36. 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.

37. Rokas, A. & Holland, P. W. (2000) Rare genomic changes as a tool for phylogenetics.*Trends Ecol. Evol* **15**, 454-459.

38. 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.

39. 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.

40. 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.

41. 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.

42. Rivera, M. C. & Lake, J. A. (1992) Evidence that eukaryotes and eocyte prokaryotes are immediate relatives.*Science* **257**, 74-76.

43. 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.

44. Sharma, R. & Gupta, R. S. (2019) Novel Molecular Synapomorphies Demarcate Different Main Groups/Subgroups of Plasmodium and Piroplasmida Species Clarifying Their Evolutionary Relationships.*Genes.* **10**, 490.

45. 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.* **10**, 739.

46. 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.

47. 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.

48. 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.

49. 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.

50. 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.

51. 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.

52. 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.

53. 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*.

54. Woo, P. C. Y., Huang, Y., Lau, S. K. P. & Yuen, K.-Y. (2020) Coronavirus Genomics and Bioinformatics Analysis.*Viruses* **2**, 1804-1820.

 doi:10.20944/preprints202006.0165.v1

55. 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.

56. Grove, J. & Marsh, M. (2011) The cell biology of receptor-mediated virus entry.*J Cell Biol.* **195**, 1071-1082.

57. 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.

58. Nowak, S. A. & Chou, T. (2009) Mechanisms of Receptor/Coreceptor-Mediated Entry of Enveloped Viruses.*Biophysical Journal* **96**, 2624-2636.