Comparative Genomics Shines Light on the Origins of Covid-19 and Probable Intermediary SARS-CoV-2 Host

Aaron Ho

Introduction

As of the current date of writing, the exact source of COVID-19 is relatively unknown. Like many coronaviruses, strong evidence points to bat related sources. However, direct human-bat interactions are quite rare. It is likely that intermediary hosts between bats and humans play a role in coronavirus transmission. In SARS-CoV the probable intermediary host was the masked palm civet. Similarly, SARS-CoV-2 is likely to have had an intermediary host.

A recent study by Zhang et al. [1] identified pangolins as natural hosts for SARS-CoV-2-like viruses and a possible intermediary. The study found Pangolin-CoV to be the second closest relative of SARS-CoV-2 after Bat-CoV RaTG13. However, some Pangolin-CoV genes are more closely related to SARS-CoV-2 than to RaTG13. Specifically, the S1 RBD subunit of Pangolin-CoV shared the highest level of similarity to SARS-CoV-2. Further studies identified specific key residues involved with human ACE2 binding that were completely conserved in Pangolin-CoV, yet highly mutated in RaTG13 [2].

We'll be testing their findings by recreating 3 key analysis. The first analysis recreated Figure 1C and checked Pangolin-CoV gene similarity in SARS-CoV-2 vs RaTG13. Our second analysis recreated Figure 2 and looked to test which CoV's were most closely related to SARS-CoV-2. Our final analysis recreated Figure 3 by comparing RBD similarity and conservation rates amongst key residues amongst CoV spike proteins.

Upon completion, all figures slightly differed from the source. Despite differences, the overall results agreed with those of the original study. When analyzing Pangolin-CoV gene identity the spike protein and overall average similarly favored SARS-CoV-2 over RaTG13. Furthermore, when doing phylogenetic analysis of CoV genomes and spike proteins, our results supported the conclusion that RaTG13 was the closest relative to SARS-CoV-2, followed by Pangolin-CoV. When comparing RBD conservation, Pangolin-CoV was indeed the most similar. Looking into key binding residues we also found that Pangolin-CoV was fully conserved while RaTG13 was not. In conclusion, our analysis agreed with the original study, and identified Pangolin-CoV as a possible intermediary host for SARS-CoV-2.

Methods

Dataset

All sequences are in nucleotides and were retrieved from the GenBank database.

Primary Genomes Referenced	
Genome	Accession
SARS-CoV-2	MN908947
Pangolin-CoV (2019)	MT121216
RaTG13	MN996532

Other Genomes Analyzed	
Genome Type	Accessions
SARS-CoV-2	MN988668, MN988669, MT019533, MT019532, MT019531, MT019530, MT019529
Pangolin-CoV (2017)	MT072865, MT072864, MT040336, MT040335, MT040334, MT040333
Bat-CoV	KF294457, MG772934, MG772933, MK211374, KY417146, KJ473815, KJ473814, KJ473813, JX993988, DQ022305, KT444582, KP886808, KF367457, DQ412042, AY278488, DQ071615
Civet-CoV	AY686863, AY572034, AY572035

Gene Identity Analysis (Figure 1)

In order to compare gene identity between Pangolin-CoV (2019) with both SARS-CoV-2 and RaTG13, we need to isolate the genes and their respective homologs. The 11 genes found in a typical betacoronavirus are *orf1a*, *orf1b*, *S*, *orf3a*, *E*, *M*, *orf6*, *orf7a*, *orf8*, and *orf10*. With the exception of RaTG13 orf10, all of these genes can be found inside their GenBank record.

RaTG13's orf10 was found using the **NCBI ORFfinder**. CoV's are positive-sense single stranded rna viruses ((+) *ssRNA*). Therefore, orfs are only found in the direction of the forward strand. Knowing that RaTG13 and SARS-CoV-2 are highly similar, genes between the two sequences are also highly similar. In SARS-CoV-2 orf10 is the last gene found after the N gene and is 38 amino acids long. The coding sequence (*CDS*) for RaTG13's N gene is identified at regions 28240-29499. With this in mind, we entered the RaTG13 genome into the ORfinder targeting orfs after index 29499 and having a minimum length of 75 nucleotides. This returned a single 38 aa orf at regions 29524-29640, which was identified as RaTG13's orf10.

After identifying genes, we used **NCBI BLASTP** to compare gene identities. After selecting the BLAST option **"Align two or more sequences"**, Pangolin-CoV is used as the query sequence while SARS-CoV-2 and RaTG13 were used as the subject sequence. This is done for every gene and results in a BLAST report comparing gene identity.

Analyzing Phylogenetic Relationship (Figure 2)

To analyze phylogenetic relationship between Pangolin-CoV's, RaTG13, SARS-CoV-2 and other previously identified *sarbecoronaviruses*, we created phylogenetic trees of both whole genomes, and spike genes.

The first step of phylogenetic analysis was aggregating the sequences we want to compare. Whole genome sequences are represented in nucleotides, while spike sequences are represented as Amino Acids. FASTA files for genomes and spike genes were retrieved from the GenBank database and aggregated into a combined FASTA file.

After aggregating the data, we performed multi-sequence alignment. The program used for this analysis was **MAFFT V. 7.45.** File input consisted of our **whole genome combined FASTA** and **spike protein combined FASTA**. File output was in the **CLUSTAL Sorted** format. The final parameter was alignment strategy, which we denoted as **auto** (defaults). After running the program, we are given multi-seq alignments for our genome and spike protein, denoted with the ***.aln** file extension.

After creating our alignment files, we proceeded with creating our phylogenetic trees. The program used for this analysis was **MEGA X (GUI)**. Creating a tree in MEGAX requires a ***.meg** alignment file. MEGAX has a built-in file converter that can convert CLUSTAL formatted files to MEGA file format. This can be done by opening the CLUSTAL file in the "**DATA**" tab.

Using the "**PHYLOGENY**" tab we created the whole genome phylogenetic tree as a **Maximum Likelihood Tree** using **Nucleotide Sequences** denoted as **Protein Coding**. We created the tree using the default parameters. Phylogeny was tested using the **Bootstrap Method** with **100** replicates. The nucleotide Substitution model was **General Time Reversible** with **Gamma Distributed** rates.

The spike protein tree was created as a **Maximum Likelihood Tree** using **Protein Sequences**. The tree was created with default parameters. Phylogeny was tested using the **Bootstrap Method** with **100** replicates. The amino acid substitution model was the **Jones-Taylor-Thornton model** with **Gamma Distributed** rates. After running both of analysis, the trees are output in ***mtsx** format and annotated in the MEGAX tree explorer.

Identifying Receptor Binding Domain Conservation (Figure 3)

Using the Genbank genome annotations for our spike proteins, the Receptor Binding Domain for SARS-CoV-2 is identified as amino acids 330-583. Using BLASTP, we entered the SARS-CoV-2 spike protein using **subrange 330-583** as the query and aligned all other CoV spike genes in our data as the subject.

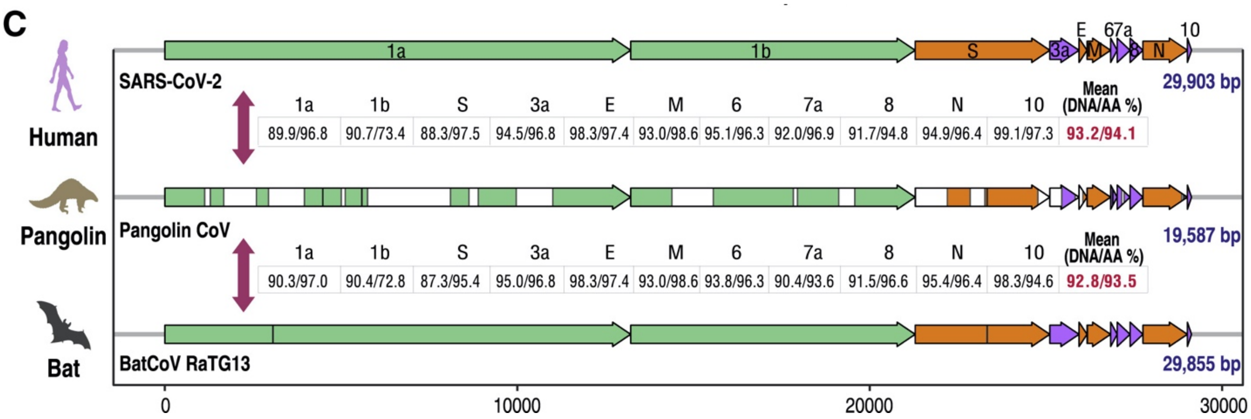
In our BLAST report, we first **filtered out genes with identities < 75%.** The alignment data was copied into a combined FASTA file. Using the BLAST built in Tree visualizer under **Descriptions->Distance tree of results,** we downloaded the distance tree in ***.nwk** format.

Our Tree was visualized using the online Tree visualization program **iTOL V. 5.5.1.** Using our nwk tree file, we can visualize the distance tree between RBD's. Using iTOL's **Multiple Sequence Alignment** **Dataset Template,** we imported our BLAST alignments to visualize our RBD conservation on our tree. Data analysis, and visualization was done using these tools, but further supplemented by hand using vector illustration techniques.

Results

Gene Identity Analysis

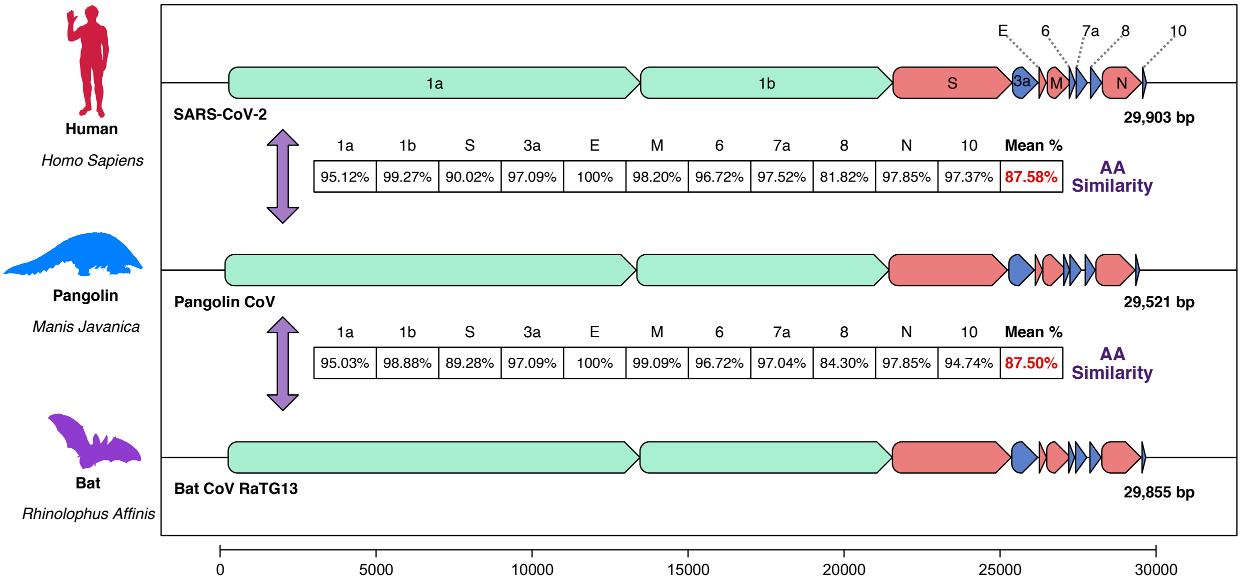
Figure 1 (reference)



1a	1b	S	3a	E	M	6	7a	8	N	10	Mean (DNA/AA %)
89.9/96.8	90.7/73.4	88.3/97.5	94.5/96.8	98.3/97.4	93.0/98.6	95.1/96.3	92.0/96.9	91.7/94.8	94.9/96.4	99.1/97.3	**93.2/94.1**

1a	1b	S	3a	E	M	6	7a	8	N	10	Mean (DNA/AA %)
90.3/97.0	90.4/72.8	87.3/95.4	95.0/96.8	98.3/97.4	93.0/98.6	93.8/96.3	90.4/93.6	91.5/96.6	95.4/96.4	98.3/94.6	**92.8/93.5**

Figure 1 (results)



1a	1b	S	3a	E	M	6	7a	8	N	10	Mean %
95.12%	99.27%	90.02%	97.09%	100%	98.20%	96.72%	97.52%	81.82%	97.85%	97.37%	87.58%

1a	1b	S	3a	E	M	6	7a	8	N	10	Mean %
95.03%	98.88%	89.28%	97.09%	100%	99.09%	96.72%	97.04%	84.30%	97.85%	94.74%	87.50%

Figure 1 Comparison

The overall conclusions of both analyses were largely similar. Our findings were consistent with the idea that some Pangolin-CoV genes showed higher AA% identity to SARS-CoV-2 than to RaTG13. Most notably, both the Pangolin-CoV spike protein and overall mean AA% identity were found to be most similar to SARS-CoV-2.

Despite reaching the same conclusions, there were many differences due to a key discrepancy. The biggest discrepancy was that our Pangolin-CoV was 29,521 bp while the original experiment had a 19,587 bp sequence. The original study performed *de novo* assembly using raw reads from Liu's study [3]. However, between the time of the original study and our analysis, Liu published a completed assembly with gene annotations [4]. Our use of updated data overhauled the values. Some comparisons like orf1a and orf10 had differing results, but the overall conclusion matched while also taking advantage of newer data.

Analyzing Phylogenetic Relationship

Figure 2 (reference)

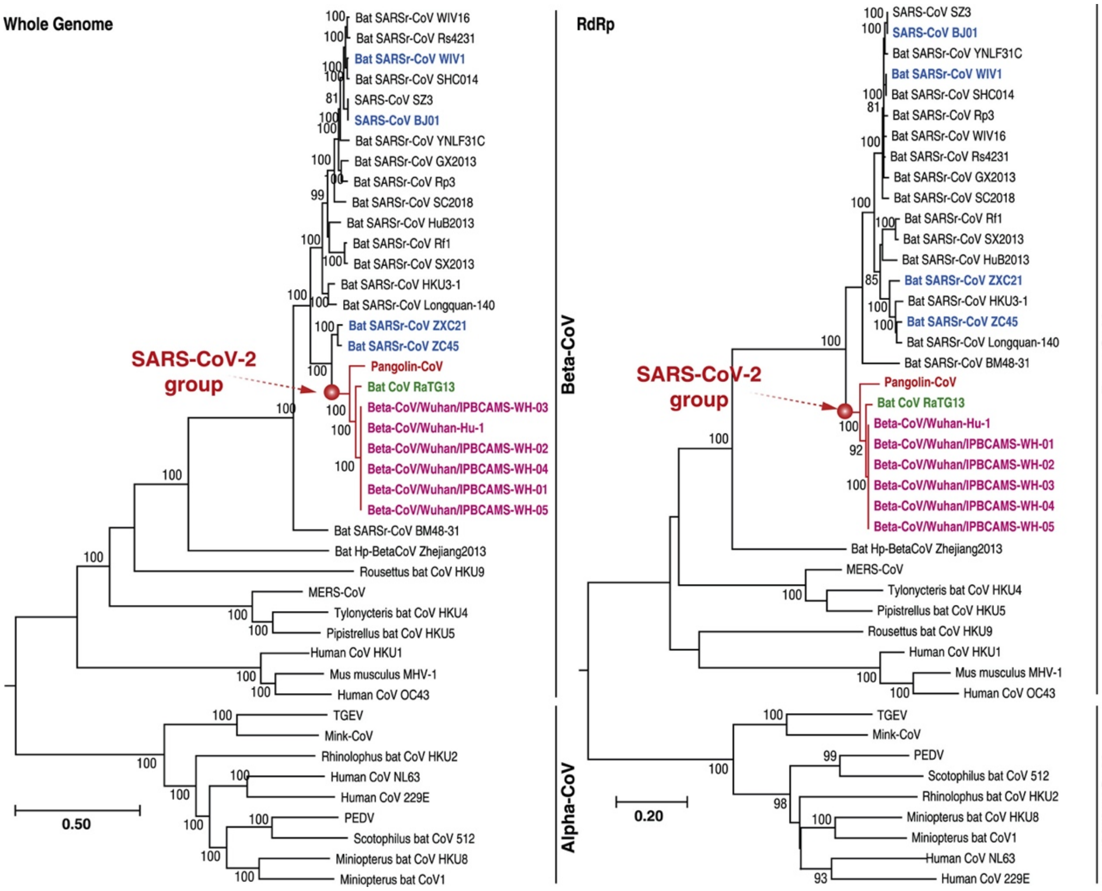
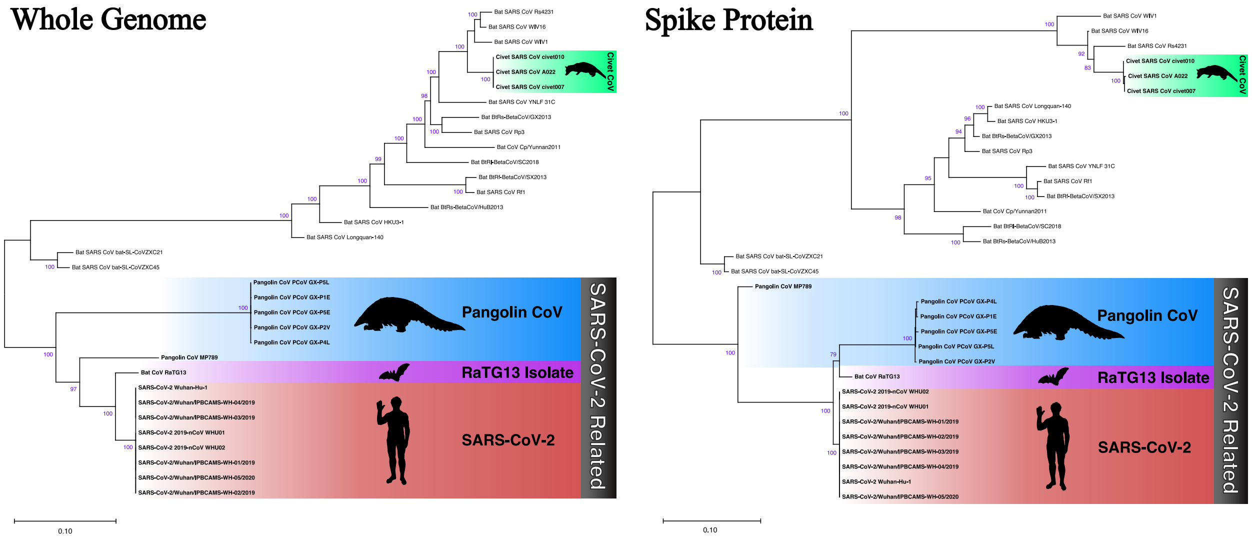


Figure 2 (results)



Whole Genome

Spike Protein

Figure 2 Comparison

Before doing our analysis, there are some key differences between the two studies. The original analysis included alphaCoV's and a variety of betaCoV's. However, to focus on possible SARS-CoV-2 intermediary hosts, we chose to only include SARS-like BetaCoV's with non-human hosts. We also included additional Pangolin-CoV's (2017). Continuing with the focus on intermediary hosts, instead of the RdRp (*Orf1a, Orf1b, S, M*), we only analyzed the spike protein (S). The spike protein is the main focus of the study, and only RdRp gene further explored. Therefore, our analysis will solely compare spike proteins.

Comparing phylogenetic analysis, our results were synonymous with the study for both the whole genome and spike protein. Both studies confidently created a SARS-CoV-2 group that clustered SARS-CoV-2, RaTG13, and Pangolin-CoV together. Our study also supported the conclusion that SARS-CoV-2 is most closely related to RaTG13, followed by Pangolin-CoV.

The addition of different Pangolin-CoV's creates some noticeable differences. The 2019 Pangolin-CoV genome was found to be more closely related to RaTG13 and SARS-CoV-2. In regard to the spike protein, Pangolin-CoV-2019 was more closely related to SARS-CoV-2, while Pangolin-CoV-2017 was more closely related to RaTG13.

Identifying Receptor Binding Domain Conservation

Figure 3 (reference)

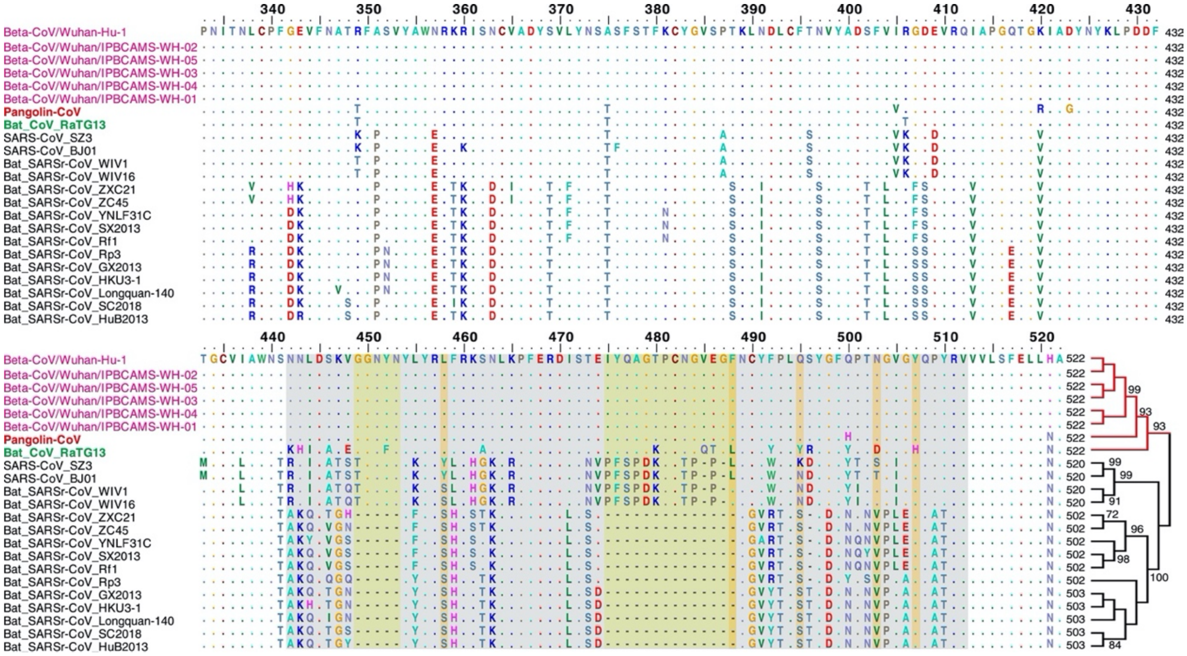


Figure 3 (results)

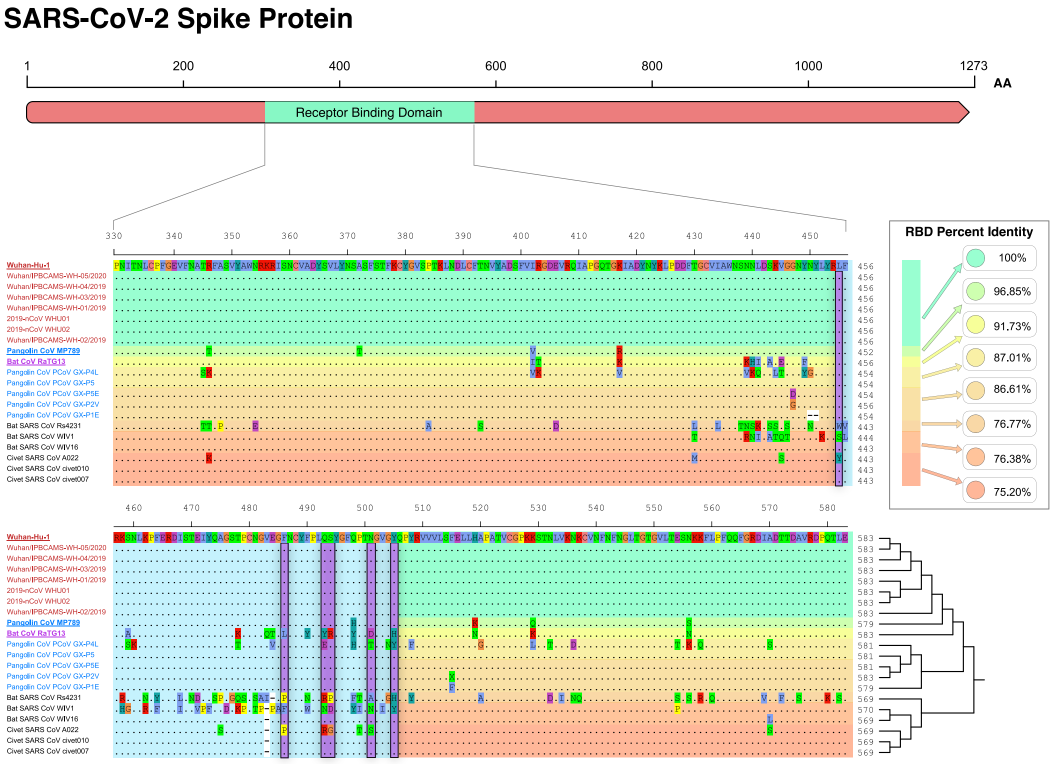


Figure 3 Comparison

Similar to previous figures, our overall conclusions agree with the findings in the original study. Amongst Wuhan SARS-CoV-2 samples, the RBD's were identical. We also concluded that Pangolin-CoV-2019's RBD was the most closely related to SARS-CoV-2. Pangolin-CoV-2019 had a 96.85% identity to SARS-CoV-2, while RaTG13 had a 91.73% identity. Furthermore, when comparing critical ACE2 binding sites, our findings were identical to that of the original study. Within the ACE2 contact region, Pangolin-CoV-2019 only had a single differing amino acid (498H), which isn't a key residue. Thus, Pangolin-CoV-2019 conserves all key residues critical to ACE2 binding. In comparison, RaTG13 only conserved a single key residue.

The conclusion and analysis were identical, but there were some slight discrepancies between the two datasets. Both studies use the same SARS-CoV-2 reference genome, which identifies the RBD region as 330-583. Despite having the same RBD and key residues, the original study uses indices that aren't consistent with the reference genome nor are they consistent with other studies that mention these key residues [5,6]. Furthermore, as some recent studies include a sixth key residue (494S) directly next to a previously established site (493Q), our analysis will also consider this residue [2,5,6].

As a bonus, the addition of 2017 Pangolin-CoV's built upon the results found in both studies. Unlike the multiple SARS-CoV-2 strains, the added Pangolin-CoV's from 2017 were quite diverse. All Pangolin-CoV-2017 strains had a lower percent identity than RaTG13, yet all but one conserved all six key residues.

Discussion

In conclusion, our findings were synonymous with the original report. Bat-CoV RaTG13 was the closest relative of SARS-CoV-2 followed by Pangolin-CoV. Despite being more closely related, we found that certain genes in Pangolin-CoV were more similar to SARS-coV-2 than RaTG13. The most important of these, was the spike protein. In addition to spike protein similarity, six critical residues involved in human ACE2 binding were identified in SARS-CoV-2's spike protein. In Pangolin-CoV these residues were fully conserved, while RaTG13 only conserved a single one of these residues. These results lead us to believe that Pangolin-CoV is better suited to binding to human ACE2 than RaTG13. Therefore, we can conclude that Pangolins hold SARS-CoV-2 like viruses and are indeed possible intermediary sources between bats and humans.

The final results of these analysis largely focused on overall gene similarities and the receptor binding domain of the spike protein. The original plan included two more sets of analysis by recreating another two figures. One regarding the S1/S2 cleavage domain, and another focusing on ACE2 similarity. Upon completion of our three analysis, we concluded that the experiments sufficiently addressed the question of intermediary SARS-CoV-2 host. Many changes seen in our analysis vs the original study was done in the context of focus. Although more figures would further support our study, it comes at the cost of focus. It wouldn't be possible do an expansive study within the concise span of a 10-page report. The final decision was to focus on three main analysis, to build upon each analysis with previous results, and to focus on the details and interpretation of what our results mean in relation to the findings of the original study.

Even if pangolins are possible intermediary host, there is no conclusive answer as to where Pangolin-CoV came from, or whether or not it's even related to SARS-CoV-2. Interestingly enough, the inclusion of Pangolin-CoV's from 2017 may have created more questions than answers. The overall similarities, yet clear differences between the two groups of Pangolin-CoV's give us insight towards the next step in finding an intermediary host. Like SARS-CoV and SARS-CoV-2, Pangolin-CoV-2017 and 2019 may represent possible viral evolution. Perhaps Pangolin-CoV-2019 directly evolved from the 2017 strain. When looking at similarities between Pangolin-CoV-2019 with both Pangolin-CoV-2017 and RaTG13, perhaps the 2019 strain is the result of a recombinant evolution, or maybe evolved from a common ancestor. Answering these questions will bring us closer to the truth of the pandemic.

No matter what the answer is, nature is not the enemy. Pangolins and bats are beautiful creatures and should not be feared nor demonized for the viruses they too suffer from. Instead, they should be treated and studied with the utmost respect. Pangolins are solitary and nocturnal creatures that mostly wish to be left alone. Looking towards the future, even if pangolins aren't intermediary SARS-CoV-2 host, ending pangolin trafficking will benefit both the pangolins, and humankind. If they are indeed the intermediary host, it was not the pangolins who brought upon the pandemic, but rather the hubris of mankind.

References

1. Zhang, T., Wu, Q. & Zhang, Z. Probable Pangolin Origin of SARS-CoV-2 Associated with the COVID-19 Outbreak. *Current Biology* **30**, 1346-1351.e2 (2020).
2. Wan, Y., Shang, J., Graham, R., Baric, R. S. & Li, F. Receptor Recognition by the Novel Coronavirus from Wuhan: an Analysis Based on Decade-Long Structural Studies of SARS Coronavirus. Journal of Virology 94, (2020).
3. Liu, P., Chen, W. & Chen, J.-P. Viral Metagenomics Revealed Sendai Virus and Coronavirus Infection of Malayan Pangolins (Manis javanica). Viruses 11, 979 (2019).
4. Liu, P. et al. Are pangolins the intermediate host of the 2019 novel coronavirus (SARS-CoV-2)? PLOS Pathogens 16, e1008421 (2020).
5. Andersen, K. G., Rambaut, A., Lipkin, W. I., Holmes, E. C. & Garry, R. F. The proximal origin of SARS-CoV-2. Nature Medicine 26, 450–452 (2020).
6. Liu, Z. et al. Composition and divergence of coronavirus spike proteins and host ACE2 receptors predict potential intermediate hosts of SARS-CoV-2. Journal of Medical Virology (2020) doi:10.1002/jmv.25726.