University of Washington – Bothell

Identifying the Origin of COVID-19: Comparative Analysis of RNA of SARS-CoV-2 and Different Animal Coronaviruses

Andy Anderson and Anna Kazanchev

Bioinformatics

Dr. Jesse Zaneveld

09 June 2022

Abstract:

The COVID-19 pandemic is one of the deadliest infectious diseases to have surfaced in recent history. Yet, details circulating where it came from remain elusive. In the present study, the origin of SARS-CoV-2 was investigated using computational methods with RNA sequence data. The Levenshtein Distance algorithm helped us in determining the theoretical mutation minimum for converting coronaviruses from potential animal sources into SARS-CoV-2. This served as a proxy for relatedness between the given coronavirus and COVID-19. The observations found that pangolin coronaviruses were consistently similar to SARS-CoV-2, making them a very likely candidate. The average number of mutations in pangolin coronaviruses was 716 and the number of mutations between different SARS-CoV-2 variants was 92. The average bat coronaviruses were the next likely candidate with 3,712 mutations. However, two strains of bat coronavirus were remarkably similar at 720 and 164 mutations, the latter even outcompeting pangolins. The other animal coronaviruses – birds, camels, cats, cows, ferrets, minks, pigs, and rats – all required upwards of 6,000 mutations. Though we cannot conclusively determine SARS-CoV-2's origin as a result of our research, we do ratify bats and pangolins as potentially the two most plausible origin vectors.

Introduction:

In December 2019, an alarmingly infectious pneumonia emerged in Wuhan, China. It is similar to the SARS coronavirus that broke out in 2003 and is now identified as severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). The disease the virus causes is named COVID-19. In the future, the risk of similar coronavirus outbreaks remains high. In addition to

controlling the COVID-19 pandemic, we must understand its origin and increase research in other potential zoonotic diseases from likely animal vectors. This will allow us to develop broad treatments before they make the jump to humans and prevent history from repeating itself.

The origins of SARS-CoV-2 have not yet been determined, but related coronavirus families can be found throughout the animal kingdom. According to a study conducted in 2020 by Shereen et al., genomic analysis revealed that SARS-CoV-2 is phylogenetically related to SARS-like bat viruses. The discovery of bats holding closely-related coronaviruses indicates that bats are natural reservoirs of these viruses. In addition, researchers in 2021 analyzed scientific evidence that may help clarify where SARS-CoV-2 came from. They concluded that the most reasonable explanation for its genesis is a zoonotic event (Holmes et al., 2021). This ties in with Shereen et al.'s piece because it supports the conclusion that COVID-19 was transmitted from animals to humans, specifically from bats. Another theory suggests SARS-CoV-2 evolved within bats which were then transferred to an intermediate host, potentially pangolins, before making their way to humans (Banerjee et al., 2021).

Peer-reviewed literature from Kuldeep Dhama et al. explores the zoonotic events learned from MERS and SARS and how they jumped the species barrier, an uncommon and complex task. Evidence of MERS, SARS-CoV, and SARS-CoV-2 cross-species jumping, or spillover from animals to humans, was discovered. Researchers explain that these viruses "undergo mutations that enhance viral entry into novel animal species, thus resulting in cross-species transmission" (Dhama et al., 2020).

Our main question is: where did COVID-19 actually come from? There are two major opposing hypotheses in the world right now and that is if it came from a lab leak or from

animals. RNA sequence data was available online to investigate the zoonotic emergence theory through different animal coronaviruses, so determining the one that is least genetically different from SARS-CoV-2 would help us find our answer. Our hypothesis was that bats would have the most similar coronaviruses to SARS-CoV-2.

Methods:

For this study, we chose to aggregate the mRNA sequences of coronaviruses of a variety of different animals we believe to be possible vectors of SARS-CoV-2's transmission to humans. While there are countless animals that have coronaviruses, in order to ensure our data and analysis is as accurate and relevant as possible, we chose animals that are known to be large coronavirus reservoirs, have sufficiently large published coronavirus mRNA sequences publicly available online, a high chance of being found near the region of the initial outbreak in Wuhan, China, as well as are likely to interface with humans to transmit the virus. Under these restraints, the comprehensive list of animal suspects we were able to study was comprised of bats, birds, camels, cats, cows, ferrets, minks, pangolins, pigs, and rats. For each species, we downloaded the complete genome of five different coronaviruses published to the National Library of Medicine's (NIH) National Center for Biotechnology Information (NCBI) database and saved them as FASTA files into their own folder. Through the use of a Python script, each FASTA file was then read, with its included mRNA sequence being sanitized and parsed into a string and added to an array. Each string was then passed into a separate function replicating the cellular process of translation within the ribosome, allowing us to convert it to a new string representing its corresponding amino acids. Next, each of these amino acid strings was compared to that of the

first published SARS-CoV-2 sequence using the Levenshtein Distance algorithm included with the Jellyfish module. The Levenshtein Distance for each pair determines the minimum number of changes necessary to convert one string into the other, allowing us to determine the minimum number of theoretical amino acid mutations that must occur for the animal coronavirus sequenced to become SARS-CoV-2. This effectively enables us to match up corresponding nucleotide pairs across sequences with a very high level of accuracy and precision, despite any base pair insertion, deletion, or substitution mutations that may have occurred. At the same time, the algorithm gives us a clear metric on the relatedness between two RNA sequences as the pair with the fewest mutations are the most similar to one another, and thus, the most likely to be closely related and originate from the same animal. Each of the Levenshtein Distances were then printed out to the console, and then exported to Microsoft Excel for further analytics and graphing. The script and FASTA files used for our project have been made publicly available on GitHub at https://github.com/AndyAnderson8/SARS-CoV-2-Origin-Study.

Results:

After processing and analyzing our data, we were able to come to a few key conclusions regarding the nature of our selected coronavirus sequences and the animals they originated from. Our primary data, as shown in Table 1, shows that the vast majority of sequences used are substantially different from SARS-CoV-2, with between 5,000-7,000 amino acid mutations necessary to match the two. While bats had three sequences that were all very high in theoretical mutations, the last two sequences were vastly lower, as shown in Figure 1. Pangolins, on the other hand, had consistently low numbers of theoretical mutations, and thus, their average

number of mutations was significantly lower than all other animals at just 716, as shown in Figure 2. When comparing the viruses we studied as a sort of control metric, we can see that MERS was the least similar as it required several thousands of mutations, followed by SARS with 1,695 mutations on average, and under 100 with other SARS-CoV-2 variants. Each of the animals we studied, with the exception of bats and pangolins, had coronaviruses that were all less similar than even MERS was. Pangolins and bats were the only two that had sequences even more similar than SARS. In particular, the fourth strain of the bat coronaviruses was the most similar to SARS-CoV-2, as it was even more similar to the strain tested than a different variant of SARS-CoV-2.

Table 1. Theoretical Minimum Amino Acid Mutations from SARS-CoV-2 by Source

	Bat	Bird	Camel	Cat	Cow	Ferret	Mink	Pangolin	Pig	Rat		MERS	SARS	SARS-CoV-2
Strain 1	5657	6235	6399	6522	6389	6424	6436	772	6542	6277		5690	1681	119
Strain 2	6229	6398	6394	6416	6400	6464	6410	456	6549	6399		5687	1688	167
Strain 3	5790	6262	6360	6518	6384	6421	6450	766	6547	6535		5691	1748	53
Strain 4	164	6256	6361	6523	6401	6422	6414	822	6548	6421		5686	1684	53
Strain 5	720	6282	6401	6377	6396	6405	6439	762	6545	6438		5694	1672	68
Average	3712	6287	6383	6471	6394	6427	6430	716	6546	6414		5690	1695	92

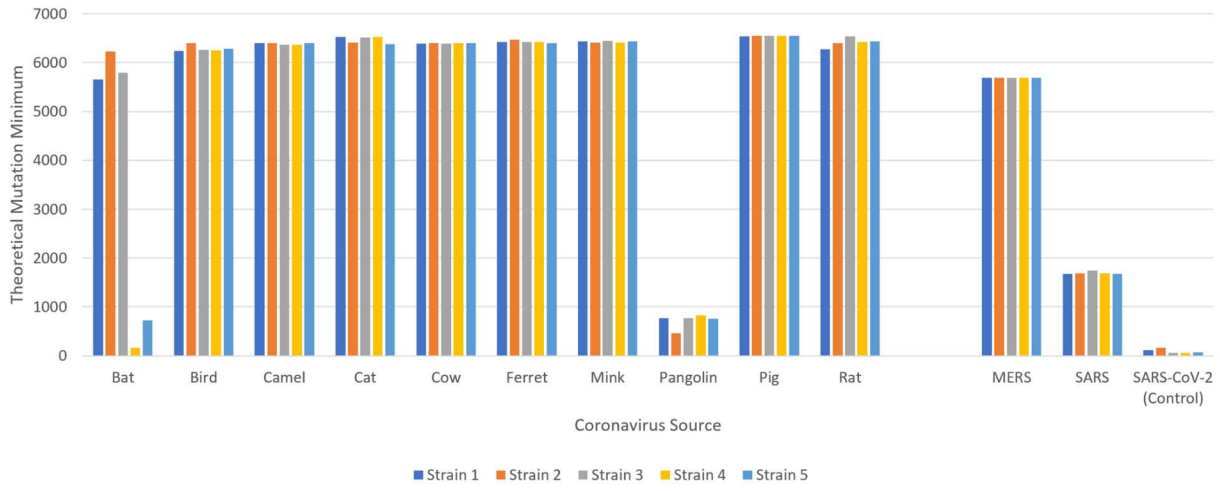

Figure 1. Theoretical Minimum Amino Acid Mutations from SARS-CoV-2 by Source. The graph illustrates the relationship between the minimum number of theoretical amino acid mutations that must occur in each respective coronavirus to convert it to SARS-CoV-2, as determined by the Levenshtein Distance algorithm, and the source from which the virus came. Each animal used met the aforementioned criteria from our methods and five coronavirus strains were aggregated from the NIH NCBI database.

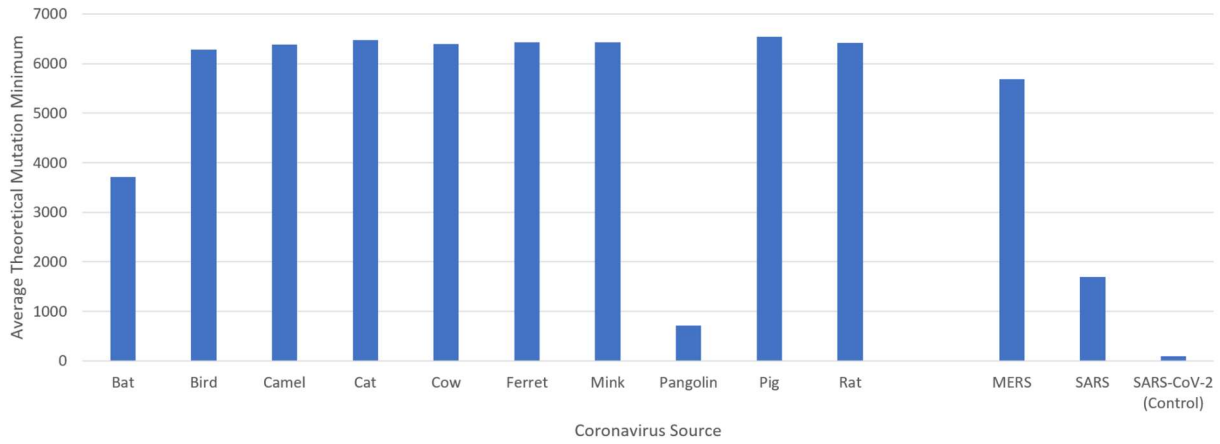

Figure 2. Theoretical Minimum Amino Acid Mutation Averages from SARS-CoV-2 by Source. The graph illustrates the averages of the various strains aggregated in Figure 1. The same data sources and procedures were used in calculations as previously described. This chart demonstrates the overall likeness of animal reservoirs as a whole without taking into account any individual sequences.

Discussion:

Ultimately, the pangolin coronaviruses we found were consistently similar to SARS-CoV-2 with an average number of mutations at 716, making them a very likely candidate for an original source. At the same time, some particular coronaviruses, specifically those from horseshoe bats, also have theoretical mutation minimums below 800, making them a possible candidate. We found no evidence to suggest any of the other animals we researched could have been responsible, as all other circulating strains appear to have upwards of 6,000 mutations. Because of pangolins' consistently low number of theoretical mutations in their coronavirus strains, our hypothesis was actually incorrect because we said that bats would have the most

similar strains on average. This also contradicts what was predicted by Shereen et al. in their study suggesting that bat coronaviruses are the most similar to SARS-CoV-2. Instead, it appears that pangolins have the most similar strains. That said, we still cannot necessarily eliminate bats as candidates for the original SARS-CoV-2 source due to how remarkably similar the two last strains were.

As a whole, our research lines up with what is generally thought to be true by the scientific community. As discussed in Wacharapluesadee et al., it appears that most current research suggests that it was likely either bat, pangolins, or interactions between the two that gave rise to SARS-CoV-2, and our data seems to indicate that as well. Though we cannot conclusively say one way or the other, our data also reaffirms Banerjee et al.'s theory that pangolins serve as an intermediate host between transmission from bats to humans. All in all, we need to do more research as a scientific community to come closer to terms with what happened and where COVID-19 came from. Our study is a good step in that direction and adds more information and evidence to suggest that these really are the species we should be exploring.

Bibliography:

Banerjee, A., Doxey, A. C., Mossman, K., & Irving, A. T. (2021). Unraveling the Zoonotic Origin and Transmission of SARS-CoV-2. *Trends in ecology & evolution*, 36(3), 180–184. https://doi.org/10.1016/j.tree.2020.12.002

Dhama, K., Patel, S. K., Sharun, K., Pathak, M., Tiwari, R., Yatoo, M. I., Malik, Y. S., Sah, R., Rabaan, A. A., Panwar, P. K., Singh, K. P., Michalak, I., Chaicumpa, W., Martinez-Pulgarin, D. F., Bonilla-Aldana, D. K., & Rodriguez-Morales, A. J. (2020). SARS-CoV-2 jumping the species barrier: Zoonotic lessons from SARS, MERS, and recent advances to combat this pandemic virus. *Travel medicine and infectious disease,* 37, 101830. https://doi.org/10.1016/j.tmaid.2020.101830

Holmes, E. C., Goldstein, S. A., Rasmussen, A. L., Robertson, D. L., Crits-Christoph, A., Wertheim, J. O., Anthony, S. J., Barclay, W. S., Boni, M. F., Doherty, P. C., Farrar, J., Geoghegan, J. L., Jiang, X., Leibowitz, J. L., Neil, S., Skern, T., Weiss, S. R., Worobey, M., Andersen, K. G., Garry, R. F., … Rambaut, A. (2021). The origins of SARS-CoV-2: A critical review. *Cell,* 184(19), 4848–4856. https://doi.org/10.1016/j.cell.2021.08.017.

Shereen, M. A., Khan, S., Kazmi, A., Bashir, N., & Siddique, R. (2020). COVID-19 infection: Origin, transmission, and characteristics of human coronaviruses. *Journal of advanced research*, 24, 91–98. https://doi.org/10.1016/j.jare.2020.03.005

Wacharapluesadee, S., Tan, C. W., Maneeorn, P., Duengkae, P., Zhu, F., Joyjinda, Y., Kaewpom, T., Chia, W. N., Ampoot, W., Lim, B. L., Worachotsueptrakun, K., Chen, V. C.,

Sirichan, N., Ruchisrisarod, C., Rodpan, A., Noradechanon, K., Phaichana, T., Jantarat, N., Thongnumchaima, B., Tu, C., … Wang, L. F. (2021). Evidence for SARS-CoV-2 related coronaviruses circulating in bats and pangolins in Southeast Asia. Nature communications, 12(1), 972. https://doi.org/10.1038/s41467-021-21240-1