EEOB 563

Final Project

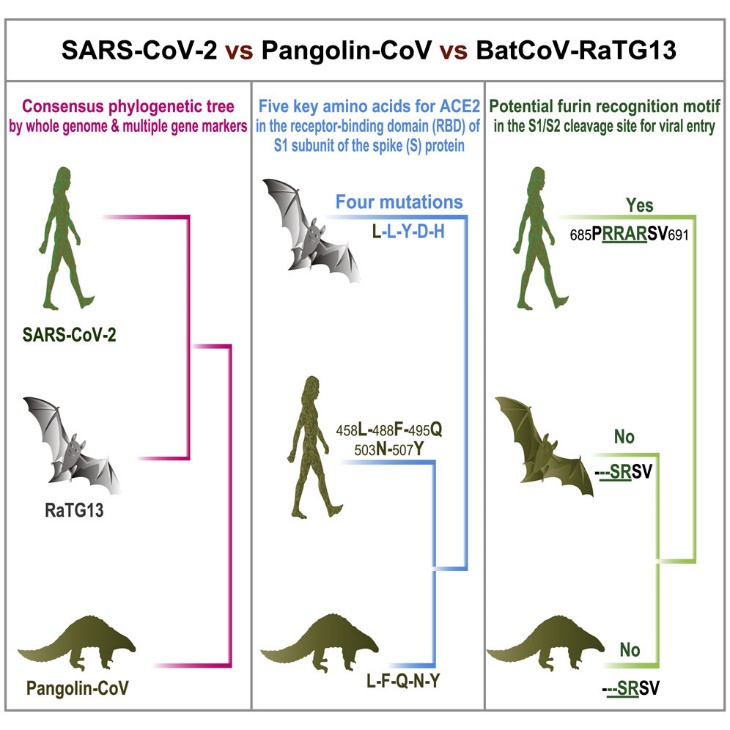
Mihir Kharate

Establishing a Phylogenetic relationship between Zoonotic Coronaviruses:

Introduction

Since last two years, Coronavirus is a topic in Kairos. There has been a lot of speculation going on with how a virus that infects animals suddenly evolved and started infecting humans. There has been some research that speculate of a possible intermediate animal host that allowed the coronavirus to mutate and adapt to a more complex organism, human. While the COVID-19 pandemic had just started, Malayan pangolins that were being smuggled into china were rescued and tested. It was found that they too were infected by a Coronavirus (which they possibly got from bats) (Bale. R., 2020). The Pangolin Coronavirus was then aligned to the SARS-CoV-2 genome and it was found that the two had 91.02% similarity, making it a possible intermediate host. However, the later match of Bat-CoV (RaTG13) strain genome proved it otherwise. In this paper I have attempted re-creating these trees and unveiling the phylogenetic relationship between these zoonotic viruses and identifying the possibility of an intermediate host. I have compared my results to the phylogenies in the Zhang et al., 2020 and Cohen., 2020 papers.

Apart from SARS-CoV-2, I have also attempted to check the phylogenetic relation of the MERS-CoV virus and its possible intermediate host, Camel. Corona viruses, a family of enveloped positive-stranded RNA viruses, can be divided into three major genera or groups, the alpha-CoVs, the beta-CoVs, and the gamma-CoVs (Li. F., 2011). MERS-CoV, SARS-CoV, SARS-CoV-2 all fall under the beta-CoV category. And hence I chose an Alpha-CoV to root my phylogenetic tree and ground the phylogenetic relationships.



The figure above taken from Zhang et al., 2020 proposes a hypothesis of a relation of the novel coronavirus (SARS-CoV-2) with an intermediate host (pangolin).

Methods

I started out by retrieving my Genomic sequences. I used Mafft (Katoh, K., 2002) to align my sequences and created my preliminary trees on conserved and gap free sequences using NJ model with a 100 bootstrap replicates. My preliminary trees only considered the reference genome coronaviruses (SARS-CoV, MERS-CoV, SARS-CoV-2, Pangolin-CoV and Bat-CoV.

From the results I inferred to add more taxa of specific strains of coronaviruses to see whether my phylogeny matches the one in Zhang et al., 2020.

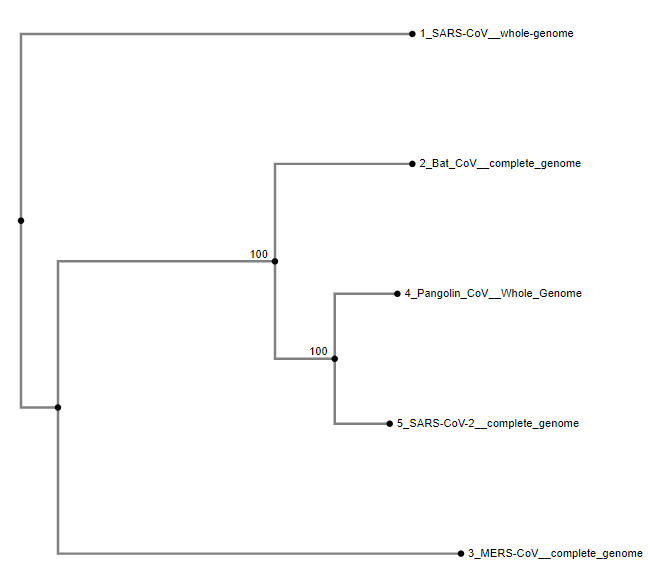
I then used my new aligned sequences to run a Maximum Likelihood analysis using RaXML-NG (Kozlov, A.M., 2019) under the GTR+G model with a 1000 bootstrap replicates. I used iTOL to visualize my trees.

The coronavirus strains I used in my paper are Bat-CoV (Reference Genome), SARS-CoV (Reference Genome), SARS-CoV-2 (Reference Genome), MERS-CoV (Reference Genome), Pangolin Co-V (Reference Genome), Bat-Cov (RaTG13 genome), Bat-CoV (HKU5 genome), Camel-CoV (Reference Genome) and Alpha-CoV (BtMs-AlphaCoV/GS2013).

All the sequences were retrieved from NCBI database.

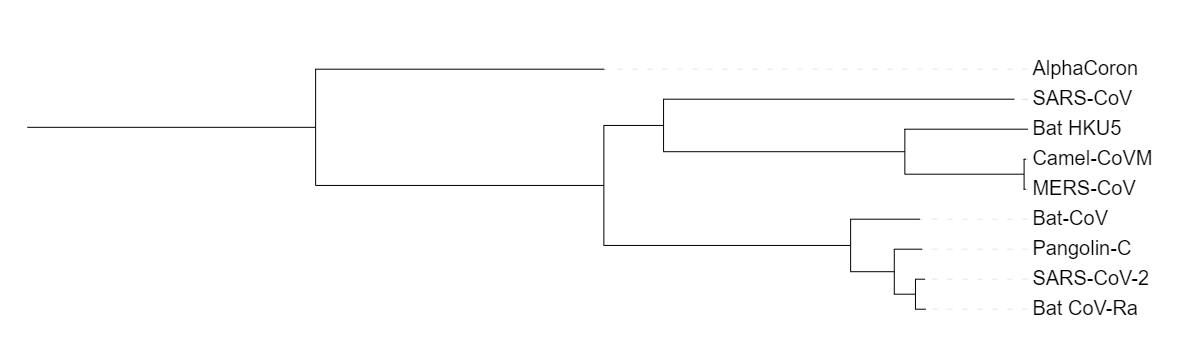
Results

My preliminary tree looked like following. It shows a close relationship between the Pangolin CoV and SARS-CoV-2 genomes, Bat-CoV genome being the outgroup to this clade.



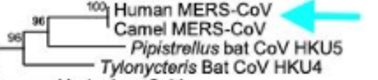
(Above is the preliminary phylogenetic tree constructed from the Multiple Sequence Alignment of whole genomes of the five different Zoonotic coronaviruses under the NJ model. The tree is unrooted).

After I added more taxa and performed the Maximum Likelihood analysis on it under the GTR+G model, the tree looked like this:

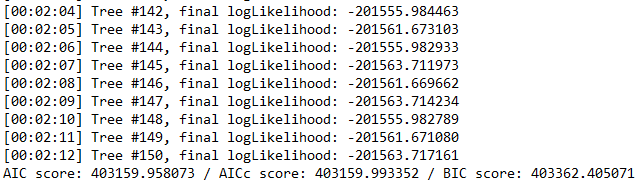


The tree was visualized in iTOL and it matches to the relationship between the taxa presented in the Zhang et. al., 2020 and the Cohen., 2020 papers.

 (Zhang et al., 2020)

 (Cohen., 2020)

The ML tree was ran under a 1000 bootstrap replicates and converged on tree 150.



Although the log likelihood values do not look too good, the convergence was on the 150th tree which I think is a good number considering 1000 replicates.

Discussion

My preliminary tree showed that SARS-CoV-2 and Pangolin-CoV are closely related to each other. However, the preliminary tree only consisted of the reference genomes. When more taxa were introduced based on specific strains of coronaviruses, the results matched the ones in the published papers by Zhang et al., and Cohen., 2020. SARS-CoV-2 was found more closely related to Bat-CoV (RaTG13) than to Pangolin-CoVs.

The Bat-CoV\_RaTG13 genome was collected in 2013 (as mentioned on the NCBI database). Although this strain is relatively older to the SARS-CoV-2 strain that originated in 2019, the alignment is still relevant as the RaTG13 genome persisted in the animal pool for over 6 years. Also, the RaTG13 had been around since 2003 and recollected in samples in 2013, indicating that the virus had been survived for at least 16 years. And the more time it survives, the more time it has to evolve and mutate to infect other hosts. This is very applicable to our knowledge of zoonotic viruses as they evolve and attack human hosts.

The phylogenies show a close relationship with RaTG13 and SARS-CoV-2 with Pangolin CoV as an outgroup. So Pangolins, however not the cause or intermediate hosts for SARS-CoV-2, could be a potential to create a new strain of SARS and this information is helpful to take preventive action so that new strains don’t arise.

My phylogenies also showed a relationship between Camel-CoV and MERS-CoV, which was previously established in many researches. It also showed the Bat-CoV-HKU5 genome to be the outgroup to the Camel and MERS-CoV clade, indicating that the intermediate hosts for MERS CoV were Camels.

Overall, my phylogenetic analysis established successful relationships between Coronaviruses that also matched to the ones previously established in Zhang et al., 2020 and Cohen., 2020.

Citation

Zhang, T., Wu, Q., & Zhang, Z. (2020). Probable Pangolin Origin of SARS-CoV-2 Associated with the COVID-19 Outbreak. *Current Biology*, *30*(7), 1346–1351.e2. <https://doi.org/10.1016/j.cub.2020.03.022>

Cohen. J., (2020, May 6). Mining coronavirus genomes for clues to the outbreak’s origins. Science | AAAS. <https://www.sciencemag.org/news/2020/01/mining-coronavirus-genomes-clues-outbreak-s-origins>

Zhao, J., Cui, W., & Tian, B. P. (2020). The Potential Intermediate Hosts for SARS-CoV-2. Frontiers in Microbiology, 11, 1–11. <https://doi.org/10.3389/fmicb.2020.580137>

Katoh, K. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Research, 30(14), 3059–3066. <https://doi.org/10.1093/nar/gkf436>

Kozlov, A. M., Darriba, D., Flouri, T., Morel, B., & Stamatakis, A. (2019). RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. Bioinformatics, 35(21), 4453–4455. <https://doi.org/10.1093/bioinformatics/btz305>

Bale, R. (2020, March 26). *Trafficked pangolins can carry coronaviruses closely related to pandemic strain*. Animals. <https://www.nationalgeographic.com/animals/article/pangolins-coronavirus-covid-possibility?loggedin=true>

Li, F. (2011). Evidence for a Common Evolutionary Origin of Coronavirus Spike Protein Receptor-Binding Subunits. *Journal of Virology*, *86*(5), 2856–2858. <https://doi.org/10.1128/jvi.06882-11>