

EBIO 3080 Lab Week 9

Nicole Mooney, Dylan Oh, Ezra Weible

UPGMA Trees

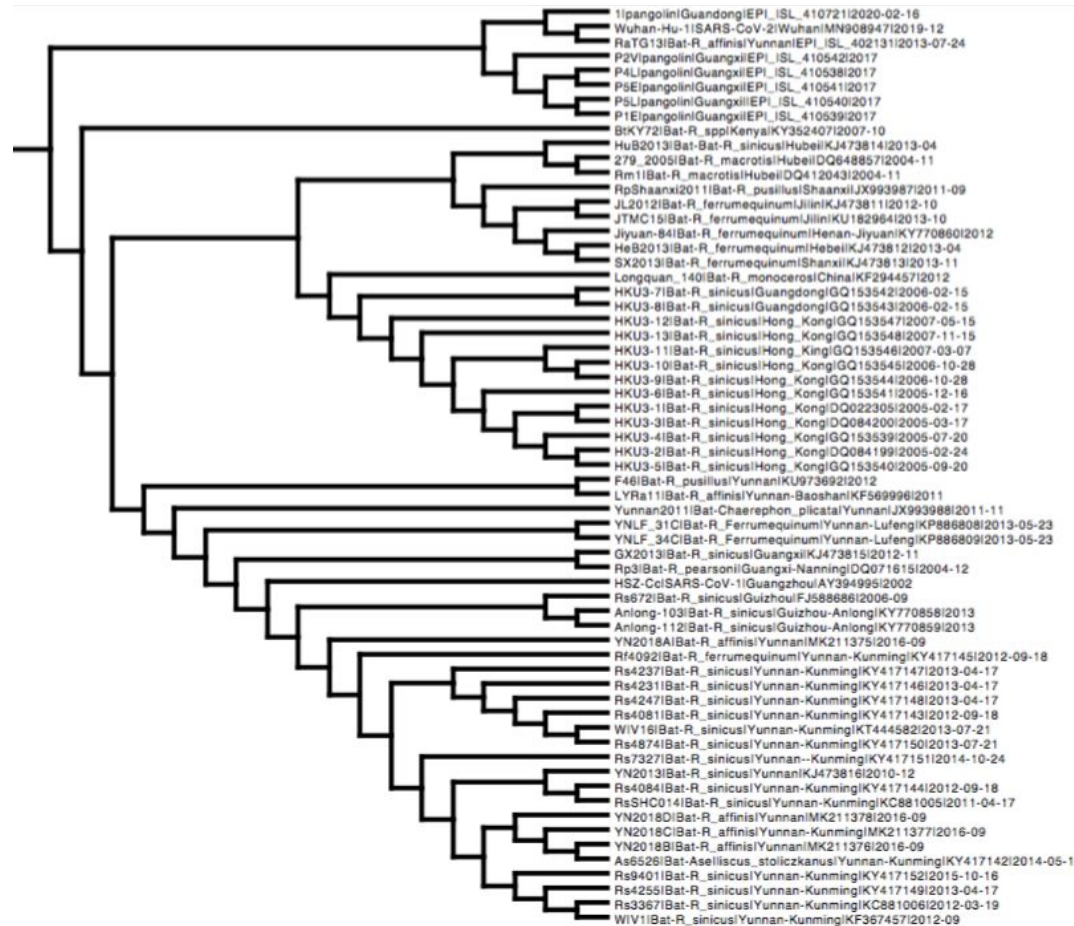


Figure 1. UPGMA-based phylogenetic tree of coronavirus sequences collected by Boni et al. The tree was constructed with the program Mesquite using the UPGMA method to estimate relationships. We used a non-recombinant region (NNR1) in order to estimate these phylogenies using 63 different coronavirus sequences. There were three different character states, where the host was either a bat, pangolin, or human. To make the tree we used fast DNA/RNA and then used standard categorical data to make a cluster analysis with the source of distance being the distance from the character matrix and more specifically the Jukes cantor distance. Lastly, the clustering method used was UPGMA.

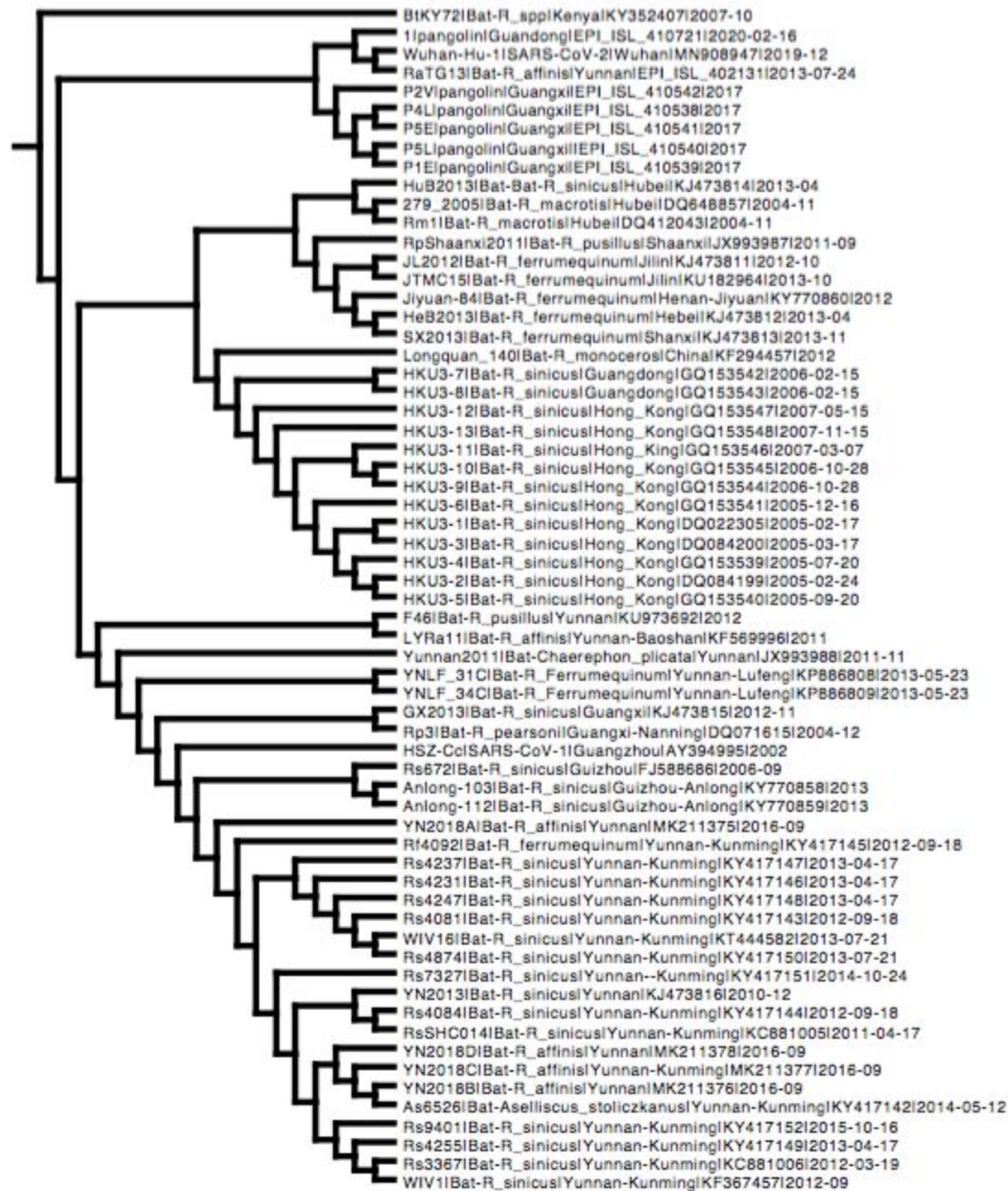


Figure 2. An edited version of our UPGMA-based phylogenetic tree of coronavirus sequences collected by Boni et al.. This tree was modified to force the top species (BtKY72|Bat-R_spp\Kenya|KY352407|2007-10, which was the same group that Boni et al. used to root the tree and is genomically external to the rest of the viruses) to be the outgroup, which altered many of the default relationships.

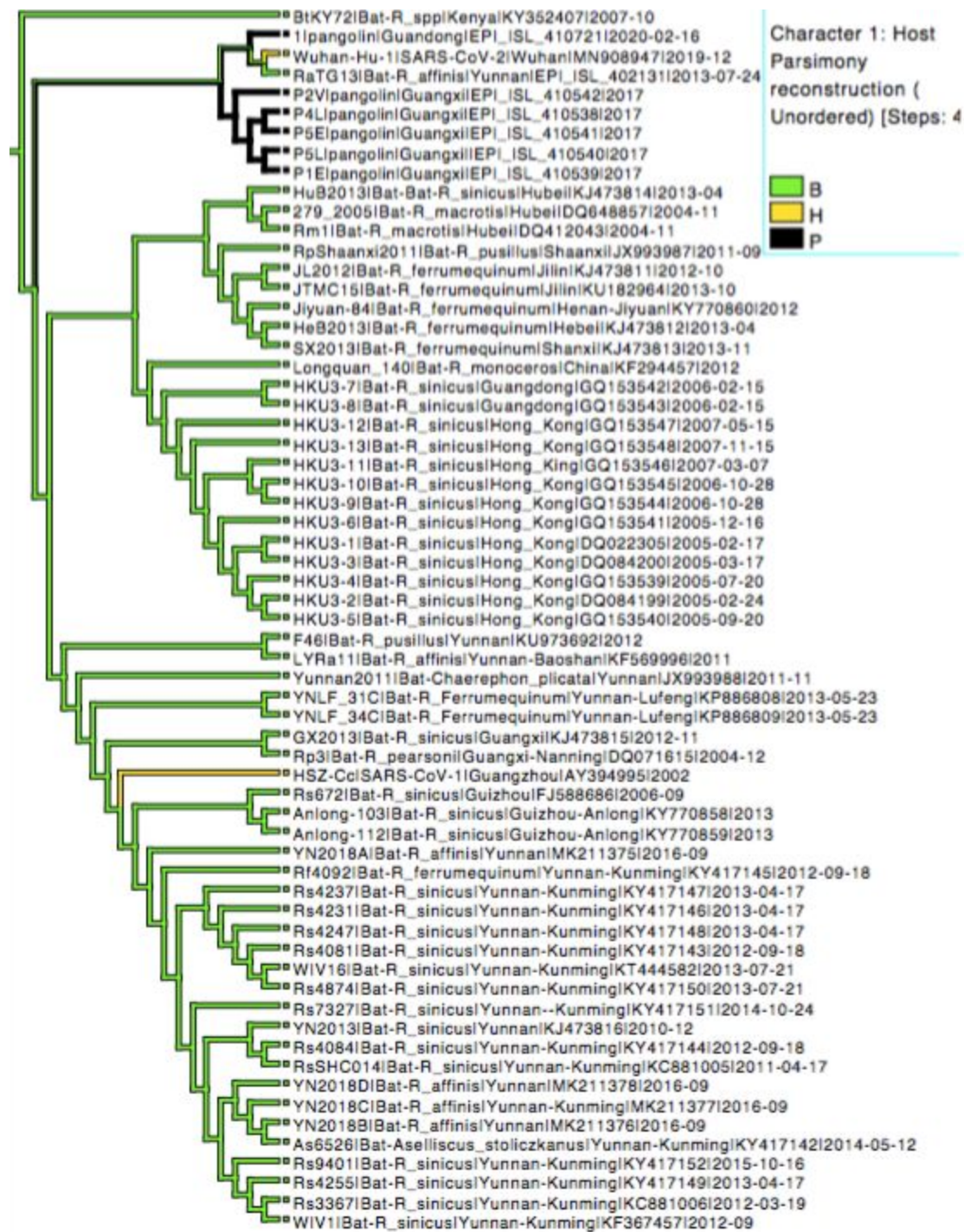


Figure 3. An edited version of our UPGMA-based phylogenetic tree of coronavirus sequences collected by Boni et al.. This tree is colour-coded to show the different host species: green for bat, black for pangolin, and yellow for human.

3. Critically review the patterns of evolution of host use.

a. Has there been convergence? Explain. 2 points.

There has been some convergence, which can be seen in the two human-host branches of the tree (yellow lines). The coronavirus phylogeny changed enough to be able to infect humans, but the two coronavirus branches that represent this are not sister taxa, nor do they share a human disease-causing coronavirus ancestor.

b. What is the host species for most coronavirus sequences? Explain. 1 point.

The host species for most coronavirus sequences is *Rhinolophus sinicus*, the Chinese rufous horseshoe bat. This particular species of bat was probably of high research interest following the initial outbreak of SARS-CoV-1 due to its geographic range relative to the outbreak as well as the reputation of bat species for being effective disease vectors, and as a result most of the sequences in the tree were taken from this host. Most of these sequences are dated between the outbreaks of SARS-CoV-1 and SARS-CoV-2.

c. What is the likely host source of the virus that caused the original SARS outbreak, SARS-CoV-1? Explain your rationale. 2 points.

The host source of the SARS-CoV-1 virus is most likely a bat. As shown by the species near SARS-CoV-1 on the tree, such as *Rhinolophus sinicus* or *Rhinolophus pearsoni*, all nearby species (including the ancestral species) were bats. Therefore, we can infer that SARS-CoV-1 evolved in order to have a human host from an ancestral virus that had a bat host.

d. What is the likely host source of the virus SARS-CoV-2, which is the cause of the current COVID-19 pandemic? Explain your rationale. 2 points.

The host source of the SARS-CoV-2 virus is most likely the bat. The parsimonious tree shows that nearly all bat-sequenced genomes in the non-recombinant regions result in the same phylogenies and characters, even compared to the Kenyan outgroup bat **BtKY72|Bat-R_spp\Kenya|KY352407|2007-10**. Moreover, the most closely related species to the human disease-causing SARS-CoV-2 is a bat:

RaTG13|Bat-R_affinis|Yunnan|EPI_ISL_402131|2013-07-24. Considering that the bat is the most common host for the coronaviruses that share an ancestor with SARS-CoV-2, it can be inferred that the pangolin species, **1|pangolin|Guandong|EPI_ISL_410721|2020-02-16**, that was infected within the same clade as SARS-CoV-2 and

RaTG13|Bat-R_affinis|Yunnan|EPI_ISL_402131|2013-07-24 evolved separately from the clade of 5 pangolin-infecting coronaviruses that share a common ancestor with the above species. Therefore, the likely host source of the SARS-CoV-2 virus was a bat.

e. To provide better resolution of the source of human disease-causing coronaviruses, provide a brief description (2-3 sentences) of what would be needed. Hint: think back to question 8a from last week's lab. 2 points.

In order to come to a stronger conclusion of the source of human disease-causing coronaviruses, you would need to include all of the host species that coronaviruses are present in. If a species is not present (or their presence is limited), the appearance of the trees could result in an incorrect conclusion due to the lack of comprehensive evidence. For example, if the bats were not present in the tree we created above, it is likely that we would have assumed that the pangolin was the original host species of the **Wuhan-Hu-1|SARS-CoV-2|Wuhan|MN908947|2019-12** virus, rather than our final assumption of the bat. Our analysis is therefore limited by the evidence we have access to.