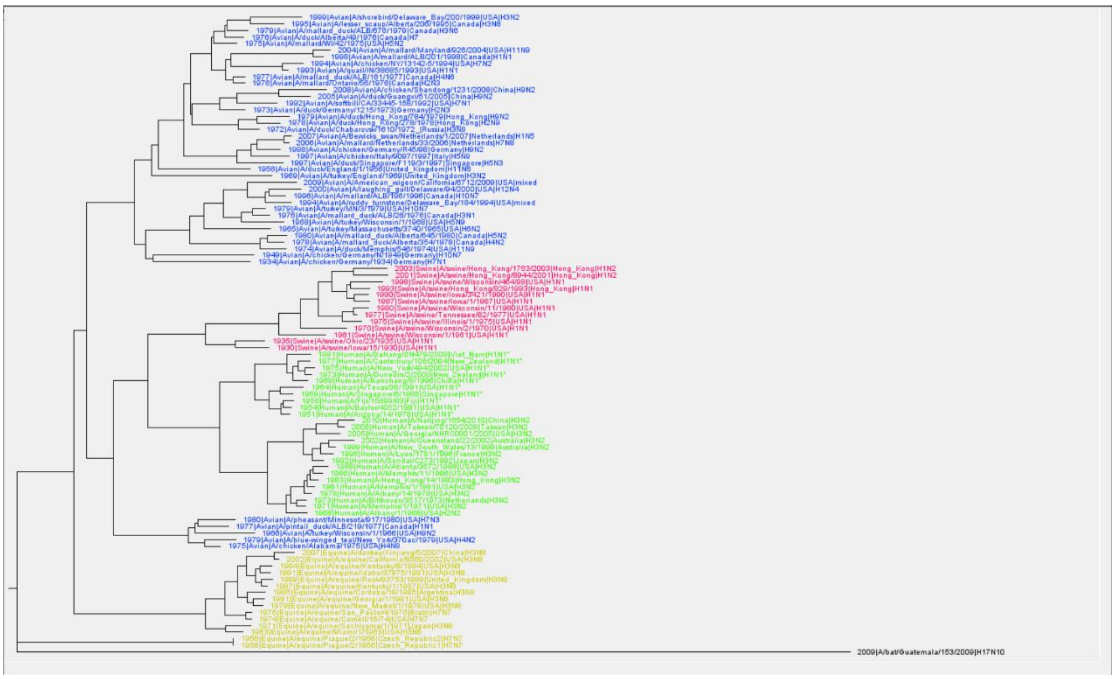
1. This is the BioNJ tree with HKY model and 1000 bootstrap replicates from the complete data set with outgroup, the H17N10 bat is isolated at the bottom, and all tips are colored by host species. Blue refers to avian, red refers to swine, green refers to human, and yellow refers to equine.



   a. **Do the viruses cluster in monophyletic clades by host? If there are such host specific clusters, is there sufficient support for this? For those that do not cluster by host, in how many clusters do they break up?**

      No, avian strains and equine strains do not cluster in monophyletic clades, they are paraphyletic.

      Swine strains and human strains look like host-specific clusters because all swine strains share a single common ancestor and all descendants in this cluster are swine strains, and it is the same to human strains. However, if there are more data with more different hosts, it is possible for these strains to be isolated into different clusters, they might be not monophyletic anymore.

      For avian strains and equine strains, both of them have two clusters.

   b. **With which host-specific cluster, if any, are the human viruses most closely related. In which hosts circulate the most diverse viruses: avian or swine? How do you conclude this?**

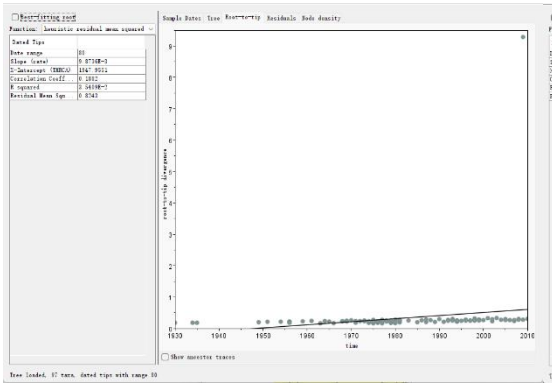
      Swine cluster is the host-specific cluster that human viruses are most closely related, because swine strains and human strains share the most recent common ancestor.

      Avian circulates the most diverse viruses, because avian viruses have a longer average pairwise genetic distance than swine viruses, which determines the diversity.

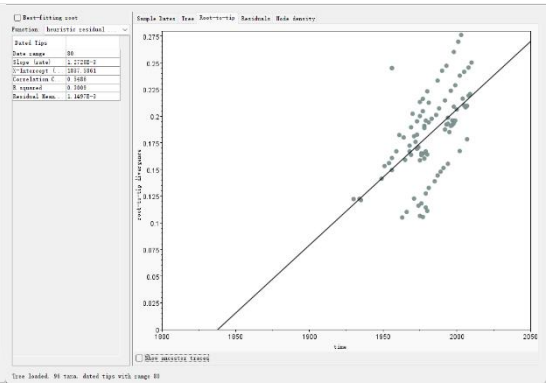
   c. **Does midpoint rooting change the direction of evolution?**

      No, when using midpoint rooting in FigTree, the direction of evolution does not change.

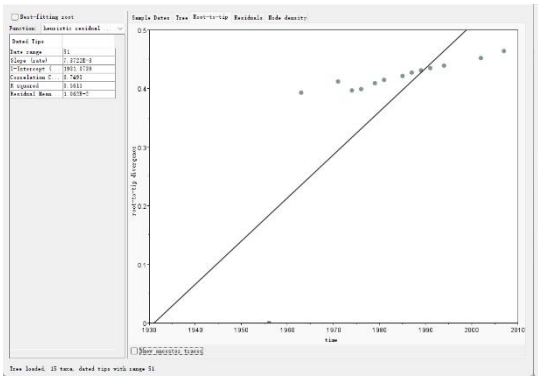
2. Here are 6 regressions of root-to-tip divergence against sampling time, in order the complete dataset with bat, complete dataset without bat, avian, equine, human, and



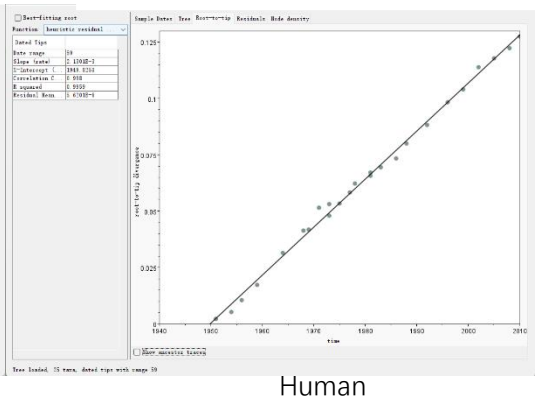
Complete Dataset with Bat



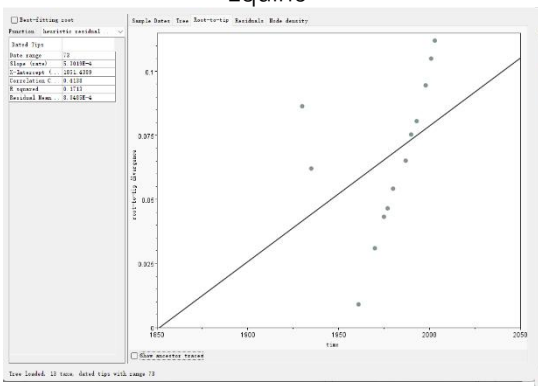
Complete Dataset without Bat



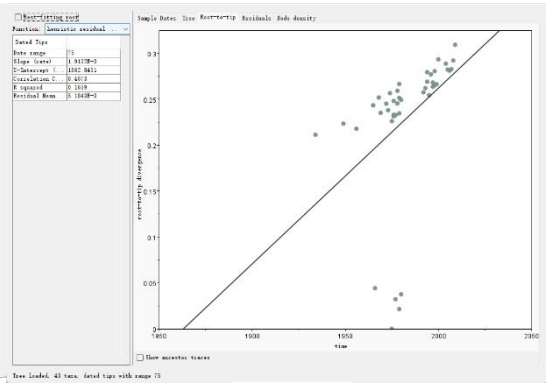
Equine



Human



Swine



Avian

swine.

There are very little correlations in some datasets like avian and swine, so best-fitting root is used to lead to a better correlation between dates and divergence

a. **Is there sufficient temporal signal in these data sets? Rank them by strongest to weakest signal**.

Only human strains show sufficient temporal signal, which has a high correlation

	Complete Dataset with Bat	Complete Dataset without Bat	Avian	Equine	Human	Swine
Slope (rate)	9.87E-03	1.27E-03	1.91E-03	7.37E-03	2.13E-03	5.30E-04
Correlation Coefficient	0.1882	0.5486	0.4073	0.7493	0.998	0.4138
	Best-fitting root					
Slope (rate)	1.46E-03	1.33E-03	1.16E-03	7.37E-03	2.13E-03	2.39E-03
Correlation Coefficient	0.6527	0.611	0.8526	0.7493	0.998	0.9991

coefficient 0.998, but when using best-fitting root, these datasets all show good correlations which are higher than 0.6.

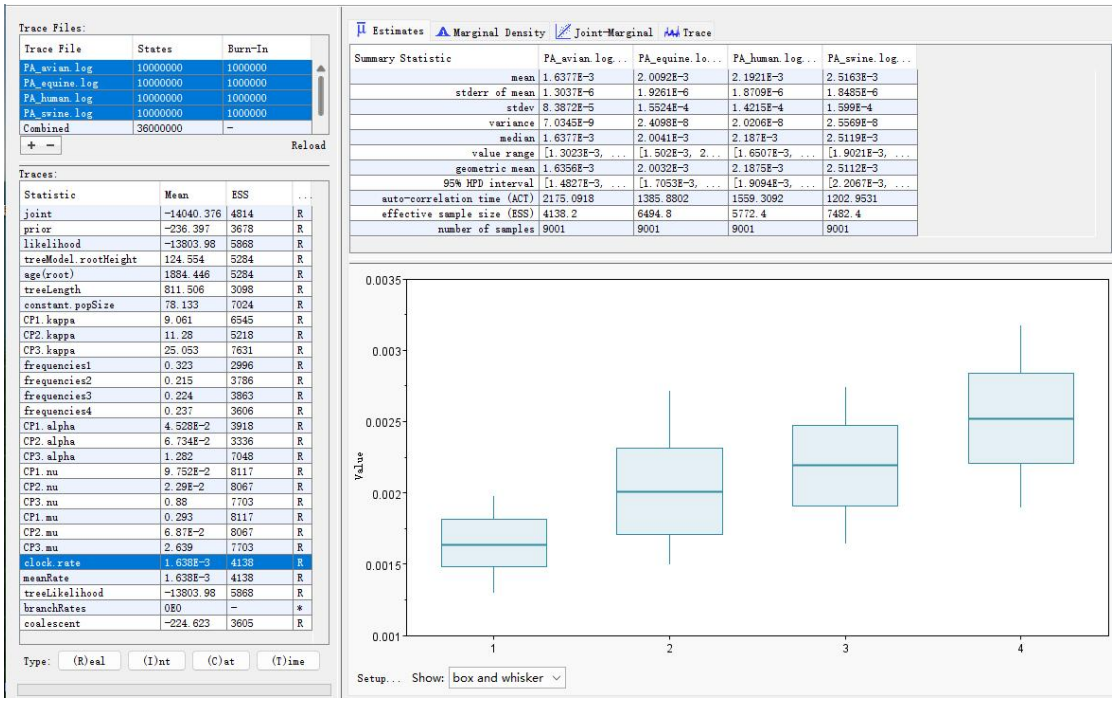
According to the correlation coefficient without best-fitting root, the signal from the strongest to the weakest is Human, Equine, Complete Dataset without Bat, Swine, Avian, Complete Dataset with Bat.

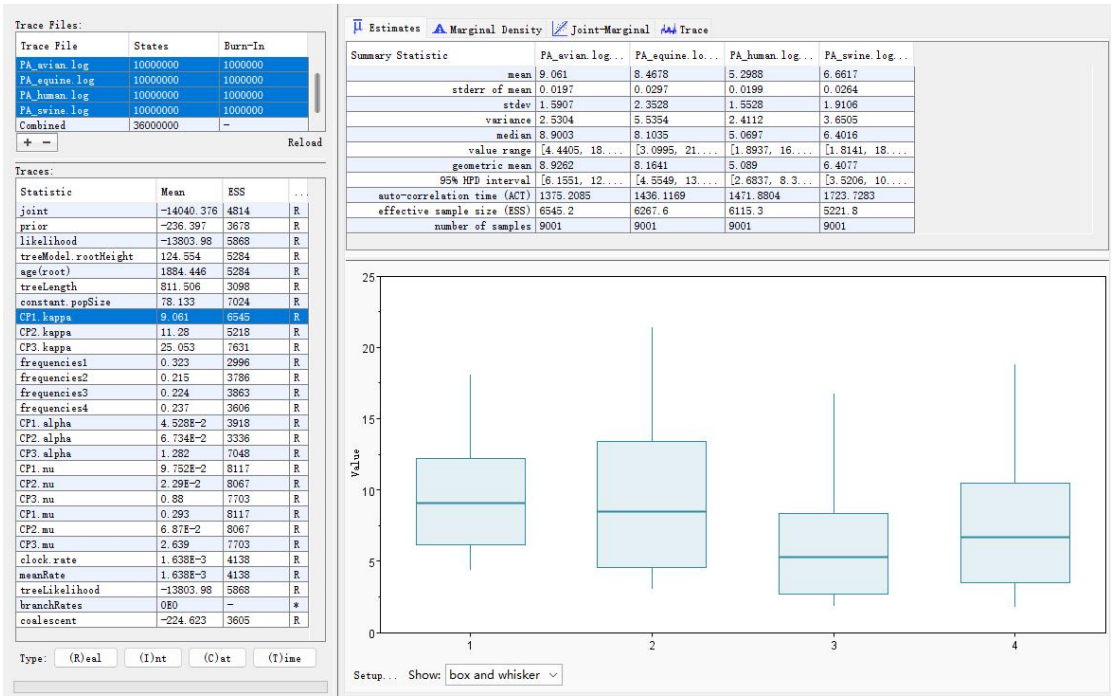
b. Based on point estimates for the rate of evolution, in which host do the viruses evolve most rapidly? Does this exploration suggest substantial variation in evolutionary rates among hosts?

The viruses evolve most rapidly in equine, the increasing divergence rate of equine is 7.37E-03.

There is a huge difference among these hosts, evolutionary rates of equine is more than 10 times larger than that of swine. After using best-fitting root, the evolutionary rate of swine increases a lot, but rate of equine is still more than 3 times larger than rate of other hosts.

3. Select the statistics called clock rate and CP1.alpha in Tracer to analyze the log file from BEAST.





		Avian	Equine	Human	Swine
clock.rate	mean	1.64E-03	2.01E-03	2.19E-03	2.52E-03
	variance	7.03E-09	2.41E-08	2.02E-08	2.56E-08
	95% HPD interval	[1.4827E-3, 1.8115E-3]	[1.7053E-3, 2.3106E-3]	[1.9094E-3, 2.4697E-3]	[2.2067E-3, 2.8363E-3]
CP1.alpha	mean	9.061	8.4678	5.2988	6.6617
	variance	2.5304	5.5354	2.4112	3.6505
	95% HPD interval	[6.1551, 12.1964]	[4.5549, 13.4247]	[2.6837, 8.3673]	[3.5206, 10.4696]

a. How do the evolutionary rate estimates compare among hosts and how do they compare to the point estimates obtained by TempEst?

Based on clock rate among hosts, swine strains have the highest mean value of clock rate 2.52E-3, which indicates swine strains have the highest evolutionary rate estimates. The second is human strains with rate 2.19E-3, the third is equine strains with rate 2.01E-3, and the last one is avian strains with rate 1.64E-3.

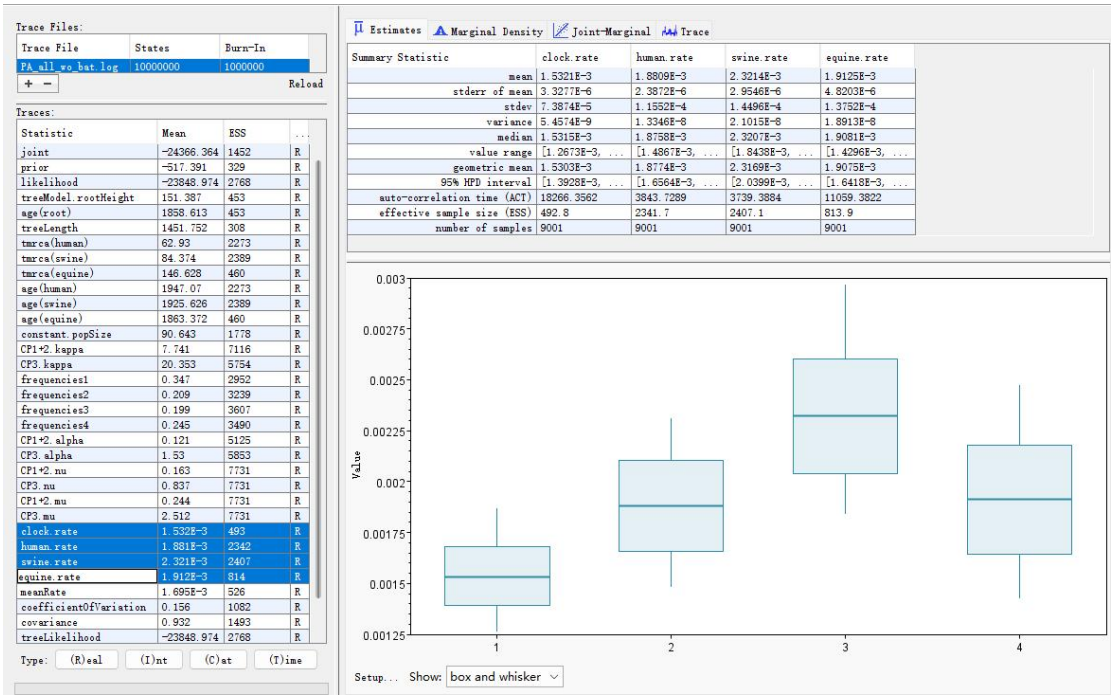
Compared to the result obtained by TempEst, the evolutionary rate of equine is smaller and the evolutionary rate of swine is larger. Besides, there is no big difference among evolutionary rates of different hosts.

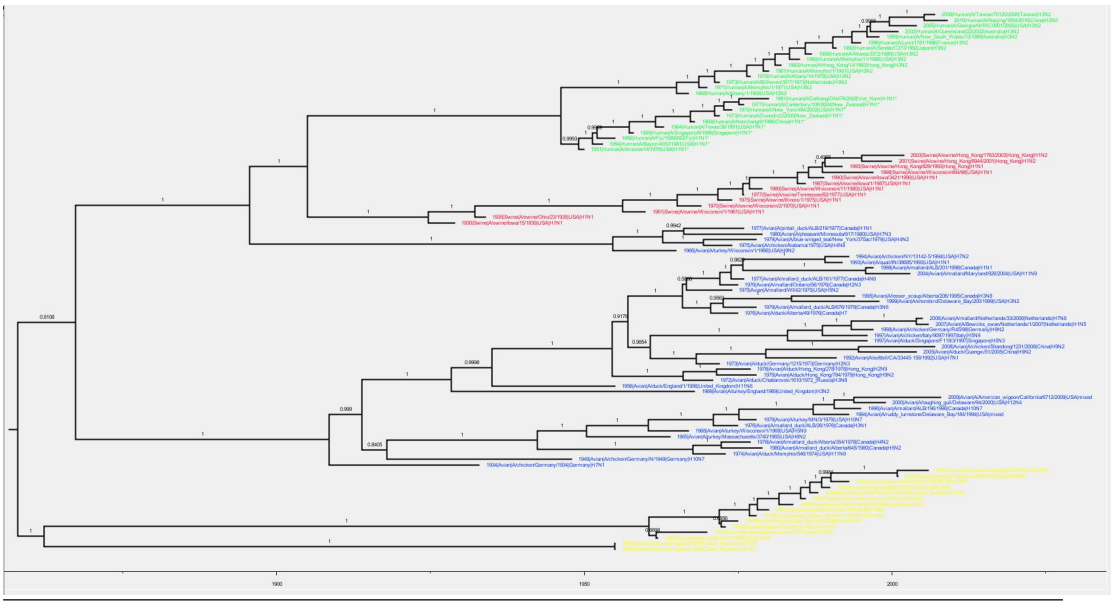
b. Which host do the viruses have the highest transition/transversion bias. In which hosts do they have the lowest rate variation among sites. Do the credible intervals overlap for the relevant parameters?

Avian viruses have the highest transition bias, which is indicated by CP1.kappa (9.061), and avian strains have the lowest rate variation, which is 7.03E-9. According

to the 95% HPD interval, credible intervals indeed overlap for CP1.kapper and clock rate.

4. A BEAST analysis for the complete data set without outgroup with a fixed local clock (FLC) model, the SRD06 substitution model and a constant population size coalescent model.





a. Is the monophyly constraint is a reasonable constraint?

Yes, it is necessary to enforce these clades to be monophyletic. Because Fixed local clock assumes a change of evolutionary rate at their most recent common ancestor, if clades are not monophyletic, several local clocks might come to overlap during the MCMC procedure, leading to estimation errors.

b. Does the fixed local clock pick up substantial rate differences among hosts?

Please discuss the estimates, also in relation to the estimates of the independent analyses.

No, there is no substantial rate difference among hosts. However, compared to the independent analyses, the rate difference among hosts become larger, and the variance is smaller.

c. **Do the same answers as those to question 1b for the BioNJ apply to this MCC tree? If not, indicate why not.**

It is still the same answer as those to question 1b. Swine cluster is still the host-specific cluster that human viruses are most closely related, and avian circulates the more diverse viruses than swine. The MCC tree does have some difference compared to BioNJ tree, especially to enforce host-specific data to be monophyletic, but it does not show the influence in this question.

d. **What are the estimates for the time of the origin of all viruses and the time of the most recent common ancestor of the human viruses.**

The estimate for the time of the origin of all viruses is 1857.8904, and the time of the most recent common ancestor of the human viruses is 1946.1675.

e. **Compare the MCC tree summary with the provided MCC tree that was inferred under an uncorrelated relaxed clock model. Is the overall clustering pattern influenced by the clock model?**

Yes, in the uncorrelated relaxed clock model, equine strains share two clusters, the reason is probably that uncorrelated relaxed clocks allow each branch of a phylogenetic tree to have its own evolutionary rate, which is different from the fixed local clock.