

# **Evolutionary and Quantitative Genetics**

## **Task 1**

Daria Plewa  
Philippe Lemey  
8 October 2023

## **List of contents**

1. Part 1
2. Part 2

### **Part 1**

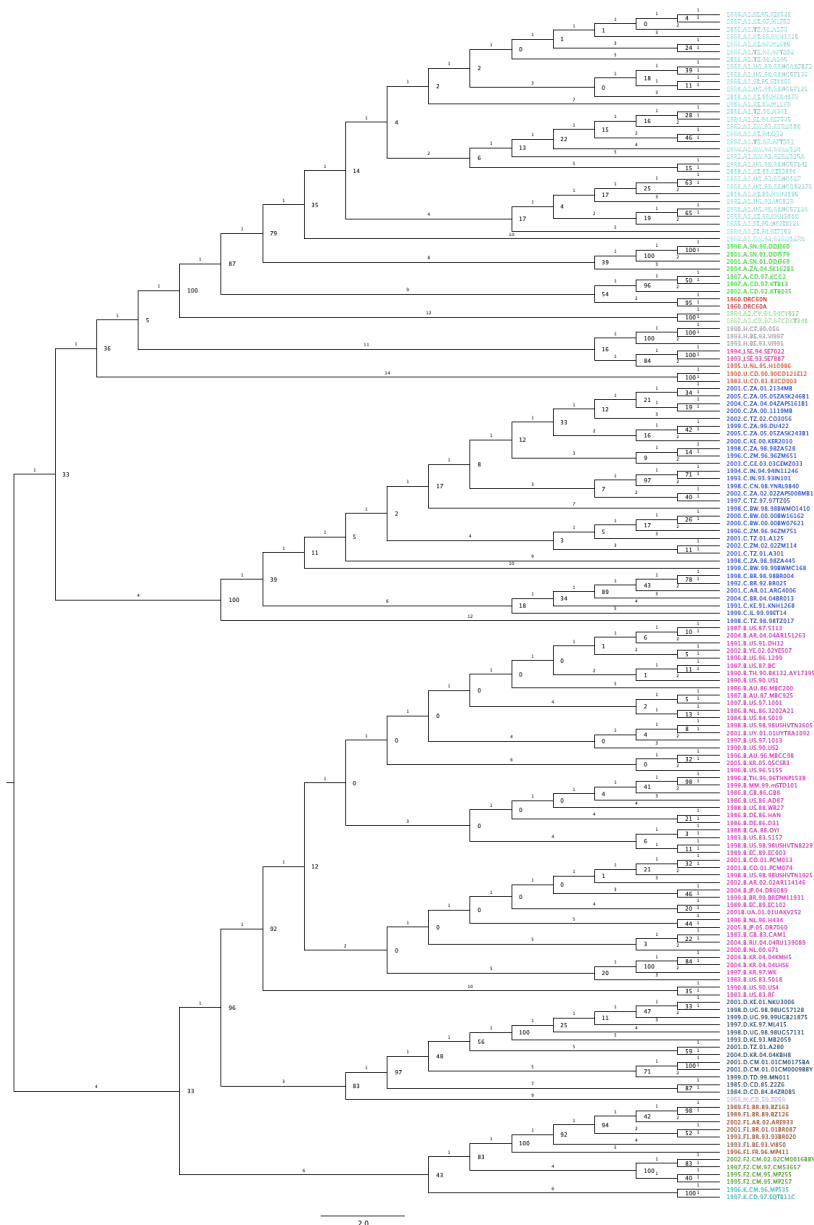
Reconstruct a maximum-likelihood tree (PhyML) tree using a GTR model of evolution and gamma-distributed rate variation among sites. Except for cluster support, most questions can be addressed using this tree, including the TempEst analysis below. To assess cluster support, repeat the same maximum likelihood inference with 100 bootstrap replicates, which will require a long run time. Root the tree using midpoint rooting in FigTree and answer the following specific questions:

Below I present the tree that I obtained from the 100 bootstrap replicates, without the bootstrap values with the PhyML. Later I will present the closer parts of the tree for better visualisation and detail.



1. Do the subtypes form monophyletic clades? If so, is there good support for this?

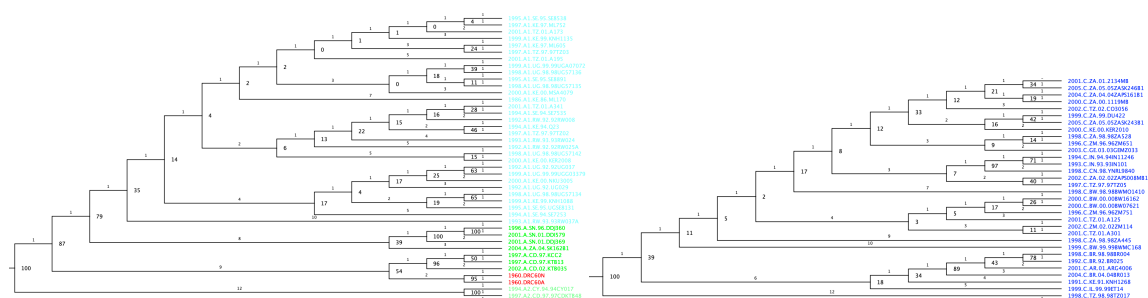
Yes most the subtypes form monophyletic clades. From one common ancestor (one particular node) come only the taxa from the same subtype. The only exceptions are the subtypes U that at some point diverged also into the J subtype and A that also diverged into the A1 and 1960 subtype. For most of the clades we see clear separation from one common ancestor to the taxa presenting one subtype. Below I present the photo with the stress on the particular subtypes.



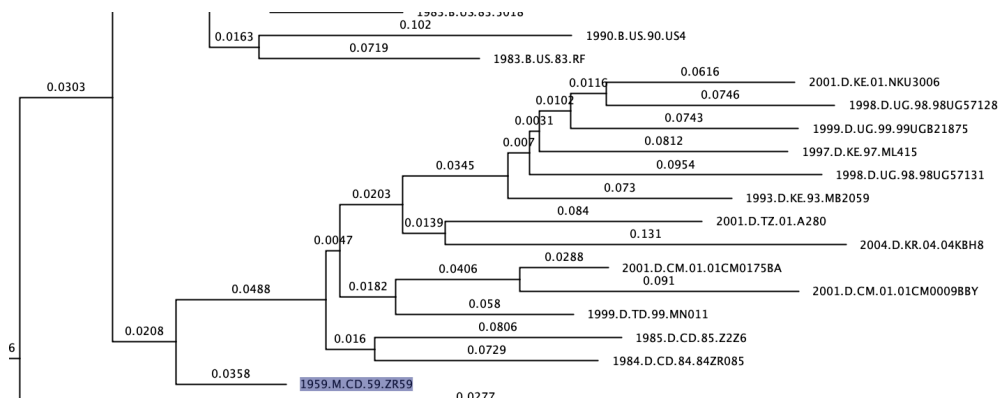
2. Without doing detailed calculations, which subtype do you expect to have the highest diversity: subtype A or subtype C? Why?

If we were looking at the subtype A of subtype A, Subtype C has more nodes than that subtype A does, however subtype A doesn't form a monophyletic clade. Yet by summing the genetic distance from the subtype C we would receive the bigger values than from the subtype A, consequently I would expect subtype C to have a higher diversity than subtype A of subtype A does.

Yet if we would look at the all subtypes of A (A1, A, A2) in comparison to subtype C, we would see that the subtype A has more taxa than C does. Additionally subtype A also diverged into strain 1960. Subtype C has more regular divergence into taxa than subtype A does, consequently the subtype A has a bigger chance of creating the subtypes of subtype - simultaneously creating the more divergent taxa.



3. Which sequence is most similar to the 1959 strain? What is the divergence between these two sequences?

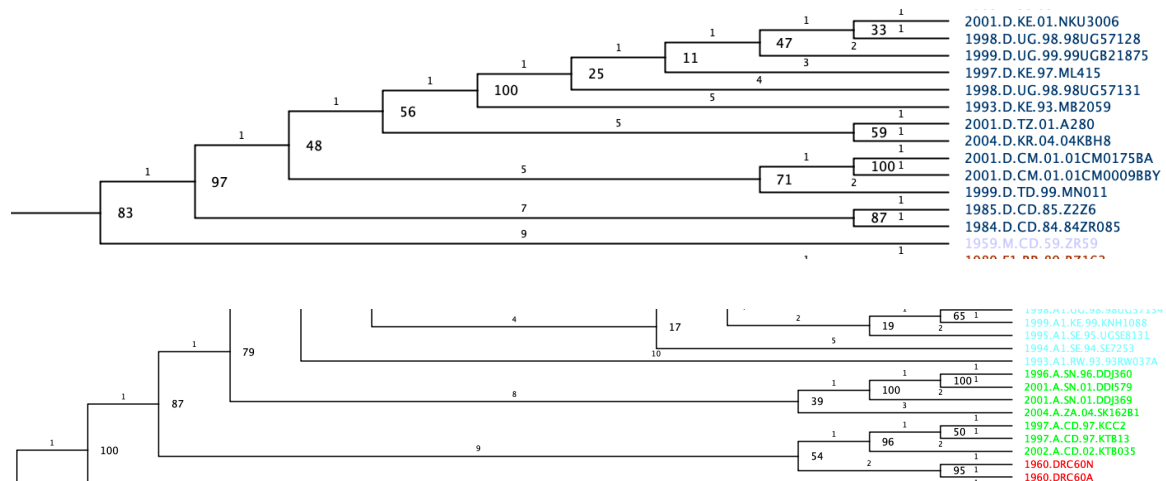


In the picture I stressed the 1959 strain. Basing on the look of the tree I would suspect one of the strains up the 1959 strain to be the most similar one. Below I prepared the table with the comparison of genetic distances the most probable strains. Basing on the results of sums of genetic distances of branches I can clearly say that the strain from 1999 year is the most similar one to the 1959 with the genetic distance (divergence) 0,1655.

Compared strains	The genetic distance (divergence)
1959 ~ 1984	0,1735
1959 ~ 1985	0,1812
1959 ~ 1999	0,1655
1959 ~ 2001	0,2391
1959 ~ 2001	0,1769

4. Do the 1959 and the two 1960 sequences fall in a subtype cluster. If so, which one? If not, to which subtype is/are the sequence(s) most closely related?

The strains 1959 and 1960 do not fall in the same subtype cluster. Below I present the screen with the evidence of that conclusion. The strain 1959 is the most related to the subtype D - they share their most recent common ancestor, the strain 1960 shares theirs most recent common ancestor with the subtype A.

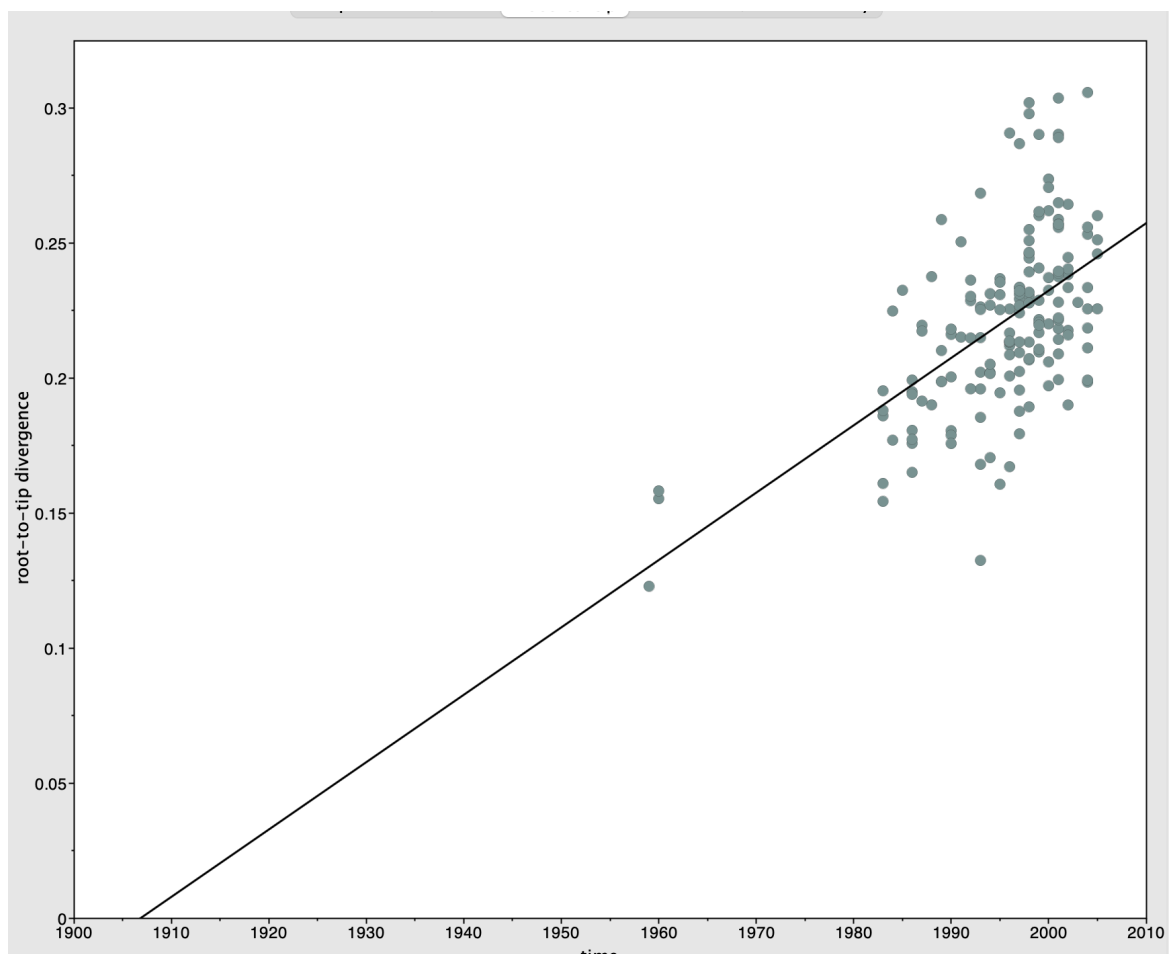


## Part 2

Use the reconstructed maximum-likelihood tree for a TempEst analysis and answer the following questions:

1. Is there a temporal signal in this HIV-1 data set?

Below I present the temporal signal in the HIV-1 data set. For the best fitting root the R is 0.3172 - this means that the signal is not as strong as we would like it. The higher R the better we can describe our data with the model. In our case the R is low consequently it doesn't explain the data correctly. Basing on the look of the linear curve we would expect the rise of the genetic distance with time.





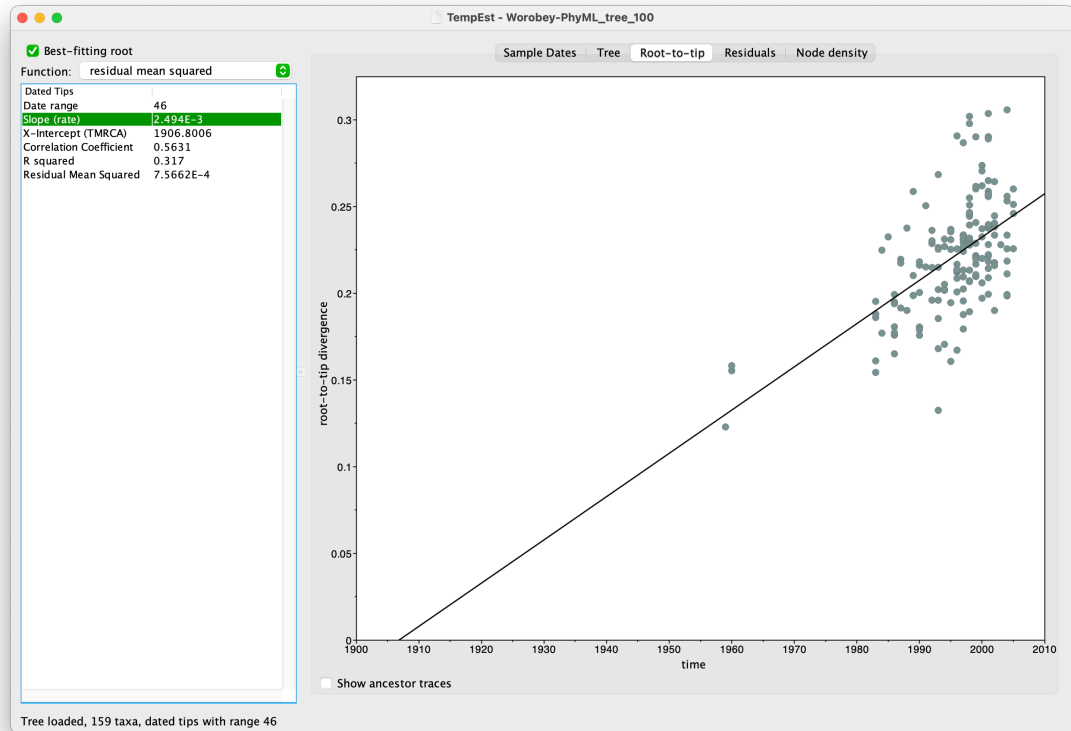
2. Does the best-fit root under the R-squared function correspond to the midpoint rooting used above?

It does not correspond. Below the comparison of appearance of subtypes in particular softwares.

FigTree	TempEst
A1	B
A	D
1960	M
A2	F1
H	F2
J	K
U	C
C	U
B	H
D	J
M	U
F1	A1
F2	A
K	1960
	A2

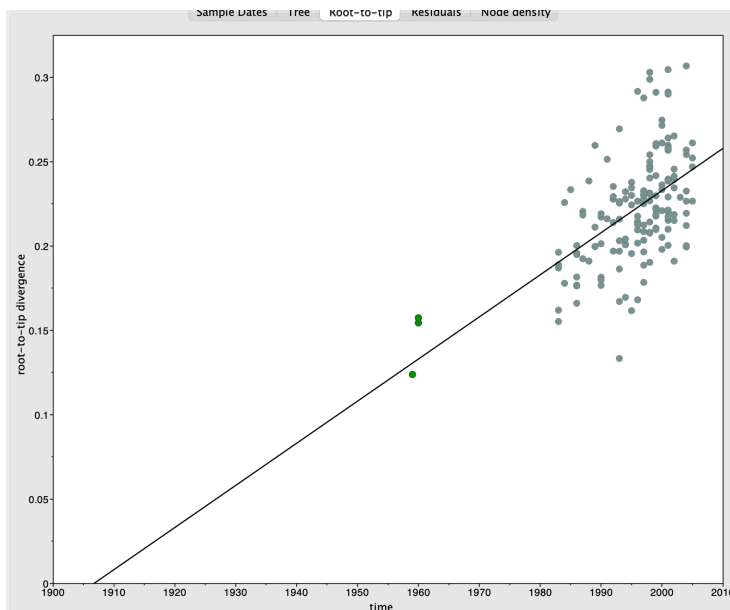
3. What would be the point estimate of the evolutionary rate based on this regression analysis?

2,4963E-3



4. Do the 1960 sequences and the 1959 sequence show more or less divergence as expected for their sampling dates based on the fitted regression line?

The tree shows bigger divergence than the model presents. 3 points on the left of the model are the strains from 1959 and 1960, when in tree they are separated by several other subtypes.



5. Should they be considered as outliers?

They could be considered as outliers if there was one strain significantly different than the others, but with 3 nearby each other we can suspect that in 60s the strains of HIV-1 were a bit different than they are nowadays. What is most important they do not seem to be contaminated (one of them should be in that case in different position, further than the others, and they are not).