Evolutionary and Quantitative Genetics

Task 1

Daria Plewa
r0976669
Philippe Lemey

8 October 2023

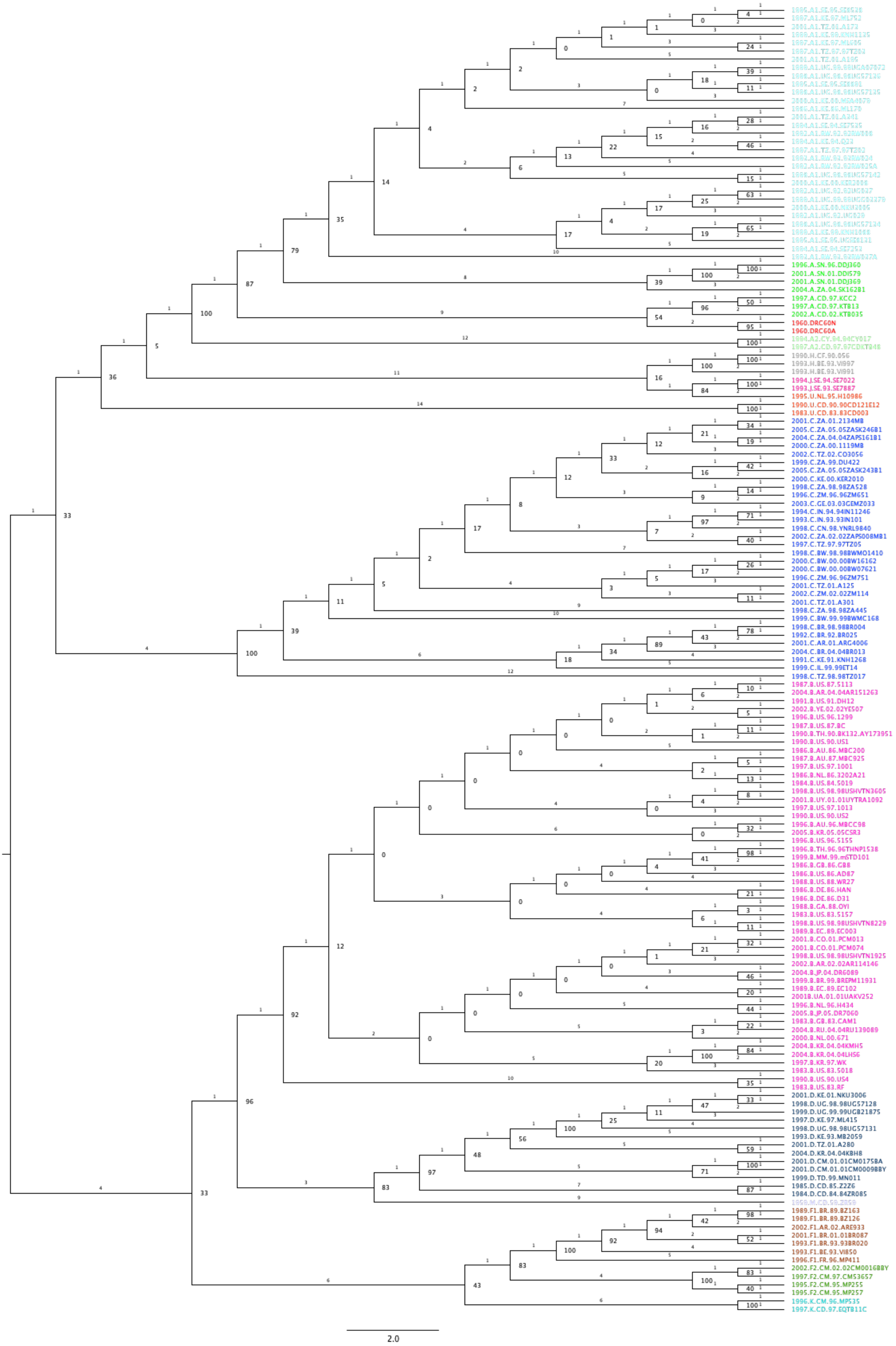
List of contents

1. Part 1
2. Part 2

Part 1

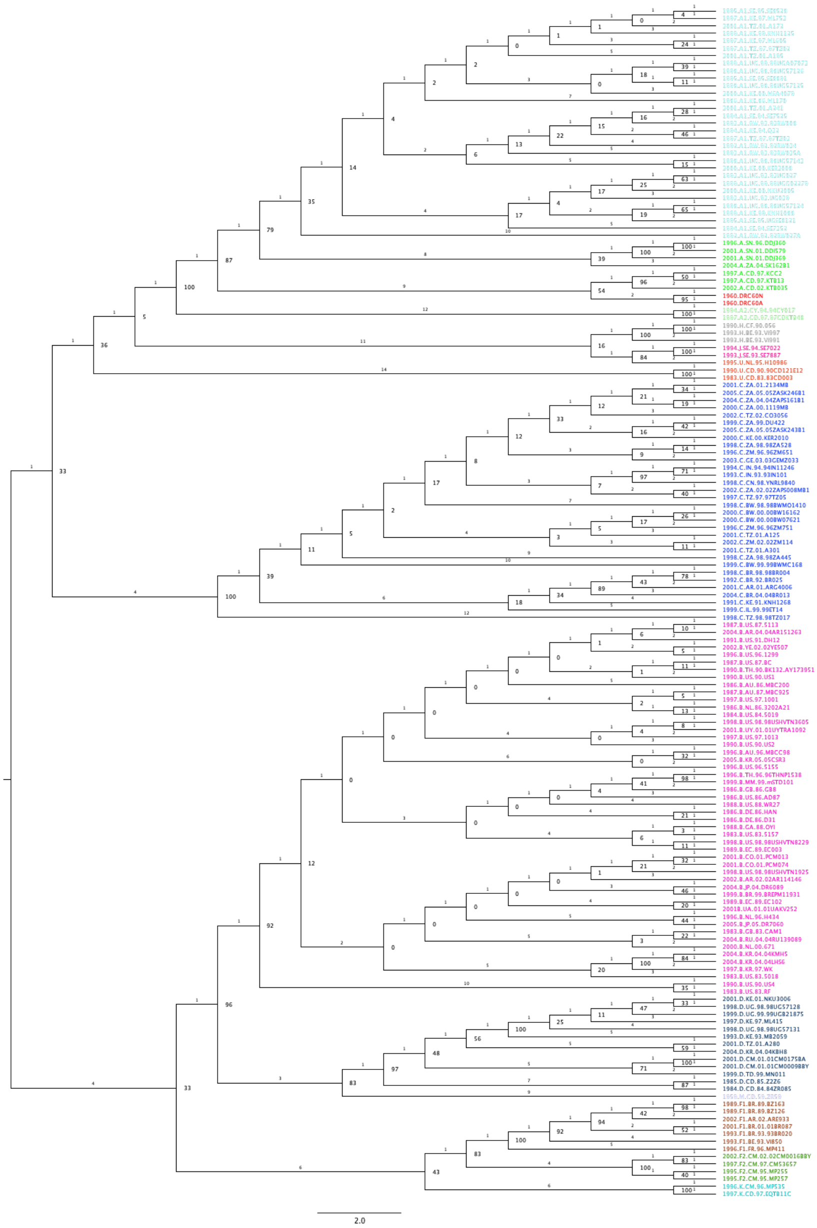
Reconstruct a maximum-likelihood tree (PhyML) tree using a GTR model of evolution and gamma-distributed rate variation among sites. Except for cluster support, most questions can be addressed using this tree, including the TempEst analysis below. To assess cluster support, repeat the same maximum likelihood 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.


1995.A1.SE.95.SE8538
1997.A1.KE.97.ML752
2001.A1.TZ.01.A173
1999.A1.KE.99.KNH1135
1997.A1.KE.97.ML605
1997.A1.TZ.97.97TZ03
2001.A1.TZ.01.A195
1999.A1.UG.99.99UGA07072
1998.A1.UG.98.98UG57136
1995.A1.SE.95.SE8891
1998.A1.UG.98.98UG57135
2000.A1.KE.00.MSA4070
1986.A1.KE.86.ML170
2001.A1.TZ.01.A341
1994.A1.SE.94.SE7535
1992.A1.RW.92.92RW008
1994.A1.KE.94.Q23
1997.A1.TZ.97.97TZ02
1993.A1.RW.93.93RW024
1992.A1.RW.92.92RW025A
1998.A1.UG.98.98UG57143
2000.A1.KE.00.KER2008
1992.A1.UG.92.92UG037
1999.A1.UG.99.99UGC03379
2000.A1.KE.00.MKU3005
1992.A1.UG.92.UG029
1998.A1.UG.98.98UG57134
1999.A1.KE.99.KNH1088
1995.A1.SE.95.UGSE8131
1994.A1.SE.94.SE7253
1993.A1.RW.93.93RW037A
1996.A.SN.96.DDJ360
2001.A.SN.01.DDI579
2001.A.SN.01.DDJ369
2004.A.ZA.04.SK162B1
1997.A.CD.97.KCC2
1997.A.CD.97.KTB13
2002.A.CD.02.KTB035
1960.DRC60N
1960.DRC60A
1994.A2.CY.94.94CY017
1997.A2.CD.97.97CDKTB48
1990.H.CF.90.056
1993.H.BE.93.VI997
1993.H.BE.93.VI991
1994.J.SE.94.SE7022
1993.J.SE.93.SE7887
1995.U.NL.95.H10986
1990.U.CD.90.90CD121E12
1983.U.CD.83.83CD003
2001.C.ZA.01.2134MB
2005.C.ZA.05.05ZASK246B1
2004.C.ZA.04.04ZAPS161B1
2000.C.ZA.00.1119MB
2002.C.TZ.02.CO3056
1999.C.ZA.99.DU422
2005.C.ZA.05.05ZASK243B1
2000.C.KE.00.KER2010
1998.C.ZA.98.98ZA528
1996.C.ZM.96.96ZM651
2003.C.GE.03.03GEMZ033
1994.C.IN.94.94IN11246
1993.C.IN.93.93IN101
1998.C.CN.98.YNRL9840
2002.C.ZA.02.02ZAPS008MB1
1997.C.TZ.97.97TZ05
1998.C.BW.98.98BWMO1410
2000.C.BW.00.00BW16162
2000.C.BW.00.00BW07621
1996.C.ZM.96.96ZM751
2001.C.TZ.01.A125
2002.C.ZM.02.02ZM114
2001.C.TZ.01.A301
1998.C.ZA.98.98ZA445
1999.C.BW.99.99BWMC168
1998.C.BR.98.98BR004
1992.C.BR.92.BR025
2001.C.AR.01.ARG4006
2004.C.BR.04.04BR013
1991.C.KE.91.KNH1268
1999.C.IL.99.99ET14
1998.C.TZ.98.98TZ017
1987.B.US.87.5113
2004.B.AR.04.04AR151263
1991.B.US.91.DH12
2002.B.YE.02.02YE507
1996.B.US.96.1299
1987.B.US.87.BC
1990.B.TH.90.BK132.AY173951
1990.B.US.90.US1
1986.B.AU.86.MBC200
1987.B.AU.87.MBC925
1997.B.US.97.1001
1986.B.NL.86.3202A21
1984.B.US.84.5019
1998.B.US.98.98USHVTN3605
2001.B.UY.01.01UYTRA1092
1997.B.US.97.1013
1990.B.US.90.US2
1996.B.AU.96.MBCC98
2005.B.KR.05.05CSR3
1996.B.US.96.5155
1996.B.TH.96.96THNP1538
1999.B.MM.99.mSTD101
1986.B.GB.86.GB8
1986.B.US.86.AD87
1988.B.US.88.WR27
1986.B.DE.86.HAN
1986.B.DE.86.D31
1988.B.GA.88.OYI
1983.B.US.83.5157
1998.B.US.98.98USHVTN8229
1989.B.EC.89.EC003
2001.B.CO.01.PCM013
2001.B.CO.01.PCM074
1998.B.US.98.98USHVTN1925
2002.B.AR.02.02AR114146
2004.B.JP.04.DR6089
1999.B.BR.99.BREPM11931
1989.B.EC.89.EC102
2001B.UA.01.01UAKV252
1996.B.NL.96.H434
2005.B.JP.05.DR7060
1983.B.GB.83.CAM1
2004.B.RU.04.04RU139089
2000.B.NL.00.671
2004.B.KR.04.04KMH5
2004.B.KR.04.04LHS6
1997.B.KR.97.WK
1983.B.US.83.5018
1990.B.US.90.US4
1983.B.US.83.RF
2001.D.KE.01.NKU3006
1998.D.UG.98.98UG57128
1999.D.UG.99.99UGB21875
1997.D.KE.97.ML415
1998.D.UG.98.98UG57131
1993.D.KE.93.MB2059
2001.D.TZ.01.A280
2004.D.KR.04.04KBH8
2001.D.CM.01.01CM0175BA
2001.D.CM.01.01CM0009BBY
1999.D.TD.99.MN011
1985.D.CD.85.Z2Z6
1984.D.CD.84.84ZR085
1959.M.CD.59.ZR59
1989.F1.BR.89.BZ163
1989.F1.BR.89.BZ126
2002.F1.AR.02.ARE933
2001.F1.BR.01.01BR087
1993.F1.BR.93.93BR020
1993.F1.BE.93.VI850
1996.F1.FR.96.MP411
2002.F2.CM.02.02CM0016BBY
1997.F2.CM.97.CM53657
1995.F2.CM.95.MP255
1995.F2.CM.95.MP257
1996.K.CM.96.MP535
1997.K.CD.97.EQTB11C
2.0

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

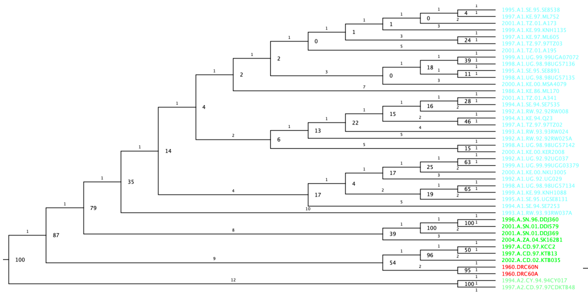
Yes most the subtypes form monophyletic clades. From one common ancestor (one particular node) come only the taxa from the same subtype. The only exceptions are the subtypes U that at some point diverged also into the J subtype and A that also diverged into the A1 and 1960 subtype. For most of the clades 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.

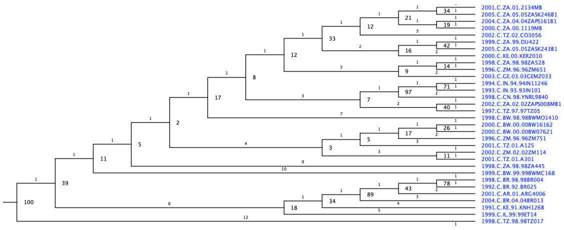


2. Without doing detailed calculations, which subtype do you expect to have the highest diversity: subtype A or subtype C? Why?

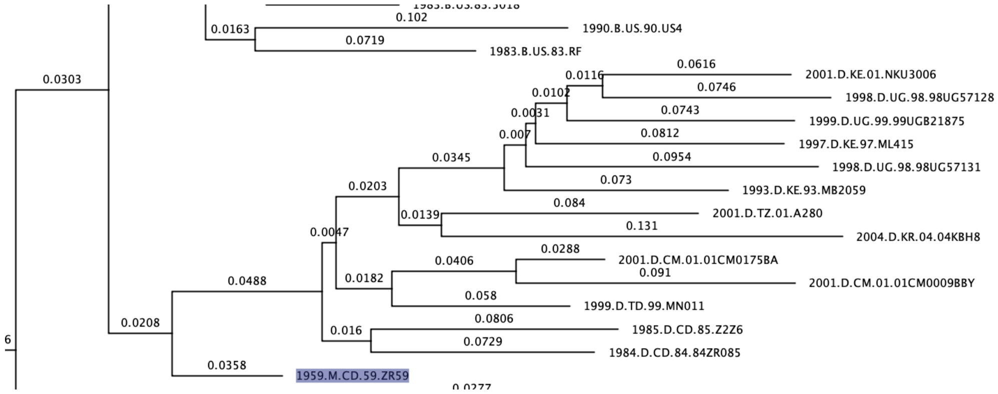
If we were looking at the subtype A of subtype A, Subtype C has more nodes than that subtype A does, however subtype A doesn't form a monophyletic clade. Yet by summing the genetic distance from the subtype C we would receive the bigger values than from the subtype A, consequently I would expect subtype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

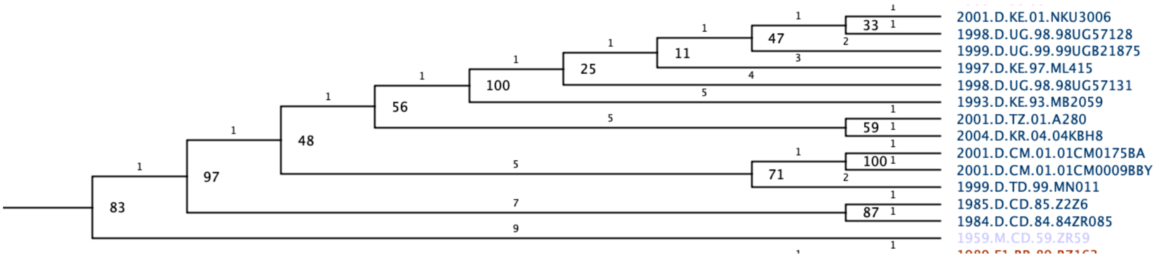


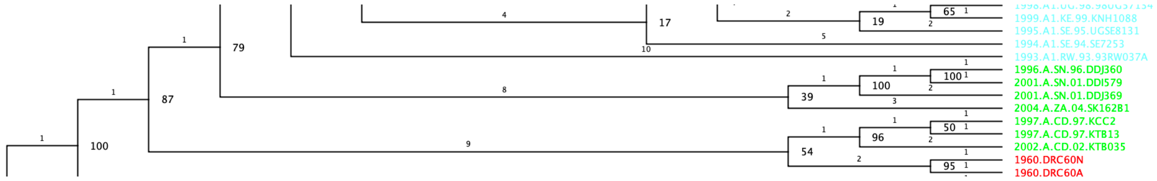
In the picture I stressed the 1959 strain. Basing on the look of the tree I would suspect one of the strains up the 1959 strain to be the most similar one. Below I prepared the table with the comparison of genetic distances the most probable strains. Basing on the results of sums of genetic distances 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.



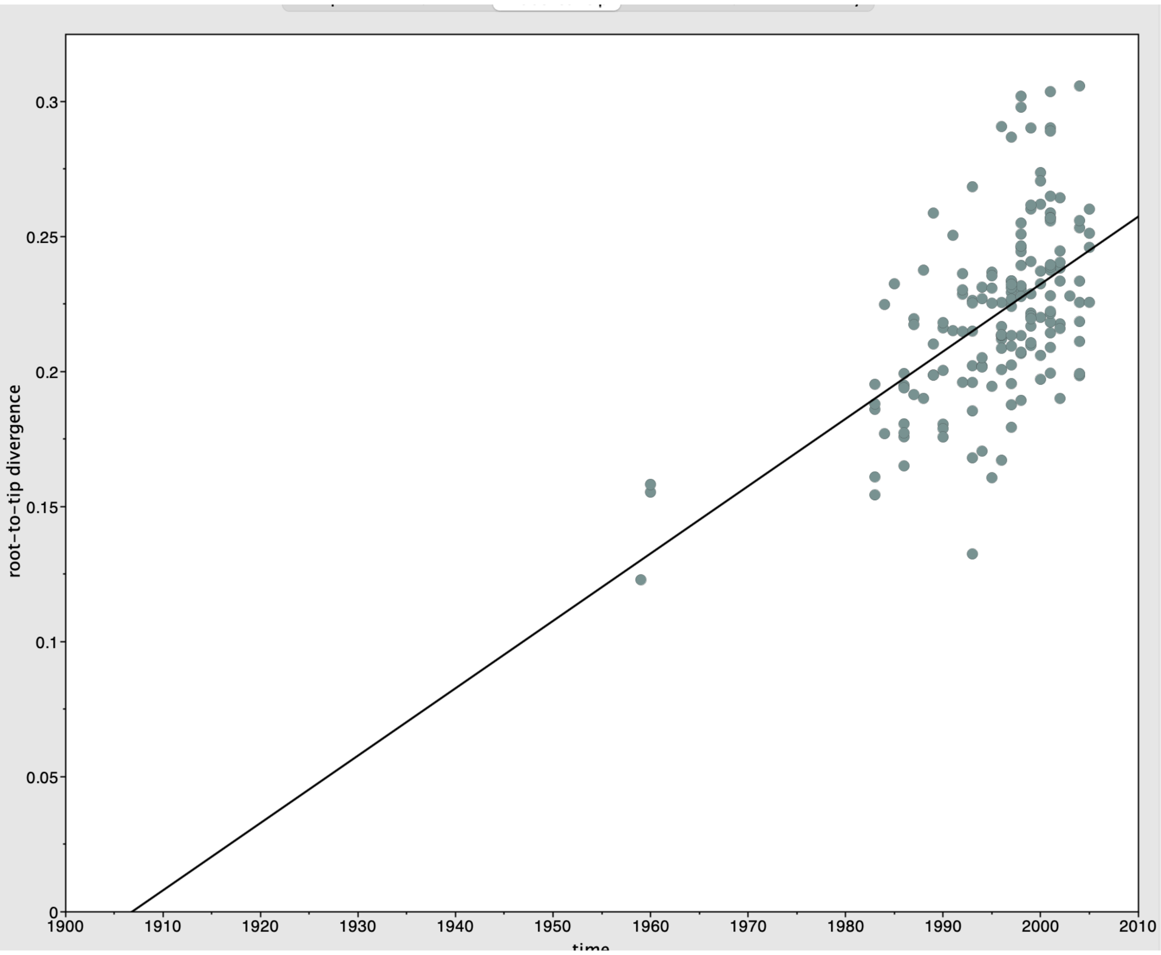


Part 2

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.



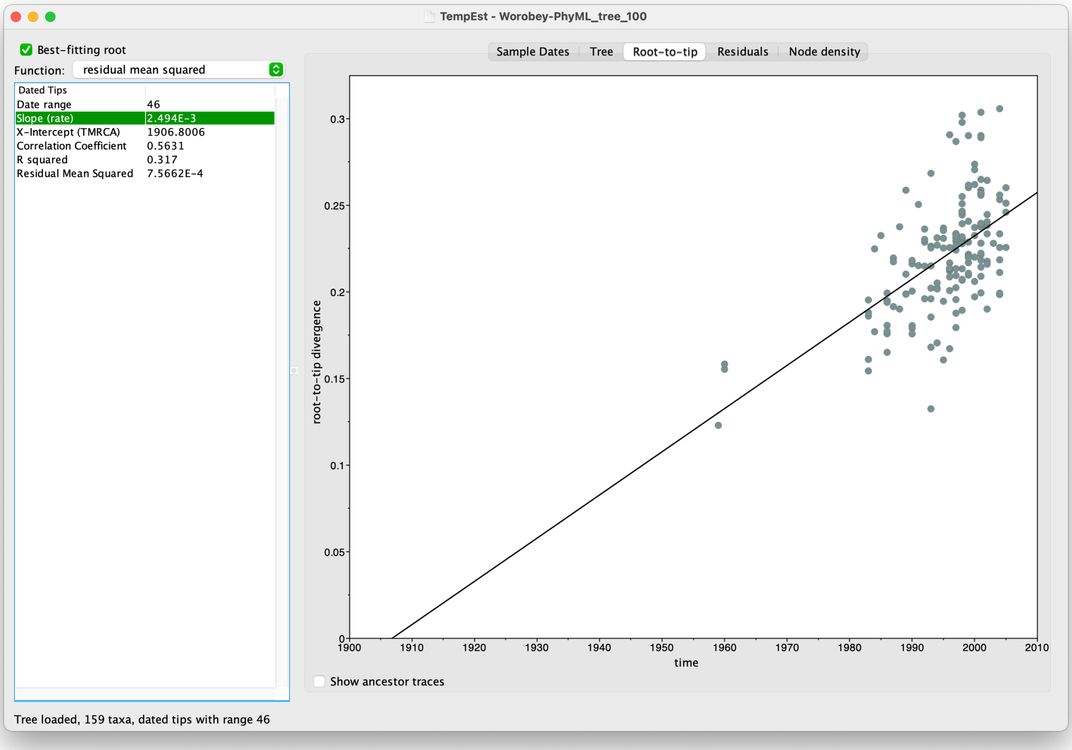
2. Does the best-fit root under the R-squared function correspond to the midpoint rooting used above?

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



☑ Best-fitting root

Function: residual mean squared

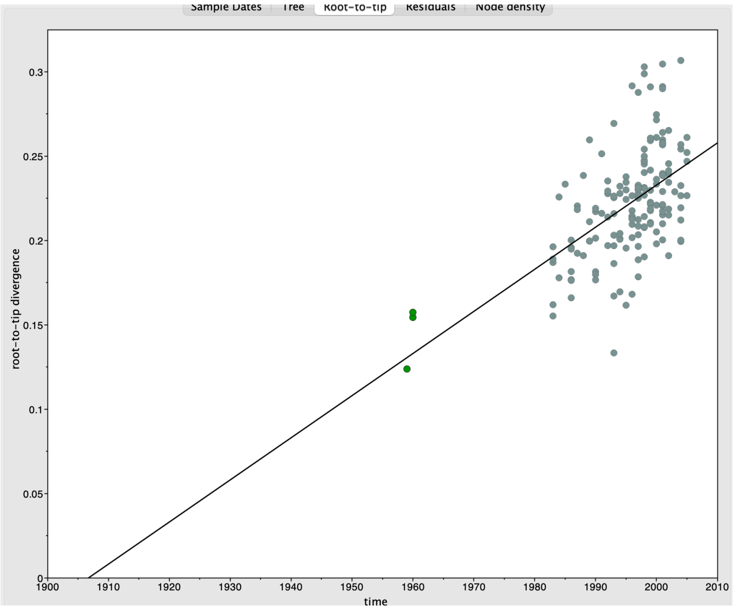
Dated Tips	
Date range	46
Slope (rate)	2.494E-3
X-Intercept (TMRCA)	1906.8006
Correlation Coefficient	0.5631
R squared	0.317
Residual Mean Squared	7.5662E-4

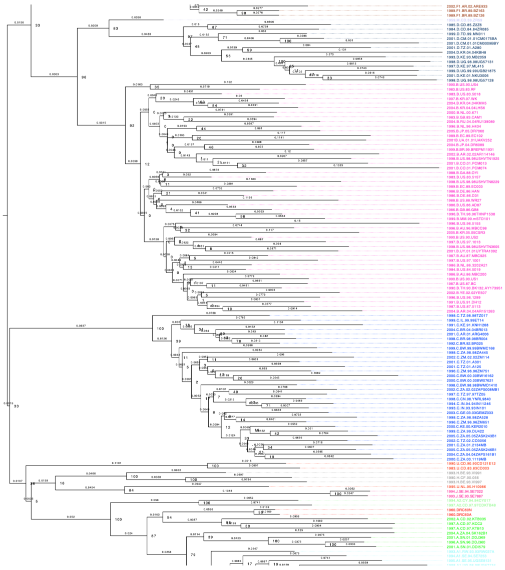
□ Show ancestor traces

Tree loaded, 159 taxa, dated tips with range 46

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 a 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).