



Tutorial: Using TempEst for data exploration

This tutorial describes the use of TempEst to examine the temporal signal of a data set and to look for problematic or erroneous sequences.

In this tutorial, we will explore the use of the interactive graphical program TempEst (tempest) (formerly known as Path-O-Gen) to examine virus sequence data that has been sampled through time to look for problematic sequences and to explore the degree and pattern of temporal signal. This can be a useful way of examining the data for potential issues before committing significant time to running BEAST (beast).

Step by step:

1. Building a non-molecular clock tree (IQ-Tree)
2. Running TempEst and loading the tree

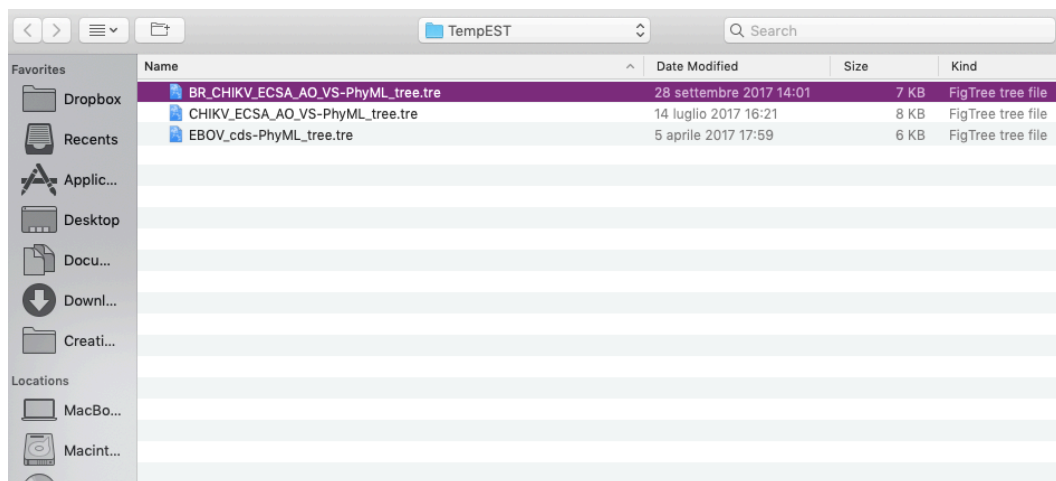
2. Running TempEst and loading the tree

Run TempEst (tempest) by double clicking on its icon.

TempEst is an interactive graphical application for examining the temporal signal in a tree of time-stamped sequences by plotting the divergence of each tip from the root against the date of sampling (a root-to-tip plot).

Once running, TempEst will look similar irrespective of which computer system it is running on. For this tutorial, the Mac OS X version will be shown but the Linux & Windows versions will have exactly the same layout and functionality.

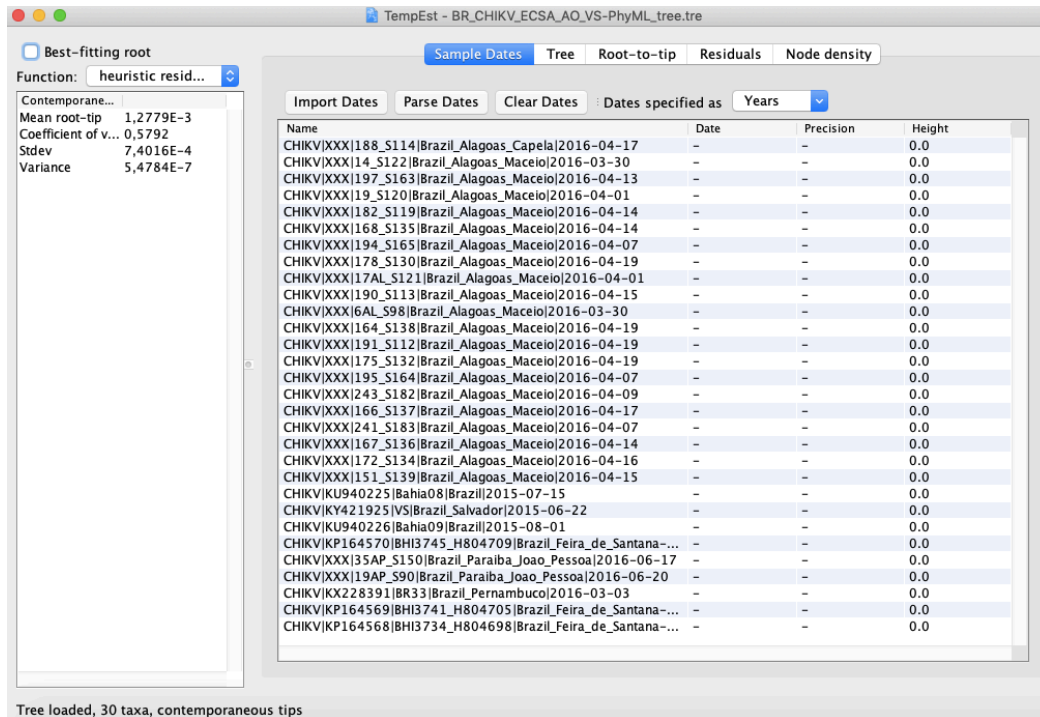
When started, TempEst will immediately display a file selection dialog box in which you can select the tree that you made in the previous section.



Go to the `/home/manager/Phylogenetics/3.Temporal_Signal/TempEST` folder, select the `BR_CHIKV_ECSA_AO_VS-PhyML_tree.tre` file and click Open .

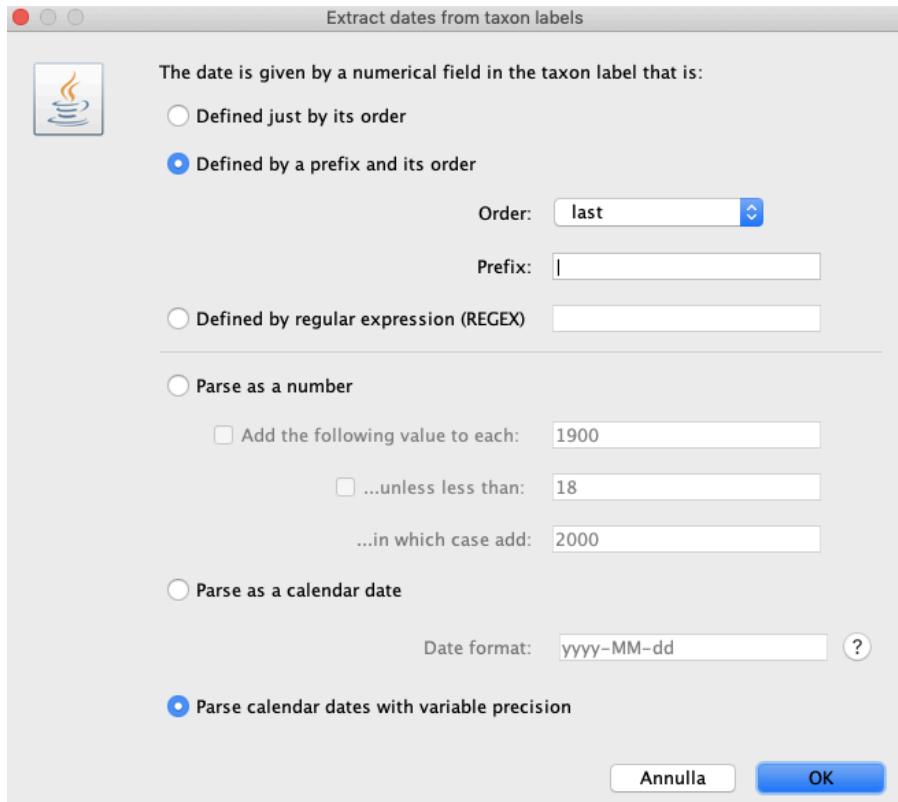
Parsing dates of sampling

Once the tree is loaded the main window will appear and look like this:



Ignore the panel on the left for the moment. The first thing that needs doing is to give the date of sampling to each of the sequences.

The actual year of sampling is given at the end of the name of each taxon. To specify the dates of the sequences in BEAUti we will use the Parse Dates button at the top of the panel. Clicking this will make a dialog box appear:



Extract dates from taxon labels

The date is given by a numerical field in the taxon label that is:

☐ Defined just by its order

☒ Defined by a prefix and its order

Order: last

Prefix: |

☐ Defined by regular expression (REGEX)

☐ Parse as a number

☐ Add the following value to each: 1900

☐ ...unless less than: 18

☐ ...in which case add: 2000

☐ Parse as a calendar date

Date format: yyyy-MM-dd ?

☒ Parse calendar dates with variable precision

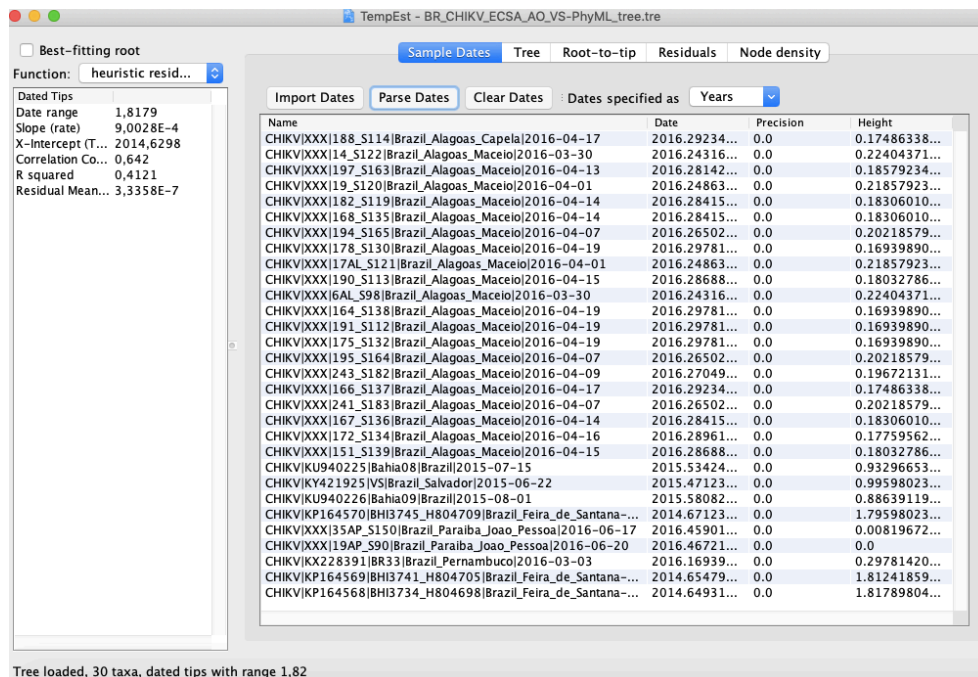
Annulla OK

This operation attempts to extract the dates from the taxon names. It works by trying to find a numerical field within each name. This dialog box is the same as that in BEAUti (beauti) and there are a wide range of options for doing this.

For these sequences you can set the options to look like the figure above: Defined just by its order
Order: last and Parse calendar dates with variable precision.

So, we can press OK .

The table will now have the year of sampling for each virus in the Dates column. Click on the Dates column header to sort the dates and check that they are all correct.

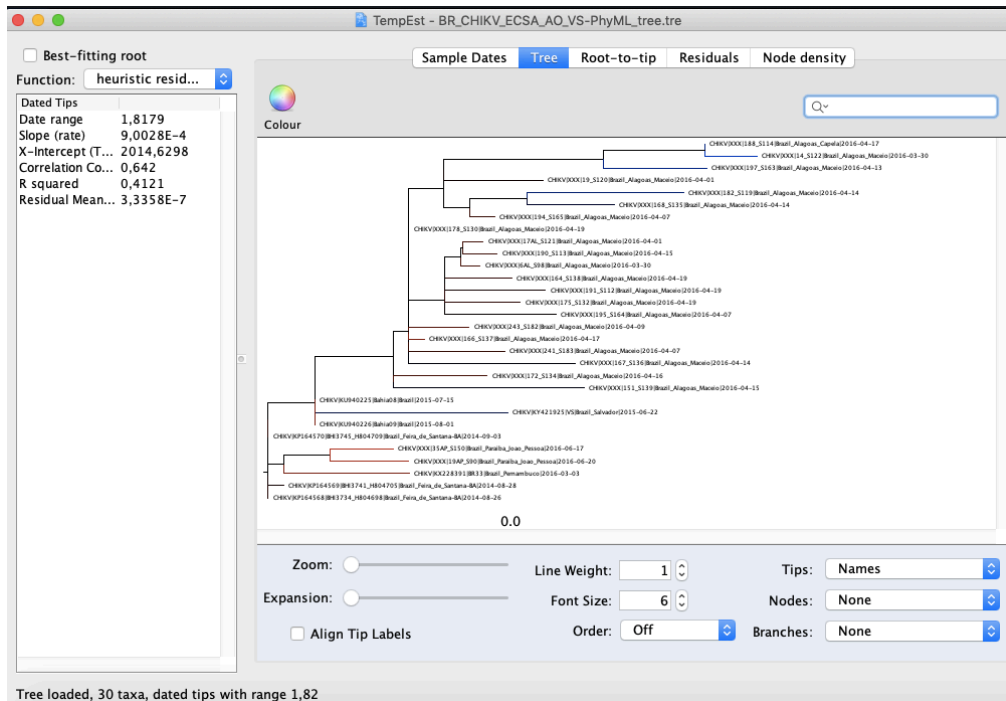


The screenshot shows the TempEst software window titled "TempEst - BR_CHIKV_ECSA_AO_VS-PhyML_tree.tre". The "Sample Dates" tab is selected. On the left, a "Function:" dropdown is set to "heuristic resid..." and a "Dated Tips" panel shows statistics: Date range 1,8179, Slope (rate) 9,0028E-4, X-Intercept (T...) 2014,6298, Correlation Co... 0,642, R squared 0,4121, and Residual Mean... 3,3358E-7. The main panel displays a table with columns: Name, Date, Precision, and Height. The table lists 30 taxa with their corresponding dates and values. At the bottom, a status bar indicates "Tree loaded, 30 taxa, dated tips with range 1,82".

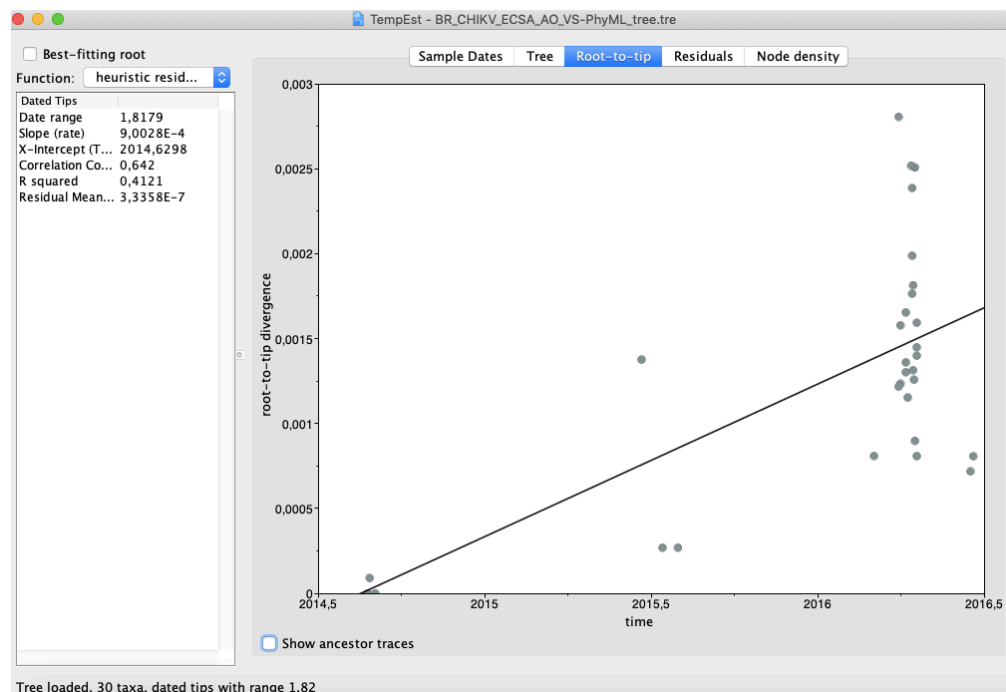
Name	Date	Precision	Height
CHIKV XXX 188_S114 Brazil_Alagoas_Capela 2016-04-17	2016.29234...	0.0	0.17486338...
CHIKV XXX 14_S122 Brazil_Alagoas_Maceio 2016-03-30	2016.24316...	0.0	0.22404371...
CHIKV XXX 197_S163 Brazil_Alagoas_Maceio 2016-04-13	2016.28142...	0.0	0.18579234...
CHIKV XXX 19_S120 Brazil_Alagoas_Maceio 2016-04-01	2016.24863...	0.0	0.21857923...
CHIKV XXX 182_S119 Brazil_Alagoas_Maceio 2016-04-14	2016.28415...	0.0	0.18306010...
CHIKV XXX 168_S135 Brazil_Alagoas_Maceio 2016-04-14	2016.28415...	0.0	0.18306010...
CHIKV XXX 194_S165 Brazil_Alagoas_Maceio 2016-04-07	2016.26502...	0.0	0.20218579...
CHIKV XXX 178_S130 Brazil_Alagoas_Maceio 2016-04-19	2016.29781...	0.0	0.16939890...
CHIKV XXX 17AL_S121 Brazil_Alagoas_Maceio 2016-04-01	2016.24863...	0.0	0.21857923...
CHIKV XXX 190_S113 Brazil_Alagoas_Maceio 2016-04-15	2016.28688...	0.0	0.18032786...
CHIKV XXX 16AL_S98 Brazil_Alagoas_Maceio 2016-03-30	2016.24316...	0.0	0.22404371...
CHIKV XXX 164_S138 Brazil_Alagoas_Maceio 2016-04-19	2016.29781...	0.0	0.16939890...
CHIKV XXX 191_S112 Brazil_Alagoas_Maceio 2016-04-19	2016.29781...	0.0	0.16939890...
CHIKV XXX 175_S132 Brazil_Alagoas_Maceio 2016-04-19	2016.29781...	0.0	0.16939890...
CHIKV XXX 195_S164 Brazil_Alagoas_Maceio 2016-04-07	2016.26502...	0.0	0.20218579...
CHIKV XXX 243_S182 Brazil_Alagoas_Maceio 2016-04-09	2016.27049...	0.0	0.19672131...
CHIKV XXX 166_S137 Brazil_Alagoas_Maceio 2016-04-17	2016.29234...	0.0	0.17486338...
CHIKV XXX 241_S183 Brazil_Alagoas_Maceio 2016-04-07	2016.26502...	0.0	0.20218579...
CHIKV XXX 167_S136 Brazil_Alagoas_Maceio 2016-04-14	2016.28415...	0.0	0.18306010...
CHIKV XXX 172_S134 Brazil_Alagoas_Maceio 2016-04-16	2016.28961...	0.0	0.17759562...
CHIKV XXX 151_S139 Brazil_Alagoas_Maceio 2016-04-15	2016.28688...	0.0	0.18032786...
CHIKV KU940225 Bahia08 Brazil 2015-07-15	2015.53424...	0.0	0.93296653...
CHIKV KY421925 VS Brazil_Salvador 2015-06-22	2015.47123...	0.0	0.99598023...
CHIKV KU940226 Bahia09 Brazil 2015-08-01	2015.58082...	0.0	0.88639119...
CHIKV KP164570 BH13745_H804709 Brazil_Feira_de_Santana-...	2014.67123...	0.0	1.79598023...
CHIKV XXX 35AP_S150 Brazil_Paraiba_Joao_Pessoa 2016-06-17	2016.45901...	0.0	0.00819672...
CHIKV XXX 19AP_S90 Brazil_Paraiba_Joao_Pessoa 2016-06-20	2016.46721...	0.0	0.0
CHIKV XX228391 BR33 Brazil_Pernambuco 2016-03-03	2016.16939...	0.0	0.29781420...
CHIKV KP164569 BH13741_H804705 Brazil_Feira_de_Santana-...	2014.65479...	0.0	1.81241859...
CHIKV KP164568 BH13734_H804698 Brazil_Feira_de_Santana-...	2014.64931...	0.0	1.81789804...

The temporal signal and rooting

We can now explore the data using the tabs at the top of the window - Tree , Root-to-tip & Residuals. If you click on the Tree tab you will see the tree as loaded from the tree file. Because we constructed this tree using a non-molecular-clock model, it will be arbitrarily rooted. If you look at the date of each virus in the tree you will see that there is no correlation with the horizontal position:



Now switch to the Root-to-tip panel. This shows a plot of the divergence from the root of the tree against time of sampling (a so-called 'Root to tip plot'):

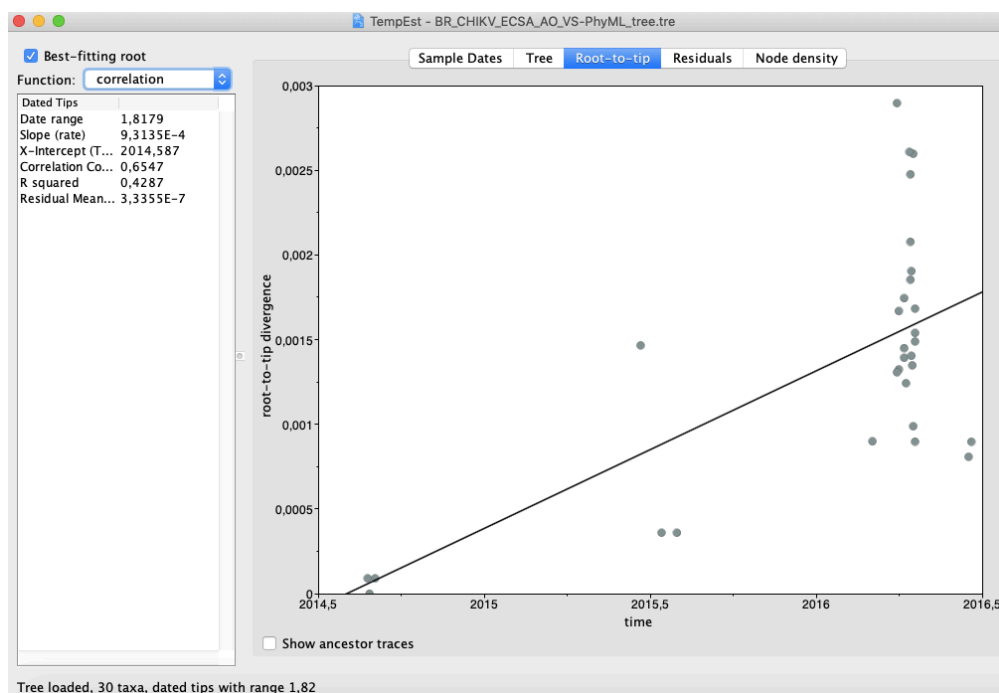


You can see that there is a quite nice correlation in this plot (the line is the best-fit regression). In the table on the left you can see the Correlation Coefficient is 0.64. Despite this, the correlation we are

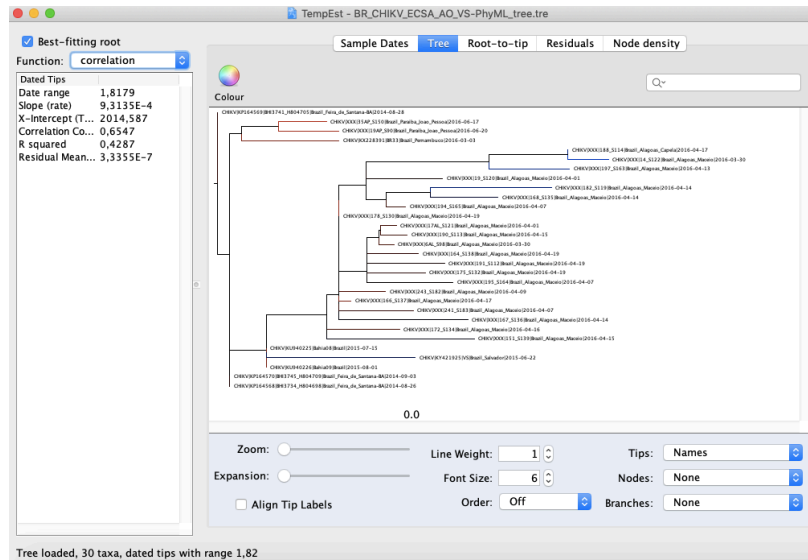
observed is not the correct one as the root is arbitrarily set by the phylogeny reconstruction software and thus divergence from root is meaningless.

TempEst can try to find the root of the tree that optimizes the temporal signal. It does this by trying all possible roots and picks the one that produces the optimal value of a range of statistics. The function it uses is selected in the menu at the top left. The options are to minimize the mean of the squares of the residuals (residual-meansquared), or to maximize the correlation coefficient (correlation) or R (R squared). These are all ad hoc procedures and no particular one is best but residual-mean-squared may be most consistent with the investigations here.

Click Best-fitting root to root the tree at the place that minimizes the mean of the squares of the residuals.

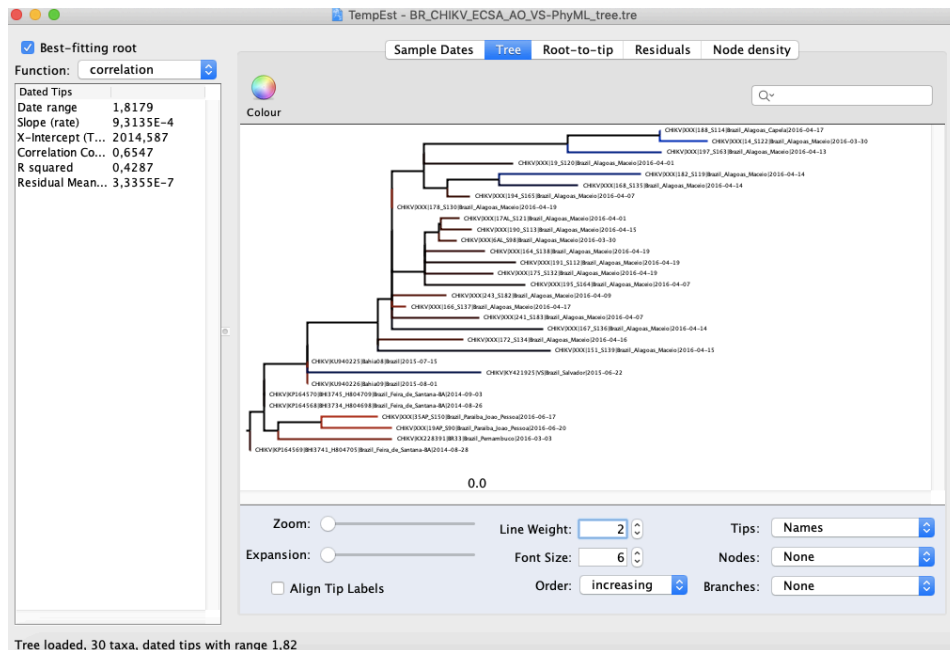


Now there is a better correlation between the dates of the tips and the divergence from the root (the correlation coefficient has nearly doubled). Return to the tree to look where the root was placed:



To make the tree easier to view, switch the Order option in the panel at the bottom to increasing .

The tree branches are coloured to show the residual with blue for tips with positive residuals (above the regression line), red for negative.



On the left hand side of the window there is a table of statistics:

Dated Tips	
Date range	1,8179
Slope (rate)	9,3135E-4
X-Intercept (T...	2014,587
Correlation Co...	0,6547
R squared	0,4287
Residual Mean...	3,3355E-7

As well as the statistical metrics (Correlation Coefficient , R squared and Residual Mean Squared) there are the following:

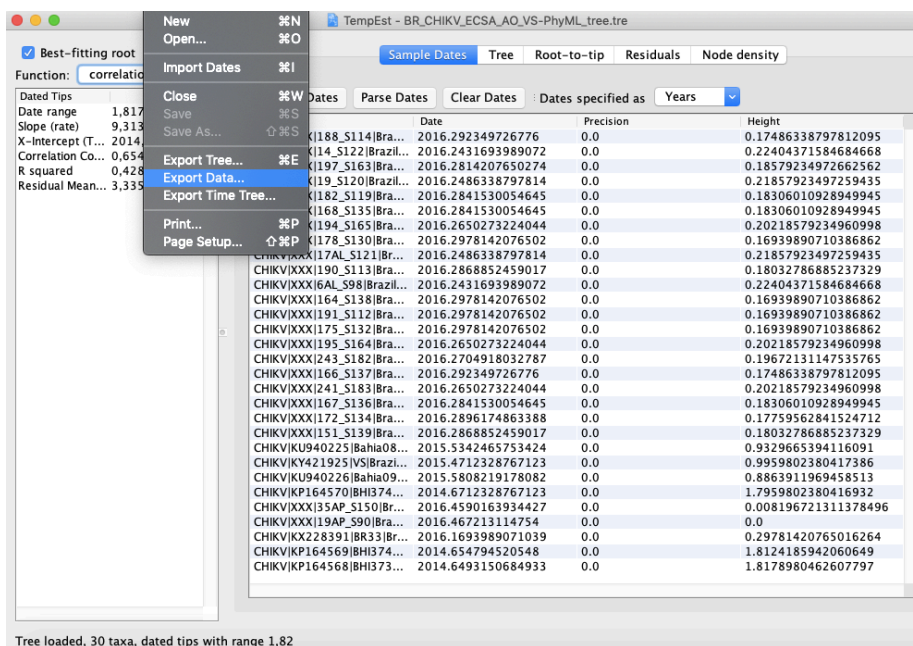
Date range: The span of dates for the viruses.

Slope (rate): The slope of the regression line. This is an estimate of the rate of evolution in substitutions per site per year.

X-Intercept (TMRCA): The point on the x-axis at which the regression line crosses. This is an estimate of the date of the root of the tree.

You can also extract the roo-to-tip data and plot is using the ggplot package in R Studio.

To do so go to “File” and them click on “Export Data”



Let's now estimate the temporal signal in all the other dataset!