



THE UNIVERSITY of EDINBURGH
Royal (Dick) School of
Veterinary Studies

Phylogenetics Workshop

Molecular clock and Tempest



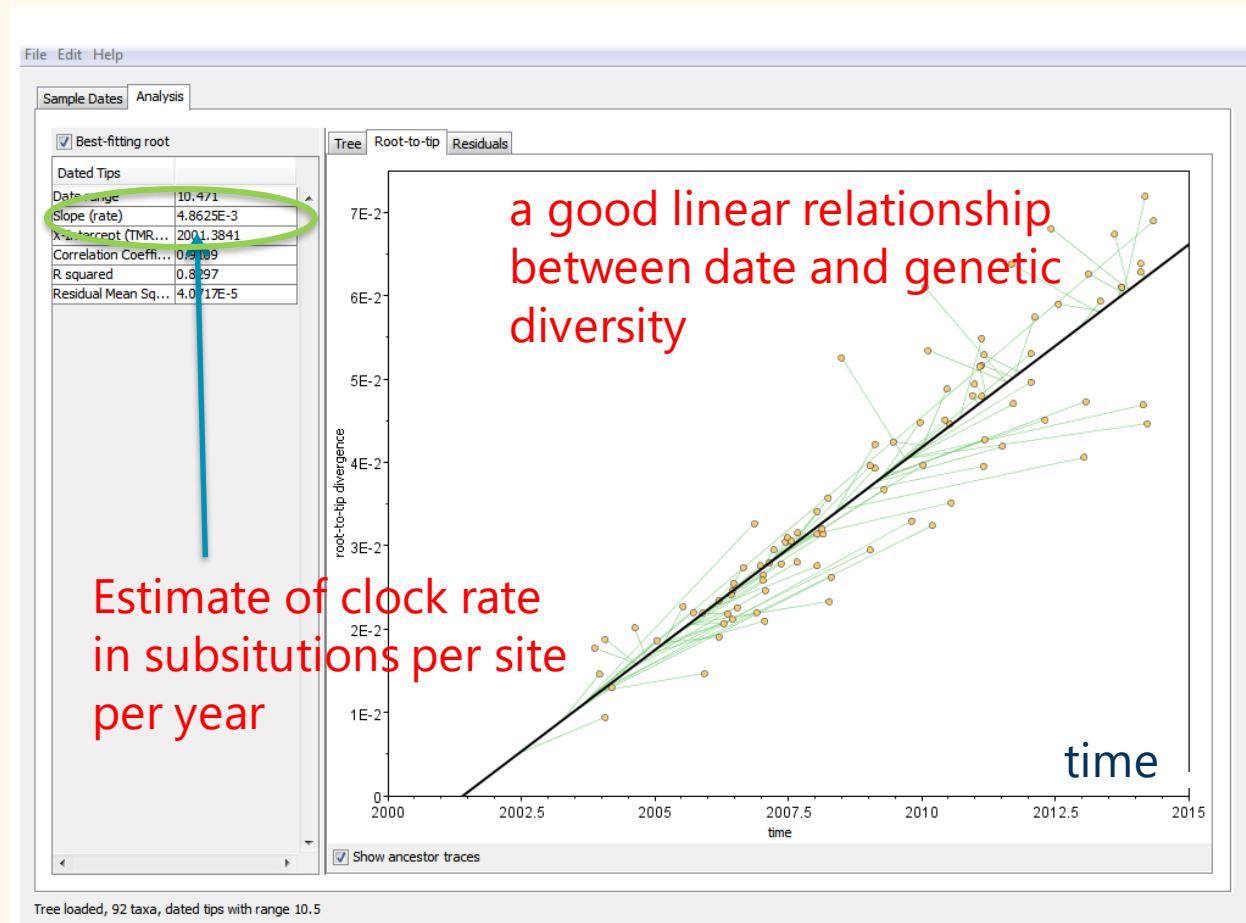
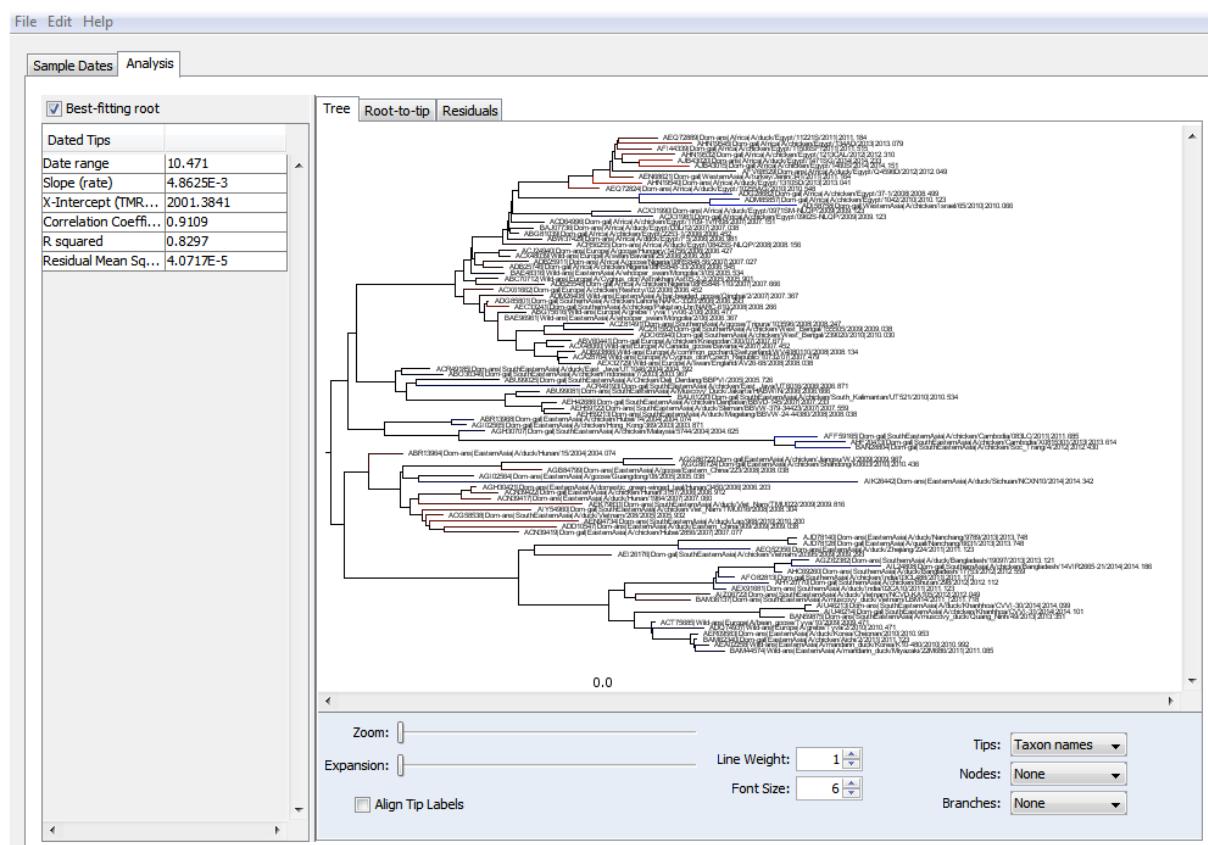
Biotechnology and
Biological Sciences
Research Council

Estimate of molecular clock using Tempest

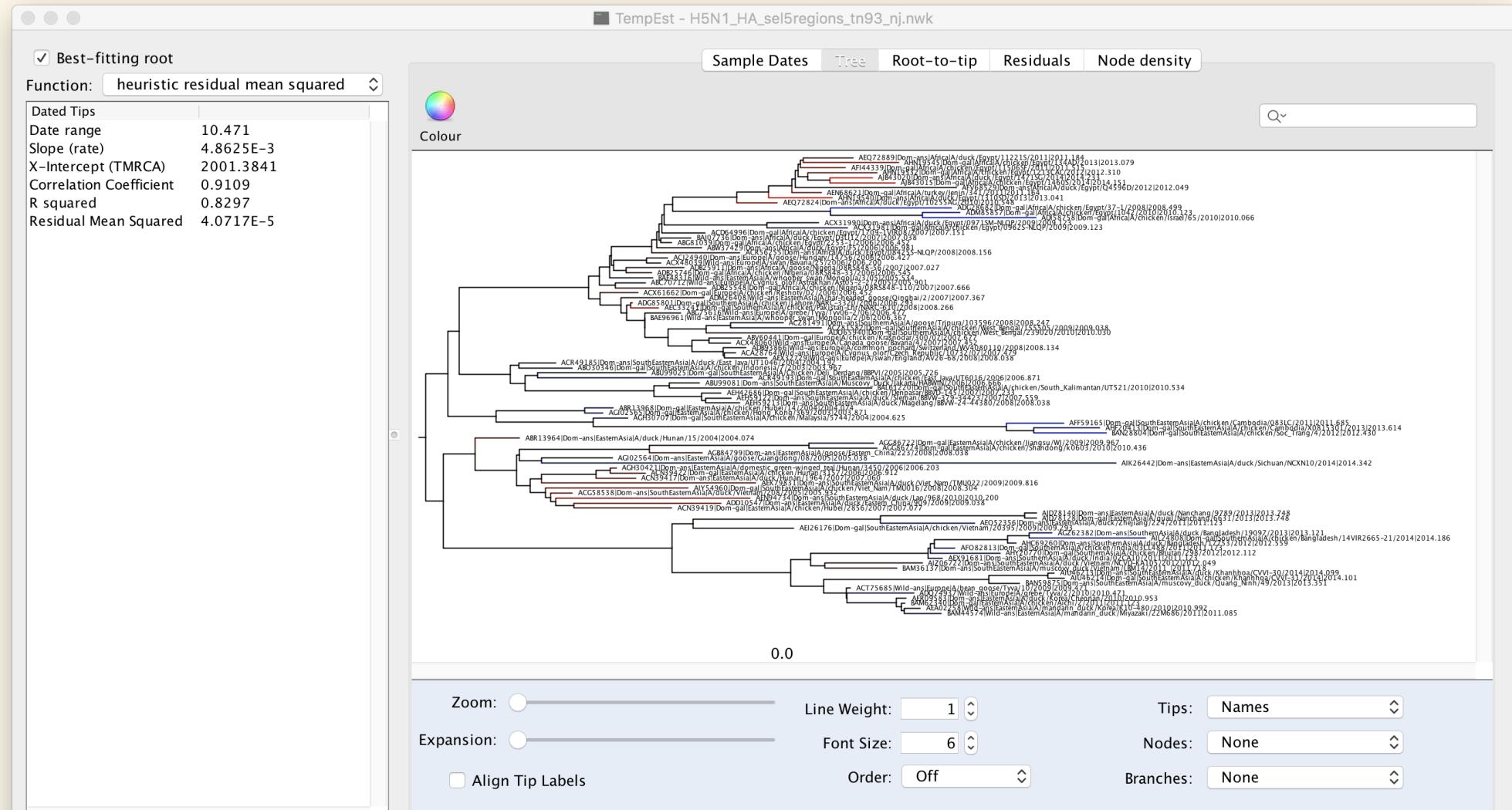
- **Required: Sequences with dates**
 - All Taxa labels should have **sampling** date
 - **i.e. the collection date from the patient**, and very much NOT the date the sample happened to be sequenced in the lab (could be days later), NOR the date where it happened to be uploaded to the database (could be years later) (!)
 - Example sequence name formats – **but do not mix date formats in the same file !!**
 - Scotland/QEUH-33C252D/2022|2022-01-20 SARS-CoV-2
 - A/California/16/2016|2016-01-26 Human influenza
 - A/duck/Hunan/15/2004|2004.074 Avian influenza with decimal date
 - Generally – good idea to include:
 - Accession number; Host; Location; Isolate Name; date txt; decimal date
- **Required: Tree in newick format (*.nwk)**
 - Use a genetic distance tree, e.g. one made in MEGA from your correctly named sequences !

General - Molecular clock estimate (using Tempest)

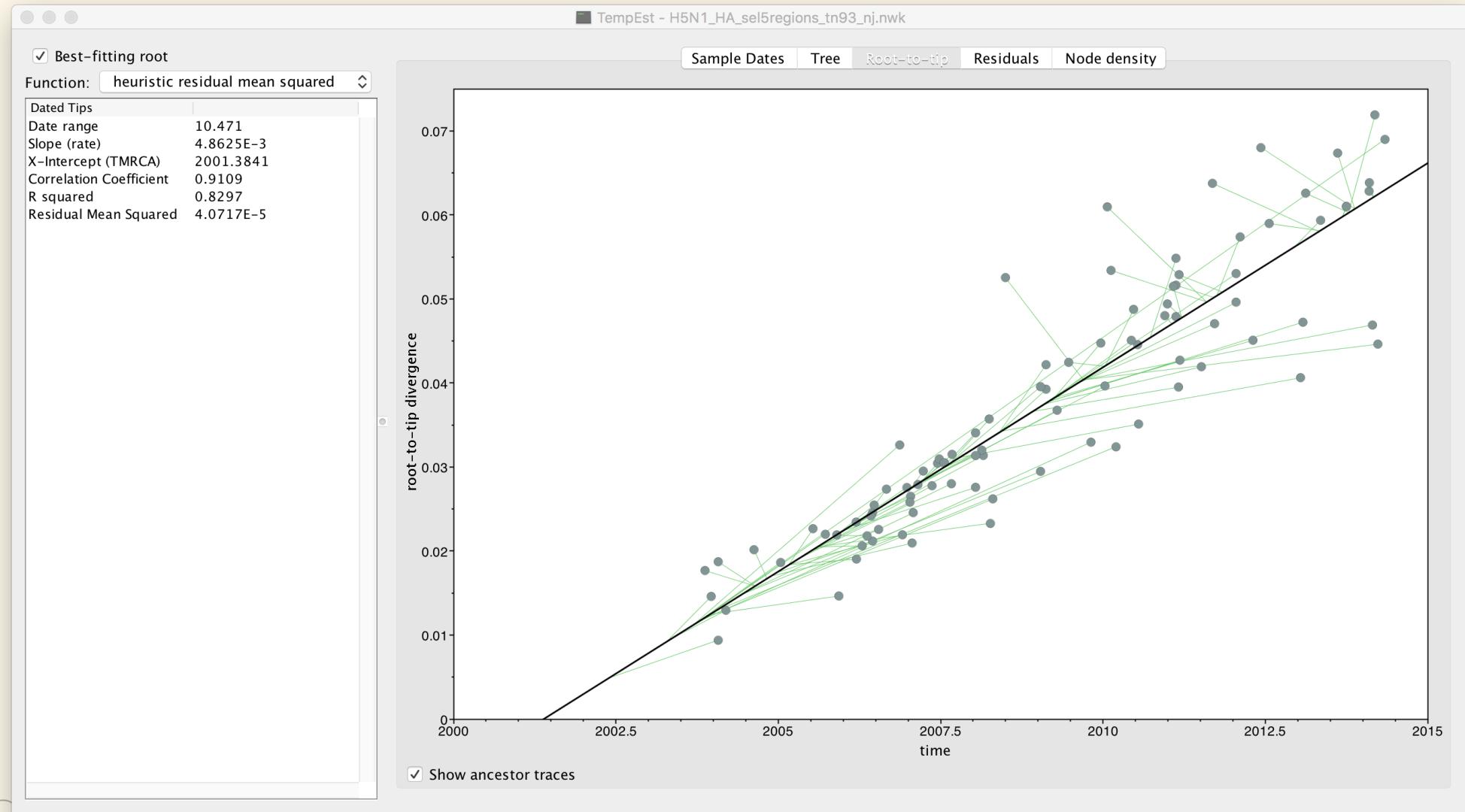
- Calculate genetic distance from root of tree to each of the tips and internal nodes
- Fit a regression line to genetic distance from root vs time
- Estimate of clock rate is the slope of the line



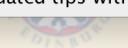
Tempest – Good signal – e.g. AIV



Tempest – Good signal – e.g. AIV



Tree loaded, 92 taxa, dated tips with range 10.5



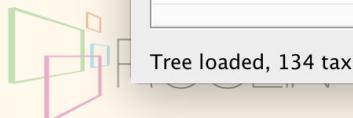
Veterinary Studies

Biotechnology and
Biological Sciences
Research Council

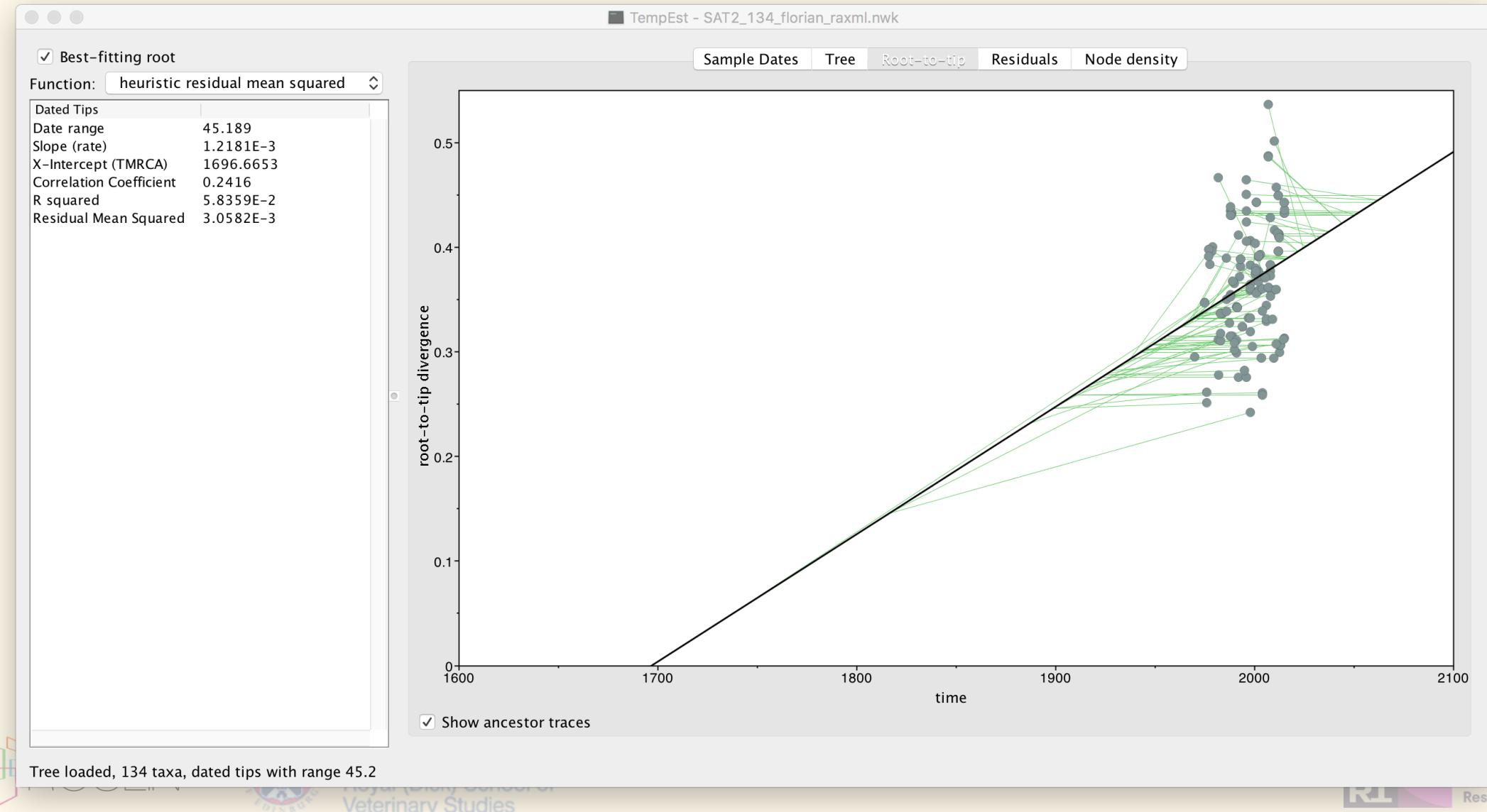
Tempest – Diverse Clades – e.g. FMDV SAT2 in Africa



Tree loaded, 134 taxa, dated tips with range 45.2

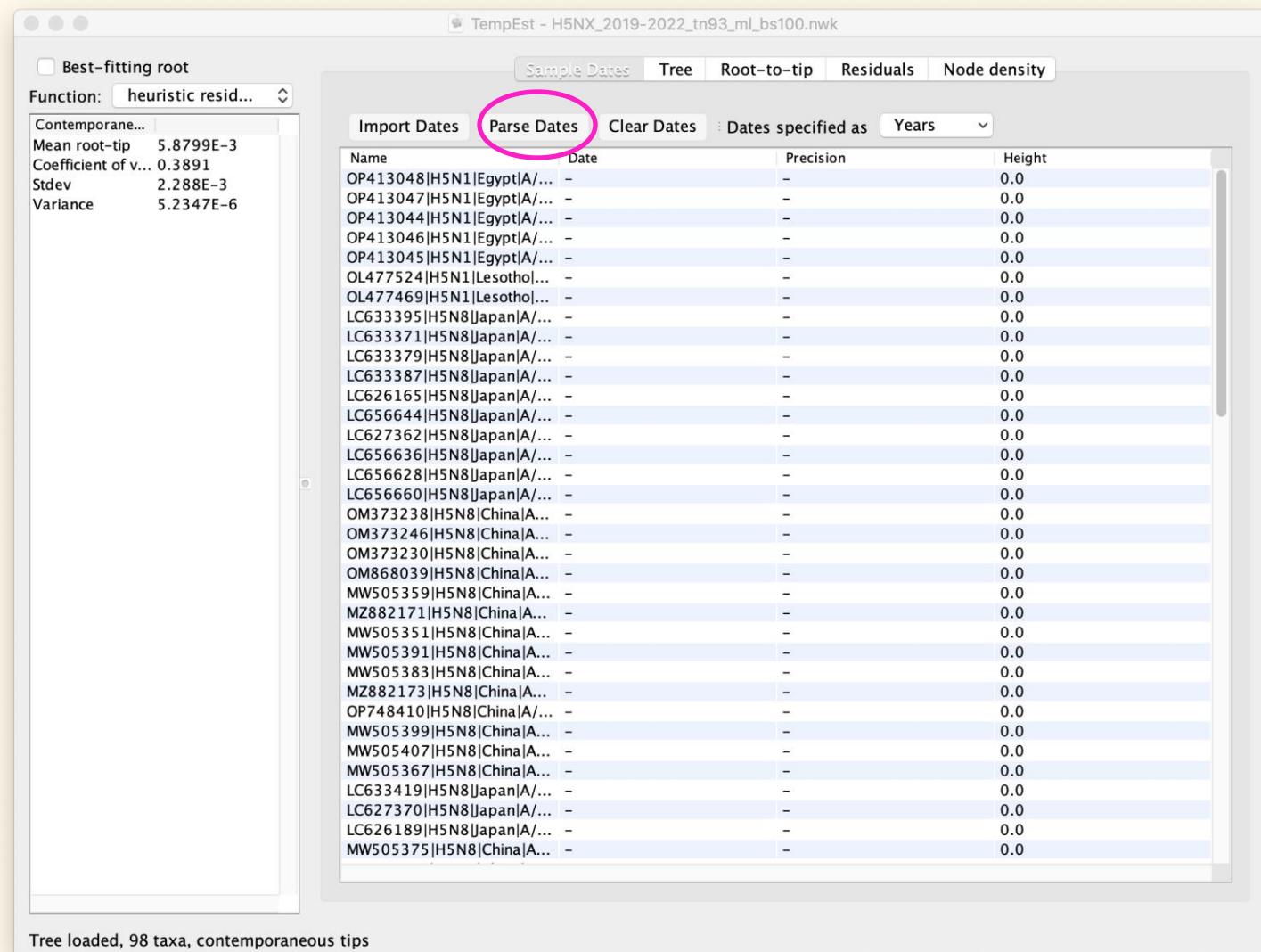


Tempest – Diverse Clades – e.g. FMDV SAT2 in Africa



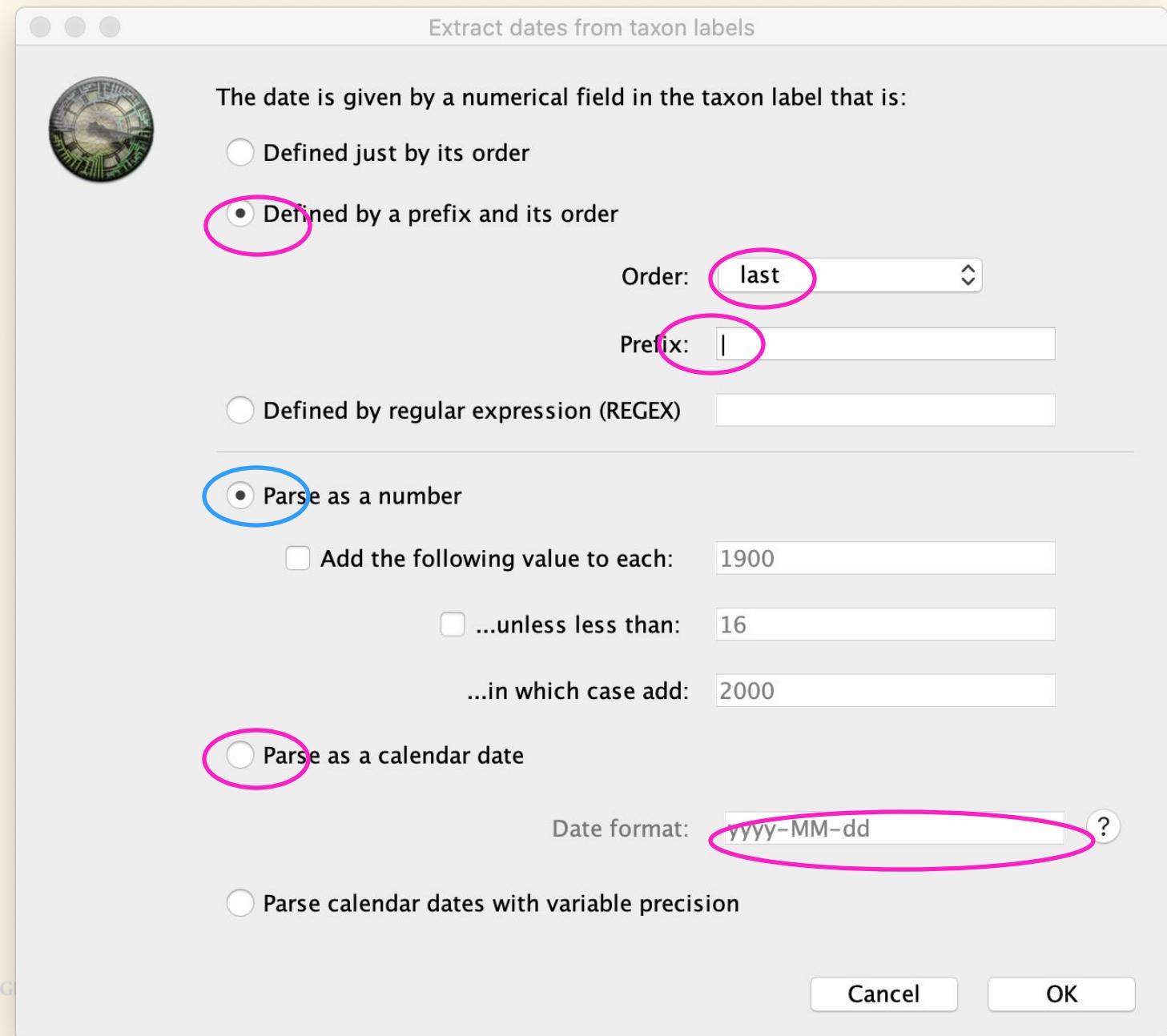
To use on your data - Open Tempest

- You will immediately be asked for the newick tree file
- Now you must tell Tempest what the dates of the sequences are
- The dates are in the sequence names
- Click Parse dates



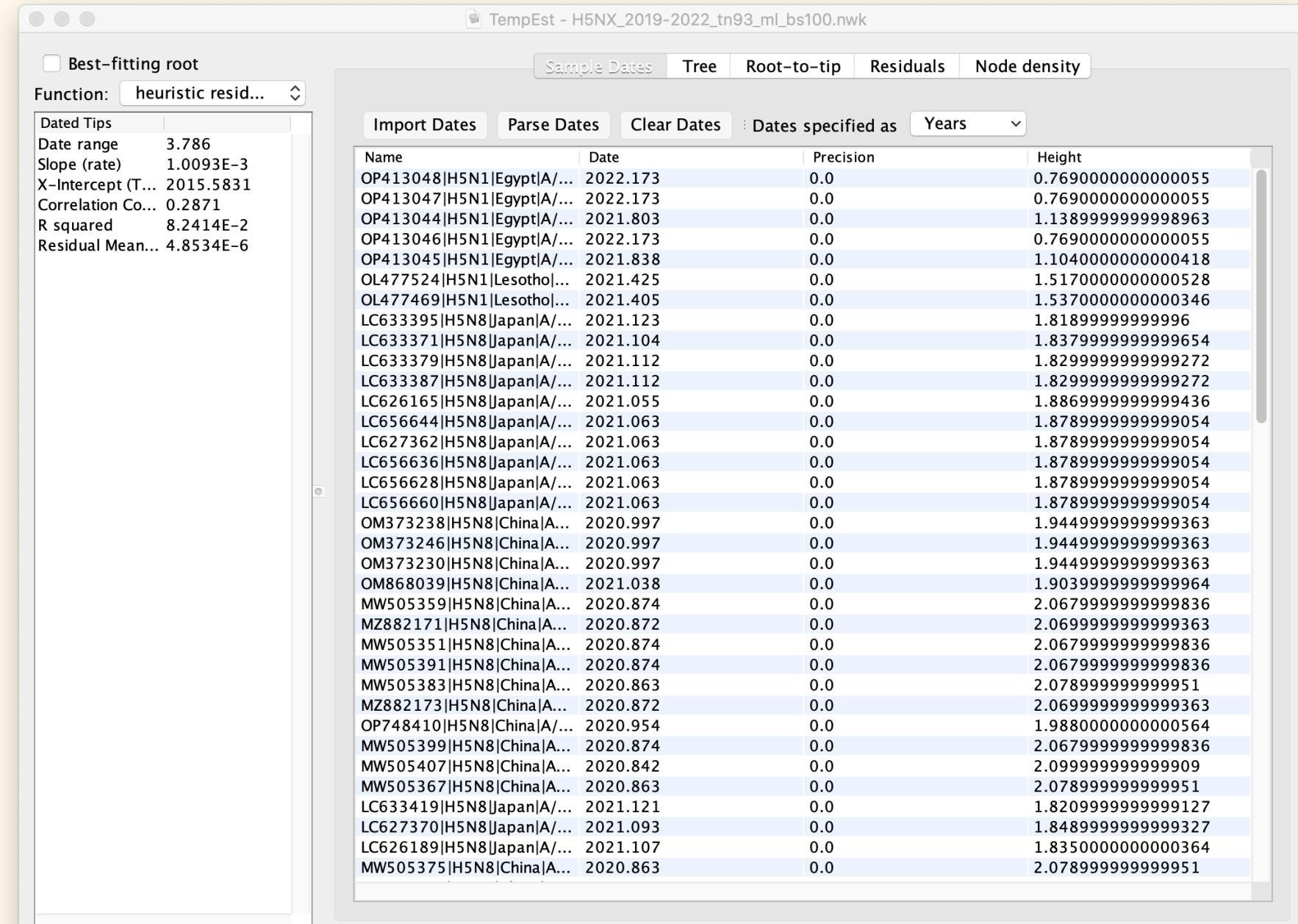
Define sequence dates from taxa names

- The dates are at the ends of the sequence name (last)
- The separator is the vertical bar (|)
- The dates are decimal dates
- The dates are calendar dates with format yyyy-MM-dd



Dates now
read
correctly

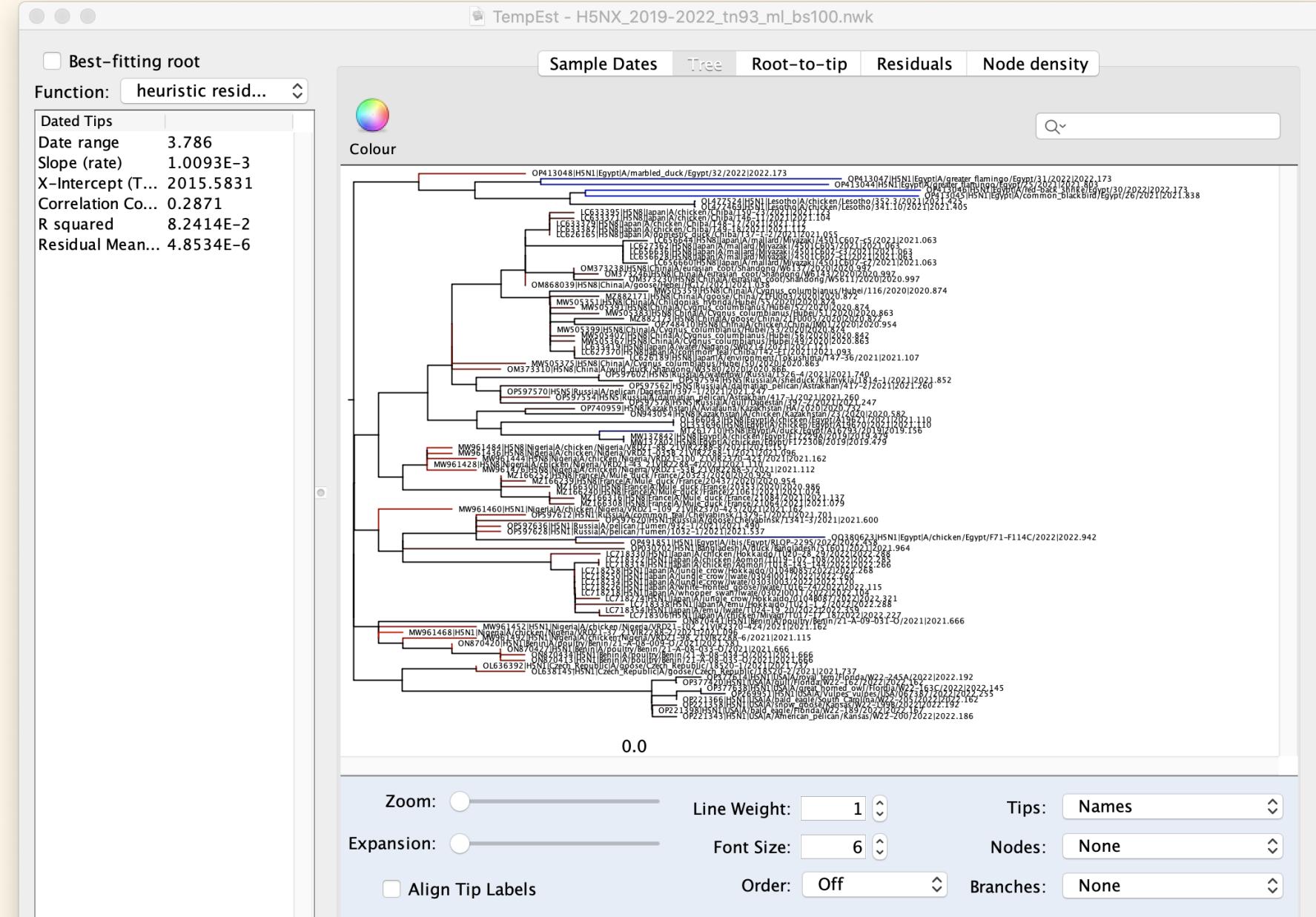
Decimal dates



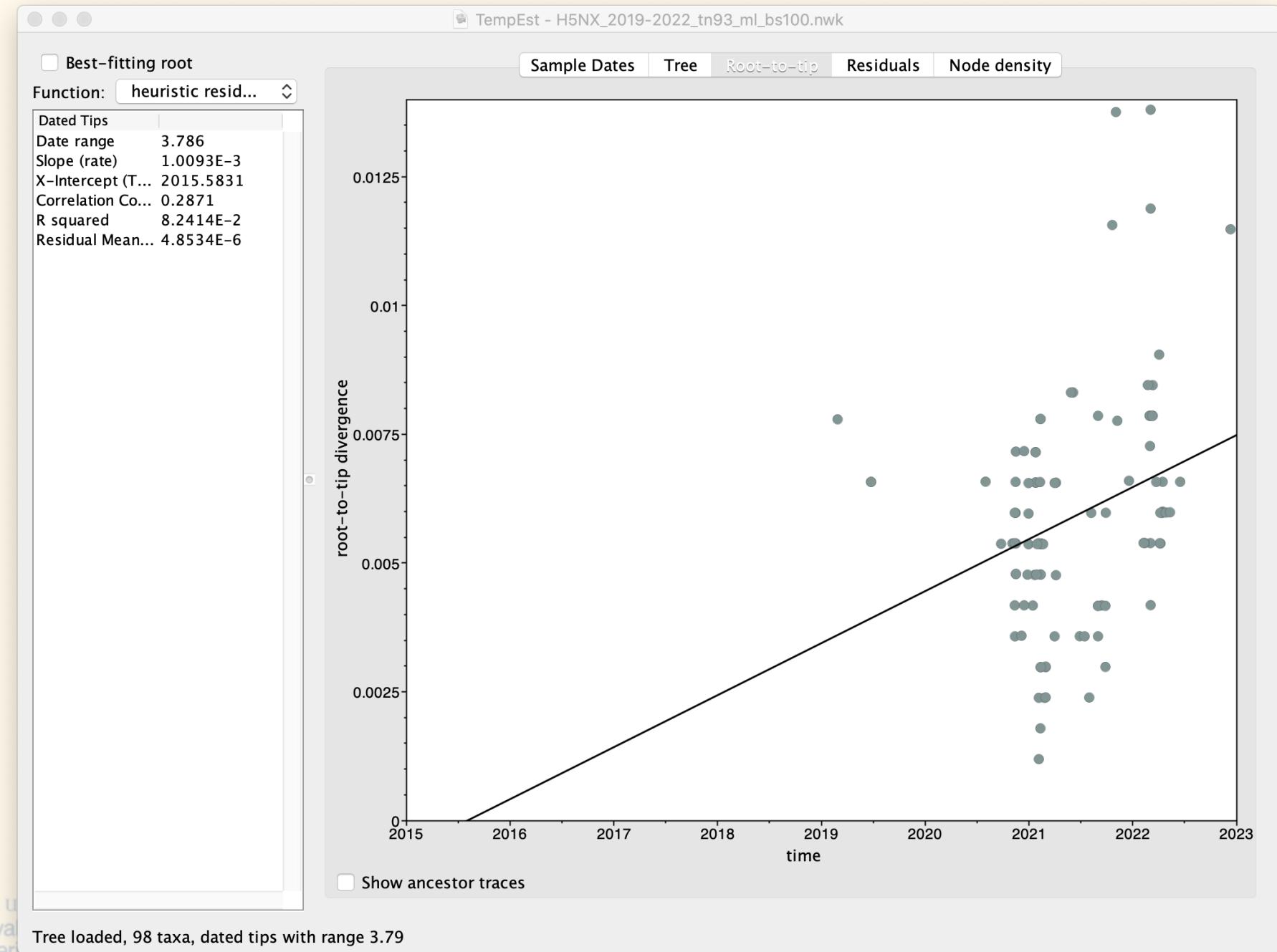
Unrooted tree



Tree loaded, 98 taxa, dated tips with range 3.79

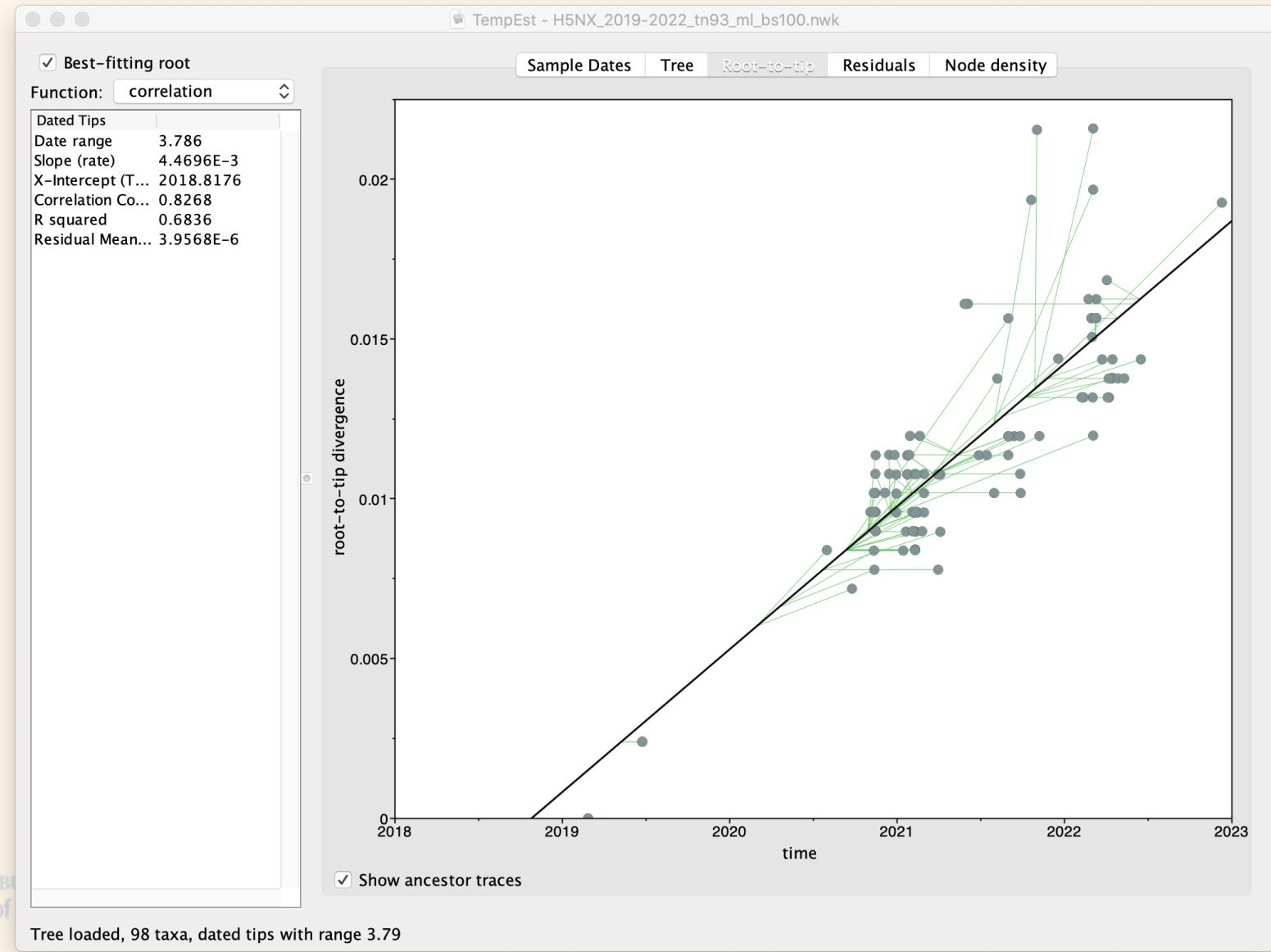


Not fitted yet – root to tip plot



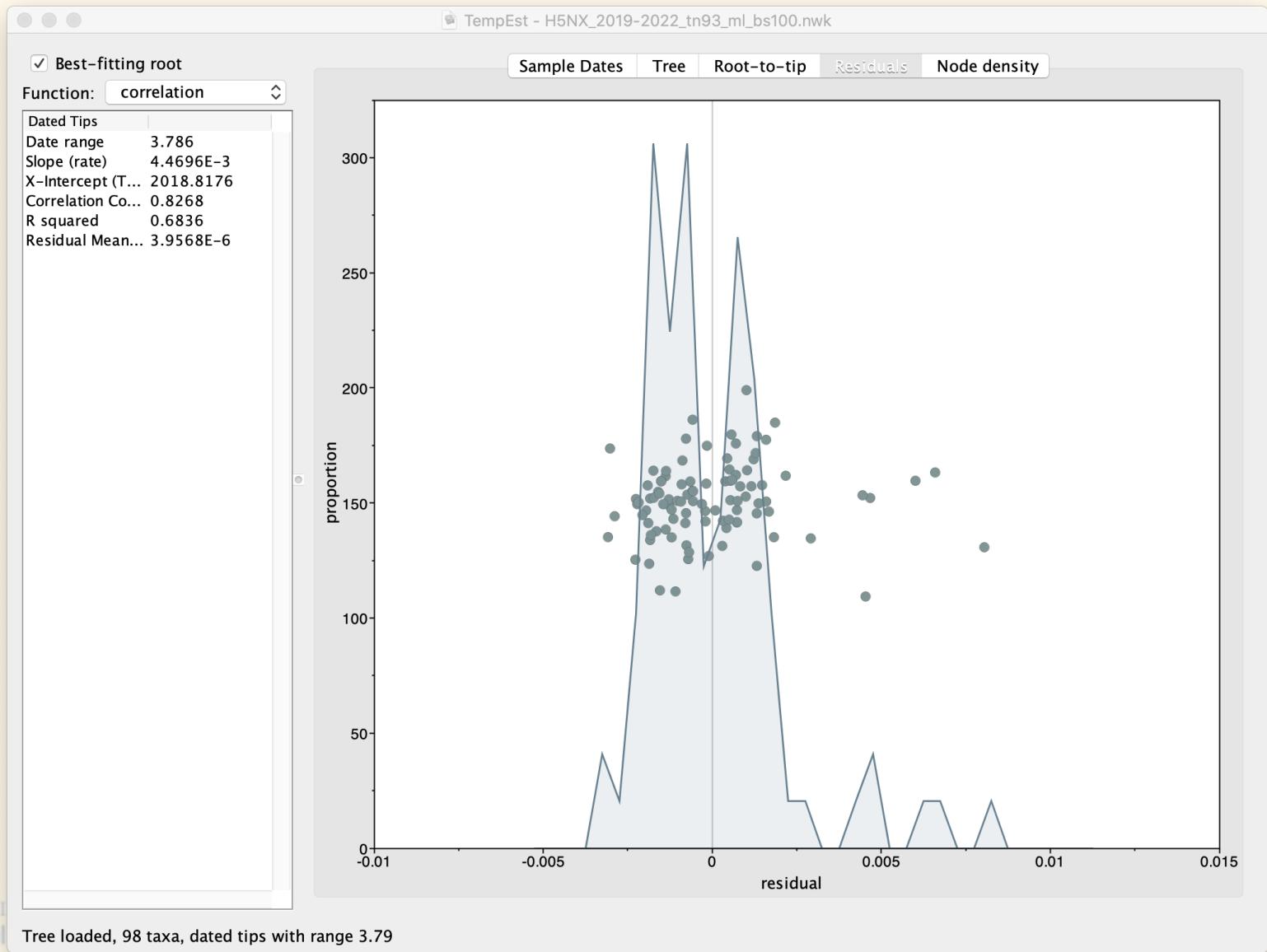
Find the best fitting root

- Click button at top left
- There are 4 ways to choose from
- Which way gives a sensible answer ?
- And are they all equally good ? (look at the correlation and R-squared)

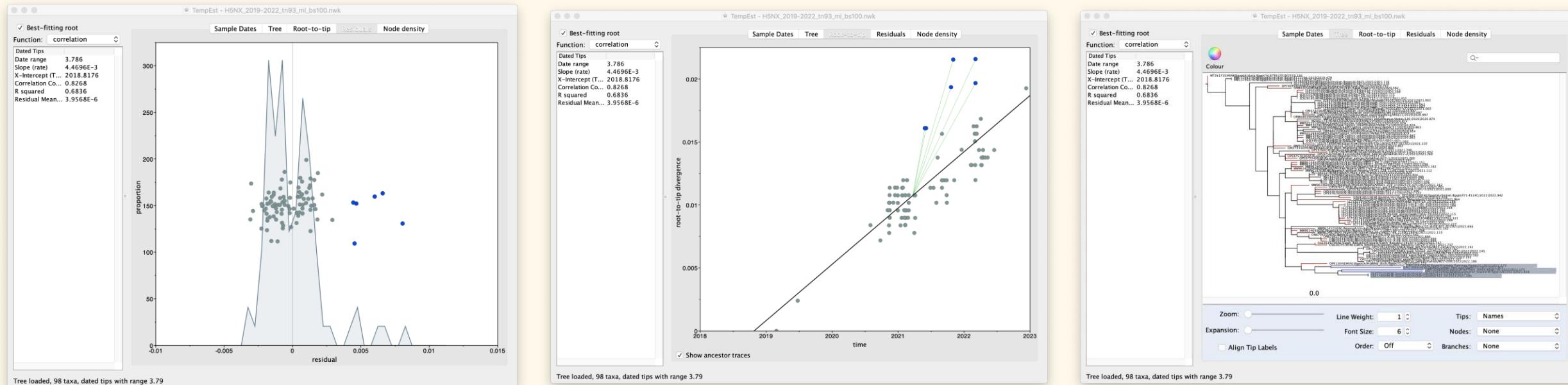


Check the residuals

- Which are the outliers ?
- Can highlight these and then see back in the tree



Outliers highlighted – to check



Questions

- What is the estimated molecular clock rate (slope on tempest) ?
- How good is are the root-to-tip fits ?
- Does the rooted tree correspond to what you think ? (you can export the tree as a nexus file which FigTree can read if wanted)
- What is the Time to Most Recent Common Ancestor of these sequences ?
 - this is when the root-to-tip diversity is zero; so it is the x-intercept

More examples for you to try

1. USA_H1N1_2011-2021.nwk
 - 1092 tips – H1N1 (HA) from USA
2. 2022-03-20_subsampled_scotland_tree.nwk
 - 1357 tips – SARS-CoV-2 from Scotland, subsample of all lineages
3. BA.2_Scotland_subsample.nwk
 - 397 tips – SARS-COV-2 from Scotland, subsample of BA.2 lineage only, restricted time-depth
4. SAT2_134_florian-duchatel_raxml.nwk
 - 134 tips, SAT2, example of diverse signal

What are you looking for ?

- Time depth of samples (& have you read the dates in correctly)
- Estimated time to most recent common ancestor (Intercept, TMRCA)
- Estimated clock rate
- Outlier sequences
- Difference between the different estimation methods ?