Submission type: Application

Phylostems: a new graphical tool to investigate temporal signal of heterochronous sequences at various evolutionary scales

Anna Doizy[12], AmauryPrin[1], Guillaume Cornu[3], Frederic Chiroleu[1] and Adrien Rieux[1*]

[1] CIRAD, UMR PVBMT, 97410 St Pierre, La Réunion, France.

[2] DoAna - statistiques Réunion, F-97480 Saint-Joseph, Reunion island, France

[3] CIRAD, Univ Montpellier,UR Forests and Societies, Montpellier, France

*Author to whom correspondence should be addressed: adrien.rieux@cirad.fr

Running title: Phylogenetic scaling of temporal signal

Abstract

1. Molecular tip-dating of phylogenetic trees is a growing discipline that uses DNA sequences sampled at different points in time to co-estimate the timing of evolutionary events with rates of molecular evolution. Such inferences should only be performed when there is sufficient temporal signal within the analysed dataset. Hence, it is important for researchers to be able to test their dataset for the amount and consistency of temporal signal prior to any tip-dating inference. For this purpose, the most popular method considered to-date has been the "root-to-tip regression" which consist in fitting a linear regression of the number of substitutions accumulated from the root to the tips of a phylogenetic tree as a function of sampling times. The main limitation of the regression method, in its current implementation, relies in the fact that the temporal signal can only be tested at the whole-tree evolutionary scale.
2. To fill this methodological gap, we introduce phylostems, a new graphical and user-friendly tool developed to investigate temporal signal at every evolutionary scale of a phylogenetic tree.
3. Phylostems allows detecting without *a priori* whether any subset of a tree would contain sufficient temporal signal for tip-based inference to be performed. We provide a "how to" guide by running phylostems on empirical datasets and supply guidance for results interpretation. Phylostems is freely available at https://pvbmt-apps.cirad.fr/apps/phylostems.

4. Considering the impressive increase in availability and use of heterochronous datasets, we hope the new functionality provided by phylostems will help biologists to perform thorough tip-dating inferences.

Keywords: Heterochronous sequence dataset, measurably-evolving populations (MEPs), phylogenetic tip-dating, R shiny app, root-to-tip regression, temporal signal.

Introduction

"Tip-dating" of phylogenetic trees is a popular and powerful type of genetic analysis aiming to make use of sequence data isolated at different points in time (i.e., heterochronous datasets) to co-estimate the timing of evolutionary events with rates of molecular evolution (Rieux & Balloux, 2016). Tip-dating requires working on measurably evolving populations (MEPs) which consist in datasets displaying detectable amounts of *de novo* nucleotide changes among the DNA sequences sampled at different timepoints (Drummond, Pybus, Rambaut, Forrest, & Rodrigo, 2003). Our ability to capture measurable amount of evolutionary change from sequence data is a factor of various parameters including the evolutionary rate per site per unit of time (μ), the width of the sampling interval (t), the number of sites in the sequences (L) and the time to the Most Recent Common Ancestor (MRCA) of all sequences (T_{MRCA}). Originally, only fast-evolving organisms such as RNA viruses were classifiable as MEPs but the recent rise in our ability to sequence DNA at high throughput from both modern and ancient material has led to a massive increase in both sequence length (L) and the timespan covered by the sequences (t), hence opening up the field of tip-dating to a variety of additional organisms (Biek, Pybus, Lloyd-Smith, & Didelot, 2015).

Phylogenetic inferences performed on such time-structured sequence data represent a powerful tool for hypothesis testing (Rieux & Balloux, 2016). They have notably been critical for *i)* dating key events in human evolutionary history (Fu et al., 2013; Rieux et al., 2014), *ii)* improving our understanding of various important pathogens emergence, spread and evolution (Bos et al., 2014; Faria et al., 2014; Eldholm et al., 2015; O'Hanlon et al., 2018;

Vanhove et al., 2019; Rambaut, 2020), *iii)* investigating the relative impacts of climatic and anthropogenic factors on the widespread extinctions of large mammals (Shapiro et al., 2004; Stiller et al., 2010), *iv)* providing meaningful information about pathogens host species jumps (Weinert et al., 2012) and *v)* estimating unknown sequence's ages in various organisms (Shapiro et al., 2011).

Inferences from tip-calibrated phylogenetic trees should only be performed when there is sufficient temporal signal within the analysed dataset (Drummond, Pybus, Rambaut, et al., 2003; Duchêne, Duchêne, Holmes, & Ho, 2015; Murray et al., 2016; Rieux & Balloux, 2016). This will for instance not be the case if the sampling period is too short for sufficient evolutionary changes to be measured, if evolutionary rates are too low or variable amongst lineages or if some samples have incorrectly been dated (Rambaut, Lam, Carvalho, & Pybus, 2016). As such it is important for researchers to be able to test their dataset for the amount and consistency of temporal signal prior to any tip-dating inference. For this purpose, the most popular method considered to-date has been the "root-to-tip regression" which consist in fitting a linear regression of the number of substitutions accumulated from the root to the tips of a phylogenetic tree as a function of sampling times (Buonagurio et al., 1986; Shankarappa et al., 1999; Korber et al., 2000; Drummond, Pybus, & Rambaut, 2003). If sampling dates are sufficiently different, then more recently sampled sequences should have undergone substantially more evolutionary change than earlier sampled sequences, which would result in a positive correlation. This method has often been used as a diagnostic of data quality and of the reliability rate estimates, where the slope coefficient corresponds to the substitution rate under the assumption of a strict molecular clock, the X-intercept is an estimate of the date of the root of the tree and R^2 indicates the degree to which sequence

104 evolution has been clocklike. However, the root-to-tip regression method is not statistically
105 suitable for proper hypothesis testing because the individual data points are not
106 independently distributed, and are instead partially correlated due to their phylogenetic
107 shared ancestry (Drummond, Pybus, & Rambaut, 2003). To overcome this limitation,
108 Navascues et al. (2010) suggested a non-parametric approach using permutations to test
109 whether the correlation is stronger than expected if the sampling dates were randomly
110 assigned. More recently, other phylogenetic approaches such as the date-randomization test
111 (Ramsden, Melo, Figueiredo, Holmes, & Zanotto, 2008; Duffy & Holmes, 2009; Duchêne et
112 al., 2015; Murray et al., 2016) or model selection/comparison (Rambaut, 2000; Murray et al.,
113 2016; Duchene et al., 2019), although way more computationally intensive, have also been
114 introduced and shown to be more robust tests for temporal signal detection and
115 characterization.

116 Despite its statistical pitfalls, the regression method remains a very helpful exploration tool
117 to quickly assess the extent of temporal signal within a dataset. It only requires a rooted
118 molecular phylogeny (whose branch lengths represent genetic distance) estimated from
119 heterochronous (dated) sequences and runs instantaneously. The regression method has
120 been implemented in the popular and interactive graphical program TempEst (Rambaut et
121 al., 2016), formerly known as Path-O-Gen. The main limitation of the regression method in
122 its current implementation relies in the fact that the temporal signal can only be tested at
123 the whole-dataset (tree) evolutionary scale. However, although a significant positive
124 correlation would indicate the presence of detectable amounts of *de novo* mutations within
125 a tree timescale, a non-positive (or a statistically non-significant) correlation does not
126 necessarily mean that no temporal signal exists at a reduced timescale, as illustrated in Fig 1.

To fill this methodological gap, we introduce phylostems, a new graphical and user-friendly tool developed to investigate temporal signal at every evolutionary scales of a phylogenetic tree. Phylostems allows detecting without *a priori* whether any subset of a tree would contain sufficient temporal signal for tip-based inference to be performed. We provide a "how to" guide by running phylostems on empirical datasets and supply insights on interpreting the outputs.

Materials and Methods

The program phylostems (Phylogenetic Scaling of Temporal Signal) is an open source, graphical Shiny based R application (Chang, Cheng, Allaire, Xie, & McPherson, 2018; R Core Development Team, 2020) built for exploring temporal signal at various scales of a phylogenetic tree. Shiny is an R package that makes it easy to build interactive web applications from R (https://shiny.rstudio.com/). Phylostems can be either used online at https://pvbmt-apps.cirad.fr/apps/phylostems/ or executed locally by downloading its source code from https://gitlab.com/cirad-apps/phylostems. A schematic representation of phylostems workflow is presented in Fig. 2.

As input, phylostems requires *(i)* a phylogenetic tree in computer-readable Nexus or Newick format with branch lengths scaled as genetic distances only, such as the ones computed using maximum likelihood approaches (e.g. Guindon et al., 2010; Minh et al., 2020; Stamatakis, 2014). In its current implementation, the online version of phylostems allows uploading trees with 1500 sequences at maximum. Larger trees will need to be processed locally by sourcing the gitlab version. *(ii)* Prior to be loaded in phylostems, the tree needs to be rooted, either at a position chosen by the user (with an outgroup) or at a most

compatible location with the assumption of a strict molecular clock (using for instance the rtt function from the ape R package (Paradis & Schliep, 2019)). When possible, we advise to use outgroup-rooted trees. Finally, *(iii)* sampling/isolation dates needs to be known for each sequences and specified within tip labels. Before-Christ (B.C) dates, sometimes required to handle sequences generated from ancient DNA data can be specified using negative values (e.g. - 400.5). Note that since missing dates are not allowed, sequences with unknown sampling years needs to be pruned out from the tree (using for instance the drop.tip function from the ape R package) prior to be uploaded in phylostems.

When a tree has correctly been loaded in phylostems, a distribution of sampling dates is plotted within the "upload" panel allowing for a visual check of sequences temporal width. At this stage, the phylogenetic tree has been loaded using the ape R package and root-to-tip distances for all sequences are recorded. Temporal signal is hence tested at every node of the input tree (including its root) meeting the following conditions required to perform a linear regression: *i)* the node must be the parent of at least n=3 tips, *ii)* there should be at least n=3 distinct combination of root-to-tip distances and sampling dates and *iii)* there should be at least n=2 different sampling dates. At each of such nodes, linear regression between sampling dates and root-to-tip distances is performed and the following parameters: (1) p-value, (2) slope, (3) adjusted R^2, and (4) intercept with the x-axis values are recorded.

Phylostems's main results are provided within the "Temporal signal" panel. First an annotated phylogenetic tree is interactively plotted by sourcing both ggtree and plotly packages (Yu, Smith, Zhu, Guan, & Lam, 2017; Sievert, 2020). On this tree, nodes with

171 temporal signal, *i.e.* nodes at which root-to-tip linear regression yielded a statistically
172 significant and positive slope, are highlighted with colours scaling to R^2 value. The default
173 threshold for the linear regression p-value has been fixed to 0.05 but the user can
174 interactively modify it using a slider bar, which enable easy investigation of nodes with
175 borderline significant trends. A table summarizing the nodes with temporal signal is also
176 displayed along with respective number of descending sequences, p-value, slope and
177 adjusted R^2 values. Most importantly, phylostems allow the user to visualize the root-to-tip
178 regressions at any chosen node of interest. To do so, the user simply needs to click on a
179 node, and the associated root-to-tip regression will be displayed. Both the tree and the root-
180 to-tip regression plots are linked, so that data points (or tree tips) selected in one plot will
181 automatically be highlighted on the other one. This enables easy investigation of outliers and
182 sequences or clades of interest.

183 Finally, when temporal signal is found at the within-tree scale, phylostems's "Make new
184 FASTA" panel allows generating a new subset sequence FASTA file that only include the
185 variant sites for the descending tips of a node of interest, a dataset suitable for further tip-
186 dating inferences.

187 In the following, we use two previously published empirical datasets to illustrate how
188 phylostems allows users exploring temporal signal at various evolutionary scales within
189 phylogenetic trees. For both datasets, we downloaded rooted-ML tree files built from non-
190 recombining genomic sequences from their original publications. The first dataset contains
191 45 strains of *Xyllela fastidiosa* (hereafter *Xf*) sampled worldwide between 1983 and 2016
192 (Vanhove et al., 2019). *Xf* is a bacterial crop pathogen of global importance, currently

threatening agriculture in various Europeans countries (Sicard et al., 2018). The second dataset comprises 98 hantaviruses isolates sampled from bank voles in Belgium between 1984 and 2016 (Laenen et al., 2019). Hantaviruses are important zoonotic viral pathogens that can cause hemorrhagic fever with renal syndrome and pulmonary syndrome, potentially life-threatening diseases in humans (Maes, Clement, Gavrilovskaya, & Van Ranst, 2004).

Results

Xf dataset

We first loaded the *Xf* rooted tree within phylostems 's "upload" panel (see Fig. 3). Looking at the plot of the sampling dates distribution, one can perform a quick visual check of sequences temporal width (here 1983-2016) to validate the data importation process. Moving to the "Temporal signal" panel, phylostems displays the *Xf* phylogenetic tree on which the structuration by the four subspecies: ssp. *pauca*, *multiplex*, *morus*, *fastidiosa* can be easily distinguished (see Fig. 4A). Visual inspection of the *Xf* tree in phylostems demonstrated a lack of strong and deep temporal signal as neither the root nor the MRCA of each subspecies displayed any significant correlation between root-to-tip distances and sampling ages, as highlighted by the absence of annotations at those nodes. Phylostems detected only one internal node (node 82) associated with temporal signal within the *Xf* tree. This node is the MRCA of a small clade containing 9 samples within the *Xf pauca ssp* clade. When clicking on this node, phylostems displays the associated root-to-tip regression plot and parameters (R^2 = 0.38, slope = 6.9E-7, P-val = 0.045, see Fig.4 B). According to phylostems's results, this small clade (N=9 sample) is the only evolutionary scale suitable for phylogenetic tip-based inferences in BEAST or other programs within the *Xf* dataset. To do

so, the “Make new FASTA” panel allows generating a new sequence file that only include the variant sites for the 9 *Xf* samples within the clade with detected temporal signal.

Hantaviruses dataset

Visual inspection of the Hantaviruses tree in phylostems demonstrated heterogenous temporal signal amongst clades, here referring to three geographical sampling areas namely Ardennes, Campine and Sonian Forest (Fig 5.A). Phylostems revealed a lack of temporal signal both at the whole tree scale and for the Sonian Forest clade. Temporal signal was observed at the MRCA of the Campine and Ardennes clades as well as within the Ardennes clade, as represented by the several highlighted nodes on the tree. A table listing all the nodes associated with temporal signal along with their associated statistics is given in Fig 5.B. When plotting the regression at the MRCA of the Campine and Ardennes clades, phylostems allows visually identifying outlier samples that are significantly deviating from the root-to-tip regression line (Fig 5.C). Here, all outliers felt within the Campine clade, suggesting that phylogenetic tip-based inferences should not be performed on both the Campine and Ardennes clades simultaneously. Possible causes for such outliers are multiple and will be argued in the discussion section.

Discussion

We introduce phylostems, a new graphical and user-friendly tool developed to investigate temporal signal within phylogenetic trees using the root-to-tip regression method. Previous implementations of this method, such as for instance in the popular and interactive graphical program TempEst (Rambaut et al., 2016) were designed to test temporal signal at the whole tree scale (i.e. at its root). Investigating temporal signal at smaller evolutionary

scales was previously doable, but this task required the user to *i) a priori* decide at which scale (i.e. on which samples) performing the test and *ii)* manually splitting or reconstructing the tree for every of such scales. The main improvement of phylostems is to allow detecting, in a single step and without *a priori*, any evolutionary scale at which temporal signal may exist within a phylogenetic tree.

Exploring the degree of temporal signal in heterochronous sequences datasets before proceeding to inference using formal molecular clock models is a crucial task (Rieux & Balloux, 2016). As illustrated by the two empirical datasets analyzed in this study, temporal signal may sometimes be heterogeneous within a tree with substantial differences between clades. In such cases, we hope that phylostems will help researchers detecting the most appropriate scales, if any, at which thorough tip-based inferences may be performed. However, because of the statistical pitfalls associated with the root-to-tip regression method (Rambaut, 2000; Rambaut et al., 2016), phylostems should rather be seen as a fast, visual and qualitative data exploration tool for temporal signal detection but should not be used to test hypotheses or undertake statistical model selection. Once temporal signal has been detected in phylostems, we advise users to make use of other available methods such as non-parametric permutations (Navascués et al., 2010), date-randomization test (Ramsden et al., 2008; Duffy & Holmes, 2009; Duchêne et al., 2015; Murray et al., 2016) or model selection/comparison (Murray et al., 2016; Duchene et al., 2019) to validate the existence of measurably evolving populations in their datasets.

Finally, phylostems can also help identifying outliers or groups of samples that substantially differ from the root-to-tip regression line and may require careful handling to avoid bias

during phylogenetic inferences. First, as illustrated by the analyse of the Hantaviruses dataset, different clades or populations in a tree may be characterized by positive but contrasted root-to-tip regression patterns that might arise from sampling bias or differences in life-history traits between clades (e.g. environmental factors, population density, evolutionary rates or epidemiological parameters). In such a case, it is suggested to perform independent phylogenetic inferences on each clade/population (Laenen et al., 2019). In other cases, outlier sequences whose sampling date is incongruent with their genetic divergence and phylogenetic position can be spotted from the regression plot (Rambaut et al., 2016). Such anomalies can reflect a problem with *i)* the sequence itself (e.g. low quality, sequencing/assembly/alignment errors, recombination or hypermutation) or *ii)* the sampling date(s) (e.g. mislabelling or biological contamination). Should the case of such outlier sequences arise, those samples should be excluded from subsequent phylogenetic inferences.

Considering the impressive increase in availability and use of heterochronous datasets, we hope the functionality provided by phylostems will help users to perform thorough tip-dating inferences. Pylostems is a dynamic application by nature. New functions will be added as new needs arise.

Acknowledgements

This work was financially supported by l'Agence Nationale pour la Recherche (JCJC MUSEOBACT contrat ANR-17-CE35-0009-01), the European Regional Development Fund (ERDF contract GURDT I2016-1731-0006632), Région Réunion and the French Agropolis Foundation (Labex Agro – Montpellier, E-SPACE project number 1504-004). We are grateful to S. Falala for his advices on building Shiny apps and CIRAD for providing hosting of the application server. We thank R. Almeida, M. Vanhove & B. Vrancken for providing access to the empirical datasets analyzed in this study. We also thank L. van Dorp, P. Campos, C.G. Crego, E. Conte, T.T CAO, F. Balloux & D. Richard for interesting discussions and testing previous versions of the app on their own datasets.

Data availability

Phylostems can be executed online at https://pvbmt-apps.cirad.fr/apps/phylostems/ but source code can also be downloaded from https://gitlab.com/cirad-apps/phylostems for local implementation. The two empirical trees used in this paper (Hantaviruses and *Xylella fastidiosa*) are accessible from the gitlab repository.

Authors contribution

A.R initially conceptualized the method. A.P generated a first version of the code. A.D improved it and converted it into a Shiny application with advices from A.R, G.C & F.C. G.C managed the online implementation & maintenance of the app. A.D & A.R wrote the first draft and all authors contributed to the final version.

References

Biek, R., Pybus, O. G., Lloyd-Smith, J. O., & Didelot, X. (2015, June 1). Measurably evolving pathogens in the genomic era. *Trends in Ecology and Evolution*. Elsevier Ltd. doi:10.1016/j.tree.2015.03.009

Bos, K. I., Harkins, K. M., Herbig, A., Coscolla, M., Weber, N., Comas, I., … Krause, J. (2014). Pre-Columbian mycobacterial genomes reveal seals as a source of New World human tuberculosis. *Nature*, *514*(7253), 494–497. doi:10.1038/nature13591

Buonagurio, D. A., Nakada, S., Parvin, J. D., Krystal, M., Palese, P., & Fitch, W. M. (1986). Evolution of human influenza A viruses over 50 years: Rapid, uniform rate of change in NS gene. *Science*, *232*(4753), 980–982. doi:10.1126/science.2939560

Chang, W., Cheng, J., Allaire, J., Xie, Y., & McPherson, J. (2018). shiny: Web Application Framework for R. *Https://CRAN.R-Project.Org/Package=shiny*. Comprehensive R Archive Network (CRAN). Retrieved from https://cran.r-project.org/package=shiny

Drummond, A., Pybus, O. G., & Rambaut, A. (2003). Inference of Viral Evolutionary Rates from Molecular Sequences. *Advances in Parasitology*. Academic Press. doi:10.1016/S0065-308X(03)54008-8

Drummond, A., Pybus, O. G., Rambaut, A., Forrest, S. A., & Rodrigo, A. G. (2003, September 1). Measurably evolving populations. *Trends in Ecology and Evolution*. Elsevier Ltd. doi:10.1016/S0169-5347(03)00216-7

Duchêne, S., Duchêne, D., Holmes, E. C., & Ho, S. Y. W. (2015). The Performance of the Date-Randomization Test in Phylogenetic Analyses of Time-Structured Virus Data. *Molecular Biology and Evolution*, *32*(7), 1895–1906. doi:10.1093/molbev/msv056

Duchene, S., Lemey, P., Stadler, T., Ho, S. Y., Duchêne, D., Dhanasekaran, V., & Baele, G. (2019). Bayesian Evaluation of Temporal Signal in Measurably Evolving Populations. *BioRxiv*. doi:https://doi.org/10.1101/810697

Duffy, S., & Holmes, E. C. (2009). Validation of high rates of nucleotide substitution in geminiviruses: Phylogenetic evidence from East African cassava mosaic viruses. *Journal of General Virology*, *90*(6), 1539–1547. doi:10.1099/vir.0.009266-0

Eldholm, V., Monteserin, J., Rieux, A., Lopez, B., Sobkowiak, B., Ritacco, V., & Balloux, F. (2015). Four decades of transmission of a multidrug-resistant Mycobacterium tuberculosis outbreak strain. *Nature Communications*, *6*. doi:10.1038/ncomms8119

Faria, N. R., Rambaut, A., Suchard, M. A., Baele, G., Bedford, T., Ward, M. J., … Lemey, P. (2014). The early spread and epidemic ignition of HIV-1 in human populations. *Science*, *346*(6205), 56–61. doi:10.1126/science.1256739

Fu, Q., Mittnik, A., Johnson, P. L. F., Bos, K., Lari, M., Bollongino, R., … Krause, J. (2013). A revised timescale for human evolution based on ancient mitochondrial genomes.

Current Biology, *23*(7), 553–559. doi:10.1016/j.cub.2013.02.044

Guindon, S., Dufayard, J. F., Lefort, V., Anisimova, M., Hordijk, W., & Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Systematic Biology*, *59*(3), 307–321. doi:10.1093/sysbio/syq010

Korber, B., Muldoon, M., Theiler, J., Gao, F., Gupta, R., Lapedes, A., … Bhattacharya, T. (2000). Timing the ancestor of the HIV-1 pandemic strains. *Science*, *288*(5472), 1789–1796. doi:10.1126/science.288.5472.1789

Laenen, L., Vergote, V., Vanmechelen, B., Tersago, K., Baele, G., Lemey, P., … Maes, P. (2019). Identifying the patterns and drivers of Puumala hantavirus enzootic dynamics using reservoir sampling. *Virus Evolution*, *5*(1). doi:10.1093/ve/vez009

Maes, P., Clement, J., Gavrilovskaya, I., & Van Ranst, M. (2004). Hantaviruses: Immunology, treatment, and prevention. *Viral Immunology*. Mary Ann Liebert Inc. doi:10.1089/vim.2004.17.481

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., & Lanfear, R. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution*, *37*(5), 1530–1534. doi:10.1093/molbev/msaa015

Murray, G. G. R., Wang, F., Harrison, E. M., Paterson, G. K., Mather, A. E., Harris, S. R., … Welch, J. J. (2016). The effect of genetic structure on molecular dating and tests for temporal signal. *Methods in Ecology and Evolution*, *7*(1), 80–89. doi:10.1111/2041-210X.12466

Navascués, M., Depaulis, F., & Emerson, B. C. (2010). Combining contemporary and ancient DNA in population genetic and phylogeographical studies. *Molecular Ecology Resources*, *10*(5), 760–772. doi:10.1111/j.1755-0998.2010.02895.x

O'Hanlon, S. J., Rieux, A., Farrer, R. A., Rosa, G. M., Waldman, B., Bataille, A., … Fisher, M. C. (2018). Recent Asian origin of chytrid fungi causing global amphibian declines. *Science*, *360*(6389), 621–627. doi:10.1126/science.aar1965

Paradis, E., & Schliep, K. (2019). ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics*, *35*(3), 526–528. doi:10.1093/bioinformatics/bty633

R Core Development Team. (2020). R: a language and environment for statistical computing, 3.2.1. *Document Freely Available on the Internet at: Http://Www. r-Project. Org*. Vienna, Austria: R Foundation for Statistical Computing. doi:10.1017/CBO9781107415324.004

Rambaut, A. (2000). Estimating the rate of molecular evolution: Incorporating non-contemporaneous sequences into maximum likelihood phylogenies. *Bioinformatics*, *16*(4), 395–399. doi:10.1093/bioinformatics/16.4.395

Rambaut, A. (2020). Phylogenetic analysis of nCoV-2019 genomes. Retrieved from http://virological.org/t/356

Rambaut, A., Lam, T. T., Carvalho, L. M., & Pybus, O. G. (2016). Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evolution*, *2*(1). doi:10.1093/VE/VEW007

Ramsden, C., Melo, F. L., Figueiredo, L. M., Holmes, E. C., & Zanotto, P. M. A. (2008). High Rates of Molecular Evolution in Hantaviruses. *Molecular Biology and Evolution*, *25*(7), 1488–1492. doi:10.1093/molbev/msn093

Rieux, A., & Balloux, F. (2016, May 1). Inferences from tip-calibrated phylogenies: A review and a practical guide. *Molecular Ecology*. Blackwell Publishing Ltd. doi:10.1111/mec.13586

Rieux, A., Eriksson, A., Li, M., Sobkowiak, B., Weinert, L. A., Warmuth, V., … Balloux, F. (2014). Improved Calibration of the Human Mitochondrial Clock Using Ancient Genomes. *Molecular Biology and Evolution*, *31*(10), 2780–2792. doi:10.1093/molbev/msu222

Shankarappa, R., Margolick, J. B., Gange, S. J., Rodrigo, A. G., Upchurch, D., Farzadegan, H., … Mullins, J. I. (1999). Consistent Viral Evolutionary Changes Associated with the Progression of Human Immunodeficiency Virus Type 1 Infection. *Journal of Virology*, *73*(12), 10489–10502. doi:10.1128/jvi.73.12.10489-10502.1999

Shapiro, B., Drummond, A. J., Rambaut, A., Wilson, M. C., Matheus, P. E., Sher, A. V., … Cooper, A. (2004). Rise and fall of the Beringian steppe bison. *Science*, *306*(5701), 1561–1565. doi:10.1126/science.1101074

Shapiro, B., Ho, S. Y., Drummond, A., Suchard, M. A., Pybus, O. G., & Rambaut, A. (2011). A Bayesian Phylogenetic Method to Estimate Unknown Sequence Ages. *Molecular Biology and Evolution*, *28*(2), 879–887. doi:10.1093/molbev/msq262

Sicard, A., Zeilinger, A. R., Vanhove, M., Schartel, T. E., Beal, D. J., Daugherty, M. P., & Almeida, R. P. P. (2018). *Xylella fastidiosa* : Insights into an Emerging Plant Pathogen. *Annual Review of Phytopathology*, *56*(1), 181–202. doi:10.1146/annurev-phyto-080417-045849

Sievert, C. (2020). *Interactive Web-Based Data Visualization with R, plotly, and shiny*. Chapman and Hall/CRC.

Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, *30*(9), 1312–1313. doi:10.1093/bioinformatics/btu033

Stiller, M., Baryshnikov, G., Bocherens, H., Grandal d'Anglade, A., Hilpert, B., Munzel, S. C., … Knapp, M. (2010). Withering Away--25,000 Years of Genetic Decline Preceded Cave Bear Extinction. *Molecular Biology and Evolution*, *27*(5), 975–978. doi:10.1093/molbev/msq083

413 Vanhove, M., Retchless, A. C., Sicard, A., Rieux, A., Coletta-Filho, H. D., De La Fuente, L., …
414 Almeida, R. P. P. (2019). Genomic diversity and recombination among Xylella fastidiosa
415 subspecies. *Applied and Environmental Microbiology*, *85*(13). doi:10.1128/AEM.02972-
416 18

417 Weinert, L. A., Welch, J. J., Suchard, M. A., Lemey, P., Rambaut, A., & Fitzgerald, J. R. (2012).
418 Molecular dating of human-to-bovid host jumps by Staphylococcus aureus reveals an
419 association with the spread of domestication. *Biology Letters*, *8*(5), 829–832.
420 doi:10.1098/rsbl.2012.0290

421 Yu, G., Smith, D. K., Zhu, H., Guan, Y., & Lam, T. T. (2017). ggtree: an r package for
422 visualization and annotation of phylogenetic trees with their covariates and other
423 associated data. *Methods in Ecology and Evolution*, *8*(1), 28–36. doi:10.1111/2041-
424 210X.12628

425

426

427

428

429

430

431

432

433

434

435

436

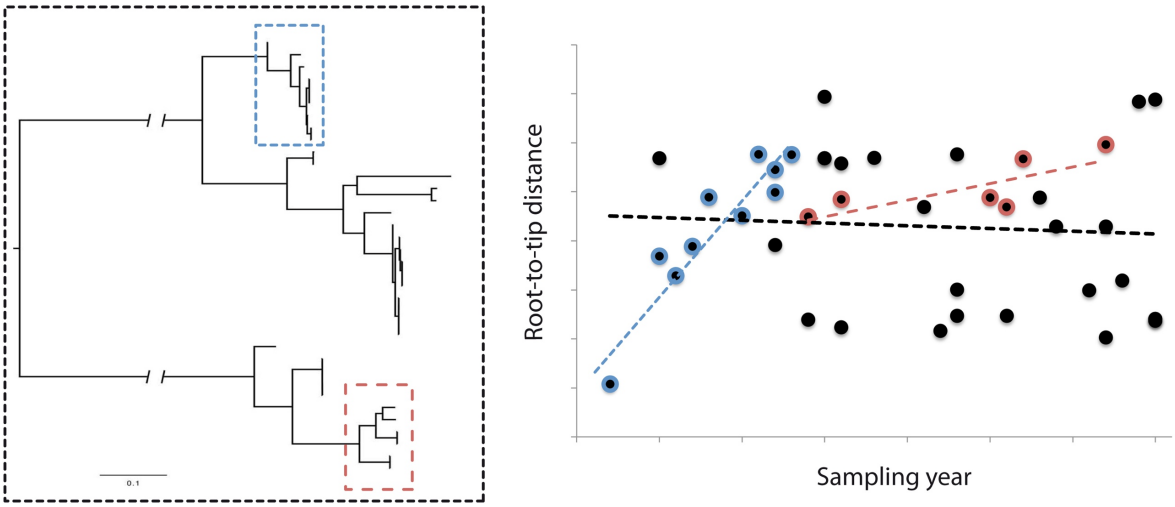


Fig. 1. Let's a tree (left panel) be constructed from a dataset of heterochronous sequences. When investigating temporal signal on the whole dataset using the regular root-to-tip regression method (right panel), no significant signal was found as the slope of the regression (black dotted line) appears to be non-positive. Hence, tip-based inferences should not be performed at the whole dataset timescale. However, as illustrated by the red and blue positive regression slopes calculated on two subsets of samples (red and blue squares on the tree), positive temporal signal exists at reduced evolutionary timescales at which thorough tip-based inferences could be performed. The main objective of phylostems is to provide the user with a graphical tool to detect without *a priori* such evolutionary clades.

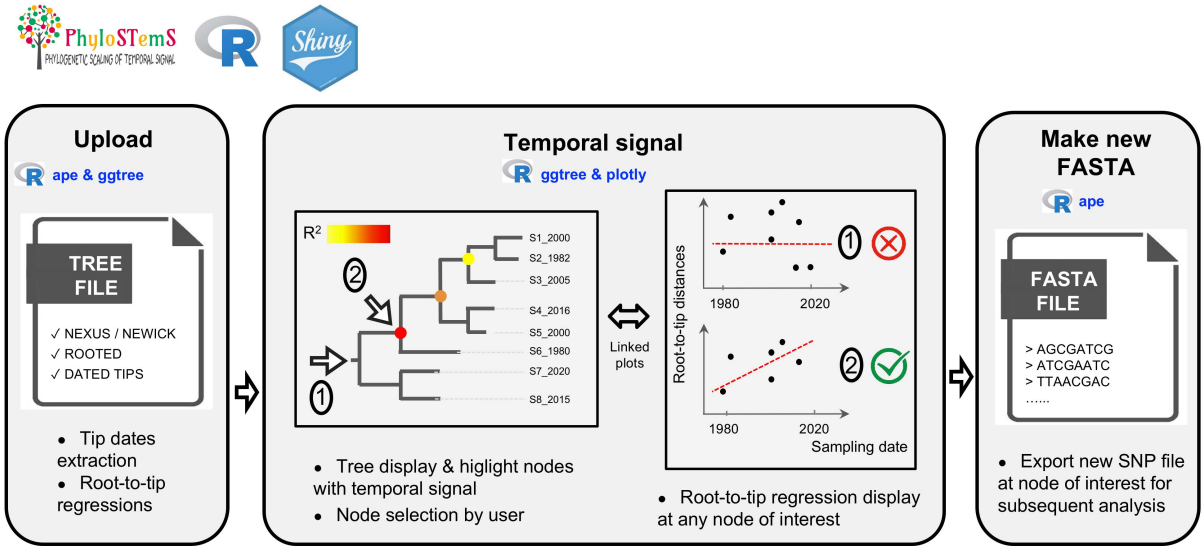


Fig. 2. Schematic representation of phylostems workflow. Main boxes ("Upload", "Temporal signal" & "Make new FASTA") represent the internal structure of the application organized in three main panels. Major tasks performed in each panel are summarized along with sourced R packages.

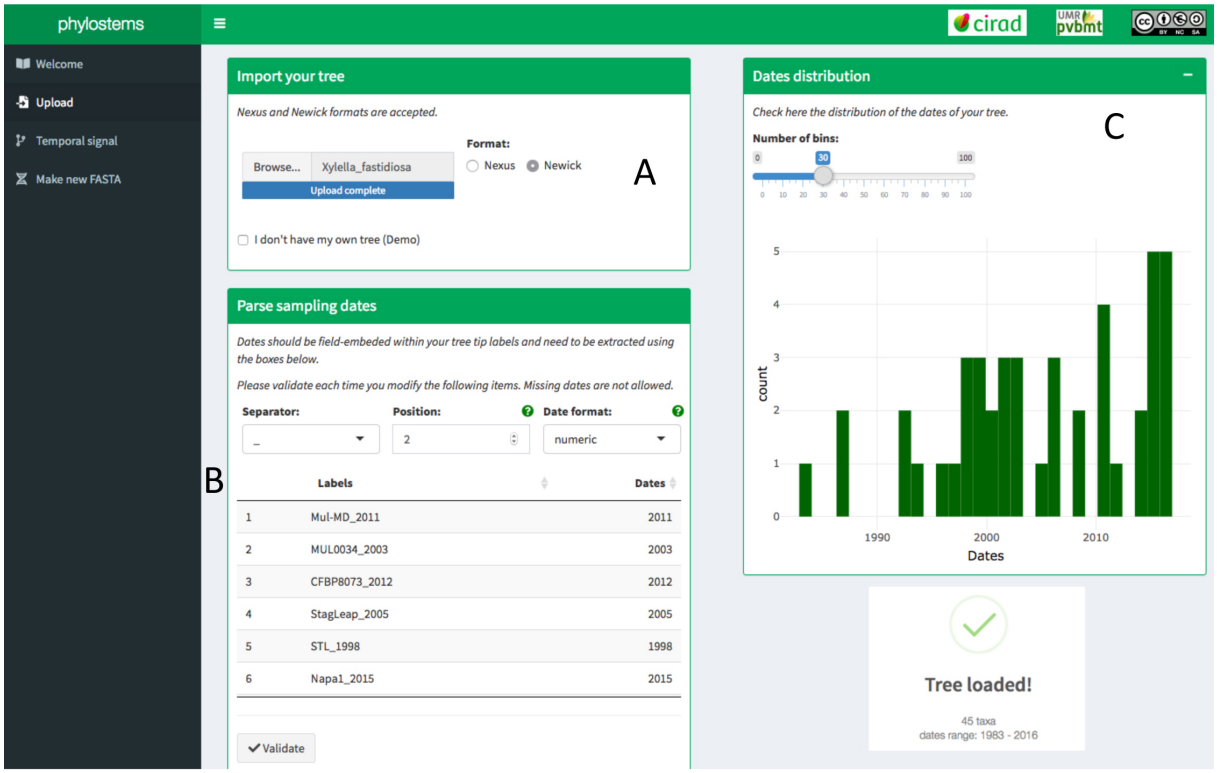
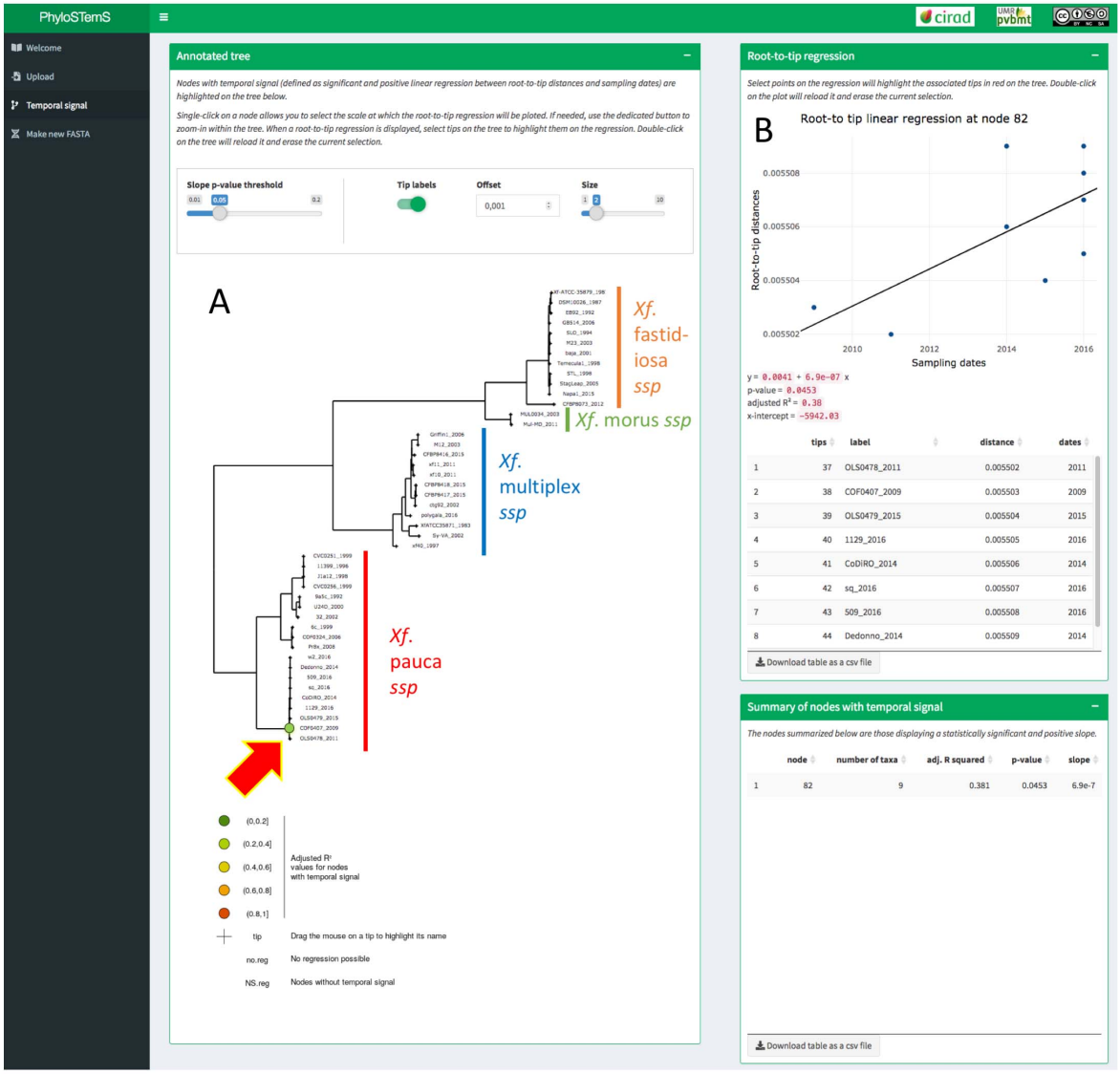


Fig. 3. Phylostems 's upload panel requesting the user to load a phylogenetic tree (A) and specify tip sampling dates from field-embedded values (B). Once loaded, a distribution of sampling dates is plotted allowing for a visual check of sequences temporal width (C).



	tips	label	distance	dates
1	37	OLS0478_2011	0.005502	2011
2	38	COF0407_2009	0.005503	2009
3	39	OLS0479_2015	0.005504	2015
4	40	1129_2016	0.005505	2016
5	41	CoDIRO_2014	0.005506	2014
6	42	sq_2016	0.005507	2016
7	43	509_2016	0.005508	2016
8	44	Dedonno_2014	0.005509	2014

	node	number of taxa	adj. R squared	p-value	slope
1	82	9	0.381	0.0453	6.9e-7

Fig. 4. Annotated phylogenetic tree of *Xylella fastidiosa* empirical dataset (A). Red arrow indicates the only node at which temporal signal was found. Root-to-tip regression at this node, along with associated parameters are plotted in (B).

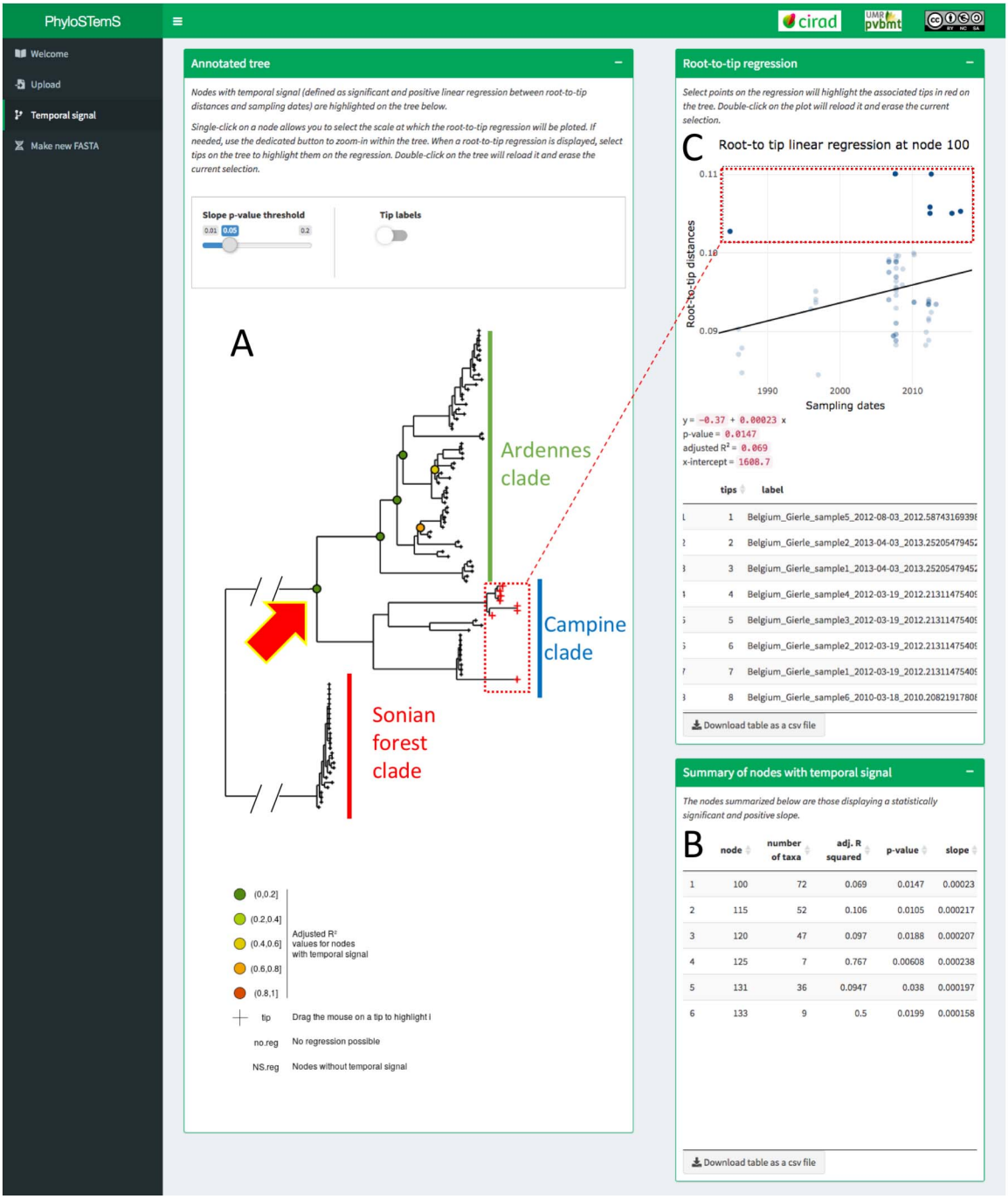


Fig. 5. Annotated phylogenetic tree for the Hantaviruses empirical dataset (A). Coloured circles indicate nodes at which temporal signal was found. A table summarizing those nodes, along with associated linear regression parameters is given in (B). Root-to-tip regression at node highlighted by the red arrow is plotted in (C). Both the tree and the regression plots are linked, so that data points (or tree tips) selected in one plot will automatically be highlighted on the other one, as illustrated by the red-dotted frames.