

Chapter 11

Phylogenetic Analysis of Poliovirus Sequences

Jaume Jorba

Abstract

Comparative genomic sequencing is a major surveillance tool in the Polio Laboratory Network. Due to the rapid evolution of polioviruses (~1 % per year), pathways of virus transmission can be reconstructed from the pathways of genomic evolution. Here, we describe three main phylogenetic methods; estimation of genetic distances, reconstruction of a maximum-likelihood (ML) tree, and estimation of substitution rates using Bayesian Markov chain Monte Carlo (MCMC). The data set used consists of complete capsid sequences from a survey of poliovirus sequences available in GenBank.

Key words Phylogenetic analysis, Maximum-likelihood, Bayesian MCMC, Poliovirus evolution

1 Introduction

Strategically, the success of the Polio Eradication Initiative (PEI) relies on intensive surveillance of cases of Acute Flaccid Paralysis (AFP) and laboratory investigations of polio samples. Virologic surveillance depends on a well-established WHO Global Polio Laboratory Network (GPLN) where polio isolation, intratypic differentiation, and genotyping are methods shared among laboratory members. Genomic sequencing is the method with the highest resolution. Due to the rapid evolution of polioviruses (PVs) [1], phylogenetic analysis of the VP1 capsid region resolves transmission pathways [2, 3]. Phylogenetic trees are invaluable tools for monitoring the progress of immunization activities as indicated by the appearance, disappearance, or reappearance of genetic lineages. PVs generally cluster geographically on a phylogenetic tree and if sequence difference is less than 15 % they are designated as genotypes. Furthermore, groups of viruses sharing less than 5 % sequence identity are designated as clusters. Single chains of transmission are phylogenetically displayed as (usually) short branches connecting sequences; growing branches correlate with expanded transmission. In this scenario, the appearance of long, isolated branches may indicate a fragmented genetic record from orphan

lineages (≥ 1.5 % sequence difference to the closest relative) indicative of possible surveillance gaps. The latest generation of rapid automated sequencers allows the routine characterization of VP1 sequences and the use of complete genome sequences for further characterization of the dynamics of PV evolution. For example, complete capsid sequences permit the use of infrequent transversion substitutions to define genetic relationships that otherwise might be obscured by saturating transition substitutions [4]. The availability of complete genomic sequences and related epidemiologic data, the development of bioinformatics and molecular evolution, and the growing computational capacity opens the door to the development of new analytical tools applied to molecular epidemiology. Here, phylodynamic methods are presented.

2 Materials

Current bioinformatics software packages are usually multiplatform; they can run on Windows, Mac, or Linux operating systems. As a recommendation, systems with at least 4 GB of RAM memory can accomplish most of the typical bioinformatics analysis described in this chapter. In order to optimize memory resources, it is best to consolidate all bioinformatics programs in a single computer and leave the rest of informatics needs (for example, word processing and email) to another computer if available. Most bioinformatics software incorporate the basic software libraries needed for running the programs. However, it is advised to follow the installation instructions provided by the software package since some may require installing or updating existing libraries (for example, Java libraries).

There are several bioinformatics packages that include phylogenetic analysis. A comprehensive list of phylogenetic resources can be found online at <http://evolution.genetics.washington.edu/phylip/software.html>. In this chapter the following software will be used: MEGA version 6 (<http://megasoftware.net/>), Geneious version 7 (<http://www.geneious.com/>), Seaview version 4.5 (<http://doua.prabi.fr/software/seaview>), BEAST version 1.8 (<http://beast.bio.ed.ac.uk/>), and FigTree version 1.4.2 (<http://tree.bio.ed.ac.uk/software/figtree/>).

2.1 General Software Packages

MEGA, Geneious, and Seaview include several approaches to phylogenetic analysis and are also useful for preparing and converting genetic data files and formats (for example, assembly and alignment). MEGA and Seaview have a long tradition of consolidating sequence editing and analysis in a single package. Geneious is a relatively new bioinformatics software package whose analytical tools extend beyond sequence and phylogenetic analysis. These programs have graphical interfaces where sequence alignments can be displayed and inspected for ambiguities and gaps. It is highly

recommended to inspect the sequence alignment before proceeding with sequence analysis. All three packages infer phylogenetic trees using different methods, including Neighbor-Joining (NJ), Maximum Likelihood (ML), and Bayesian approaches.

2.2 *Phylodynamics Software*

BEAST (Bayesian Evolutionary Analysis Sampling Trees) is a complete software package to study viral phylodynamics [5]. For example, it has been used to infer the dates and the population dynamics of multiple emergences of circulating vaccine-derived polioviruses (cVDPV) [6]. BEAST v1.8 includes four programs: (1) BEAUTi, (2) BEAST, (3) Tree annotator, and (4) Log combiner. There are two additional programs that need to be installed separately: TRACER and FigTree. Succinctly, BEAUTi prepares the XML file that contains the sequence alignment and model specifications. BEAST takes as input the XML file and generates two outputs; a log file and a tree file. The log file can be read using TRACER while the tree file is processed using Tree annotator, which will generate a Maximum Clade Credibility (MCC) tree. The annotated MCC tree is displayed using FigTree.

3 Methods

Estimation of genetic distances is the basic building block in phylogenetic inference. Nucleotide substitutions are modeled following Markov models and estimated using substitution matrices. For in-depth study, Allman [7] and Yang [8] provide excellent background information on the statistical models used in estimation of genetic distances. Poliovirus evolution is characterized by (a) having a high substitution rate, and (b) a differential rate of accumulation of transition (substitutions between two purines $A \leftrightarrow G$ or two pyrimidines $T \leftrightarrow C$) and transversion (substitutions between a pyrimidine and a purine $T, C \leftrightarrow A, G$) nucleotide changes. Genetic distances estimated using the Kimura 2-parameter model correct for both multiple hits at the same site and the differential rate of transition and transversion changes. Parameter-rich models of evolution incorporate additional substitution dynamic nuances. Fit of a particular dataset to a model of evolution can be investigated using MEGA or Modeltest [9]. In addition, substitutions are not homogeneously distributed along the genome. For example, in the protein coding gene VP1, the majority of substitutions occur at the third codon position. The observed variability of substitutions across sites can be incorporated into the models of evolution by using a gamma distribution. Most phylogenetic programs incorporate this function and the user is asked to either provide the shape parameter alpha (α) or letting the program estimate it. Estimates of α for poliovirus VP1 sequence alignments gravitate around 0.3.

3.1 Genetic Distances

Dataset: FASTA alignment of WPVI sequences available at GenBank. Accession numbers EF374000–EF374030.

1. Once the FASTA file is open in MEGA, different types of substitution models are available under the Compute Pairwise Distance menu. Three separate runs will provide estimated distances by choosing p -distance, Kimura 2-parameter model, and Tamura–Nei model (with gamma distributed rates among sites) respectively. Gamma distributed rates is selected by switching from Uniform rates to Gamma distributed in the Rates among Sites option.
2. Estimated distances are shown in a matrix for all pairwise comparisons and exported in tabular CSV format for use in a spreadsheet (*see Note 1*).
3. Comparison of genetic distances can be estimated using three different substitution models by combining each file into a single worksheet. Parameter-rich models capture increased genetic distance when sequences under comparison become more divergent. From the example dataset, estimated distances between closely related sequences (p -distance, 0.014 substitution per site [s/s]) have a very slight increase compared to the K2P (0.015 s/s) and TN+G (0.015 s/s) models. When moderately divergent sequences are compared (p -distance, e.g., 0.078 s/s), the estimated genetic distance increased by about 30 % by applying the TN+G model (~ 0.103 s/s).

3.2 Maximum Likelihood Tree

Dataset: FASTA alignment of WPVI sequences available at GenBank. Accession numbers EF374000–EF374030.

1. Import the fasta alignment in Geneious v7.
2. Select Tools \rightarrow Tree or click on the Tree icon.
3. The ML algorithm implemented in Geneious is PhyML [10]. It is not installed by default and should be installed as a downloadable plugin following directions in Tools \rightarrow Plugins.
4. Once the PhyML plugin is installed, it will be available in the Tree menu. Click on the PhyML tab.
5. The options displayed in the PhyML menu (Fig. 1) include choice of substitution model and its parameters, branch support, and topology strategy and optimization.
 - (a) In this example, the GTR model of nucleotide substitution is used (*see Note 2*). Next, to allow the substitution rate to vary among sites the option Number of substitution rate categories (N) should be >1 . We set N to 4. The value of the Gamma distribution parameter can be fixed or estimated by the program. We set this option to Estimated. An additional option allows the proportion of invariable

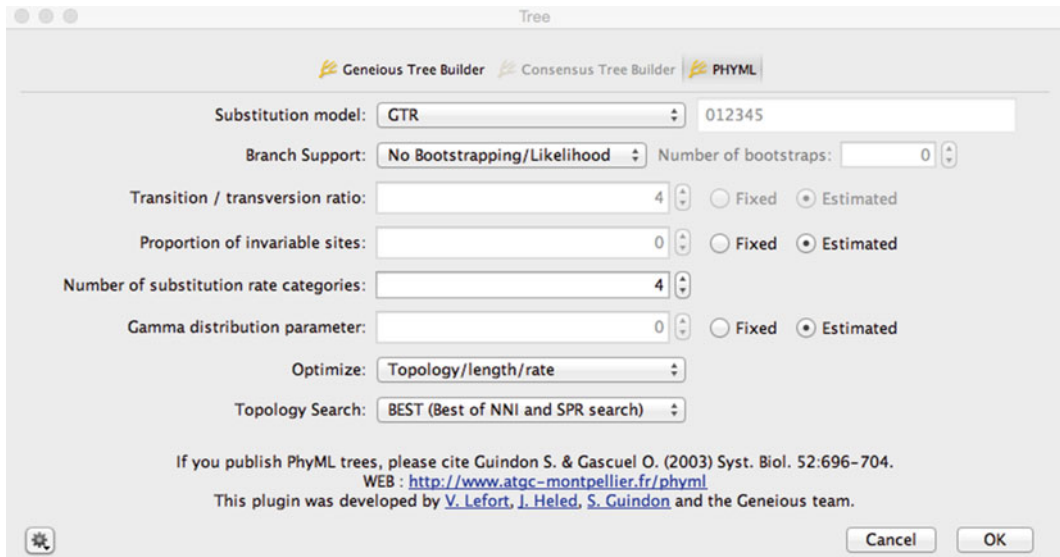


Fig. 1 PhyML window menu displayed in Geneious v7. The PhyML plugin for Geneious is available from Geneious. Model parameters and phylogenetic options for the ML run are set in this window

sites to be fixed or estimated (fixing the value to zero results in canceling this option). We set Proportion of invariable sites to Estimated.

- (b) There are four options under Branch support, the most used of which is Bootstrap. This method is based on resampling with replacement from the original nucleotide sites of a sequence and inferring a new tree from each sample. Comparison of the tree from the original sequence with those arising from the resampling will indicate the level of statistical support for the branching topology. This method is computationally expensive. In this example, no bootstrap values are computed (*see Note 3*).
- (c) Under Optimize, we set this option to Topology/length/rate.
- (d) Last, the option Topology search contains three choices. We choose BEST, which combines two alternative topology strategies.
- (e) After the run, PhyML will generate a tree file located in the same Geneious folder that contains the alignment under analysis. The tree file is visualized in Geneious. The tree visualization within Geneious allows further editing of the tree, including branch swapping, rooted/unrooted and circular layouts, and expansion of the tree for better visualization of trees with numerous sequences (Fig. 2). The tree can be exported in Nexus or Newick formats for further editing in other software, for example FigTree.

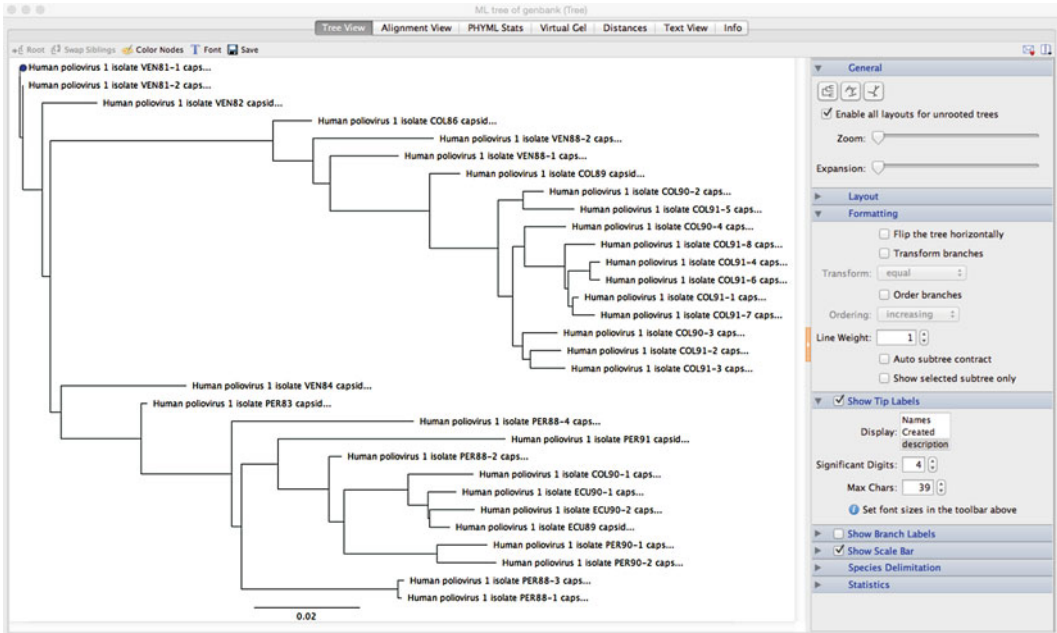


Fig. 2 Visualization within Geneious of the phylogenetic tree reconstructed by PhyML. The tree was rooted to sequence VEN-81 by selecting the sequence in the tree and pressing Root in the options shown in the *left hand corner* above the tree

3.3 Molecular Clock Analysis

This section describes the methods used to estimate substitution rates from a set of serial samples of poliovirus sequences. Rates of evolution are generally put into practice for estimating divergence times. For example, by assuming a constant rate of evolution (molecular clock) among viral lineages, it is possible to infer the dates of ancestral or source infections. Bayesian Markov chain Monte Carlo (MCMC) methods are used to infer substitution rates and are implemented in the software BEAST. In general terms, Bayesian phylogenetics are based on prior assumptions about parameters; the optimum values are obtained from a continuous distribution sample set a particular state and then proposing new states from sliding windows that will be accepted according to acceptance ratios. Development of Markov chain Monte Carlo (MCMC) algorithms provided the methods for achieving Bayesian computation.

Dataset: FASTA alignment of WPV1 sequences available at GenBank. Accession numbers EF374000–EF374030.

1. The first step is incorporating the temporal data associated with each sequence in the sequence name (tip dates). Source of temporal data includes specimen date or onset date. BEAST can recognize several date formats. In this example, a decimal format is used. For example, when the date is 1 July 1981, we

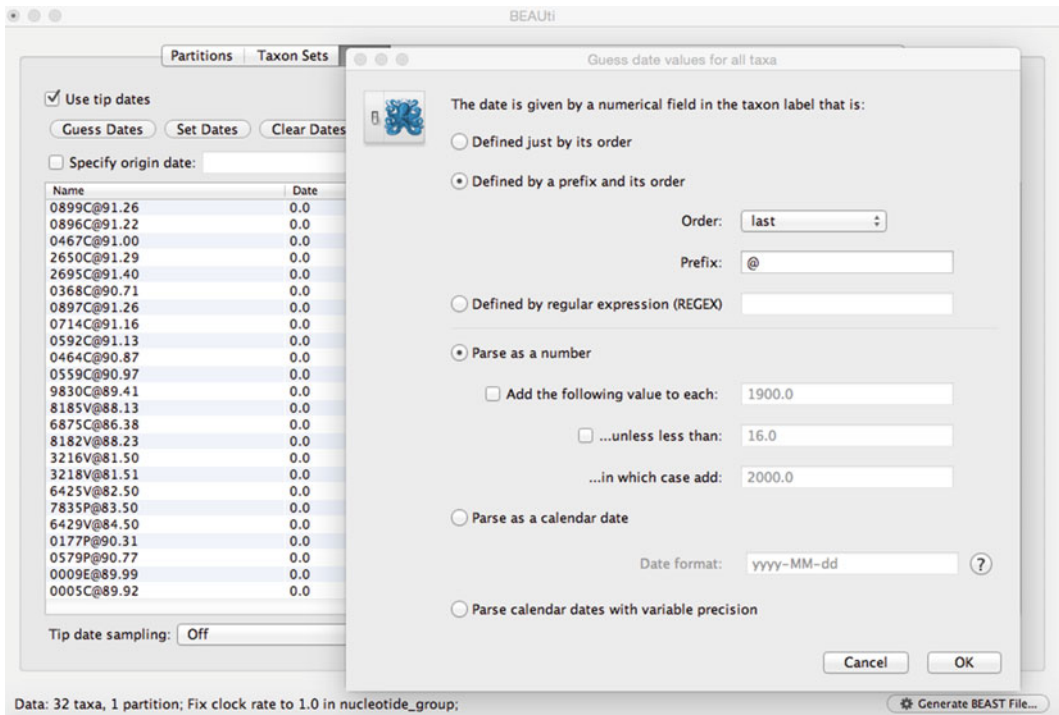


Fig. 3 BEAUTi window menu showing the options for incorporating numerical fields (tip dates) from the sequence name in the analysis

- search the day of the year for 1 July (in this example, the 182nd day) and then divide it by 365. The corresponding decimal format is 1981.49. This process can be easily obtained using Excel.
- Once all sequence names are labeled incorporating temporal data, the corresponding alignment (in Nexus format) is imported into BEAUTi.
- BEAUTi prepares an XML file for BEAST. Once the file is imported in BEAUTi, choose the Tip tab and check Use tip dates option. Pressing Guess dates will open a new menu for recognizing the numerical field for each sequence name (Fig. 3) (*see Note 4*). If successful, the Date column will be populated with the corresponding numerical values.
- The next tab to be modified is the Site tab. Under Substitution model, choose GTR and under Site heterogeneity, choose Gamma + Invariant sites.
- In the clocks tab, select Strict clock and Estimate. Alternatively, the rate can be fixed to a value consistent with the units of time.
- For the purpose of this example, the tab Trees can be left at default values (*see Note 5*).

7. The Priors tab shows every parameter of the model selected and its corresponding prior distribution. The priors that appear in red should be set. In this example, `clock.rate`. A prior selection window appears after clicking on the parameter. Select uniform distribution (*see Note 6*).
8. The Operators tab is set to Auto Optimize.
9. The MCMC tab tunes the MCMC chain. Length of chain depends on the size of the data set and the models chosen. For initial tests, choose a value of 1,000,000. The Log parameters option specifies how often a sample is recorded. It is recommended a final sampling of no more than 10,000 samples. The value in this field can be calculated as Length of chain/10,000. The option Echo state is related to the amount of information displayed on the screen and it is recommended to follow the value calculated from the Log parameters option. The remaining options can be set as default or adjusted per user requirement.
10. Click on Generate BEAST file. If any parameter has improper priors, it will be shown in the window before saving the XML file.
11. Run BEAST. The XML file generated in BEAUTi can be chosen from the dialog box. Click on Run. Progress of the run will be displayed on the screen. Once finished, two files will be generated; log (file with extension `.log`) and tree files (file with `.trees` extension).
12. The log file is analyzed using the program Tracer.
 - (a) Import the log file into Tracer.
 - (b) The mean substitution rate estimated from BEAST is displayed under `clock.rate`. Confidence intervals around the mean are displayed in the Estimates tab as 95 % HPD (highest posterior density) intervals (Fig. 4).
 - (c) Assessment of sample autocorrelation is checked in the ESS (Effective Sample Size) column. Low ESS values are displayed in red or yellow and generally are indicative of short runs and the chain length needs to be adjusted accordingly.
13. Processing of the tree file is performed in TreeAnnotator. The tree file contains all sampled trees (in this example, 10,000 trees). TreeAnnotator summarizes in a single tree the information from all sampled trees.
 - (a) Burnin (as the number of trees) is the number of excluded trees from the summary. In general is set to 10 % of the total number of sampled trees (in this case, 1000).
 - (b) Nodes are annotated according to the Posterior probability limit set in TreeAnnotator. If set to zero, all nodes will incorporate a summary of the annotations stored during the BEAST run.

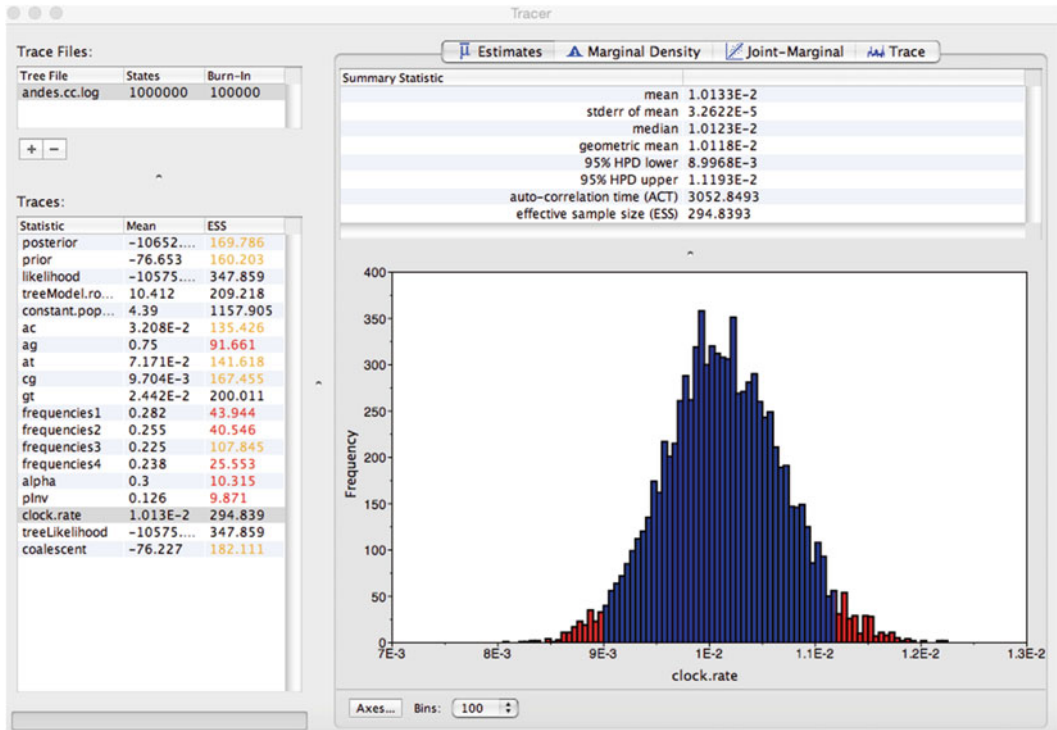


Fig. 4 Inspecting the log file in TRACER. The *left hand side* column displays the traces of each parameter including the mean and ESS values. On the *right*, statistics about parameters (including 95 % HPD) are shown in the *upper* window and a graphical visualization is displayed in the *lower* window. Each sampled value during the MCMC run can be graphically inspected in the tab labeled Trace

- (c) Select Maximum clade credibility (MCC) tree in the Target tree type option for obtaining a tree that has the highest product of the posterior probability of all of its nodes. Node heights can be summarized as Mean or Median values.
 - (d) Choose the tree file generated by BEAST and give a file name for the resulting tree.
 - (e) The MCC tree generated by TreeAnnotator can be visualized in FigTree.
14. In order to generate a time scale in FigTree, open the Time Scale tab and in the Offset by option enter the date of the most recent sample (as displayed in the sequence name) and in the Scale factor option change the default to -1 (negative one). In the Scale axis tab, check the option Reverse Axis. Check the Node Labels tab and select Node Ages from the Display option. All nodes will display estimates of the divergence dates in time scale values (Fig. 5). 95 % HPD values can be summarized in bars across nodes by checking Node Bars and selecting the parameter of interest for display.

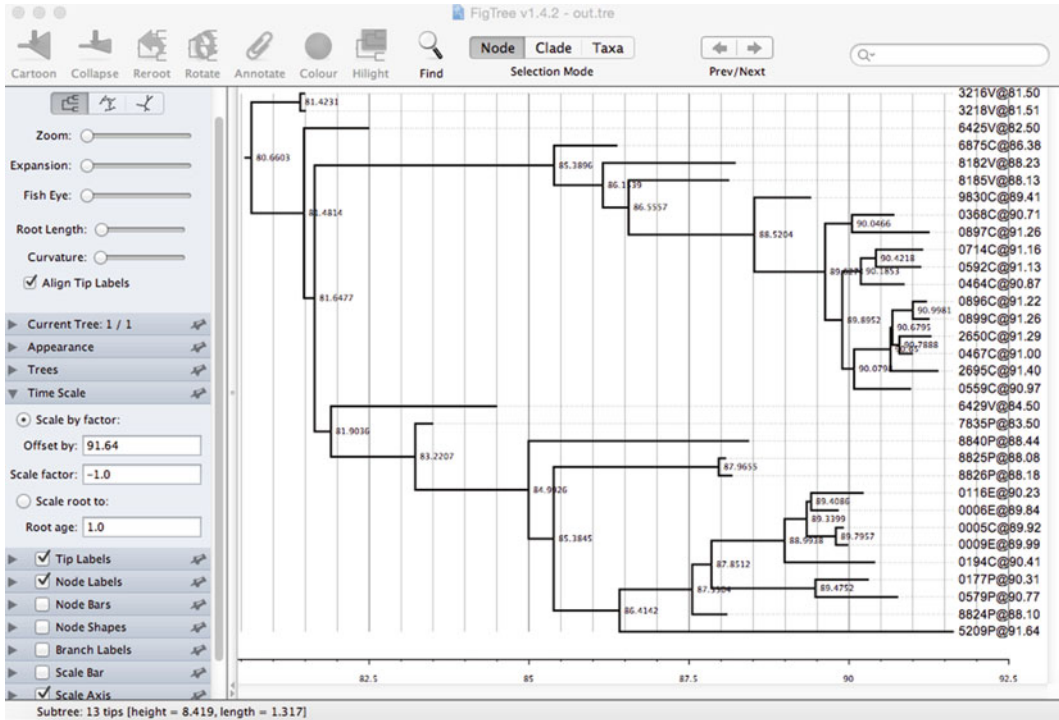


Fig. 5 The program FigTree edits and displays the annotated MCC tree generated by TreeAnnotator. The time scale on the tree is displayed by setting the Time scale and Scale axis options as shown in the figure

4 Notes

1. Pairwise distances displayed in MEGA can be exported in Excel and CSV formats. In addition, the order of sequences can be changed before exporting the matrix by holding the sequence name and moving it up or down the column.
2. In order to fix the Transition/Transversion ratio to a specific value (for example, fixed to 10), choose TN93, HKY85, or K80 models of evolution.
3. The number of bootstraps is dependent on the size of the data set. It is recommended choosing a minimum of 100 replicates when selecting this option.
4. It is recommended to include a tag (prefix) before the date for easier detection of the numerical field in the sequence name. For example, VEN81-1@1981.49 When a date is unknown, for example the sequence from Sabin vaccine, BEAST can estimate the date by choosing Tip date sampling (sampling with individual priors) in the Tips menu. The sequence is first chosen in the Taxon sets tab and it becomes available in the Apply to taxon set option.

5. There are several models for investigating population size changes under coalescent models, including coalescent exponential growth and Bayesian skyline. There are numerous sources of information dealing with specific coalescent models, including the BEAST-users mailing list (<https://groups.google.com/forum/#!forum/beast-users>).
6. For polioviruses, a reasonable initial value for clock.rate is 0.011 and upper and lower values in the range of 0.1 and 0.001 respectively. Alternatively, the Gamma distribution can be set as prior distribution.

References

1. Jorba J, Campagnoli R, De L, Kew O (2008) Calibration of multiple poliovirus molecular clocks covering an extended evolutionary range. *J Virol* 82:4429–4440
2. Kew OM, Mulders MN, Lipskaya GY, da Silva EE, Pallansch MA (1995) Molecular epidemiology of polioviruses. *Semin Virol* 6:401–414
3. Kew OM, Pallansch MA (2002) The mechanism of polio eradication. In: Semler BL, Wimmer E (eds) *Molecular biology of picornaviruses*. ASM Press, Washington, DC, pp 481–491
4. Al-Hello H, Jorba J, Blomqvist S, Raud R, Kew O, Roivainen M (2013) Highly divergent type 2 and 3 vaccine-derived polioviruses isolated from sewage in Tallinn, Estonia. *J Virol* 87:13076–13080
5. Volz EM, Koelle K, Bedford T (2013) Viral phylodynamics. *PLoS Comput Biol* 9:e1002947
6. Burns CC, Shaw J, Jorba J, Bukbuk D, Adu F, Gumede N, Pate MA, Abanida EA, Gasasira A, Iber J, Chen Q, Vincent A, Chenoweth P, Henderson E, Wannemuehler K, Naem A, Umami RN, Nishimura Y, Shimizu H, Baba M, Adeniji A, Williams AJ, Kilpatrick DR, Oberste MS, Wassilak SG, Tomori O, Pallansch MA, Kew O (2013) Multiple independent emergences of type 2 vaccine-derived polioviruses during a large outbreak in northern Nigeria. *J Virol* 87:4907–4922
7. Allman ES, Rhodes JA (2004) *Mathematical models in biology: an introduction*. Cambridge University Press, New York
8. Yang Z (2006) *Computational molecular evolution*. Oxford University Press, Oxford
9. Posada D, Crandall KA (1998) MODELTEST: testing the model of DNA substitution. *Bioinformatics* 14:817–818
10. Guindon S, Dufayard JF, 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. *Syst Biol* 59:307–321