PhyloGeoTool
User Manual

EWOUT VANDEN EYNDEN, PIETER LIBIN, KRISTOF THEYS,
GUY BAELE
Rega Institute for Medical Research, KU Leuven

May 2017

Contents

This document explains how to use the PhyloGeoTool software. PhyloGeoTool is a web application that is installed on a web server, and can be accessed via a web browser. Most users will access a publicly accessible instance of the PhyloGeoTool (i.e. an instance that was already deployed on a web server) and have no need to install an instance themselves. An example of a publicly available instance of the PhyloGeoTool that allows to explore a large HIV cohort (i.e. EuResist [2]) can be found at `http://phylogeotool.gbiomed.kuleuven.be/euresist/`. It is however also possible to install your own instance of the PhyloGeoTool; for this, we refer to the installation manual [1]. Such an instance can be used to investigate a dataset on a local computer or to host your own public PhyloGeoTool instance.

1 Getting started

PhyloGeoTool implements a visual method to explore large phylogenetic trees and to depict characteristics of strains and clades, including their geographic context, in an interactive way. The tool also provides the possibility to insert new virus strains into the existing phylogenetic tree, allowing users to gain insight in the placement of such new strains without the need to reconstruct the phylogeny de novo.

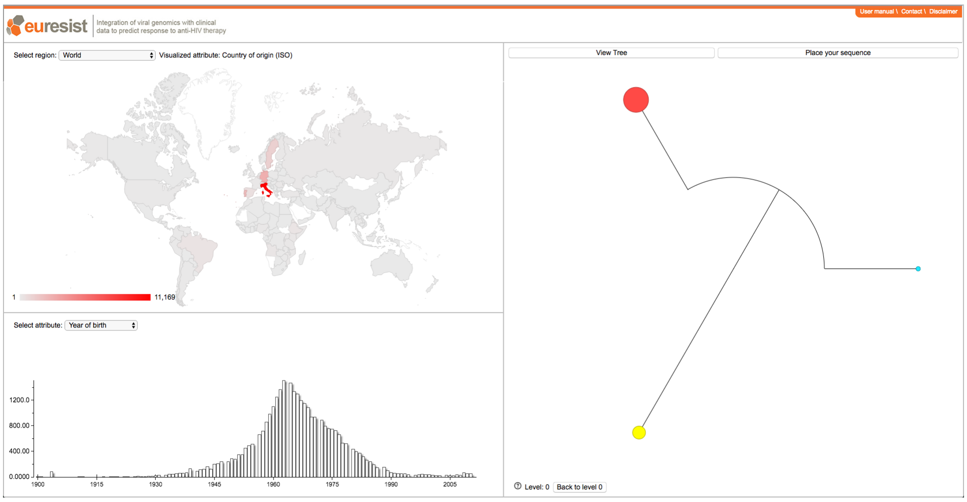
A particular PhyloGeoTool instance can be used by navigating the browser to the URL at which this instance has been deployed. For example, the aforementioned public EuResist instance of the PhyloGeoTool is deployed at `http://phylogeotool.gbiomed.kuleuven.be/euresist/`, and thus, you should navigate your browser to this exact URL. Another instance of the software has been set up in the context of Dengue virus, using genetic sequence data that is publicly available in Genbank. For this instance, the E gene was used to build a phylogenetic tree including all four serotypes and additional information associated with each sequence is visualized at `http://phylogeotool.gbiomed.kuleuven.be/dengue/`.

Note that any browser (Chrome, Firefox, Internet Explorer) should be able to visualize the PhyloGeoTool application.

If you experience any problems, please contact us: phylogeotool@kuleuven.be.

[1] https://github.com/rega-cev/phylogeotool

2 Core functionality and user interface


Select region: World

Visualized attribute: Country of origin (ISO)

Select attribute: Year of birth

euresist
Integration of viral genomics with clinical data to predict response to anti-HIV therapy
User manual \ Contact \ Disclaimer
View Tree
Place your sequence
1
11,169
1200.0
800.00
400.00
0.0000
1900
1915
1930
1945
1960
1975
1990
2005
Level: 0
Back to level 0

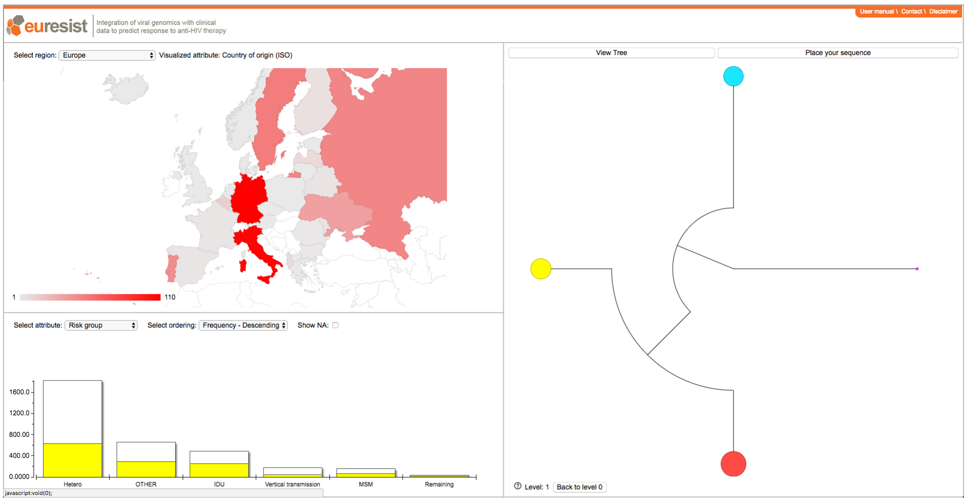
Figure 1: Start and main view of the PhyloGeoTool. The panel on the upper left shows a map that depicts the geographical distribution of the sequences of the partitioned phylogeny; the lower left panel shows a bar chart that depicts the value distribution for one of the sequence attributes; and the right panel shows a top-level view of the clustering of the tree, as determined by our clustering algorithm.

The PhyloGeoTool is divided into different panels (see Figure 1):

- The top left panel shows a map where each country is colored according to a gradient, where a darker color signifies that more sequences are originating from this country. The maximum value of the gradient bar shown denotes the highest number of sequences in the dataset for a single country. A drop-down box allows you to select the geographic region on which you want to focus (e.g. Europe, North America, . . .).
- The bottom left panel shows a histogram or bar chart depending on the chosen attribute. If the attribute is represented by a continuous dataset, a histogram will be shown while a discrete dataset is displayed by a bar chart plot. The histogram or bar chart shows a certain attribute (i.e. characteristic) from all sequences currently visualized in the tree panel on the right. A drop-down box allows you to select the attribute to be plotted on the bar chart. The attributes available for selection depend on the attributes made available by the PhyloGeoTool instance.
- The right panel shows a top-level view of the clustering of the tree (initially the root of the tree), as determined by our clustering algorithm.We provided a detailed explanation about the working of our

clustering algorithm in a manuscript that is currently being reviewed. This panel allows you to visualize the phylogenetic tree in its entirety, colored according to the shown clusters, and to add new sequences to the visualized phylogeny.

2.1 Hovering and clicking cluster nodes



Select region: Europe

Visualized attribute: Country of origin (ISO)

Select attribute: Risk group

Select ordering: Frequency - Descending

Show NA: □

Level: 1

javascript:void(0);

Figure 2: Hovering over a node in the cluster phylogeny generates an overview (in the panels on the left) of the characteristics of that cluster. The bottom left panel shows these characteristics relative to the entire data set.

When the user hovers the mouse pointer over a node in the clustered phylogeny, both the attribute chart and the map are updated. The attribute chart will be extended (i.e. extra bars will be shown on top of the original chart for each value) with the data that is located in the cluster over which the mouse pointer hovers. The map, on the other hand, will only show the geographic distribution of the sequences found in the cluster over which the mouse pointer hovers.

When you click on a node in the clustered phylogenetic tree, the PhyloGeoTool will navigate to the clustering of this particular node (i.e. will go down one level in the node on the clustered phylogenetic tree). A counter on the bottom left of the cluster panel keeps track of the depth of the navigation. As you can visit a cluster node by clicking it, you can go back to into your navigational path by using the browser's 'Back' button. The depth of the navigation is also shown when you hover over the '?' symbol which is shown next to 'Level: '. Navigating directly to the root of the tree can be done modifying the URL that it ends with 'root' or by pressing the button label 'Back to level 0'.

Additionally, the URL encodes the level of descent in the clustered phylogenetic tree. A URL can hence be bookmarked or shared to go back at the exact location in the clustered phylogenetic tree.

2.2 Selecting a geographic region

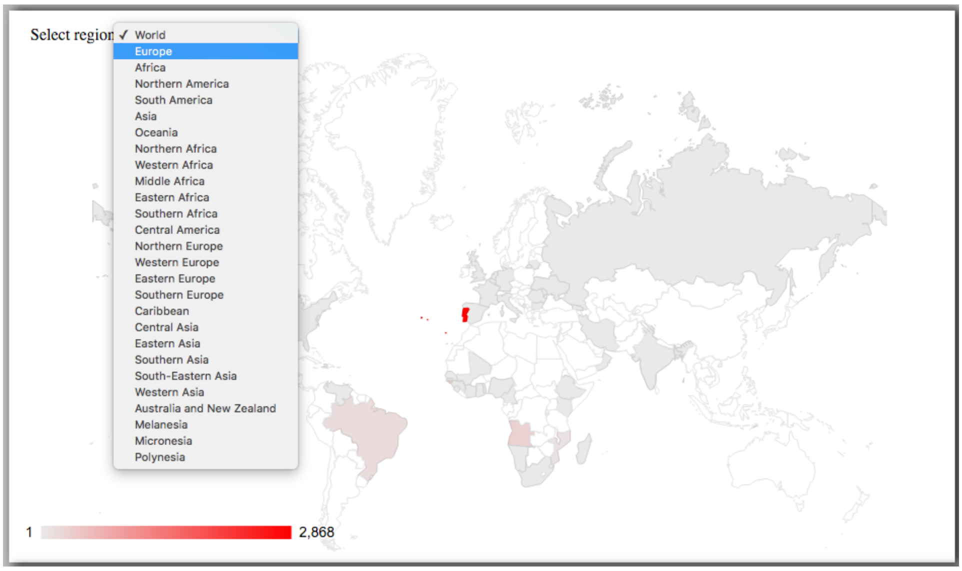


Figure 3: Change the region in the drop down box

You can select your geographical region of interest in the drop-down box on top of the map (see Figure 3).

2.3 Selecting a sequence attribute

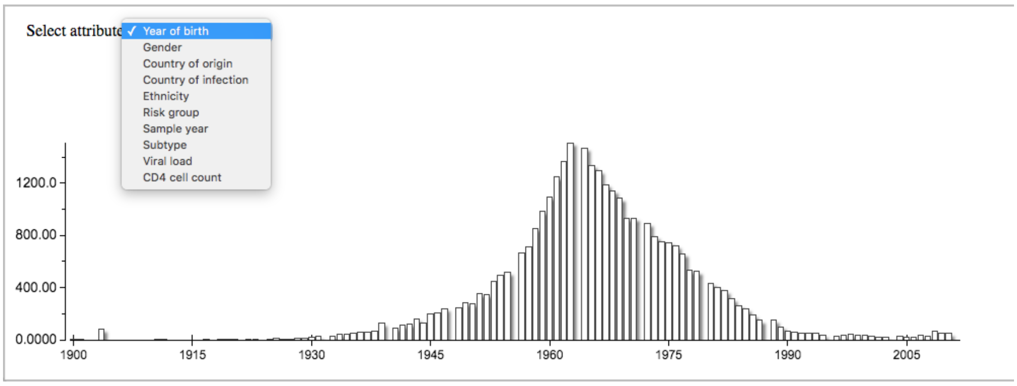


Figure 4: Change the attribute in the drop down box

You can select an attribute of interest in the drop-down box on top of the attribute chart (see Figure 4).

2.4 Exporting the phylogenetic tree

2.4.1 Private database

When the 'View Tree' button (located at the top of the clustered phylogeny panel) is clicked for a private database being explored using the PhyloGeoTool, a pop-up window is shown (Figure 5) that visualizes the phylogeny as a radial phylogenetic tree. The colors in the radial tree correspond to the colors of the clusters. PhyloGeoTool uses a FigTree (`https://github.com/rambaut/figtree`) jar (Java ARchive) file to visualize the radial tree with its corresponding colors.

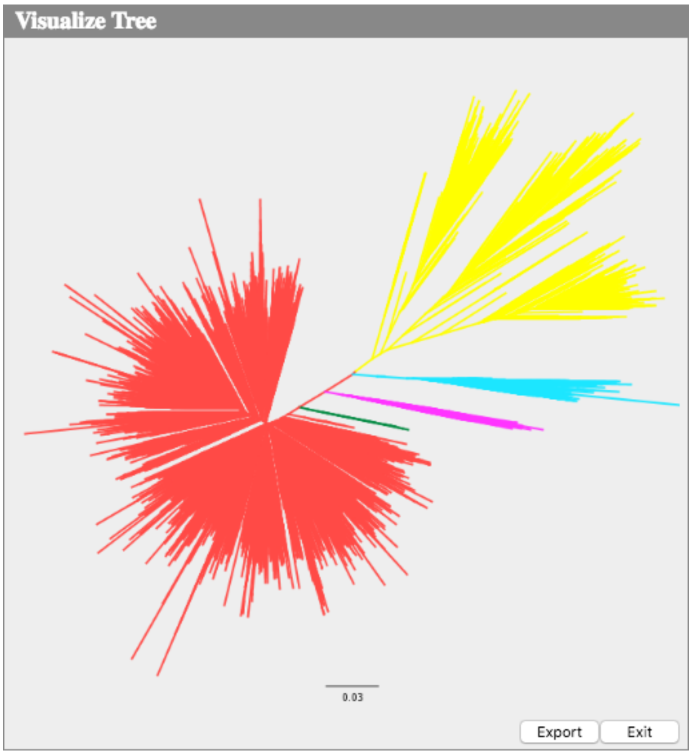
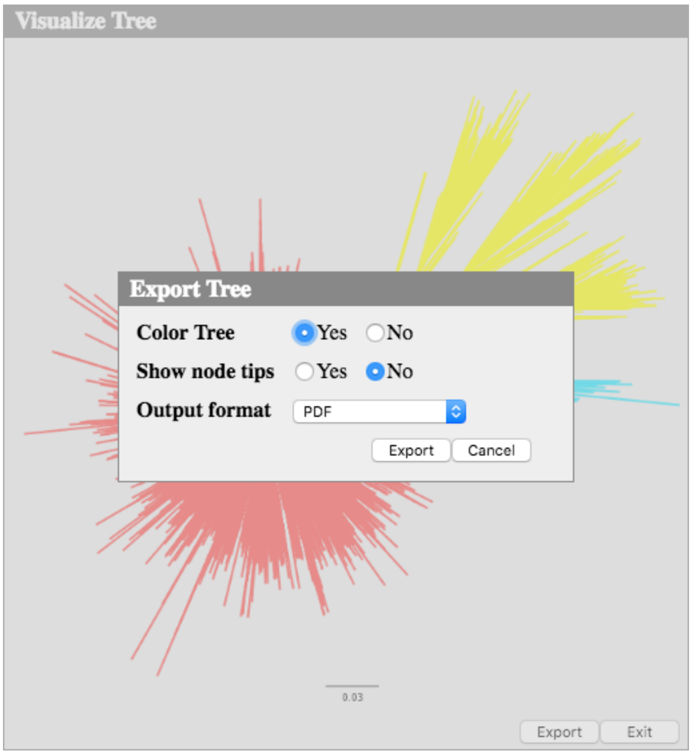


Figure 5: Visualizing the phylogenetic tree, for a private sequence database, colored according to the clustering scheme.

On the bottom of the pop-up window, there are two buttons: 'Export' and 'Exit'. The button with label 'Export' will show a window similar to the one shown in Figure 6. There are three options available that allows the customization of the export:

- Color Tree: enable/disable the coloring of tree based on the clusters' color.
- Show node tips: include/exclude leaf names in the phylogenetic tree.
- Output Format: the format in which you want your tree to be downloaded (i.e. PDF, NEXUS, SVG, PNG). [2]



Color Tree ● Yes ○ No

Show node tips ○ Yes ● No

Output format PDF

Figure 6: Form displaying the preferred settings to export the shown phylogenetic tree.

2.4.2 Public database

When the 'View Tree' button (located at the top of the clustered phylogeny panel) is clicked for a public database being explored using the PhyloGeoTool, a pop-up window is shown (Figure 7) that visualizes the phylogeny as a radial phylogenetic tree. In the case of a public database, sequence data can be downloaded from the PhyloGeoTool, which will typically not be allowed for private database deployments. The other features remain the same as in section 2.4.1.

[2] Exports in the NEXUS file format can be visualized in programs such as FigTree (`http://tree.bio.ed.ac.uk/software/figtree/`).

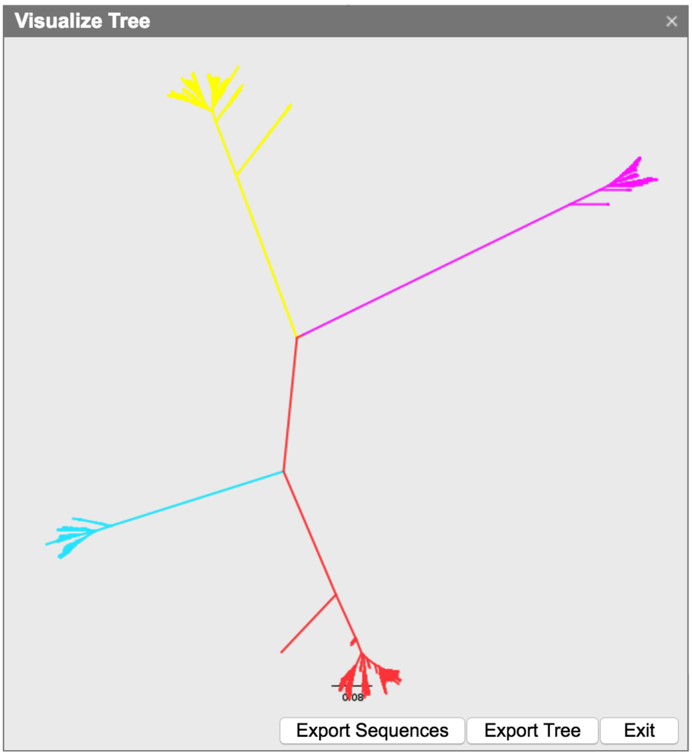
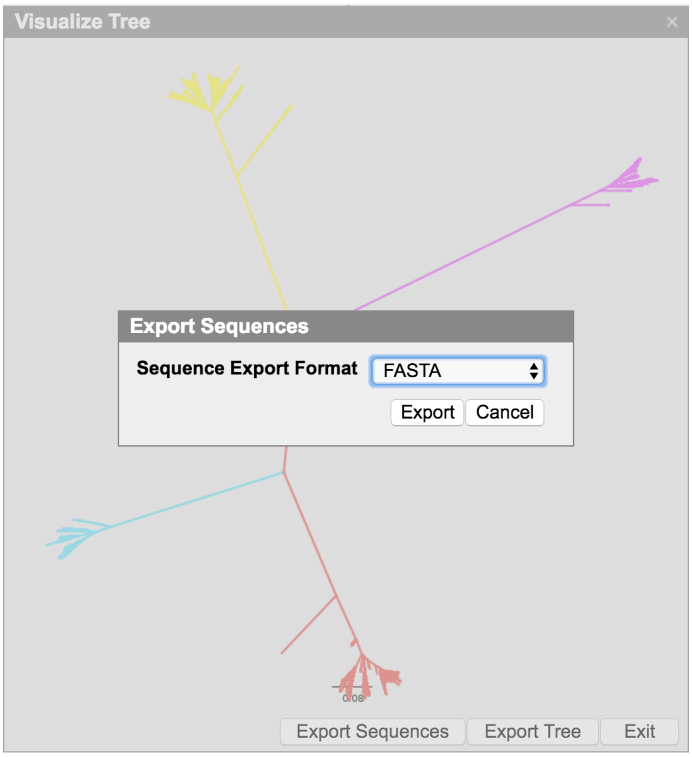


Figure 7: Visualizing the phylogenetic tree, for a public sequence database, colored according to the clustering scheme.

On the bottom of the pop-up window, there is now an extra button: 'Export Sequences'. This feature allows to export sequence data for the cluster currently being explored, into a FASTA file that can be downloaded (Figure 8).



Sequence Export Format: FASTA

Figure 8: Form displaying the preferred settings to export the sequence data of the clusters shown.

2.5 Automated phylogenetic placement of sequences

Phylogenetic placement enables the placement of unknown query sequences onto an existing phylogenetic tree, without the need to re-compute the entire phylogeny (which is a time-consuming process, especially when a large number of sequences is involved). In the PhyloGeoTool, we use phylogenetic placement to position new sequences onto the phylogeny, in a reasonable amount of time (i.e. minutes). We implement this placement using the well-known pplacer software package [1].

To place your sequences on the cluster phylogeny, click the 'Place your sequence' button in the clustering window. A window (see Figure 9) will pop up that enables you to place your sequences.

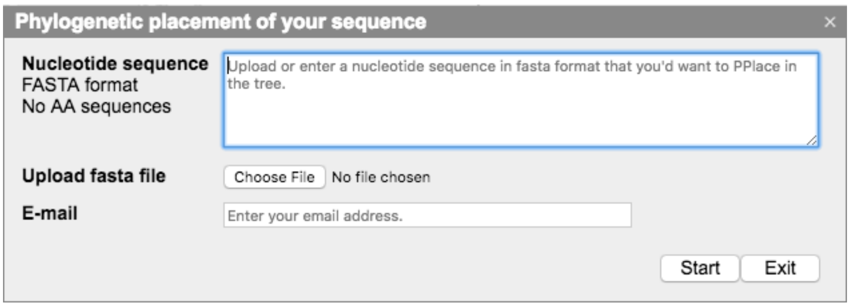
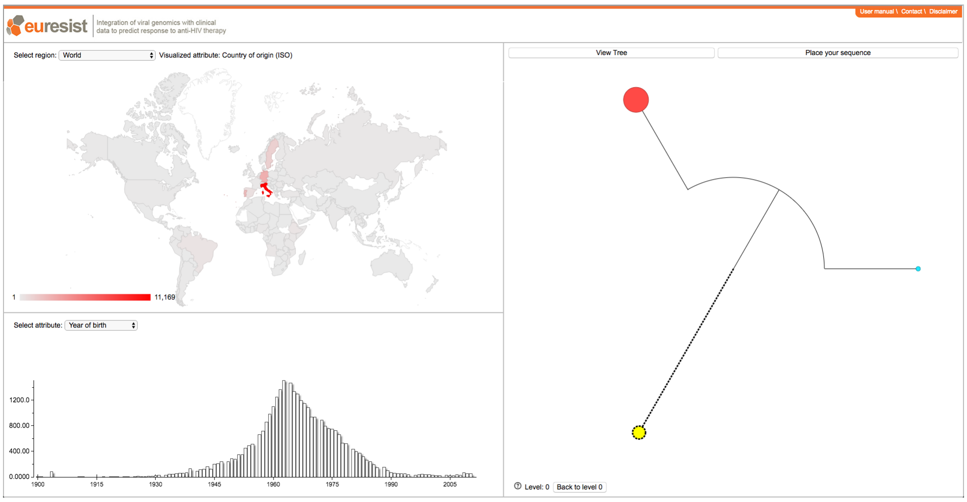

Phylogenetic placement of your sequence
×
Nucleotide sequence
FASTA format
No AA sequences
Upload or enter a nucleotide sequence in fasta format that you'd want to PPlace in the tree.
Upload fasta file
Choose File
No file chosen
E-mail
Enter your email address.
Start
Exit

Figure 9: Phylogenetic placement window.

There are two ways to add a new sequence to the tree:

- Paste your sequence in the 'Nucleotide sequence' field.
- Upload a FASTA file containing the nucleotide sequence using the file chooser.

Currently, we only support the submission of 1 sequence at a time. If you submit more than 1 sequence to be added to the phylogenetic tree, a warning will be shown. Before Pplacer starts working, a BLAST search is performed on the submitted sequence. If BLAST returns a low score (if, for example, the sequence is too short, too different from the represented dataset, . . .), the sequence is rejected and an e-mail is sent to the user to inform him/her of the low quality sequence data. Once the Pplacer process has finished you will receive an e-mail with a link to show the placement result. You will be taken to the original PhyloGeoTool web-page (such as described in Section 2) with the difference that the cluster, which contains the pplaced sequence, is surrounded by a dashed line (see Figure 10).


Select region: World

Visualized attribute: Country of origin (ISO)

Select attribute: Year of birth

Level: 0

euresist
Integration of viral genomics with clinical data to predict response to anti-HIV therapy
User manual \ Contact \ Disclaimer
View Tree
Place your sequence
1
11,169
1200.0
800.00
400.00
0.0000
1900
1915
1930
1945
1960
1975
1990
2005
Back to level 0

Figure 10: Start and main view of the PhyloGeoTool with a pplaced sequence.

3 Example: exploring a large HIV-1 dataset

As mentioned earlier, we host a publicly available instance of the PhyloGeoTool that uses all HIV-1 data that is currently available in the EuResist Integrated Data Base [2]. This database contains virus genotypes, clinical responses and epidemiological markers of more than 66.000 patients from 12 different countries. We have made a YouTube video available where all features we discussed in this user manual are demonstrated using the EuResist PhyloGeoTool instance: `https://www.youtube.com/watch?v=xiYODenyEIQ`.

4 Contact information and support

The PhyloGeoTool is developed and maintained by the Rega Institute KU Leuven. You can contact us on phylogeotool@kuleuven.be

References

1. F. A. Matsen, R. B. Kodner, and E. V. Armbrust. pplacer: linear time maximum-likelihood and bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics*, 11:538, 2010.

2. M. Zazzi, F. Incardona, M. Rosen-Zvi, M. Prosperi, T. Lengauer, A. Altmann, A. Sonnerborg, T. Lavee, E. Schülter, and R. Kaiser. Predicting response to antiretroviral treatment by machine learning: the euresist project. *Intervirology*, 55(2):123–127, 2012.