

VirPhyKit: An integrated toolkit for viral phylogeographic analysis

A quick guide to VirPhyKit

Sept, 2025

Yuqi Yin & Fangluan Gao

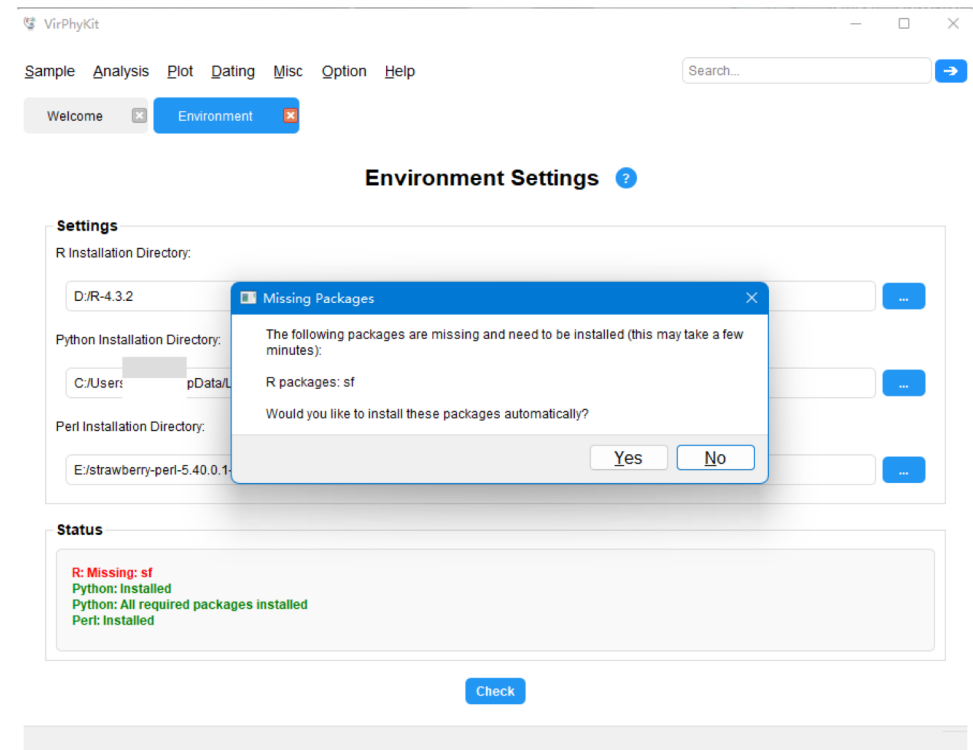
1. Installation

System requirement:

- Python 3.6 or later
- R 3.5 or later
- Perl

Prior to using VirPhyKit, users must specify the **installation paths for R, Python, and Perl** in the 'Options > Environment Settings ' menu and click [Check]. The toolkit will then automatically:

- Verify required Python and R package dependencies
- Identify any missing packages
- Prompt users with options to automatically install missing dependencies (on Linux, missing packages must be installed manually).



2. Modules in VirPhyKit

Menu	Module	Brief description
Sample	SeqHarvester	Retrieves viral isolates from GenBank, displays comprehensive metadata for filtering and download
	SeqIDRenamer	Provides a simple workflow for batch processing of sequence IDs using custom naming rules
	SeqGrouper	Groups viral sequences into customizable categories (e.g., geographic region, host, or user-defined criteria), with automated distribution reports for research and surveillance
	VirSpaceTime	Visualizes spatial and/or temporal distributions of viral isolates
Analysis	GeoSubsampler	Generates balanced sequence subsets by either geographic region or sample size, with integrated support for bootstrap analysis
	RRT	Detects sampling bias in phylogeographic analyses using a statistical approach
	TempMig	Reconstructs and visualizes spatial diffusion of viral pathogens through time
Plot	RSPP-Viz	Visualizes root state posterior probabilities inferred from trait-annotated MCC trees or MultiTypeTree
	BSP-Viz	Visualizes the export results of Bayesian skyline plots
Dating	TreeTime-RTT	Assesses clock-like evolutionary patterns through linear regression of root-to-tip distances against sampling dates in TreeTime
	TreeDater-LTT	Conducts a lineage-through-time (LTT) analysis using Treedater
Misc	MJRM Generator	Automatically constructs Markov jump and reward matrices from user-defined discrete traits (e.g., geographic location or host species) for phylogeographic analysis

3. Usage

SeqHarvester

Step 1: Enter the full name of the virus for retrieval or upload a text file containing the **accession number**.

Step 2: When retrieving by virus name, the results will display **the complete metadata** for all available isolates from GenBank in the '**Results**' panel.

Step 3: From the '**Genome Segments**' dropdown menu, select the **required segment**.

Step 4: Select the **output directory**, then click [**Download**] to initiate sequence retrieval.

The screenshot shows the SeqHarvester (Sequence Metadata Harvester) interface within the VirPhyKit application. The interface includes a menu bar (Sample, Analysis, Plot, Dating, Misc, Option, Help), a search bar, and a 'SeqHarvester' tab. The main content area is titled 'SeqHarvester (Sequence Metadata Harvester)' and includes a description: 'SeqHarvester retrieves viral isolates from GenBank, displaying comprehensive metadata for filtering and download.'

The interface is divided into several sections:

- Settings:** Includes a 'Virus Name' input field (annotated with a red circle 1), an 'OR' label, a 'Virus name or Accession File' dropdown menu (annotated with a red circle 2), and an 'Accession File' input field.
- Options:** Includes an 'Output Directory' input field, a 'Genome Segments' dropdown menu (annotated with a red circle 3), and a 'Download' button (annotated with a red circle 4).
- Status:** Includes a text area for 'Enter the full virus name to retrieve sequences, or select an accession number file to download sequences directly.'
- Results:** A table with columns 'Type', 'Number', 'Percentage', and 'Describe'.

Red arrows point from the numbered annotations to the corresponding elements: from circle 1 to the 'Virus Name' field, from circle 2 to the 'Virus name or Accession File' dropdown, from circle 3 to the 'Genome Segments' dropdown, and from circle 4 to the 'Download' button. A red box highlights the 'Results' table.

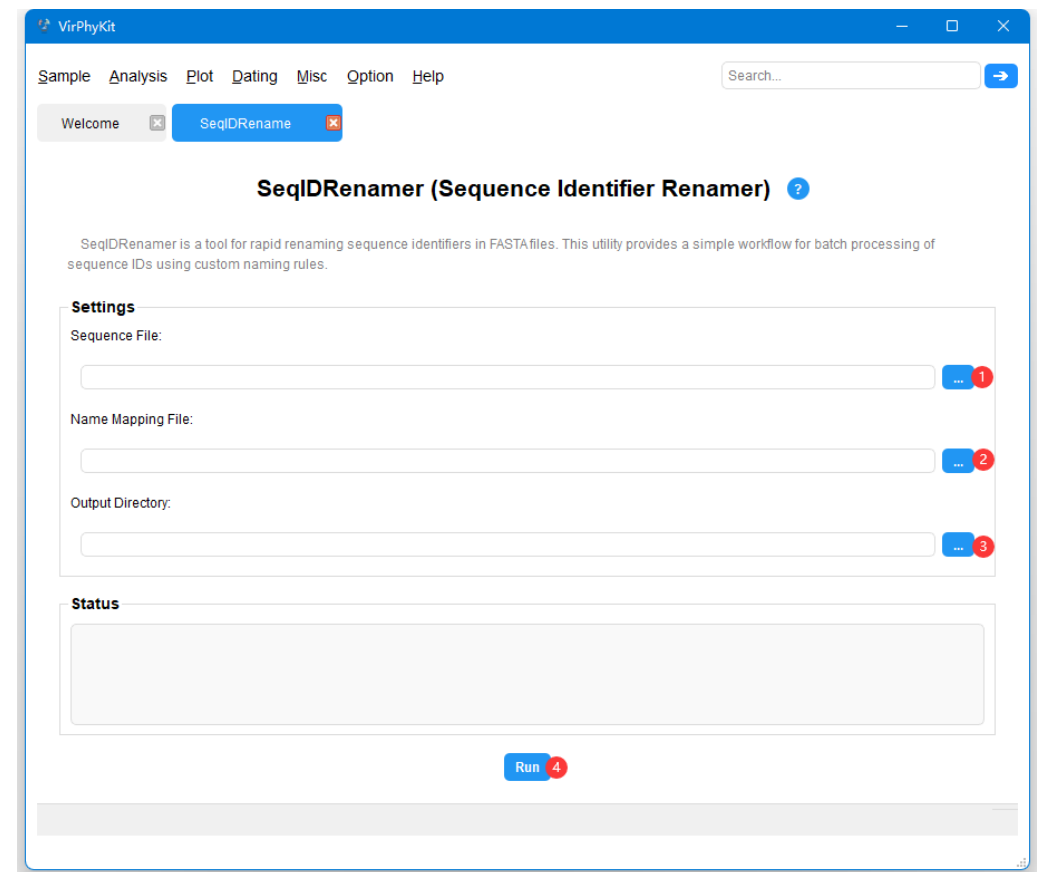
SeqIDRenamer

Step 1: Select a sequence file to rename.

Step 2: Upload a **tab-delimited mapping text file** ('OrigName\tNewName').

Step 3: Select the output directory to save the renamed file.

Step 4: Click the **[Run]** button to start the renaming process.



SeqGrouper

Step 1: Upload **accession numbers** in a text file (.txt) or upload a **GenBank file** (.gb).

Step 2: After **checking 'Enable'**, you can choose to use **default** grouping rules (by region) or upload a **custom** grouping table.

Step 3: Click **[Show Table]** to display viral isolate information (You can double-click any information to modify it).

Step 4: Select the columns that need to be grouped and click **[View]** to view the group distribution. **Note** that the grouped columns will be added to the table.

VirPhyKit

Sample Analysis Plot Dating Misc Option Help

Search...

Welcome SeqGrouper

SeqGrouper (Sequence Grouping Tool) ?

SeqGrouper is a specialized tool for grouping viral sequences into customizable categories (e.g., geographic region, host, or user-defined criteria), with automated distribution reports for research and surveillance.

Settings

Accession File (.txt):
E:/AC.txt

OR
GenBank File (.gb):

Category Mapping Table (.txt):
☒ Enabled
Built-in default mapping table (grouped by region)

Group of Sequences by Geo Location (Using Mapping Table)

Sequence Information Select a column to group

	Isolate	ID	Organism	Length	Host	Geo Location	Collection Date
1	MRS-PV3p-...	GQ427066.1	Pepper mild mottle virus	207	Homo sapiens	France	2008-01-01
2	Scr	LC793741.1	Pepper mild mottle virus	6356	Scrophularia buergeriana	South Korea	2022-05-26
3	CH-15-AB-...	OP723372.1	Pepper mild mottle virus	417	Capsicum annuum	Saudi Arabia	2022-01-10
4	PRO54348	MT385868.1	Pepper mild mottle virus	6356	Capsicum annuum	Chile	2019-05-03

2 Show Table View Save as CSV 5

VirSpaceTime

Step 1: Ensure R is properly configured (in 'Option-Environment').

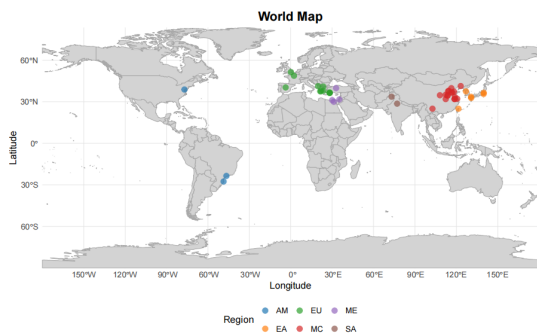
Step 2: Select **spatial or temporal** visualization mode.

Step 3: Upload a **tab-delimited text file**, containing **geographic coordinates** for spatial analysis or **population sample sizes over time** for temporal analysis.

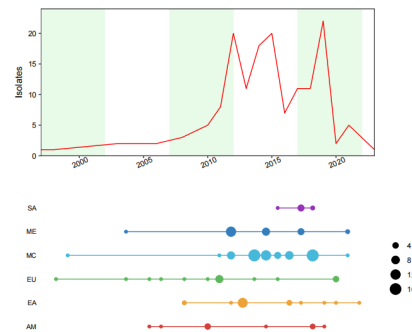
Step 4: Select the **output directory** and **filename**.

Step 5: Click [**Generate**] to create visualization.

Spatial mode



Temporal mode



A screenshot of the VirPhyKit VirSpaceTime web interface. The interface includes a menu bar (Sample, Analysis, Plot, Dating, Misc, Option, Help) and a search bar. The main content area is titled "VirSpaceTime (Viral Space-Time Visualizer)" and includes a description: "VirSpaceTime is a tool for visualizing spatial and/or temporal distribution patterns of viral isolates." The "Settings" section contains a "Visualization mode" dropdown (Temporal selected), an "Input File (.txt)" field, an "Output Directory" field, and an "Output Filename" field (set to "output_plot"). The "Status" section shows "R environment ready". A "Generate" button is at the bottom right. Red numbered circles (1-5) highlight key elements: 1. R environment ready status, 2. Visualization mode dropdown, 3. Input File field, 4. Output Filename field, 5. Generate button.

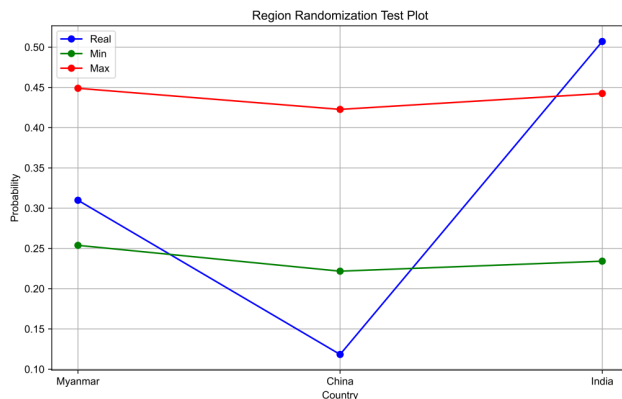
RRT

Step 1: Upload the **original MCC tree** and the **region-randomized MCC tree**.

Step 2: Select an **output directory** for the posterior probability csv table.

Step 3: Click **[Run]** to conduct region-randomization tests. The test **results** will be displayed in the **Status panel**.

Step 4: Specify the **output directory** for saving results.



The screenshot shows the VirPhyKit software interface for the RRT (Region Randomization Test). The window has a blue header with the title 'VirPhyKit' and a menu bar with options: Sample, Analysis, Plot, Dating, Misc, Option, Help. Below the menu bar is a search bar. The main content area is titled 'RRT (Region Randomization Test)' with a help icon. A brief description of the RRT method is provided. The 'Settings' section contains three input fields: 'Original MCC Tree' (with a file path and a blue button), 'Randomized MCC Trees' (with a file path and a blue button), and 'Export Data (CSV)' (with a file path and a blue button). Red arrows with numbers 1, 2, and 3 point to these buttons. The 'Status' section shows a message: 'Randomized file uploaded. Randomized file uploaded. Randomized file uploaded. RRT SUCCESSFULLY PASSED! Click 'Preview' to view the results.' Below this message are two buttons: 'Run' (with a red arrow and number 3) and 'Preview' (with a red arrow and number 4).

TempMig

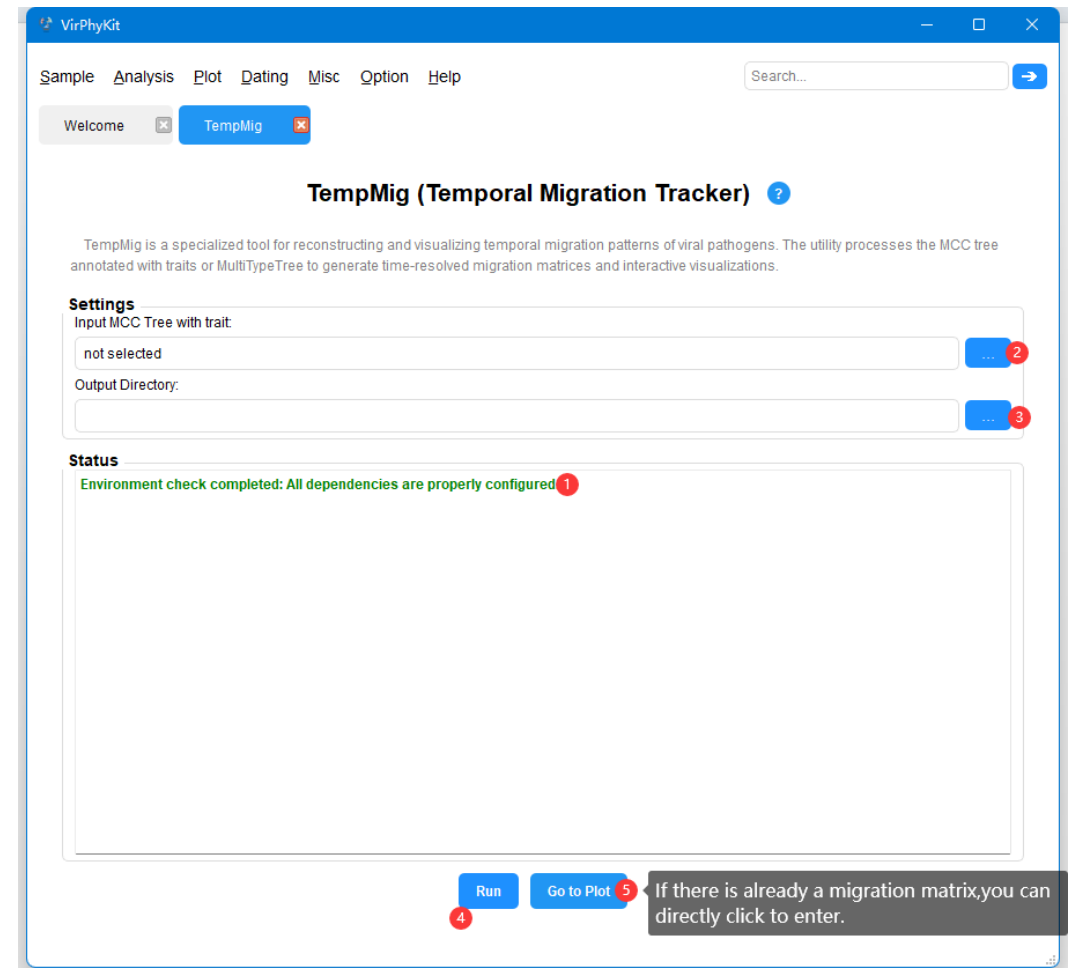
Step 1: Ensure **Python, Perl and R** installation directories are configured in the '**Option-Environment**' menu.

Step 2: Upload an **MCC tree** annotated with traits or a **MultiTypeTree**.

Step 3: Specify the output **directory** for the migration matrix.

Step 4: Click **[Run]** to process the MCC tree/MultiTypeTree and **generate migration matrix**.

Step 5: Click **[Go to plot]** to move to the '**TempMig Plotter**' tool and **visualize the results** using the migration matrix.



TempMig Plotter

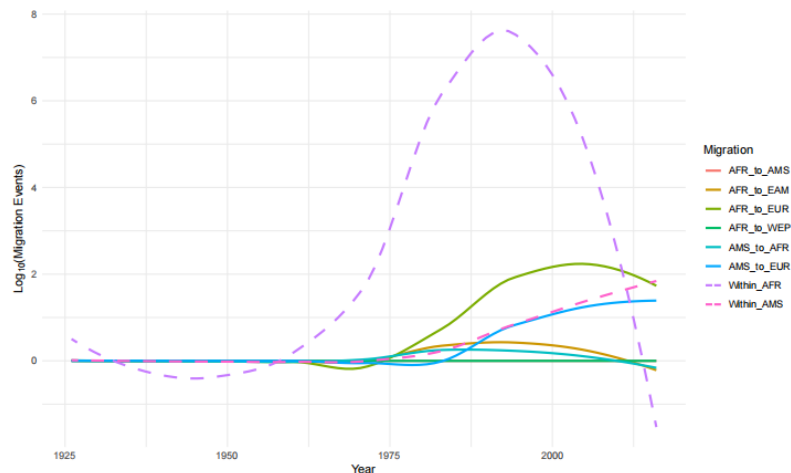
Step 1: Upload a **migration matrix file** (output from 'TempMig').

Step 2: Select the **migration directions** for **visualization**.

Step 3: Specify the output **directory**.

Step 4: Select a **visualization style** (Normal or Smooth).

Step 5: Click **[Run]** to generate **temporal migration patterns**



TempMig Plotter

Settings

Input migration matrix file: TempMig/migration_matrix.txt

Selected	From	To	Within
<input type="checkbox"/>	AFR	AFR	Within_AFR
<input type="checkbox"/>	AFR	AMS	No migration events
<input type="checkbox"/>	AFR	EAM	
<input type="checkbox"/>	AFR	EUR	
<input type="checkbox"/>	AFR	SEA	No migration events
<input type="checkbox"/>	AFR	WEP	No migration events

Output Directory:

Line styles: ☒ Normal ☐ Smooth

Status

R: Installed, including required packages (tidyr, ggplot2)

Plot

GeoSubsampler

Step 1: Upload a **FASTA sequence file** to be subsampled. Ensure each sequence name **includes its geographic region (marked with ‘_region’)**.

Step 2: (Optional) Enable the ‘Bootstrap subsampling mode’ option and set the number of replicates.

Step 3: Specify the desired **sample size** and/or **specific region** for the subset.

Step 4: Select an **output directory** to save the resulting sequences.

The screenshot shows the VirPhyKit application window with the GeoSubsampler tool active. The interface includes a menu bar (Sample, Analysis, Plot, Dating, Misc, Option, Help) and a search bar. The main panel is titled "GeoSubsampler (Sequence Geographic Subsampler)" and contains a description: "GeoSubsampler is a specialized tool for region-stratified subsampling of FASTA-formatted sequence datasets. It generates balanced sequence subsets by either geographic region or sample size, with integrated support for bootstrap analysis." Below this is a "Settings" section with the following fields: "Sequence File:" (with a file selection button and a red notification bubble with the number 1), "Bootstrap subsampling mode" (checkbox), "Replicates:" (a dropdown menu set to 10 with a red notification bubble with the number 2), "Number of Sequences (Ignored in Bootstrap subsampling mode):" (a text input field with a red notification bubble with the number 3), "Region (Optional, Ignored in Bootstrap subsampling mode)" (a text input field), and "Output Directory:" (with a file selection button and a red notification bubble with the number 4). At the bottom of the settings is a "Run" button with a red notification bubble with the number 5. A "Status" section at the bottom contains a warning message: "Warning: The original file may be overwritten in region-specific mode. Please backup before running!".

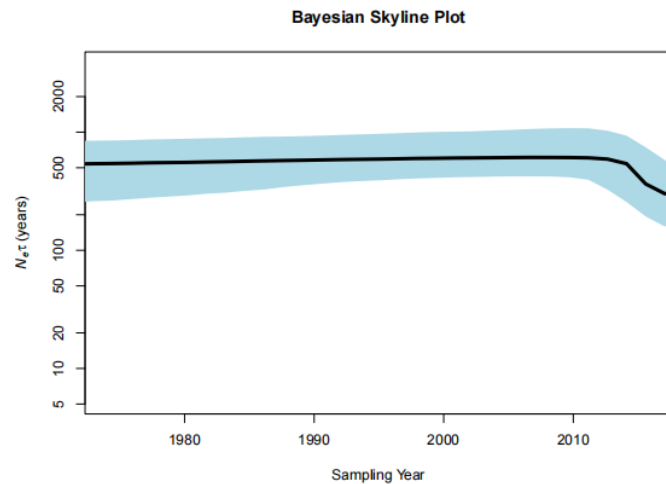
BSP-Viz

Step 1: After ensuring that the **R environment** is configured correctly, select the table file (you can click the **[Batch]** button to draw in batches later).

Step 2: Select the **time axis direction** and whether to add a **secondary axis**.

Step 3: Specify the **output directory** and click **[Hue Harmony]** to adjust color schemes.

Step 4: Click **[Plot]** to draw.



VirPhyKit

Sample Analysis Plot Dating Misc Option Help

Welcome BSP-Viz

BSP-Viz (Bayesian Skyline Plot Visualizer) ?

BSP-Viz is a specified tool for visualizing Bayesian Skyline Plots (BSP), supporting both single-file or batch processing modes.

Settings

.tsv File(s): 1

Time Axis Direction: Forward ☐ Include Secondary X-Axis

Output Directory: 3

Status

R: Installed with required packages (tidyr, ggplot2)

Hue Harmony Plot 4

RSPP-Viz

Step 1: Select the **plot style** you want to generate (pie or histogram).

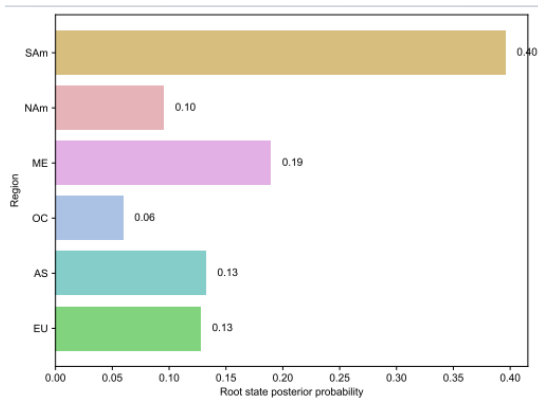
Step 2: Select the **MCC tree** (annotated with traits) or MultiTypeTree for analysis (batch processing is available in **Batch mode**).

Step 3: Specify the **output directory**.

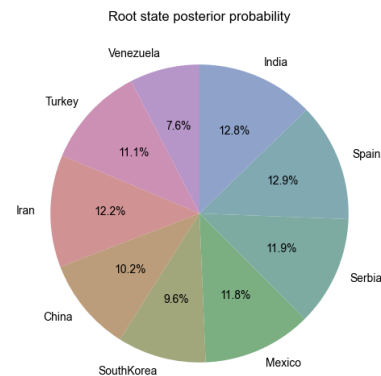
Step 4: Specify the **plot size**.

Step 4: Use **[Hue Harmony]** to adjust color schemes, and click **[Plot]** to create the plot.

Histogram



Pie



The screenshot shows the RSPP-Viz web application interface. The title bar is 'VirPhyKit'. The main menu includes 'Sample', 'Analysis', 'Plot', 'Dating', 'Misc', 'Option', and 'Help'. A search bar is located in the top right. The 'Welcome' tab is active, and the 'RSPP-Viz' tab is also visible. The main heading is 'RSPP-Viz (Root State Posterior Probability Visualizer)'. Below this, a description states: 'RSPP-Viz is a specialized tool for visualizing root state posterior probabilities inferred from MCC trees (annotated with traits) or MultiTypeTree.' The 'Settings' section includes: 'Plot Style' (radio buttons for 'Pie charts' and 'Histogram', with 'Histogram' selected), 'MCC Tree File' (a text input field with a file selection button and a 'Batch' button), 'Output Directory' (a text input field with a file selection button and a 'View' button), and 'Plot Size (pixels)' (input fields for 'Width' (800) and 'Height' (600)). The 'Status' section is empty. At the bottom, there are buttons for 'Hue Harmony' and 'Plot'.

TreeTime-RTT

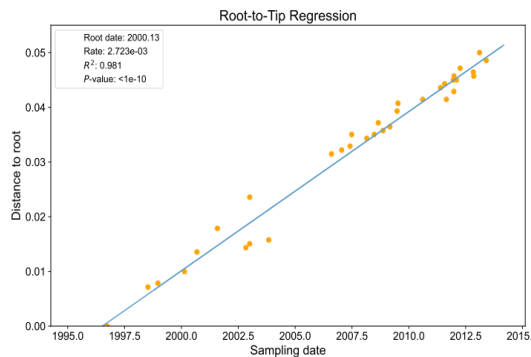
Step 1: Upload a **FASTA sequence** file, a **Newick tree** file, and a **metadata** file (.csv).

Step 2: If the third column of the metadata file is a **“region”** (or other **trait**) column and a **trait mapping file** is uploaded, isolates sharing the same trait will be **color-coded** in the plot. Otherwise, only the **standard RTT regression** plot will be generated.

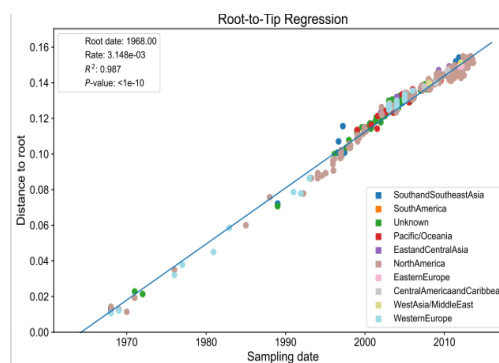
Step 3: Select the **output directory**.

Step 4: Click **[Run]** to perform the RTT analysis.

standard



color-coded



VirPhyKit

Sample Analysis Plot Dating Misc Option Help

Welcome TreeTime-RTT

TreeTime-RTT (Root-to-tip Regression by TreeTime)

TreeTime-RTT is a tool for assessing temporal signal (clock-like evolutionary patterns) through linear regression of root-to-tip distance against sampling dates in TreeTime.

Settings

Fasta File (Aligned):

Newick Tree File:

Metadata File:

Mapping File (Optional):

Built-in default mapping table (grouped by region)

Output Directory:

Status

Run

The screenshot shows the VirPhyKit web interface for the TreeTime-RTT tool. The interface includes a navigation bar with tabs for Sample, Analysis, Plot, Dating, Misc, Option, and Help. A search bar is located in the top right. The main content area is titled "TreeTime-RTT (Root-to-tip Regression by TreeTime)" and provides a brief description of the tool. Below the description, there is a "Settings" section with four input fields: "Fasta File (Aligned)", "Newick Tree File", "Metadata File", and "Mapping File (Optional)". Each field has a "..." button to the right, and red arrows with numbers 1, 2, and 3 point to these buttons. The "Mapping File (Optional)" field has a dropdown menu showing "Built-in default mapping table (grouped by region)". Below the settings, there is an "Output Directory" field with a "..." button. At the bottom of the settings section, there is a "Run" button with a red arrow and the number 4 pointing to it. A "Status" section is located at the bottom of the interface, but it is currently empty.

TreeDater-LTT

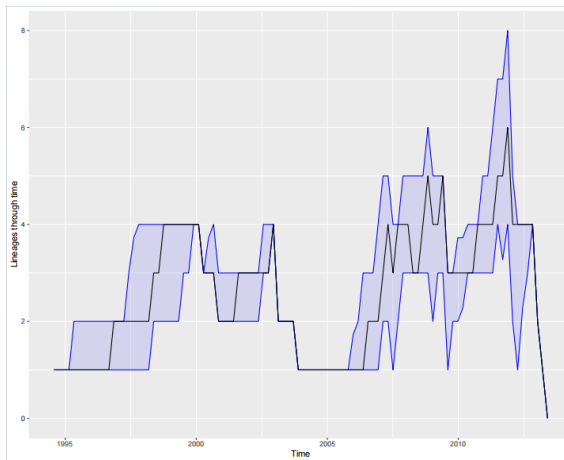
Step 1: Select a **Newick tree** file (.nwk).

Step 2: Upload the **metadata file** (.csv) with sampling dates.

Step 3: Enter sequence **length** and enable the '**LTT**' option to generate a **lineage-through-time plot**.

Step 4: Specify the **output directory** for saving results.

Step 5: Click **[Run]** to perform the LTT analysis using TreeDater, and view the results in the **Status** panel and output images.



Try it yourself !