

VirPhyKit: An integrated toolkit for viral phylogeographic analysis

A quick guide to VirPhyKit

Sept, 2025

Yuqi Yin & Fangluan Gao

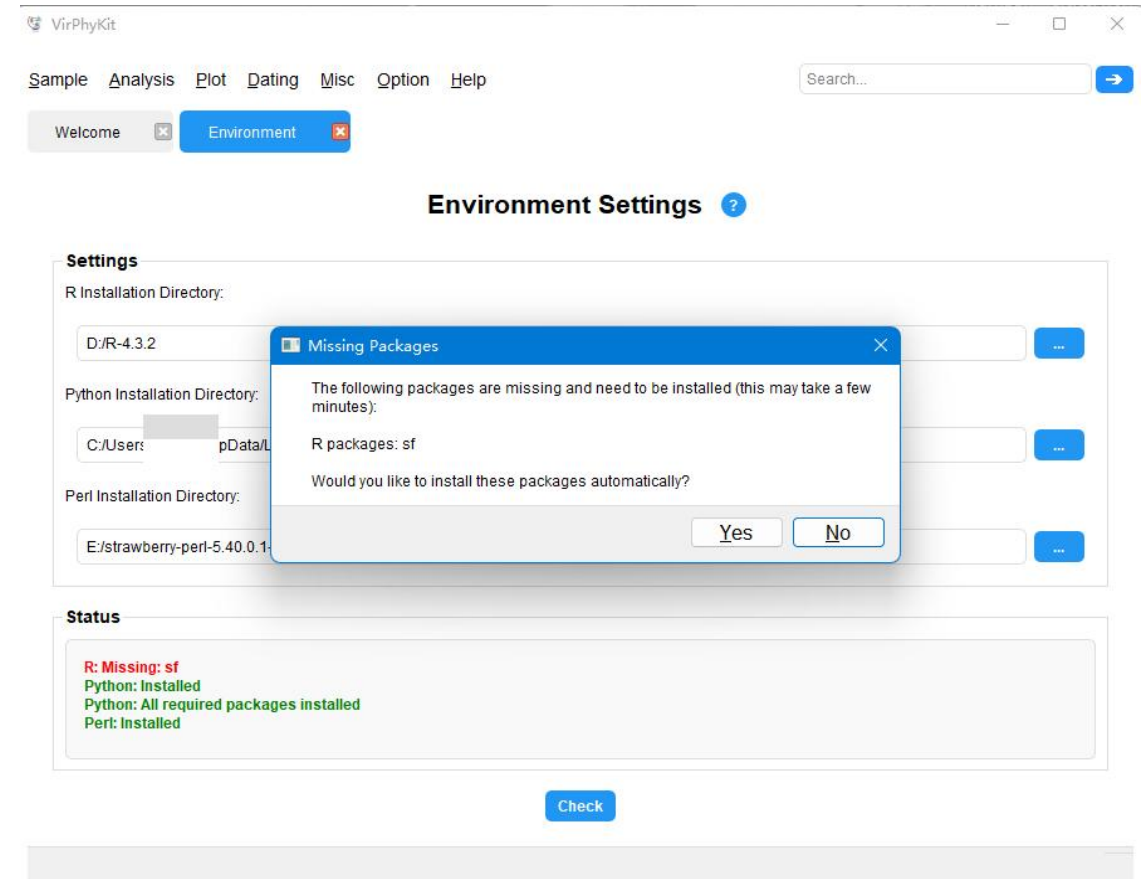
1. Installation

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 web application interface. The interface includes a navigation bar with links: Sample, Analysis, Plot, Dating, Misc, Option, and Help. A search bar is located in the top right corner. 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, Options, Status, and Results. The Settings section contains fields for "Virus Name:" and "Accession File:", with a "Search" button next to the "Virus Name:" field. The Options section contains fields for "Output Directory:" and "Genome Segments:", with a "Download" button next to the "Genome Segments:" field. The Status section contains a text area for "Enter the full virus name to retrieve sequences, or select an accession number file to download sequences directly." and a "Retrieval progress: 0/0" indicator. The Results section contains a table with columns: Type, Number, Percentage, and Describe. Red arrows and numbers (1, 2, 3, 4) highlight specific elements: 1 points to the "Virus name or Accession File" input field, 2 points to the "Search" button, 3 points to the "Genome Segments" dropdown menu, and 4 points to the "Download" button.

VirPhyKit

Sample Analysis Plot Dating Misc Option Help

Welcome SeqHarvester

SeqHarvester (Sequence Metadata Harvester) ?

SeqHarvester retrieves viral isolates from GenBank, displaying comprehensive metadata for filtering and download.

Settings

Virus Name: **Search**

OR **1**

Accession File:

Options

Output Directory:

Genome Segments: **3**

Status

Enter the full virus name to retrieve sequences, or select an accession number file to download sequences directly.

Retrieval progress: 0/0

Results

Type	Number	Percentage	Describe
------	--------	------------	----------

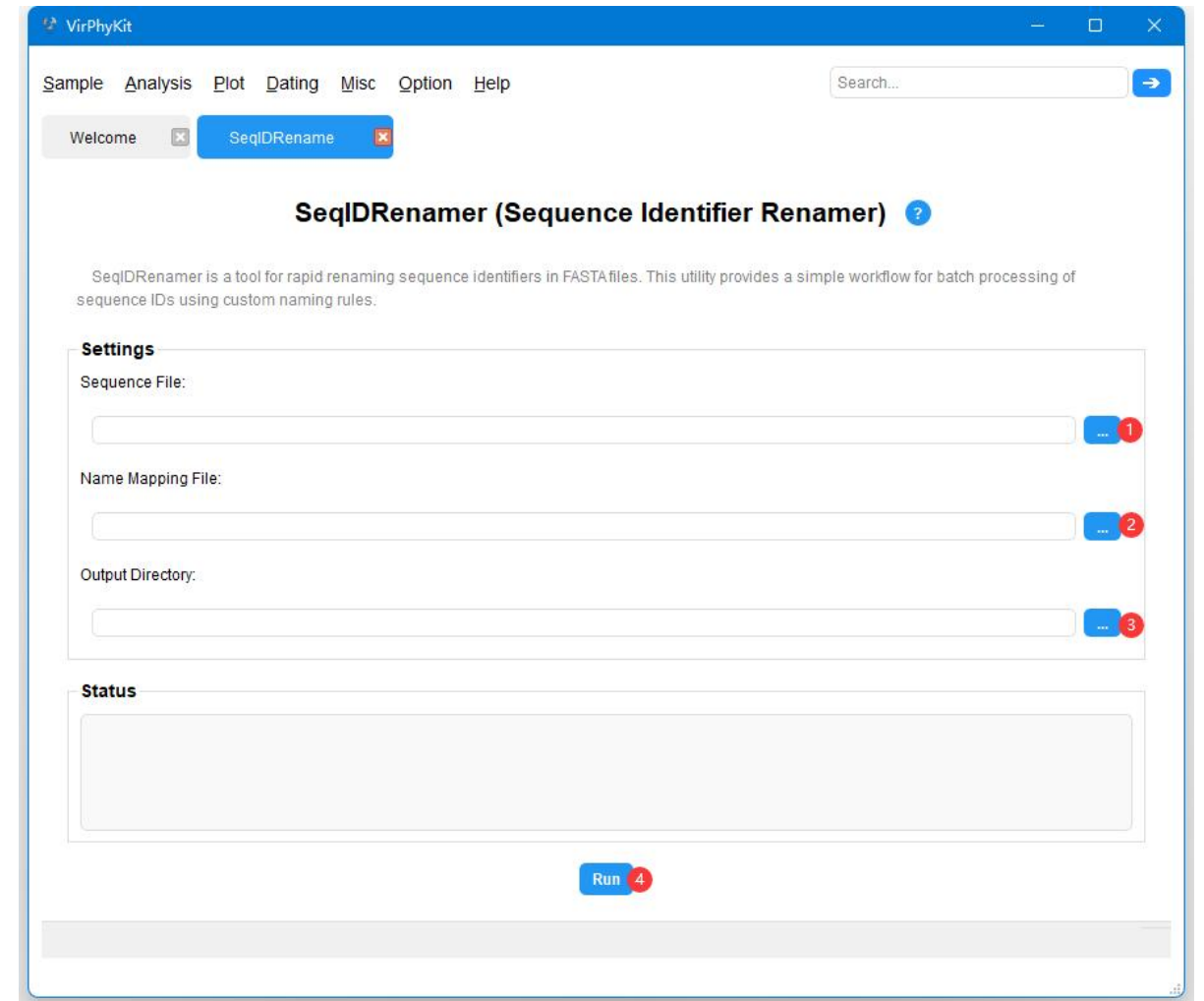
SeqIDRenamer

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

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').

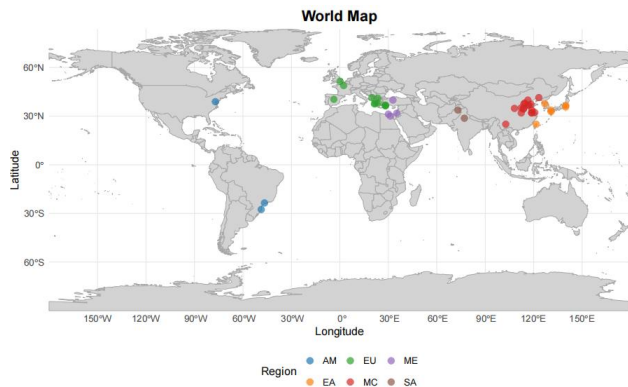
Step 2: Select **spatial** or **temporal** visualization mode.

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

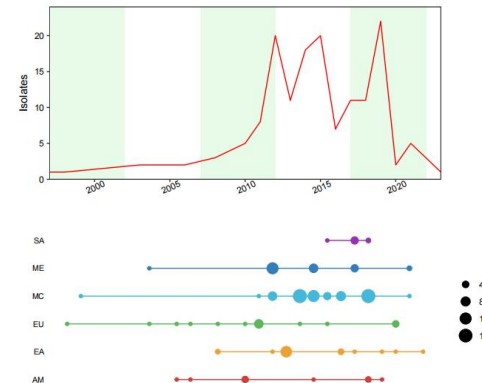
Step 4: Select the **output directory** and **filename**.

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

Spatial mode



Temporal mode



The VirPhyKit interface for VirSpaceTime visualization. The interface includes a menu bar (Sample, Analysis, Plot, Dating, Misc, Option, Help) and a search bar. The main panel displays the "VirSpaceTime (Viral Space-Time Visualizer)" tool. The "Settings" section includes a "Visualization mode" dropdown (Temporal selected), an "Input File (.bt)" field, an "Output Directory" field, and an "Output Filename" field (output_plot). The "Status" section shows "R environment ready". A "Generate" button is located at the bottom right.

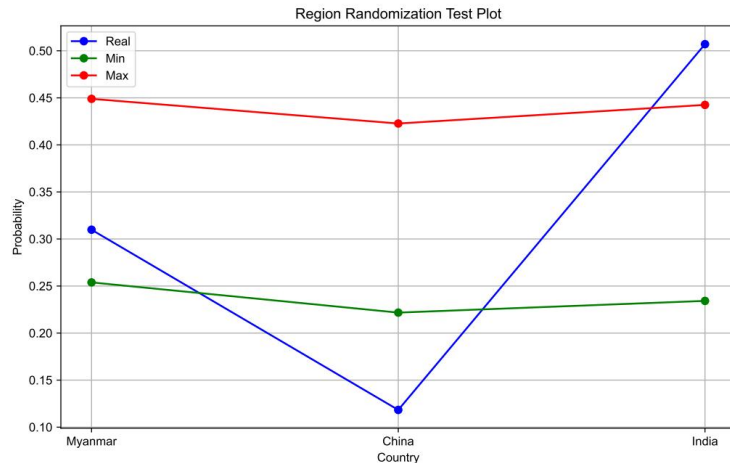
RRT

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.



VirPhyKit

Sample Analysis Plot Dating Misc Option Help

Welcome RRT

RRT (Region Randomization Test) ?

RRT is a statistical method that detects sampling bias in phylogeographic analyses. It compares ancestral root state probabilities between the original dataset and location-permuted scenarios to correct for systematic overestimation of viral origins in oversampled regions.

Settings

Original MCC Tree:

C:/Users/.../Desktop/Example/RRT/Strict BSP MCC.tre

Randomized MCC Trees:

.../RRT/R8 mcc.tre, C:/Users/.../Desktop/Example/RRT/R9 mcc.tre, C:/Users/.../Desktop/Example/RRT/R10 mcc.tre

Export Data (CSV):

C:/Users/.../Desktop/output.csv

Status

Randomized file uploaded.
Randomized file uploaded.
Randomized file uploaded.
RRT SUCCESSFULLY PASSED! Click 'Preview' to view the results.

Run Preview

TempMig

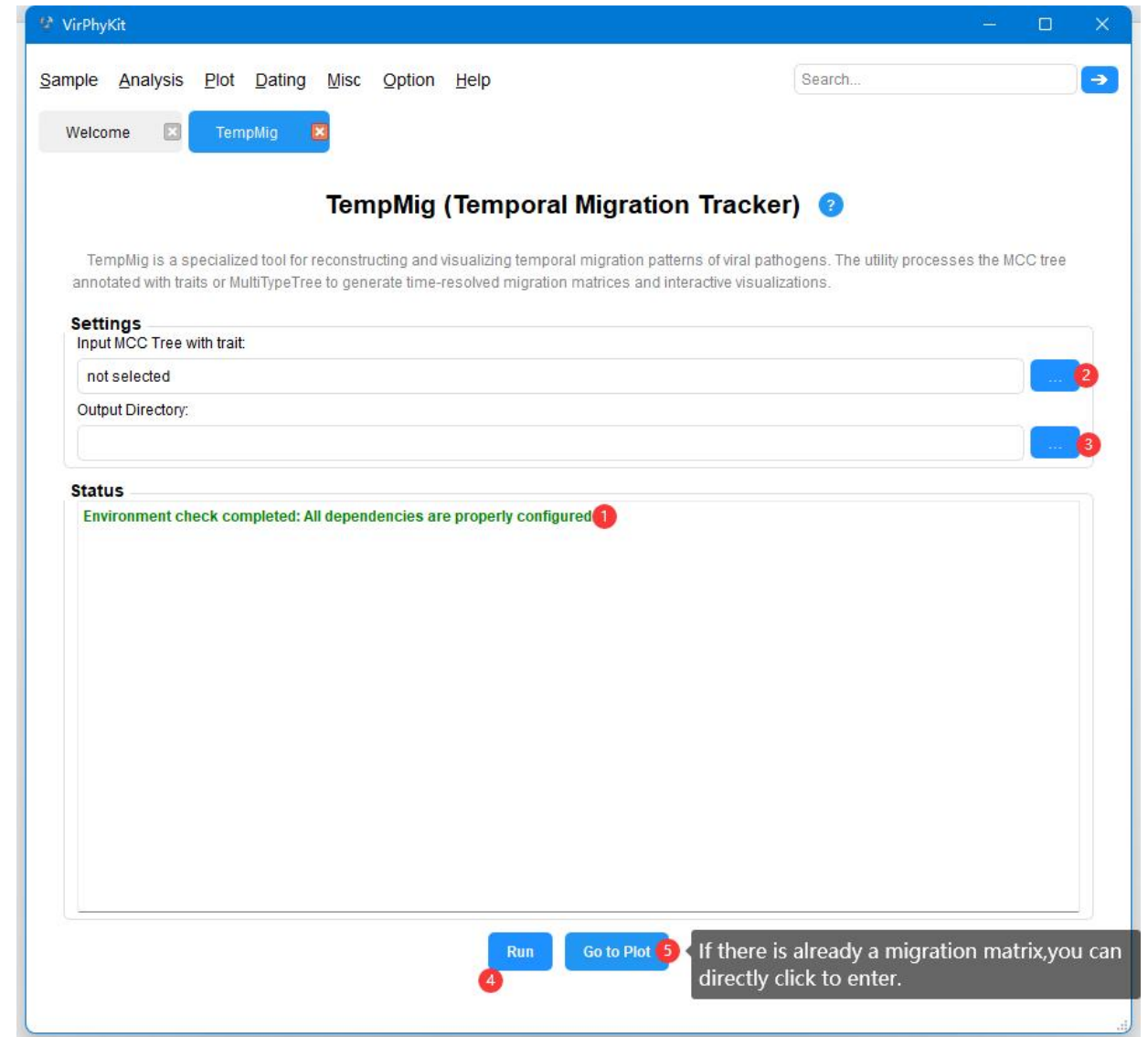
Step 1: Ensure **Python, Perl and R** installation directories are configured in the '**Option-Environment**' menu.

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

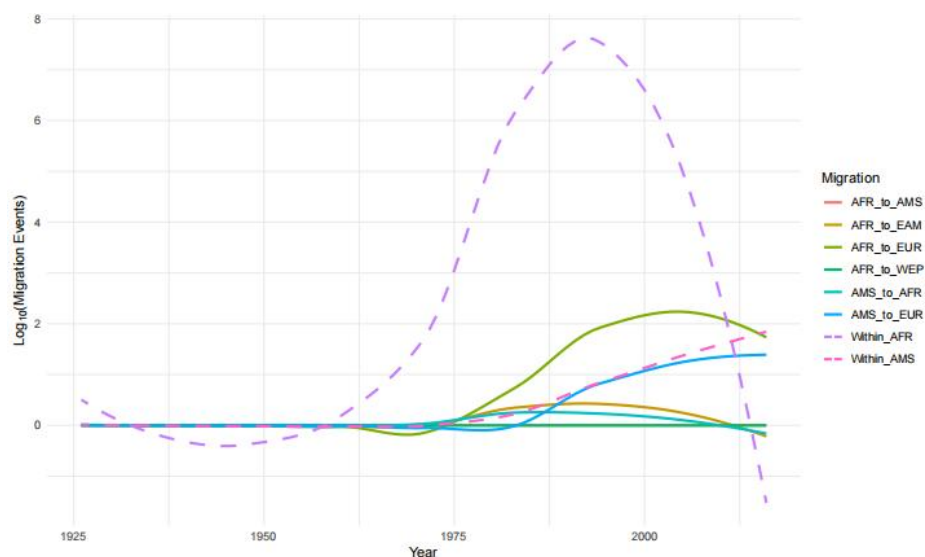
Step 1: Upload a **migration matrix file** (output from 'TempMig').

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

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

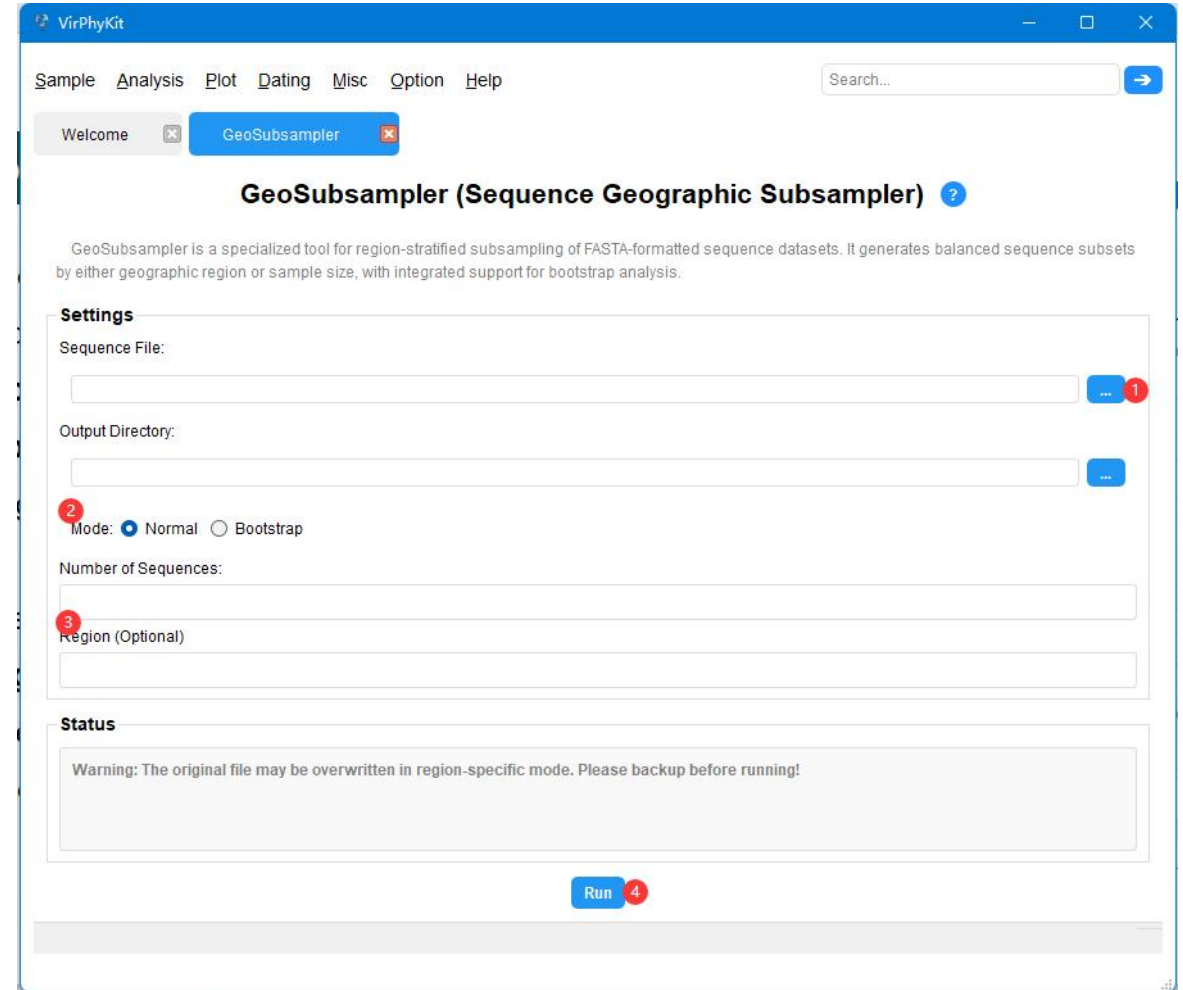
GeoSubsampler

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 web interface with the GeoSubsampler tool active. The interface includes a navigation bar with links for Sample, Analysis, Plot, Dating, Misc, Option, and Help. A search bar is located in the top right. The main content area is titled "GeoSubsampler (Sequence Geographic Subsampler)" and includes a brief 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."

The "Settings" section contains the following fields and options:

- Sequence File:** A text input field with a blue "..." button and a red notification bubble (1).
- Output Directory:** A text input field with a blue "..." button.
- Mode:** Radio buttons for "Normal" (selected) and "Bootstrap". A red notification bubble (2) is next to the "Mode" label.
- Number of Sequences:** A text input field.
- Region (Optional):** A text input field. A red notification bubble (3) is next to the label.

The "Status" section displays a warning message: "Warning: The original file may be overwritten in region-specific mode. Please backup before running!". At the bottom right, there is a blue "Run" button with a red notification bubble (4).

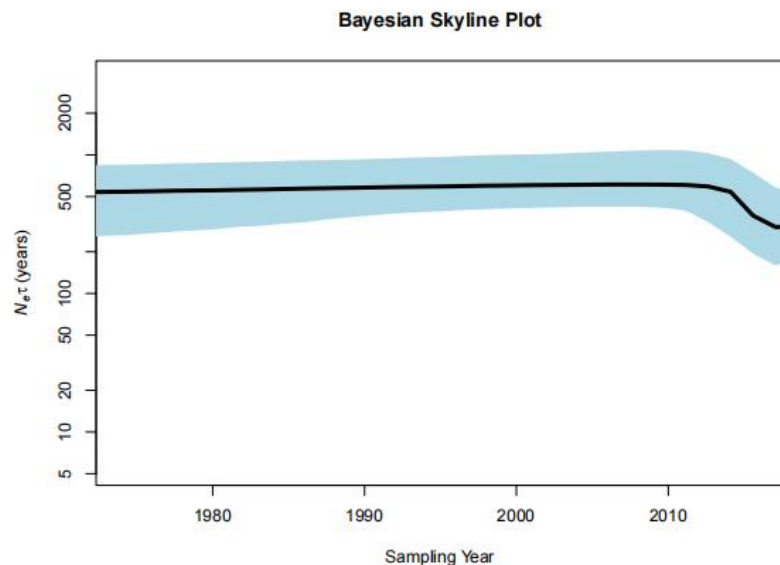
BSP-Viz

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

Search...

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).

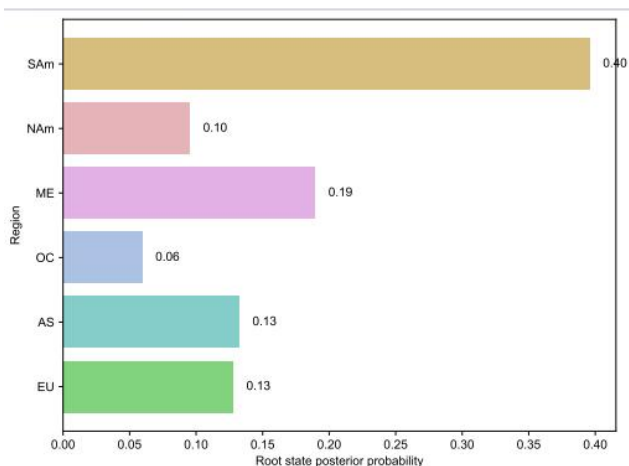
Step 2: Select the **MCC tree** (annotated with traits) or MultiTypeTree for analysis (batch processing is available in **Batch mode**).

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

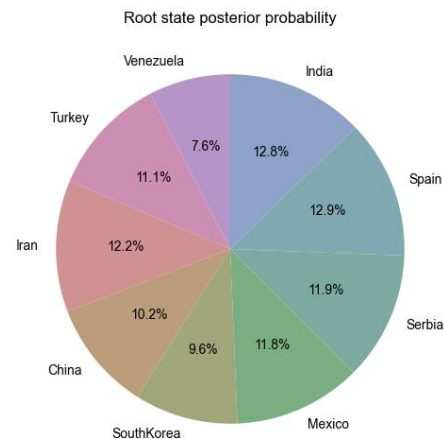
Step 4: Specify the **plot size**.

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 on the 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

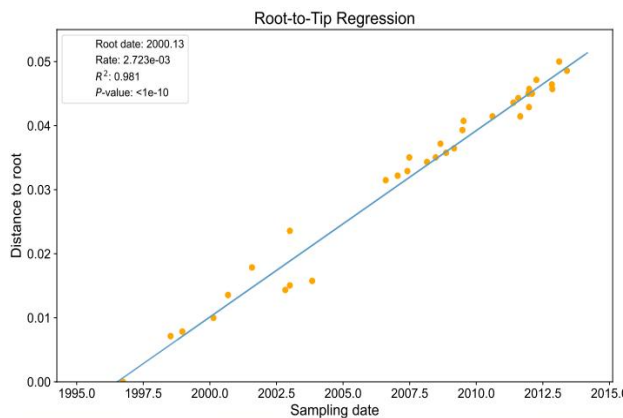
Step 1: Upload a **FASTA** sequence file, a **Newick tree** file, and a **metadata** file (.csv).

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.

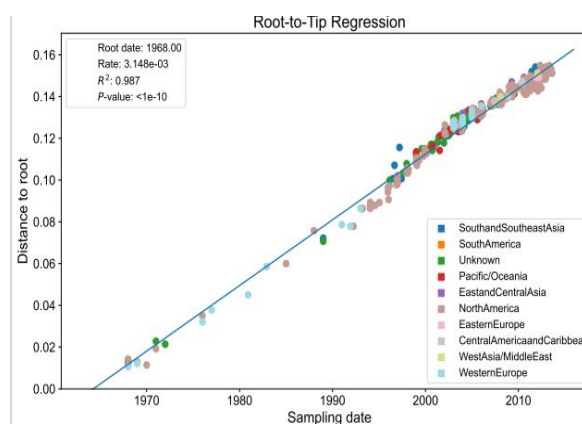
Step 3: Select the **output directory**.

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 image shows the VirPhyKit web interface for the TreeTime-RTT tool. The interface is divided into several sections: a top navigation bar with links to Sample, Analysis, Plot, Dating, Misc, Option, and Help; a main header area with a search bar and a 'TreeTime-RTT' tab; a settings section with input fields for Fasta File (Aligned), Newick Tree File, Metadata File, Mapping File (Optional), and Output Directory; and a status section at the bottom. Red arrows and numbers (1, 2, 3, 4) are overlaid on the interface to indicate the steps: 1 points to the Fasta File input, 2 points to the Metadata File input, 3 points to the Output Directory input, and 4 points to the 'Run' button.

TreeDater-LTT

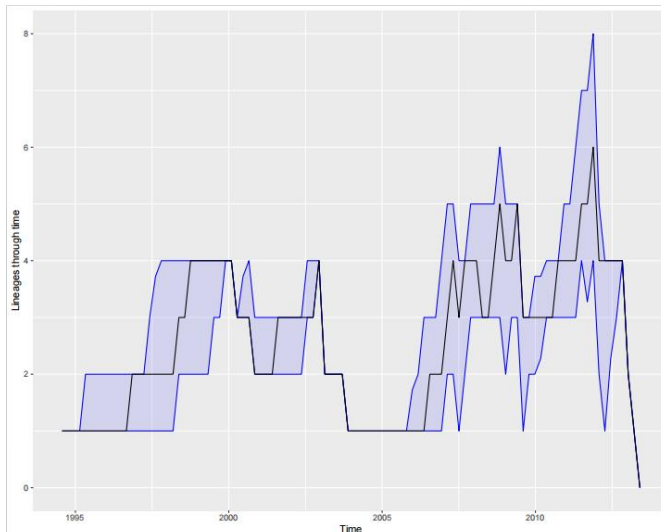
Step 1: Select a **Newick tree** file (.nwk).

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.



The screenshot shows the VirPhyKit web interface for the TreeDater-LTT tool. The interface includes a navigation bar with tabs: Sample, Analysis, Plot, Dating, Misc, Option, and Help. A search bar is located in the top right. The main panel is titled 'TreeDater-LTT (Lineage Through Time with Treedater)' and contains a description of the tool. Below the description is a 'Settings' section with four input fields: 'Tree File (.nwk)', 'Metadata File (.csv)', 'Sequence Length', and 'Output Directory'. Each field has a blue button with a red number (1, 2, 3, and 4 respectively) indicating the step number. A red double-headed arrow connects the 'Sequence Length' field and the 'Plot Lineage Through Time (LTT)' checkbox, with a red number 3 in the middle. The 'Plot Lineage Through Time (LTT)' checkbox is currently unchecked. At the bottom of the 'Settings' section is a 'Status' panel. A blue 'Run' button with a red number 5 is located at the bottom right of the interface.

Try it yourself!