VirPhyKit: An integrated toolkit for viral phylogeographic analysis

A quick guide to VirPhyKit

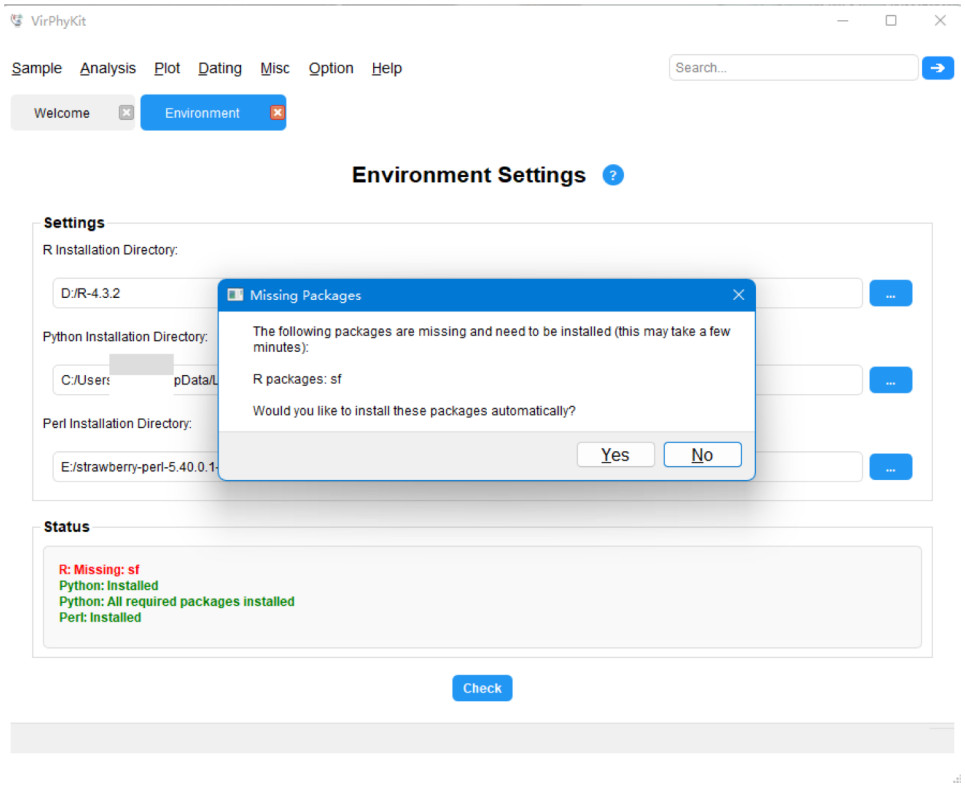
Sept, 2025

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

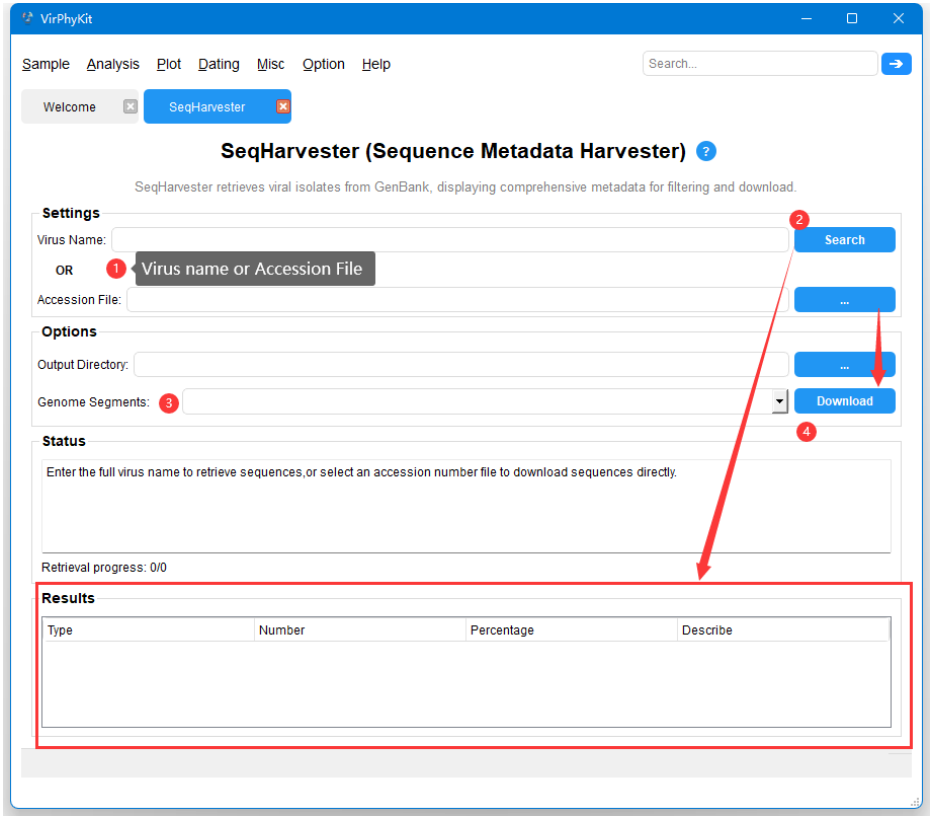
SeqHarvester

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.



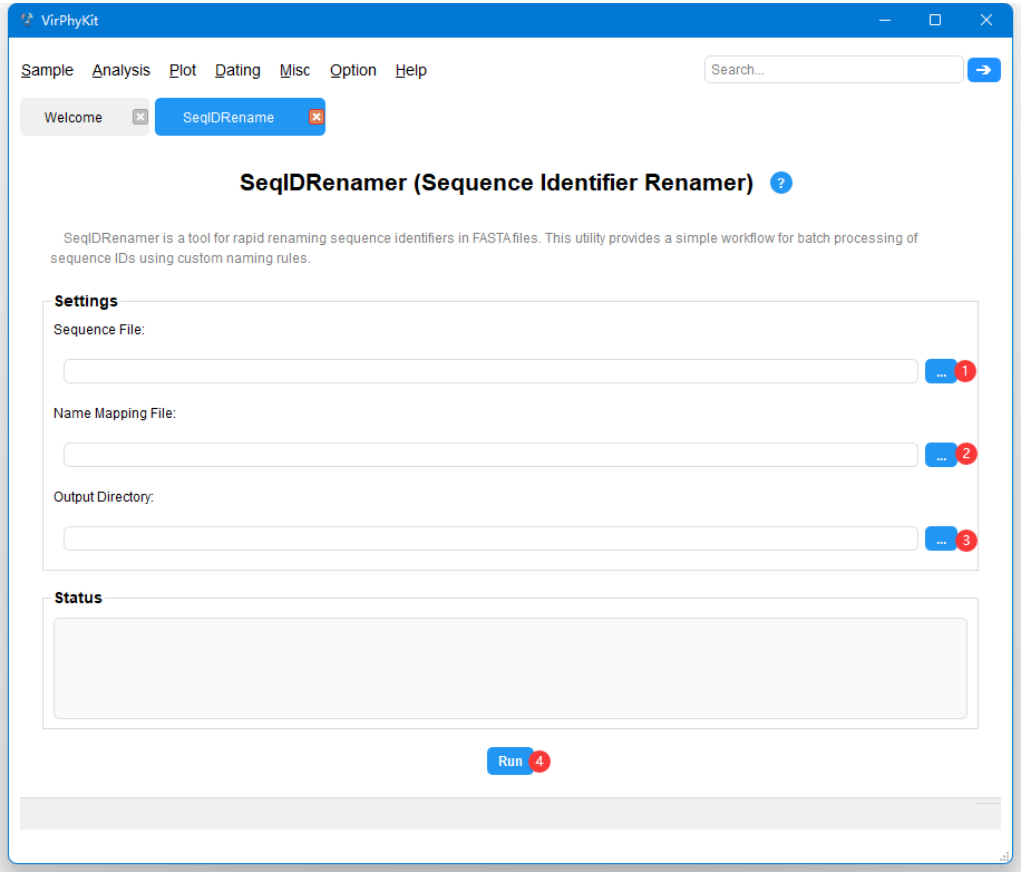
SeqIDRenamer

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.



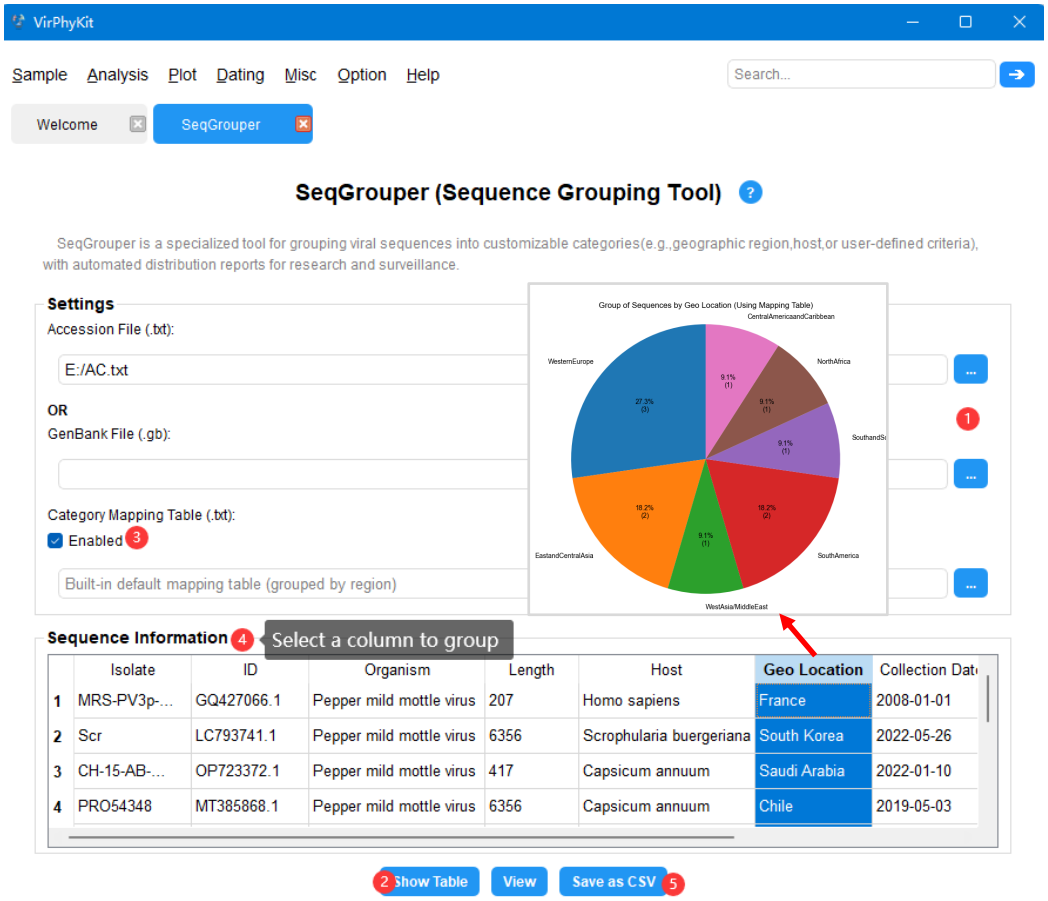
SeqGrouper

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.



VirSpaceTime

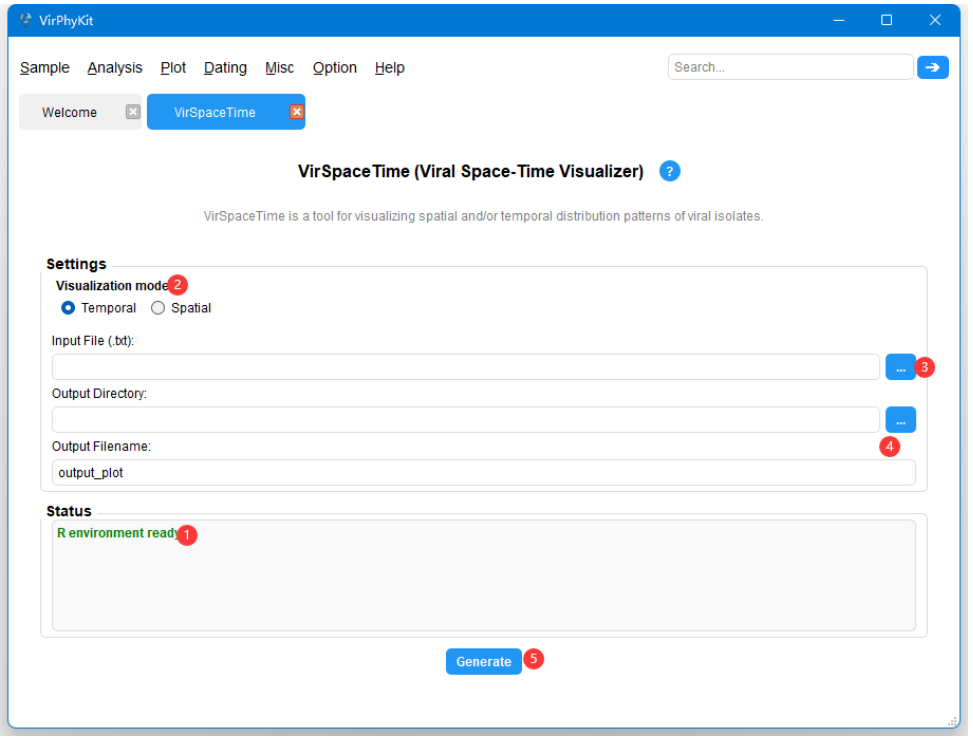
Step 1: Ensure R is properly configured (in ‘Option-Environment’).

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

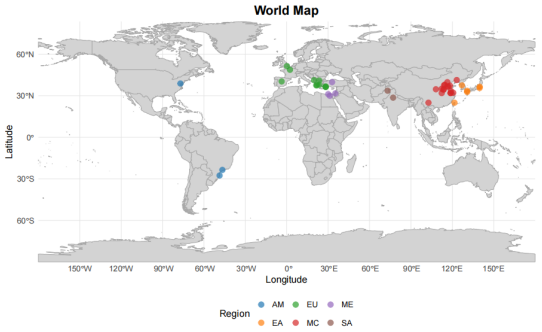
Step 3: Upload a **tab-delimited text file**, containing **geographic coordinates** for spatial analysis or **population sample sizes over time** for temporal analysis.

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

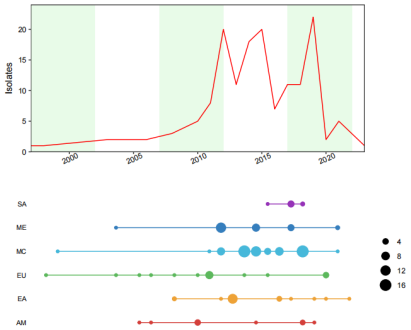
Step 5: Click **[Generate]** to create visualization.



Spatial mode



Temporal mode



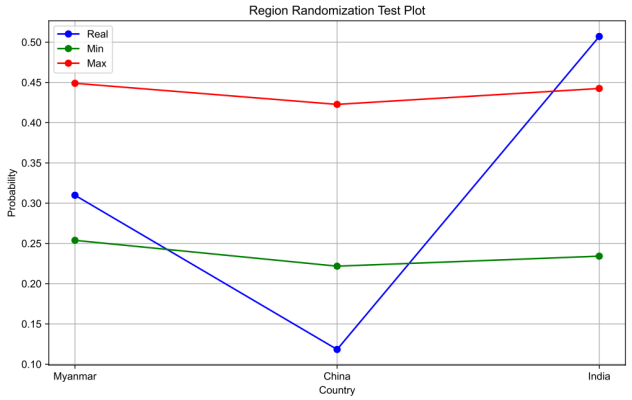
RRT

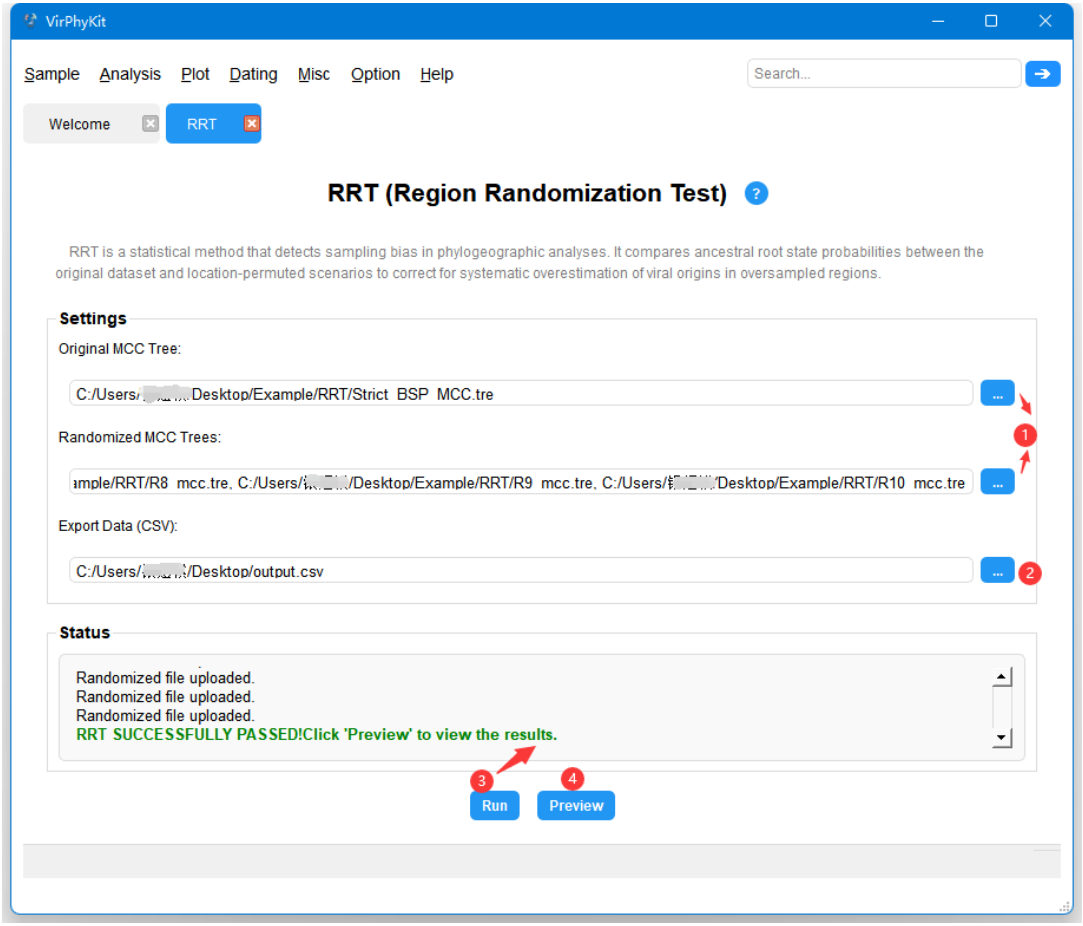
Step 1: Upload the **original MCC tree** and the **region-randomized MCC tree**.

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.





TempMig

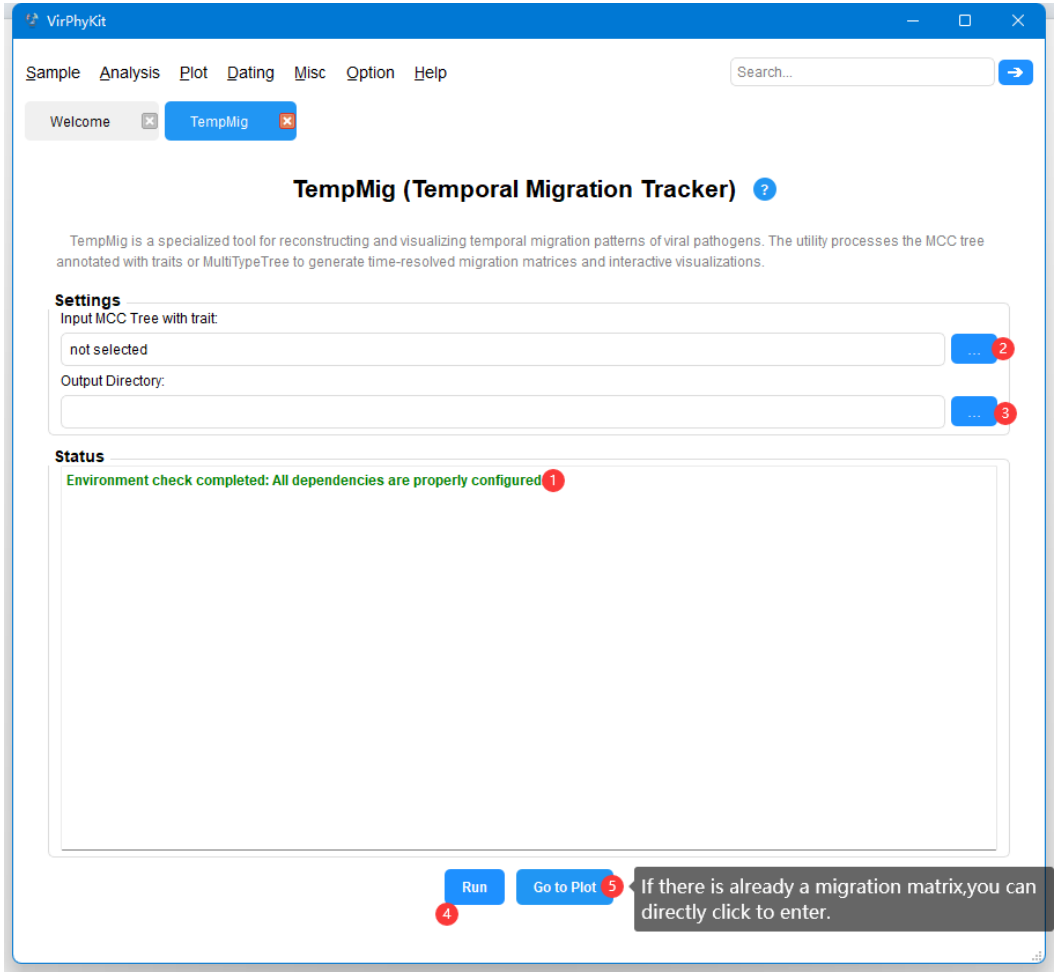
Step 1: Ensure **Python, Perl and R** installation directories are configured in the **'Option-Environment'** menu.

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

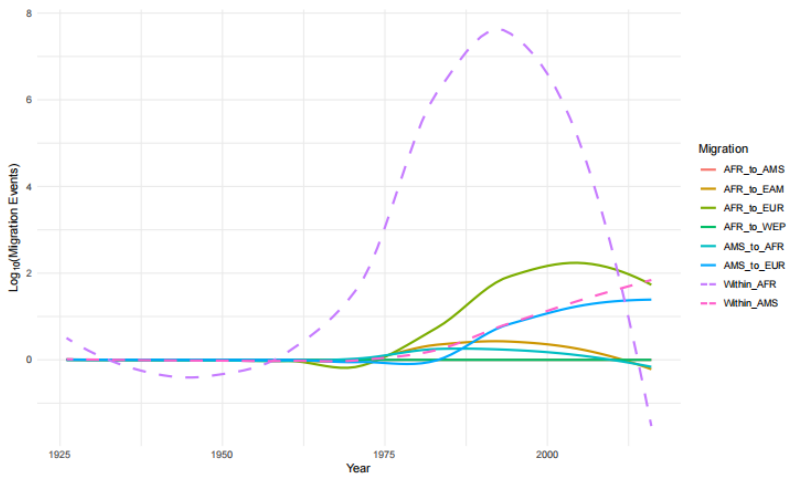
Step 1: Upload a **migration matrix file** (output from ‘TempMig’).

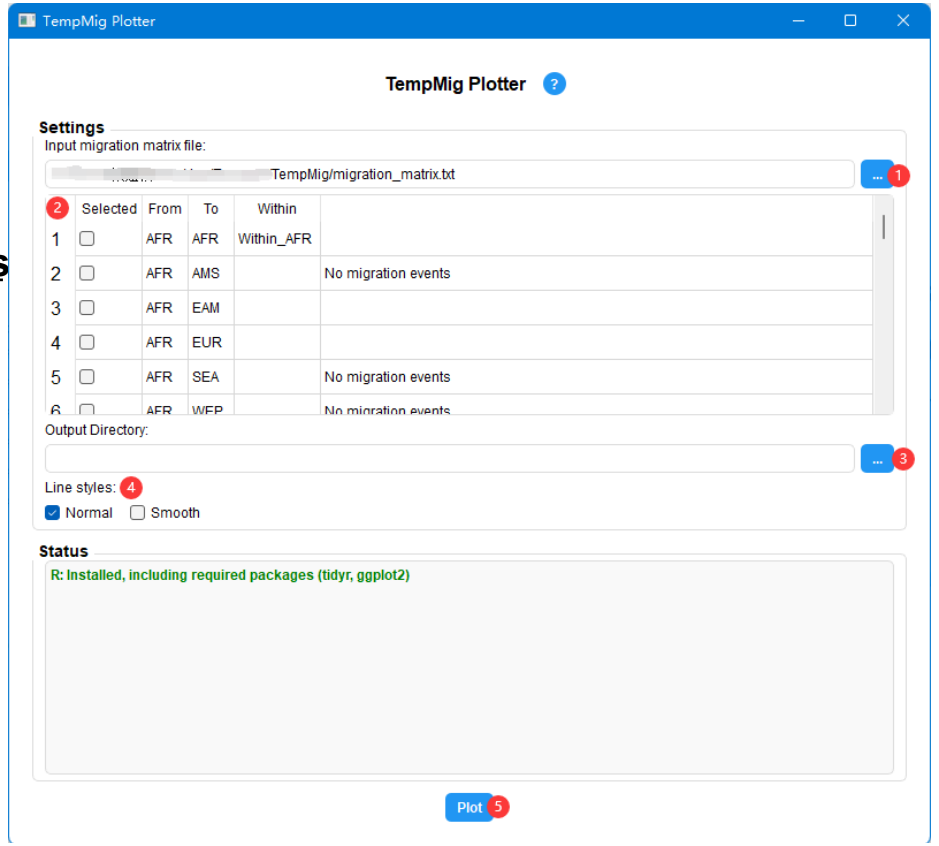
Step 2: Select the **migration directions** for **visualization.**

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

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

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





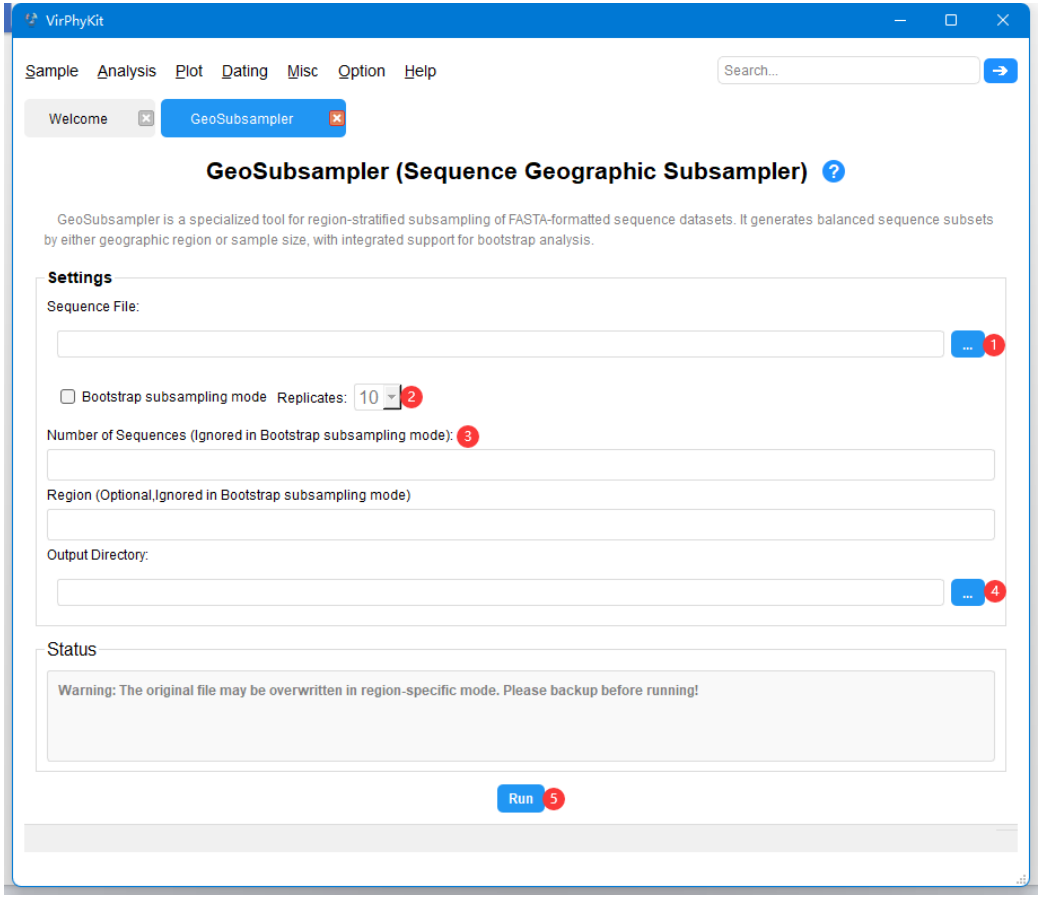
GeoSubsampler

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.



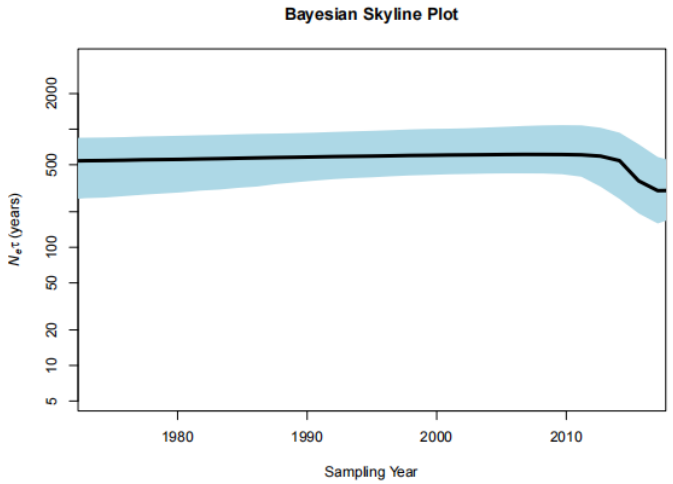
BSP-Viz

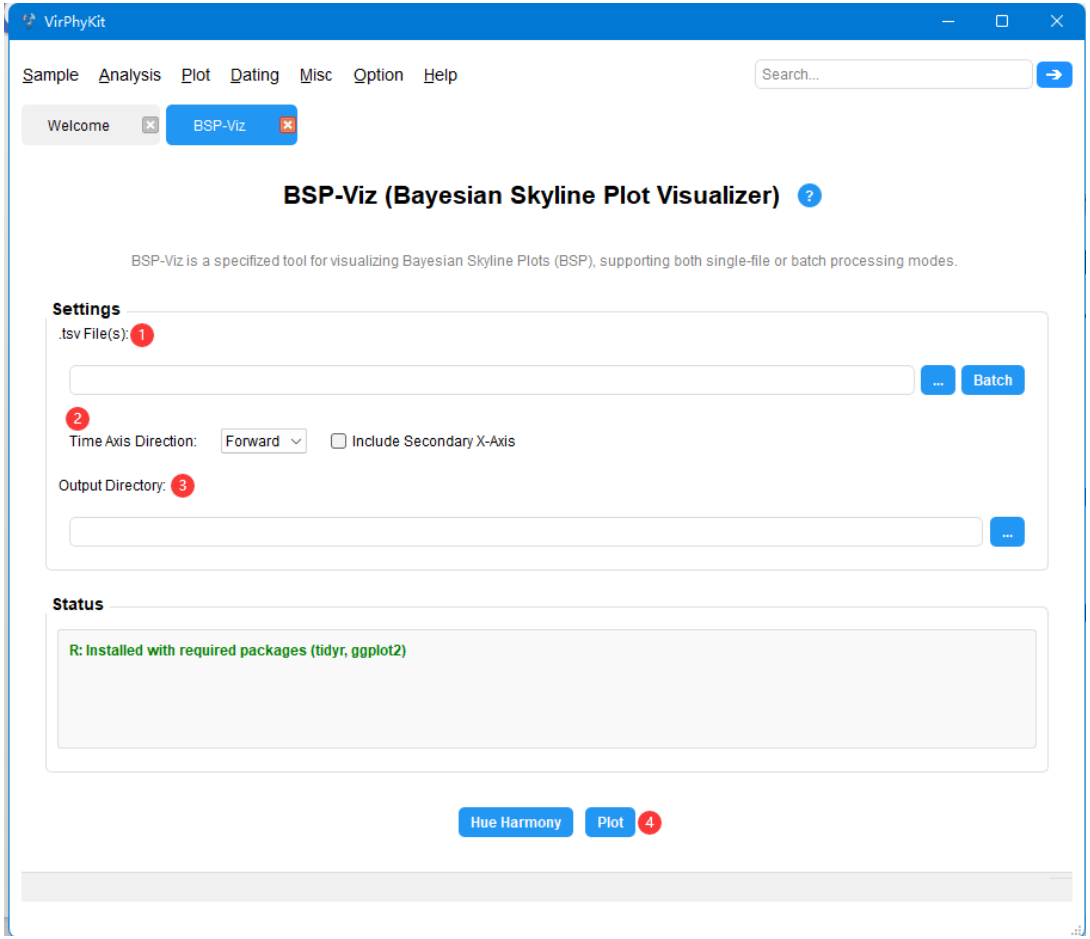
Step 1: After ensuring that the **R environment** is configured correctly, select the table file (you can click the **[Batch]** button to draw in batches later).

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.





RSPP-Viz

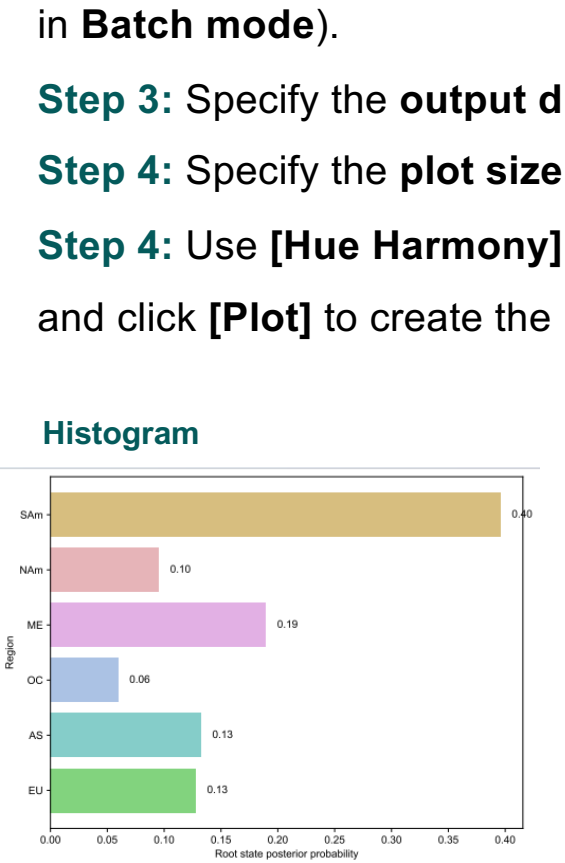
Step 1: Select the **plot style** you want to generate (pie or histogram).

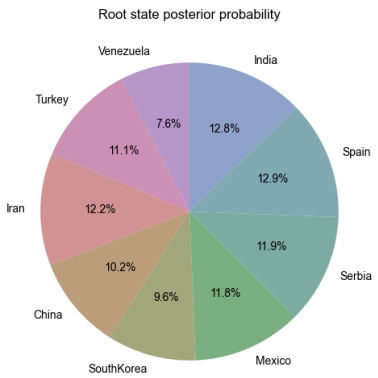
Step 2: Select the **MCC tree** (annotated with traits) or MultiTypeTree for analysis (batch processing is available in **Batch mode**).

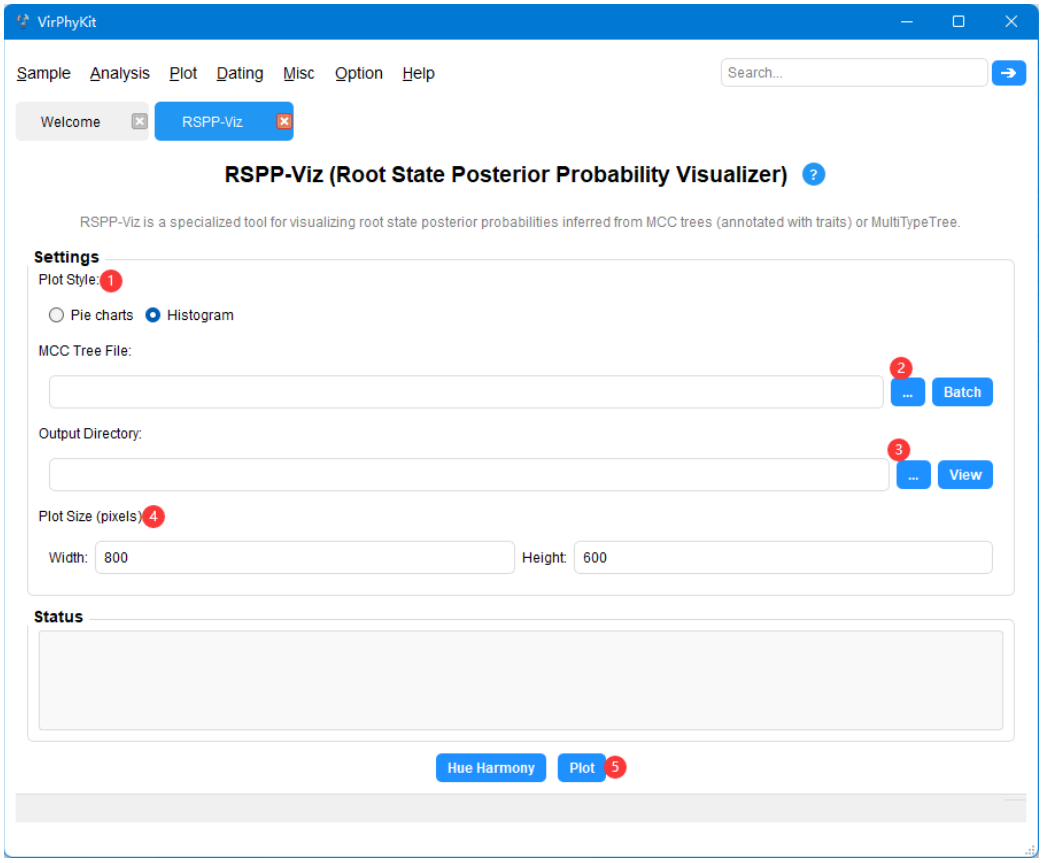
Step 3: Specify the **output directory**.

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

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







TreeTime-RTT

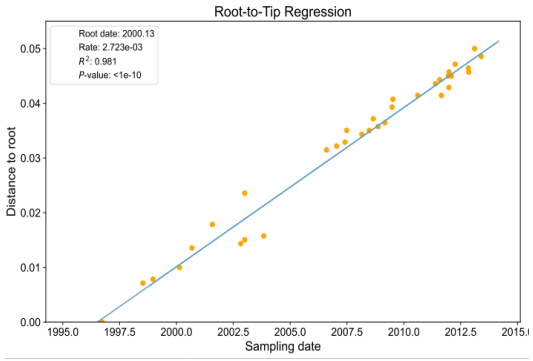
Step 1: Upload a **FASTA sequence** file, a **Newick tree** file, and a **metadata** file (.csv).

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.

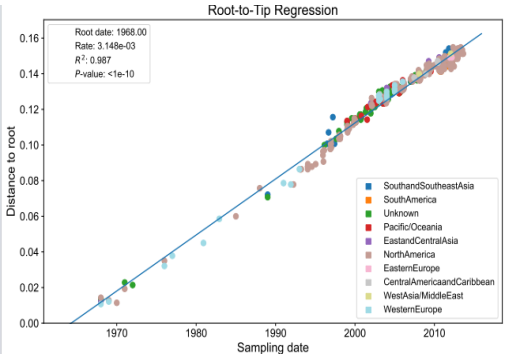
Step 3: Select the **output directory**.

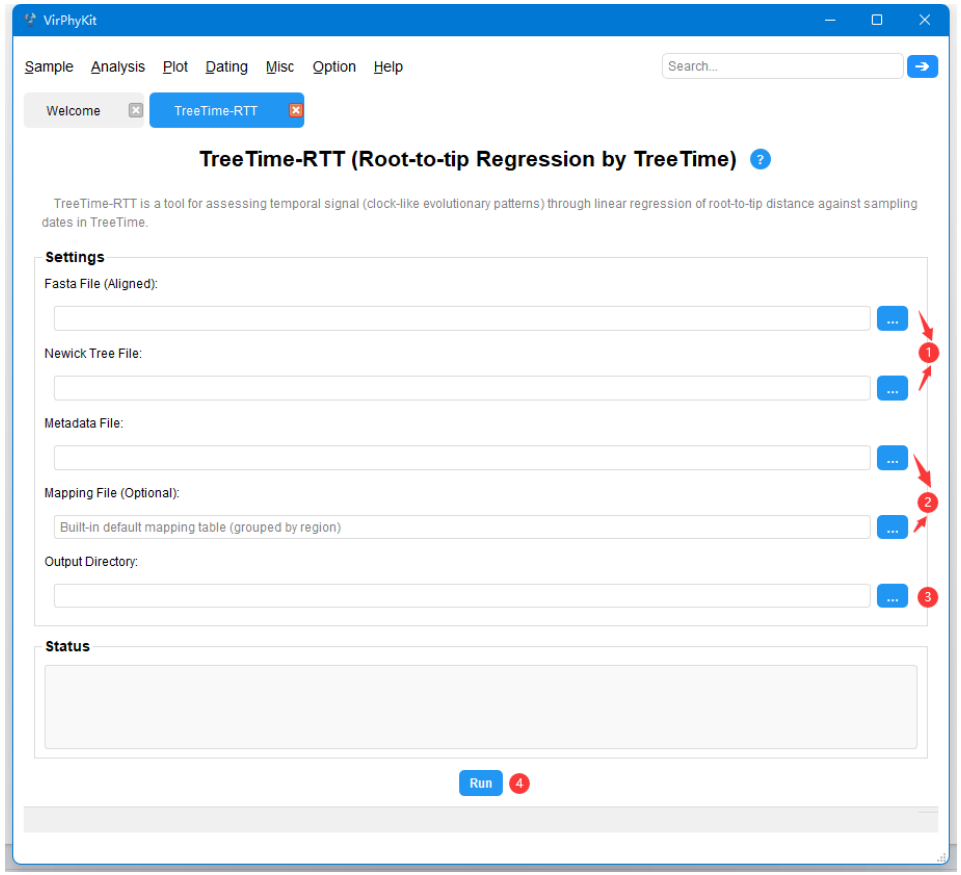
Step 4: Click **[Run]** to perform the RTT analysis.

standard



color-coded





TreeDater-LTT

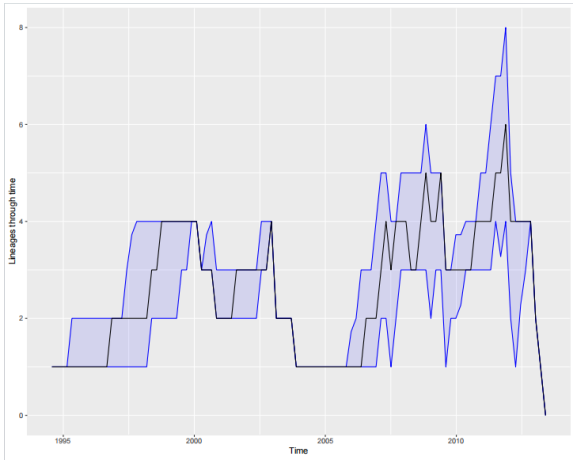
Step 1: Select a **Newick tree** file (.nwk).

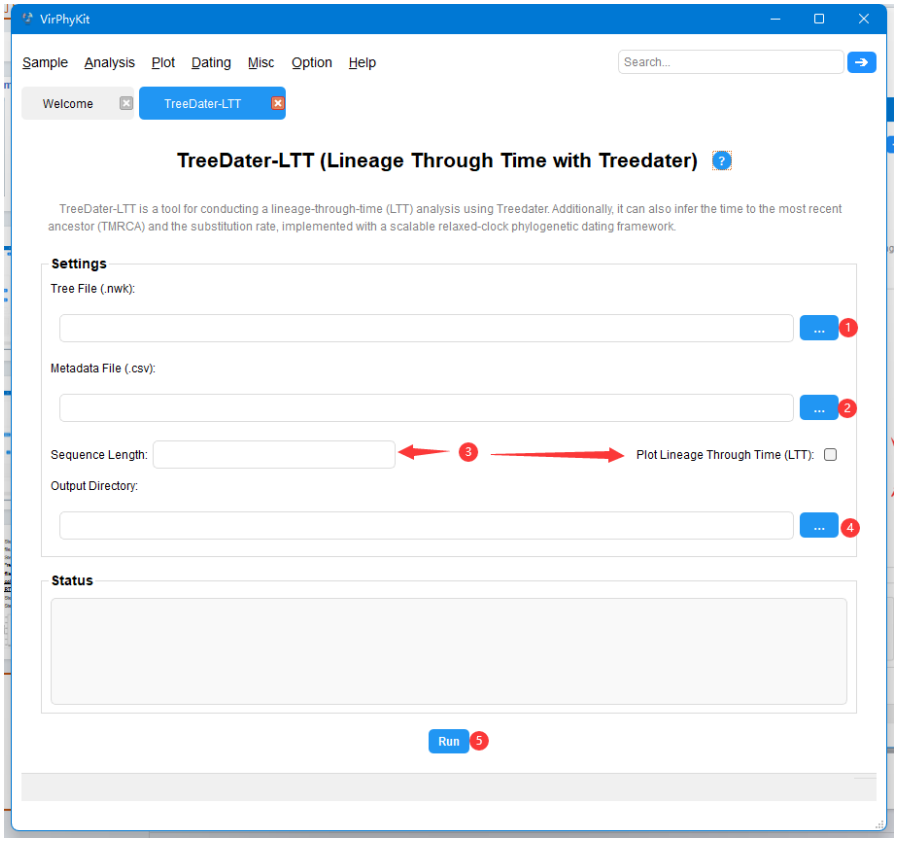
Step 2: Upload the **metadata file** (.csv) with sampling dates.

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 !